CLASSIFICATION OF CPG ISLAND AND PROMOTER REGIONS USING RARE K-MER MOTIFS

by

EZZEDDIN KAMIL BIN MOHAMED HASHIM

**Thesis submitted in fulfilment of the
requirements for the Degree
of Doctor of Philosophy**

July 2015

ACKNOWLEDGEMENT

“In the Name of Allah, the Most Gracious, the Most Merciful. Praise be to Allah, the Lord and the Sustainer of the worlds. The Most Gracious, the Most Merciful”. (Al-Fatihah: 1-3)

First and foremost, I would like to express my deepest gratitude and appreciation to both of my supervisors, i.e. Professor Rosni Binti Abdullah and Encik Karim Hercus, for their priceless guidance, invaluable encouragement, and relentless support in managing almost everything related to my study. My sincerest thank to all of my research collaborators i.e. Dr. Arif, Dr. Stephen, and Dr. Laurence for helping me with the planning and analyses of my research.

To my dearest family members, my beloved father Mohamed Hashim, my beloved mother Sabariah, my beloved teacher Ustaz Aswadi, my beloved wife Siti Faridah, and my beloved kids Lutfi and Aina’. My PhD. journey is impossible without your love, prayers, sacrifice, and supports which I am truly indebted forever to all of you.

Not forgotten to mention, those who are directly and indirectly assisted me in my study which are my fellow colleagues at the Parallel and Distributed Computing Centre (PDCC) i.e. Hadri, Najihah, Adilah, Nadzrin, Mubarak, Atheer, Muhannad, and Ahmed; and my fellow colleagues at the MGRC/Synamatix offices i.e. Poh, Zayed, Maarif, Ali, Tay, Sui Meng, Su, and others.

Last but not least, I would like to extend my gratitude to Universiti Sains Malaysia (USM) and Malaysian Genomic Research Centre (MGRC) for funding my study for all these years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3’ UTR	3’ Un-Translated Region.
5’ UTR	5’ Un-Translated Region.
A, C, G, T, U	Adenine, Cytosine, Guanine, Thymine, and Uracil.
AC	Accuracy (for the evaluation).
AW	All derivations of a *k*-mer (of *k*=8-, 9-, or 10-mers).
ANN	Artificial Neural Network.
BRE	B (TFIIB) Recognition Element (binding site).
CAGE	Cap Analysis of Gene Expression.
CAP	CxxC Affinity Purification.
CC	Correlation Coefficient (for the evaluation).
CGI	CpG Island.
ChIP-seq	Chromatin Immuno-Precipitation Sequencing.
CPE	Core Promoter Element (binding site).
DBTSS	DataBase of Transcription Start Site.
DHS	DNaseI Hypersensitive Site.
DNA	Deoxy-ribo-Nucleic Acid.
DP	Dynamic Programming.
DPE	Downstream Promoter Element (binding site).
EPD	Eukaryotic Promoter Database.
EST	Expressed Sequence Tag (parts of cDNA).
FN	False Negative (for the evaluation).
FP	False Positive (for the evaluation).
GMM	Gaussian Mixture Model
GSP	Gene Start Position.
GSR	Gene Start Region (overlapping of GSPs).
GTF	General Transcription Factor (binding site).
H3K4me	Histone 3, K(lysine) 4, methylation.
HGP	Human Genome Project.
HMM	Hidden Markov Model.
HMP	Hypo-Methylated Promoter.
INR	INitiatoR (binding site).
IUPAC	International Union of Pure and Applied Chemistry.
KDD	Knowledge Discovery in Databases.
KL	Kullback-Lieber
LDF	Linear Discriminant Function.
LRS	Log Ratio Score.

MeDIP	Methylated DNA Immuno-Precipitation.
mRNA	Messenger RNA.
MSA	Multiple Sequence Alignment.
mtDNA	Mitochondrial DNA.
MTE	Mobile Ten Element (binding site).
NBC	Naïve Bayes Classifier
NCBI	National Centre for Biotechnology Information.
ncRNA	Non-Coding RNA.
NN	Neural Network.
O8MD3	Other 8-Mers from Distribution 3 (of the human spectrum).
PCA	Principle Component Analysis.
PCR	Polymerase Chain Reaction.
PhastCons	Phylogenetic Conserved elements.
PIC	Pre-Initiation Complex (of transcription machinery).
PPV	Positive Predictive Value (for the evaluation).
PSO	Particle Swarm Optimization
PSWM	Position Specific Weight Matrix or PWM.
PSSM	Position Specific Scoring Matrix.
PWM	Position Weight Matrix, interchangeable with PSWM.
QDF	Quadratic Discriminant Function.
R8MD1	Rare 8-Mers from Distribution 1 (of the human spectrum).
R8MD2	Rare 8-Mers from Distribution 2 (of the human spectrum).
Repbase	REPeat dataBASE.
RKM	Rare *k*-mer.
RM(1-3)	Research Methodology Number 1, 2, or 3.
RMC	Rare Pattern Cluster (overlapped of 8-, 9-, and 10-mers).
RNA	Ribo-Nucleic Acid.
RVM	Relevance Vector Machine
RW	Rare-Word, in general, particularly rare 8-, 9-, or 10-mers.
RWC	Rare-Word Clustering.
RWCD	Rare-Word Central Distribution.
RWML	Rare-Word Maximum Likelihood.
SAGE	Serial Analysis of Gene Expression.
SCD	Sum of (absolute) Central Deviation.
SN	SeNsitivity (for the evaluation).
SNP	Single Nucleotide Polymorphism.
SP	SPecificity (for the evaluation).
SV	Structural Variation.

