Computational Epigenetic Micromodel - Framework for Parallel Implementation and Information Flow

Karthika Raghavan and Heather J Ruskin
Center for Scientific Modeling and Complex Systems Modeling (Sci-Sym)
kaghavan@computing.dcu.ie

Following the sequencing of the human genome, attempts to characterise the human epigenetic profile and determine gene expression have focused on the roles played by key elements, such as dynamic histone modifications, and stable DNA methylation. In this paper, we elaborate on the development of a novel micro model, which aims to predict molecular events leading to quantification of gene expression. This bottom-up approach facilitates low-level information processing and enables an overview of the information exchange between scales, whereby gene expression controls the phenotype or physical appearance of an organism at higher level. The focus here is testing and refinement of the model for a large dataset, in order to assess epigenetic change influence at genome level for abnormal or disease conditions. Evolution of several independent gene blocks, based on molecular interactions across the human chromosome 21, is analysed and impact discussed. Parallel implementation of the micro-model, (using MPI) to handle the large data load, is also described.

1 Introduction

Epigenetics is a relatively recent study, initially stimulated by work on the Human Genome and expanded through subsequent efforts [1] to quantify molecular events. The epigenetic changes in the genome are characterized by chemical
alterations to molecules which include DNA methylation and Histone protein modifications. Within each human genome, the control over gene expression is achieved through a well-established co-ordination between these two elements; (Epigenetic Layer = Histone Modification $\leftrightarrow$ DNA Methylation) [17]. Efforts to discover the significance and effect of specific modifications, especially with respect to onset of cancer, has led to a pool of largely unconnected epigenetic information [1]. This has initiated attempts to trace underlying networks and model the events in order to understand the complex interactions that lead to normal and abnormal gene expression. Here we present improvements on a prototype model, developed, [12] which can stochastically simulate significant histone modifications, DNA methylation, quantify their interdependencies and hence define gene expression (result of the ongoing complex interactions). The microlevel model, which simulates molecular events, has been tested for human chromosome 21, in order to help comprehend the flow of information across multiple scales of genomic organisation. Parallelization strategies have been utilised to split the large chromosome 21 dataset and perform appropriate simulations.

2 Background

The Genome of a biological cell is the central processing unit which controls and co-ordinates the cellular activities within the organism. With the aid of chemical switches or changes to DNA $^1$ and Histone proteins, the genome is able to control the location, level and amount of gene expression. Early work, mentions that phenotype or visible changes in an organism is influenced by three main factors [19]: the surrounding environment, genotype (genetic content) and a phenomenon, known as epigenetics, in which gene contents can be altered without changing the DNA sequence; the influence on phenotype can be expressed as $P=G*E*EpiG$, where the symbol $*$ implies both simple additive effects and interdependent terms. A classical example used to explain the differential working of the same genome is that of identical twins, while originating from the same zygote or stem cell, individuals express different phenotypic traits or physical appearances [1]. These are fundamentally attributed to the control exercised by Histone modifications (HM) and DNA methylation (DM) inside each living cell.

DM refers to the modification of DNA by addition of a methyl group to the cytosine base (C) and is a well conserved epigenetic change, introduced and maintained [18, 13] by a family of enzymes called DNA Methyl Transferases (DNMT) [4]. DNA methylation is mostly associated with gene suppression and more than 90% of methylation in the human genome occurs in CG dinucleotides$^2$, located in repeat regions (non coding), in CpG islands (long regions with high content of GC and CG dinucleotides [15]) and in genes. The level of DM spread across the sequences also vary based on the sequence pattern contained within.

---

$^1$ basic molecular unit of Genomes made of 4 types of bases - A,T,G,C.

$^2$ a di-nucleotide is made of 2 base pairs, in this case C is followed by G
Literature states that inherent CG patterns are differentially methylated by the DNMT enzymes with a spectrum of DM profile observed in the human genome [11].

