Computational Analysis of Epigenetic Information in Human DNA Sequences

Karthika Raghavan*, Heather J Ruskin1 and Dimitri Perrin1
1Centre for Scientific Computing and Complex Systems Modeling. (Sci-Sym)
Dublin City University
Dublin, Ireland
*E-mail: kaghavan@computing.dcu.ie

Abstract— Over the last few years, investigations of human epigenetic profiles have identified key elements of change to be Histone Modifications, stable and heritable DNA methylation and Chromatin remodeling. These factors determine gene expression levels and characterise conditions leading to disease. In order to extract information embedded in long DNA sequences, data mining and pattern recognition tools are widely used, but efforts have been limited to date with respect to analyzing epigenetic changes, and their role as catalysts in disease onset. Useful insight, however, can be gained by investigation of associated dinucleotide distributions. The focus of this paper is to explore specific dinucleotides frequencies across defined regions within the human genome, and to identify new patterns between epigenetic mechanisms and DNA content. Signal processing methods, including Fourier and Wavelet Transformations, are employed and principal results are reported.

Keywords: Epigenetics, DNA sequences, Signal Processing Analysis.

I. INTRODUCTION

Epigenetics is a relatively recent study, initially stimulated by work on the Human Genome and expanded through subsequent efforts, [1]. This research has highlighted key cellular influences on gene expression levels, such as dynamic histone modifications, stable DNA methylation and chromatin remodeling. Computational epigenetic micro-models that mimic Histone modifications under standard DNA methylation levels have been recently developed, but incorporating useful information based on sequence patterns may provide important insight on epigenetic mechanisms [2]. In particular, DNA methylation, (methylation of a cytosine base), is responsible for many cellular processes [3] and is highly influenced by DNA patterns in the genome as well as environmental factors. The methylating enzymes (DNA Methyl Transferases - DNMT) control location, and level of DNA methylation for, all types of cells (i.e. DNMT3a and DNMT3b – de novo, DNMT1 –maintenance methylation) [4]. Specific sequence impact is still being investigated, but recent evidence [5] has shown that distinct distribution patterns of the CG di-nucleotide within regions of the human genome facilitate accessibility for these enzymes. More than 90% of methylation in the human genome occurs in CG dinucleotides, located in repeat regions (non coding) and CpG islands (long regions with high content of GC and CG dinucleotides) [6], associated with silenced genes. In particular, knowledge of the distribution and location of CG dinucleotides is sought, together with the implications for DNA methylation. In order to link these patterns with epigenetic events, we investigate both global periodicities of CG in characteristic regions of chromosome 21 using Fourier analysis (FT) and locations of these frequencies within sequences, using Wavelet Transformations (WT).

II. METHODS

The first method involves analysis of the frequency domain for CG dinucleotide from DNA sequences using Fourier analysis. For this, the auto correlation of CG distribution is calculated, so that background noise is removed (or e.g. contribution of other nucleotides in the sequence) and frequencies of the desired components are highlighted in the power (Fourier) spectrum. Positional Autocorrelation for each lag/interval ‘k’ and the power spectrum of autocorrelation, that provide information on global periodicities, were calculated as described, [7]. DNA sequences obtained using Map viewer, NCBI database (www.ncbi.nlm.nih.gov) and UCSC genome browser (http://genome.ucsc.edu) were classified into 3 sets (20 Genes, non-coding regions near the genes and all CpG islands in chromosome 21), for Fourier analysis (details in Table I).

Maximal Overlap Discrete Wavelet Transformation (MODWT) was adapted to check for the location of CG distribution along a given input sequence. This transformation indicates loci of high and low frequencies of CG, not detected by Fourier transformation [8]. MODWT
consists of the application of linear filters that transform a given input into coefficients related to variations in the input signal over a set of scales (using non orthogonal basis vectors in this case). Similar to Discrete Wavelets, multi-scaling is done with the help of mother and father wavelets, (Debaucies in our case). The wavelet representation of a discrete signal \( f(t) \) in \( L^2(\mathbb{R}) \) or Hilbert space \([9]\) is given by:

\[
f(t) = \sum_{k} s_{J,k} \phi_{J,k}(t) + \sum_{k} d_{J,k} \phi_{J,k}(t) + \ldots + \sum_{k} d_{1,k} \phi_{1,k}(t)
\]

for, \( 1 < k < L \), with \( L \), the length of filter used \((L = 4 \text{ here})\). The coefficients \( s_{J,k} \) and \( d_{J,k} \) are the smooth and detail components respectively \([8]\). MODWT, allows the input data to have any length, and also retains the downscaled (high frequency component) values for each level of decomposition \( J \), in the above equation. For MODWT analysis, DNA strings were converted to numerical sequences as described in literature, \([10]\). (E.g. \text{CGAATCG} can be represented as \text{1000010}, where 1 is recorded only for a position where C is followed by G; 0 otherwise). Raw and assembled contiguous DNA sequences in chromosome 21 were utilized for this analysis using map viewer in NCBI database. (Table I)