Histones are proteins that protect DNA from restriction enzymes and also act as bolsters in chromosome condensation [7]. Nine histone proteins forming a “Histone Core”, are attached to 146-148 bp of DNA molecules. The histone core contains two sets of four types (H2A and H2B) and (H3 and H4) of histones that pair with one another respectively and a 5th type, H1 that binds DNA to the histone for chromosome condensation [17]. A combination of modifications (such as acetylation, methylation, phosphorylation, ubiquitination and sumoylation carried out by many types of enzymes), within specific amino acids in each histone type leads to gene expression or inactivation [10]. It has also been found recently, that these histone modification patterns, unlike the stable DNA methylation, are dynamic in nature and are completely recreated during events like DNA replication, transcription 3 and cell division [1].

Knowledge on which patterns are set during gene expression, and the way in which these stimulate transcription by activating or deactivating certain elements, is not precise to date. However, work has been reported on the significance of individual histone modifications [8, 3]. The literature also suggests that activation of one change leads to successive modifications of other amino acids [9]. Even though new findings with regard to the impact of several histone modifications have been reported, inconsistency of precise information with regard to how a network of histone modifications communicate within themselves and are orchestrated during a specific event, is a major challenge. Despite this insufficiency, we do know for sure that the interactions between histones and DNA methylation are disrupted at some stage, during the onset of cancer. Identification of specific factors has contributed to these circumstances but the picture is incomplete. Hence, a computational model that can simulate the networks and mimic the working of epigenetic mechanisms should improve understanding of how such events can lead to malignancy. The focus here is testing and refinement of the model for a large dataset of all CG islands in chromosome 21, in order to assess epigenetic change influence at genome level for abnormal or disease conditions. This bottom-up approach facilitates low-level information processing between histone types in nucleosomes within gene blocks, enables to understand how several genes communicate and finally aid to comprehend how gene expression controls the phenotype or physical appearance of an organism at higher level.

3 Transcription is the process of copying information stored in a DNA strand into a complementary strand of RNA (messenger molecule or mRNA) with the aid of biological catalysts.
3 Methods

The two main (DM and HM) model components are key to gene expression and the framework here is designed to address and simulate the natural behaviour of these constituent elements. The whole model connecting these two components has been split into two implicit modules - EP1 and EP2 (epigenetic modules 1 and 2). EP1, the main component explains modeling histone modifications, which are concisely represented as active nodes, and randomly sampled/visited based on associated probabilities over a time line. This sampling, which is influenced by DM levels (simple values of range $\in [0,1]$) accounts for the stochastic modeling of all dynamic histone modifications. Dynamic graphs/tables are utilized to store levels of modification, as nodes are sampled and hence used to express output state of the model. Details of this component has been extensively discussed in [12] and further developments are shown in Figure 3. In EP2, human DNA sequences were analysed using time series methods such as Fourier and Wavelet Transformations to check for inherited and stable sequence patterns, containing CG dinucleotides that decide methylation levels. Elaboration on the methods and datasets used to analyze the pattern sequences are described in detail in [11]. Here chromosome 21 is categorised into ‘groups’ with different/distinctive DNA pattern sequences, associated with DM variation. Simulations are performed to ascertain preferred characteristic histone modifications for these distinct regions.

3.1 Basis of EP1 - Epigenetic Module 1

Epigenetic Module–1 here models the choice of multiple histone modifications preferred at a time step in the model using random selection. Here, main algorithm is based on Stochastic modeling as mentioned above, which is a powerful technique of presenting data or predicting outcomes that takes into account a certain degree of randomness, or unpredictability. This type of modelling, used as a prediction tool differs from a deterministic approach, (such as that of top-down, differential equation system specification), in that more than one solution, based on known attributes or values, may be considered. This component is based on simplified abstraction and incorporation of individual epigenetic events and their inter-dependencies to predict the behaviour of histone evolution under stipulated and controlled conditions. The novelty of our approach lies in encoding possible modifications together in each histone type as numeric strings [12]. These encodings which refer to the selected histone state, in terms of modifications, is based on probabilities of transition and DM values. The histone objects contain modification graphs/tables that are constantly updated with dynamic changes redefining the objects current state after each time step or random state selection. Building on earlier versions, [12], the DM values here are associated with the distribution of specific CG patterns in the DNA sequence (Figure1).
3.2 Basis of EP2 - Epigenetic Module 2

In case of Epigenetic Module–2, the DNA sequences were analyzed for CG dinucleotides patterns to correlate it with the spread of DM. Literature [14] mentions that DNA methylation is highly influenced by DNA patterns in the genome and environmental factors. In particular, knowledge of the distribution and location of CG dinucleotides (at CG islands, genes and non-coding regions), were sought, to understand the biological significance associated with them with regard to determining level of DM (based on distributions). Figure 2 explains (EP2 component) in detail about the results and conclusions on spread of CG dinucleotide in different sequence regions. Unique CG patterns are associated with specific persistence of DM across the genome [5, 11]. Since literature and results from our sequence analyses reveal the importance of CG spread and establishment of associated HM attachment, the sequences in chromosome 21 were grouped based on their sequences contents and hence checked for histone protein modifications (details of categorization are explained in detail below).

The mutual interdependencies in the epigenetic biology (as in Fig. 1) are illustrated in Fig. 3.

The stochastic model is thus designed to explore how the dynamics associ-
ated with histone modifications can control DNA methylation, as well as utilising additional sequence information to that end. Real system analogies may be found in cancer-associated conditions, (such as over expression of cancer genes or silencing of tumour suppressor genes). Hence the initial and simple model accommodating Histone Modifications, along with information on DNA sequences is refined to closely resemble the real system and and can be used to study cancer associated conditions (such as over expression of cancer genes or silencing of tumor-suppresor genes). Our initial attempt is based on mimicking the biological epigenetic structure as illustrated in Figure 1 – a simplified construction of the model. The status of epigenetic profile in the model is defined by DNA Methylation and Histone Modifications and by their ongoing mutual interactions (Figure 3) during model execution.

3.3 Epigenetic Interdependency

A simple yet strong and well defined inter-dependency between histone evolution, transcription rate and level of DNA methylation inside each computational Block of object. There are 3 main interactions in our model, that take place between the 3 elements – Histone Modifications (controlled by probabilities of transition), DNA Methylation values and Transcription. The mathematical relation between these elements is well described in Figure 3. The model utilizes probabilistic cost function to set the interdependencies between variables (HM and DM based patterns), random function to populate possible solutions and calculate the final output or rate of transcription (T) using exponential equations \( T = e^x \times e^y \times k \), “x” and “y” being variables and “k” a constant. Ultimately, along a given time line, our micromodel can predict or forecast a possible network of molecular events in a simple and strong manner that occur during specific cellular events such as gene expression and suppression. This basic model is tested for several hundred blocks of CG islands categorized based on their sequence
patterns. i.e. Each group of islands contain a specific CG pattern or distributions. For example, group 3 contains all island blocks with CG dinucleotides maximum separated by 3 nucleotides among other distances of separation. Usually 3 nucleotides/base-pair spacing between CG is observed in coding sequences while 2/4/8/bp spacing is most associated with non coding regions.

MPI is one of the most famous and powerful interfaces designed to initiate and maintain communications between processors. This API has been utilized here to split the model simulation for each group of blocks (with common sequence patterns) per cluster node. The details on utilization of parallel strategies is described as follows.

4 Parallelization Strategies

Parallel computing is a form of simultaneous computation that is used to handle large problems (in terms of high complexity/data or both) using more than one processor [6]. Parallelization, [6] and strategies for its implementation, become more complicated as the number of communications between processors increases, leading to large amounts of data transfer. Here, the data-driven nature of the model makes parallelization a natural choice, with data load distributed across processors. The HPC platform, (Dual Intel Xeon Quad core of...
ram 2.66GHz) was used to run simulations for each histone object. One node with 8 processors is used for each group of CG islands, and data is split per histone type (H2A/H2B/H3/H4 types) which is assigned to each processor. Literature states that each nucleosome consist of a pair of 4 types of histones and several nucleosomes put together form a block of gene sequence (146 bp of DNA sequence wraps around a nucleosome). Hence the number of objects are dynamically created based on sequences within Chromosome 21. Inside each node, efficient distribution and optimized load balance is carried out to minimize computational time for simulation. Data transfer occurs at each time step in the model implementation, [12], where a specific processor in each node computes output at each time step.