**TABLE I. DETAILS ON THE TYPE OF SEQUENCE DATA USED FOR EACH TIME SERIES METHOD.**

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Method</th>
<th>Input Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Fourier Transformation</td>
<td>Disease associated Genes &amp; the Introns located near the genes - PRSS7, IFNGR2, KCNE1, MRAP, IFNAR2, SOD1, CRFB4, KCNE2, ITGB2, CBS, FTCD, PFKL, RUNX1, COL6A1, COL6A2, PCNT2, CSTB LIPI, TMPRSS3, APP</td>
</tr>
<tr>
<td>2.</td>
<td>Maximal Discrete Overlap Wavelet Transformation.</td>
<td>Contiguous Sequences with Accession No. - NT_113952.1, NT_113954.1, NT_113958.2, NT_113953.1, NT_113955.2 &amp; NT _029490.4</td>
</tr>
</tbody>
</table>

**III. RESULTS AND DISCUSSION**

A. **Fourier Analysis**

The Fourier analysis calculated from auto correlation of a DNA segment, (of any length containing CG with lag or spacing “k” between occurrences), serves to detect the characteristic global frequencies for each of the three regions. The results are shown in Figure 2 and Table II.

**TABLE II. DISTRIBUTION OF CG IN DEFINED REGIONS OF CHROMOSOME 21, HUMAN GENOME.**

<table>
<thead>
<tr>
<th>Sequence No.</th>
<th>Regions</th>
<th>Characteristic Periodicities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Genes</td>
<td>3</td>
</tr>
<tr>
<td>2.</td>
<td>CpG Islands</td>
<td>3,7</td>
</tr>
<tr>
<td>3.</td>
<td>Non-coding</td>
<td>4.8, 10-11, 24.25, 26</td>
</tr>
</tbody>
</table>

Here, separate sequences were extracted containing 20 genes (involved in diseases), Introns close by and all CpG islands in chromosome 21, with the help of online tools, (UCSC genome browser and NCBI map viewer). Figure 2 represents amplitudes of the power spectrum for all values of CG periodicities possible: (periodicity here refers to appearance of CG after a lag – or interval- of “n” bp). Gene coding regions show an apparent peak at 3-bp, which might be expected due to the codon bias, \([7]\). (In FT, periodicity and frequency, where the former is the inverse of the latter, are commonly used to describe signal components).

CpG islands, (throughout chromosome 21), also contribute to the peak at a periodicity of 3-bp since these are present near the promoter regions, (w.r.t. coding theory triplets and translation to amino-acids). They also have 7-bp spacing due to repeats containing CG, that meet the conditions: length of DNA sequence >200 bp, \((G+C)\) content > 50%, observed /expected ratio of CG > 60%, as reported, \([6]\).
a CG island, over and above percentage CG content in any input sequence. One of the more prominent and interesting features, however is observed in the non-coding regions, which display patterns (between 24 and 26-bp). Research has indicated that 8-bp intervals, (and also 4-bp, which corresponds to satellite repeats, [5]), between CG dinucleotides, is a pattern that is, most amenable to the attraction of DNA methylation complexes. In fact, genes silenced in germ cells by the de novo methylation mechanism, have 8-bp spaced CG near their promoters, which is preferentially recognized by the methylating enzymes complexes of DNMT3a, [5]. Another peak, observed in Figure 2, between 10 to 11-bp periodicity may support genomic structural condensation [5]. Other peaks, at periodicity of 15 and 20-bp, are less persistent and disappear when average wave amplitude versus periodicity over all 22 chromosomes for the 3 regions was plotted, with those at 8-bp and 24 to 26-bp only being consistent throughout. The hitherto unreported 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. One major clue, which has been noted previously, [12], is the presence of several million repetitive 25-mers in the human genome. Although not uniform throughout, this is known to be high, on average, in chromosome 21. Further, in [13], the authors explain that piRNA or Piwi protein associated iRNA, which is significantly involved in cellular processes and propagation of de novo DNA methylation is usually of length 24 to 26 nucleotides. This piRNA, (bound with small Interference RNA, i.e. components of RNA-induced silencing complex), is involved in silencing the retrotransposon regions in specific cells, through RNA interference mechanisms [13]. It appears that this specific periodicity of the CG dinucleotide as a human genome marker could help in understanding the role that iRNA associated DNA Methylation plays in gene expression. These interesting facts are only the tip of an iceberg in the sequence analysis of humans, especially with respect to differential gene expression, as controlled by epigenetics. The average wave amplitude over the 3 regions for each periodicity, (shown as the dotted line in Figure 2) also shows peaks at frequency 4, 8 and 24 to 26-bp. This is due to the fact that non-coding regions form more than 50% of Human genome, (with Genes and CG Islands occurring in between) and hence the contribution of Intron regions to periodicities and nucleotide periodicities is relatively high of wave amplitude (contributing a large component to wave amplitude).

B. Wavelet Analysis

A summary of MODWT results is shown in Fig. 3

Figure 3. High frequency components from MODWT analysis (Distance in bp Vs MODWT coefficients) of Contig-1, (Accession No: NT_113952.1) at scale 11. Each vertical coloured bar marks specific distributions indicated by Fourier analysis in Introns (8,24,25 and 26 bp), genes and CG Islands.