The algorithm steps and data division for load balance obey the following:

1. **Read and Store Inputs**
   (a) Histone Data - The possible combinations of Histone modification as described above are read and stored in the model. These include string of histone states and the probabilities of shift between the states [12].
   (b) DNA sequences are read and stored along with information on CG positions throughout sequences (CG position in CG Islands and those that follow a CG distribution).
   (c) User Selected Values are provided – DNA methylation per a Block in a specific time-step, total number of iterations(or time-steps).

2. **Create Objects**
   (a) Based on the size of each Island and CG patterns, as many blocks of CG islands – Nucleosomes and histone objects in each are created. The blocks are then categorized based on their CG spacing.

3. **Simulate**
   (a) For each iteration, evolution of each histone type is simulated and the new state information is sent to the master processor. The master node receives the updated modification tables and recomputes output parameters, such as new DM and transcription, while probabilities of shift are amended, based on the previous time step DNA value (ref. Figure 3 for interdependencies). The DM value, (as noted previously), is associated with level of methylation for CG islands. The load balancing of epigenetic objects and simulation of specific histone types is achieved, as described by the algorithm extract below.

```c
/* ********************** */
For each Group of Islands
QUARD=4; //value set to no. of histone types.
HALFQUARD=2; //node has 8 proc hence "2" here.
for 1:max_iteration (time-step)
{...
```

for i=1:no_of_Blocks
// (inside each nucleosome)
if (( rank_of_processor / QUARD ) == (i% QUARDSET ))
for j=2:9 (Histone type numbering in System)
if ( ( rank_of_processor % QUARD )+1 ==j/2)
SIMULATION_OF_HISTONE_TYPE
if (! Master Proc ) MPI_Send all mod_tables to Master Proc.
else MPI_Receive Mod_tables;
Calculate Output parameters;
// the modifications are sent & received once
as a compressed matrix,
to avoid communication overhead.
/*********************/

(b) Continue simulation and communications till maximum no. of iterations are reached. (10,000 time-steps here)

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 the number of times each state is visited in all 8 histones
   for each nucleosome.

In order to investigate the system behaviour, all contiguous sequences in chromosome 21 were combined and analysed (total sequence length being 35,660,412 bp). Chromosome 21 has 464 CG islands, present in all contig sequences, which were extracted based on criteria [15] using a perl script before the micromodel was executed. These islands were segregated into 30 groups based on the CG patterns observed in each island. As mentioned in earlier, the number of nucleosomes is decided based on the length of an island sequence (number = length of sequence/148 bp) and histone evolution was observed for 10,000 iterations. Outputs namely, Histone states, in every nucleosome for each CG group were reported. Considerable evidence from literature show that histone types H3 and H4 play a direct role in cancer onset [1] and hence follow a cascade of occurrences during transcription. Also, CG islands with 3bp spacing are usually differentially expressed based on the cell type or disease conditions [5].

Hence the H4 histone evolution for group 3 alone is reported and discussed in detail.

5 Results

The simulation was carried for each group of CG islands separately, group 3 being reported here, and the simulation time on an average for each group, took 2.5 computational hours. (As individually, time varied for each group based on the number of objects created and DNA sequences). The DNA methylation was set to a range of values between ∈[0.1,0.9] and detail of this analyses can be found in [12]. For initial values less than 0.2 of DM, the system was consistent in maintaining the same values and hence this encouraged more acetylation and
lesser methylation modifications in Histones. Another range of values greater that 0.7 and less that 1.0 were consistent and forced system to remain in a fully methylated state. This allowed an inverse state of the system such as visiting Histone states with more methylation modification and less acetylation. Hence to observe histone evolution we discuss in detail two sets of results observed under Low DM (<0.15) and High DM (>0.85). These simulations demonstrate effective emulation of the biological process of transcription of genes (e.g. onco-genes) for low DNA methylation levels and reverse case of high DNA methylation and gene suppression (e.g. silencing of tumor suppressor/control genes).

Figure 4: A Comparison between the average preferences of H4 states for high and low DNA Methylation Levels for all CG Island in Group 3 (where maximum CG spacing for these islands is 3 bp)

Figure 4 shows the preferences of H4 states for high and low DNA Methylation levels. Under low DM levels (initially set by user), acetylated amino acids states, such as the 11th, 35th and 47th predominated i.e. states containing acetylated amino acids such as K5, K8 and K12 were highly visited (refer [12] for histone states encoding). Irrespective of the probability assigned to the three preferred states the system preferred these over several time steps. Such consistent results demonstrate the ability of our model to reproduce the presence of these modifications during transcription, (as reported [20, 16] in particular, during expression of oncogenes). For higher levels of DNA methylation (>0.85, Figure 4) during the simulation, the preference is more towards choosing methylated histone states leading to reduced transcription rate. During this high methyla-
tion 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. Interestingly, there is also an appearance of serine phosphorylation (states 39 and 35 in Figure 4 that contain phosphorylation of serine) during both conditions 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 [2] and aid in structural condensation. Details on the evolution of H3 histone type can be found in [12]. Such modifications for defined DM levels, can plausibly be argued to either initiate or aid in establishing an epigenetic control picture for gene expression. The simulations readily scale up and can be used to test massive amounts of data, as in this case study, so that molecular genetic events and their role in complex epigenetic control of the system as a whole can be explored.

6 Conclusion & Future Work

Our Epigen micromodel, iterates between possible histone modifications that can be associated with definite DM range. DM fluctuations over specific time intervals are associated with distinct CG dinucleotides in the DNA sequence. Local or intra-node parallelisation can be applied to a set of genes, involved in common disease, where the methylation profile and histone modifications can be monitored at chromosome level. We report on the effect of DM level associated with CG island groups (specifically 3bp-spaced nucleotides). The stochastic model can also be extended to investigate other CG groups. As DNA methylation controls the direction of histone evolution, high and low DM levels lead to different preferred levels of histone state selection. Nevertheless, system tolerance to initial selection of histone states is good, with specific 'marker' states consistently chosen over execution. This consistency in predicting characteristic histone modifications under defined DNA methylation levels, emphasizes the model's capability to mimic the real system accurately and provides an effective tool for investigating epigenetic events.

A general model, which can simulate conditions of any epigenetically-associated disease, is clearly the ideal target. The current model is robust and sufficiently powerful to predict recognised molecular events for high and low DM , (corresponding e.g. to suppression of highly-expressed genes and aberrant expression e.g. of oncogenes). 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 and especially larger datasets. The novelty of our approach lies in modeling (for all histone types simultaneously) based on assimilated information such as increasing acetylation modifications during gene expression and vice versa. Our model also has the advantage of a readily-expandable structure,
which can be further developed to accommodate additional modifications and multiple sequence patterns for large datasets.

Ultimately, the aim of the predictive model must be association of the molecular events and epigeentic mechanisms with aberrant effects at the macro-scale, such as onset of disease. Currently, the focus is to expand the model structure to simultaneously accommodate all CG island groups, categorized by sequence pattern and to allow for multiple-gene groups, associated with known disease conditions. This expansion will make our model most efficient in terms of being more realistic as reported in literature [1]. In the long-term, the aim is to establish a comprehensive model framework to encompass epigenetic mechanisms, together with DNA sequence pattern analysis. This, with a view to determining their influence on gene groups and, ultimately, on alterations to the phenotype of an organism, implicated in disease.

7 Acknowledgment

The model has been developed using C++ language and MPI library. 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.

Bibliography


[2] Barber, C. M., F. B. Turner, Y. Wang, K. Hagstrom, S. D. Taervera,
S. Mollah, B. Ueberheide, B. J. Meyer, D. F. Hunt, P. Cheung,
and C. D. Allis, “The enhancement of histone H4 and H2A serine 1
phosphorylation during mitosis and s-phase is evolutionarily conserved”,

modification: Patterns and paradigms”, Nature Review Genetics 10 (2009),
295–304.

Rabiena, A. Behn-Krappaa, and G. Orenda, “Eukaryotic DNA methy-

[5] Glass, J. L., M. L. Fazzari, A. C. Ferguson-Smith, and J. M. Gre-
ally, “CG di-nucleotide periodicities recognized by the dnm-3a-dnm-3l