Contig sequences (i.e. a set of overlapping DNA segments derived from a single genetic source by DNA sequencing techniques), in chromosome 21 were analyzed to help locate and interpret important regions that contained CG dinucleotide. We investigated each of the contig sequences (1,2 and 3 being reported here), for specific CG patterns using all values of scale allowed. (Maximum scale allowed is the logarithmic value of input length). We focus on high frequency components for specific scales, denoted by J in equation (1), whose range of values are from 1 to 16. (The maximum scale allowed for all 3 contigs was 16 due to length of input sequence). In general, a scale range of 9 to 12 was found to be suitable to identify regions of interest, (genes, islands and introns) and to extract information on CG content. The most suitable scale however, appeared to be 11, which had a smoothed representation of the coefficients and reduced noise when MODWT results and Fourier analysis were compared. (Elaborated in next subsection)

Figures 3, 4 and 5 show the MODWT coefficients, (blue line) and colored bar strips indicate presence of specific periodicities of possible interest along the sequences.

Figure 4. Representation of high frequency components from MODWT analysis, (Distance in bp Vs MODWT coefficients) of Contig-2 (Accession No: NT_113954.1) at scale 11. Each vertical coloured bar marks specific distributions indicated by Fourier analysis in Introns (8,24,25 and 26 bp), genes and CG Islands.
Several of the peaks in Figure 3, 4 and 5, (corresponding to MODWT coefficient > (+0.03) - values obtained as a product of basis vectors and input signal/DNA sequences) contain CG spacing patterns, which reinforce those found by Fourier analysis: (these include the 8, 24 to 26-bp features, here indicated by the coloured bar strips). The peaks also reflect regions, (as indicated in the three figures) of genes and CpG islands. Information for each contig sequence gives location of putative Islands and genes present in these, (taken from the NBCI database). This emphasizes, as discussed in the previous subsection, that the long non-coding or retrotransposon regions, (with possible CG spacing at 24 to 26 bp), span gene and CG Island regions which are common features in the complex human genome, (dark/light blue and green bars as in Figure 3, 4 and 5).

Such high peaks, relating to contributions from different component regions, (genes, islands and non coding regions), in a given DNA sequence, help to characterize features of the genome. Hence the coefficients (> +0.03) indicate the presence (and overlap) of multiple CG patterns, associated strongly with the intron regions, (which predominate in the genome), but also reinforced by several genes and by island concentrations. In general however these specific distributions are indicated by peaks i.e. Wavelet coefficients greater than (+0.005). This suggests that wavelet analysis is extremely sensitive to CG distribution change for a given contig sequence although elements of that contribution are not explicitly indicated.

The exploratory analysis, using the wavelet transformation, highlights regions of high CG and its distribution across DNA sequence loci. In particular, such patterns are seen to occur between loci, 20k-30k and 130k-140k in Contig 1 (Fig. 3), between 55k-65k loci in Contig 2, (Fig. 4) and 0-5k, 50k-60k, 80k-90k in Contig 3, (Fig. 5). Methods to analyze significance of wavelet coefficients computed for multi-component (and overlapping) data have been reported before for other applications, [8,14]. These include calculation of threshold (based in median absolute deviation or MAD) for wavelet co efficient and variance analysis, which is being applied for current work, i.e. CG distributions that form peaks, to try to distinguish contributions of specific frequencies of genes, Islands and Introns. Similar analyses can be applied to other dinucleotides, such as AA/TT/AT, which are known to be associated with chromatin remodeling [7,15].

IV. CONCLUSION

Two signal decomposition approaches have been applied to DNA sequences, with a view to investigating patterns of CG dinucleotide in relation to epigenetic mechanisms. Fourier analysis highlighted characteristic frequencies, present in CpG islands and in non-coding regions, but provided no location information. Wavelet analysis gave further detail on frequencies, but also on loci along the DNA sequence at which major effects occurred. Periodicities, (content repeats in bp) were related to known biological features in some instances, (notably bp spacing) corresponding to codons, to de novo methylation and others, but also to other unconfirmed potential methylation mechanisms at 24-26bp spacing. Concentration of CG dinucleotide patterns at specific locations was also reported, although discriminating between contributions from gene, intron and non-coding regions was not possible here. Improved understanding of how pattern components, and their sequence locations affect methylation may help us understand how epigenetic changes are stimulated. Further, in providing input to models of histone modification, reported in prototype, [2]), such analyses may help to describe interdependencies of epigenetic events in terms of differential gene expression and initiation of disease.

Wavelet analysis was implemented in Matlab, using MODWT package, released by University of Washington. (http://www.atmos.washington.edu/~wmtsa/).

ACKNOWLEDGMENT

We gratefully acknowledge financial support from Science Foundation Ireland, project 07/RFP/CMSF724, in the early stages of this work and, subsequently, of an EC Complexity-Net /IRCSET pilot award. We thank ICHEC, (Irish High End Computing Centre) for providing access to major computational facilities, required for background work. We would also like to thank Dr. Jose-Abilio Matos, University of Porto, Portugal for his valuable suggestions during our collaborative discussions.

REFERENCES