SVM	Support Vector Machine.
TAF	TBP-Associated Factor (binding site).
TBP	TATA-Binding Protein.
TC	Tag Cluster.
TF	Transcription Factor.
TFBS	Transcription Factor Binding Site.
TLRS	Total Log Ratio Score.
TN	True Negative (for the evaluation).
TP	True Positive (for the evaluation).
TPC	True Predictive Cost (for the evaluation).
TSR	Transcription Start Region (or region containing TSSs).
TSS	Transcription Start Site.
UCSC	University of California Santa Cruz.
UMR	Un-Methylated Region.
UTR	Un-Translated Region.

PENGKELASAN PULAU CPG DAN KAWASAN PENGGALAK MENGGUNAKAN MOTIF K-MER JARANG

ABSTRAK

Analisis empirikal ke atas *k*-mer DNA telah terbukti berkesan untuk mencari elemen-elemen berfungsi dalam genom manusia. Di antara kajian empirikal tersebut, *k*-mer jarang (RKM) adalah subjek yang sangat menarik untuk dikaji disebabkan ciri jujukan unik yang ada pada mereka. RKM telah dirujuk sebagai *k*-mer (*k*=7 ke 11) DNA berfrekuensi rendah dalam taburan frekuensi *k*-mer genom pelbagai mamalia, tetapi banyak pula variasinya yang terkumpul di spektra rendah *k*-mer. Objektif pertama kami ialah menemui motif-motif RKM dalam genom manusia; mengenalpasti aplikasi-aplikasi pengiraan RKM yang berpotensi dalam biologi; dan mentaabirkan perwakilan aplikasi-aplikasi yang dikenalpasti. Secara ringkasnya, matlamat pertama tersebut dicapai dengan menggunakan strategi perbandingan dan beberapa alatan bioinformatik (iaitu pelayar-pelayar UCSC, EpiGRAPH, dan Galaxy) yang telah mengkolerasikan RKM dengan beberapa unsur genom iaitu pulau-pulau CpG (CGIs), penganjur, rantau-rantau tidak diterjemahkan 5’ (5’ UTR), dan rantau kromatin terbuka; dan dengan menggunakan beberapa pendekatan perlombongan rentetan intrinsik yang telah mengenalpastikan beberapa ciri RKM yang unik iaitu topologi, komposisi, dan gugusan dalam unsur-unsur genom yang telah dikolerasi. Penemuan-penemuan tersebut dirumuskan sebagai sumbangan pertama (berkenaan biologi) dan mereka dianalisa bersama-sama untuk mentaabirkan tiga isyarat perkataan jarang (RW) iaitu gugusan (RWC), kecenderungan maksima (RWML), dan taburan berpusat (RWCD) untuk mewakili CGI dan penganjur (iaitu aplikasi-aplikasi pengiraan yang telah dikenalpasti). Objektif kedua pula ialah pemodelan CGI menggunakan isyarat-isyarat RW yang telah ditaabirkan dan memasukkan mereka ke dalam tiga teknik RW (iaitu RWC, RWML, dan RWCD). Seterusnya, teknik-teknik RW ini dioptimumkan menggunakan algoritma Pengoptimuman Kumpulan Zarah (PSO) yang standard, dilatih menggunakan beberapa kromosom, diuji menggunakan pengesahan silang 5 lipatan, dan akhirnya diuji menggunakan skala genom manusia. Penilaian CGI dibuat menggunakan empat data pengesahan iaitu Illingworth, Weber, elemen-elemen PhastCon, dan ulangan Alu dan protokol penilaian Hackenberg yang sepadan. Objektif ketiga adalah pemodelan penganjur menggunakan prosedur-prosedur yang sama seperti pemodelan CGI. Perbezaannya adalah data-data pengesahannya (oleh Carninci dan RefSeq), protokol-protokol penilaian Abeel yang sepadan, dan ketepatan ramalan yang lebih rendah daripada ramalan-ramalan CGIs. Walaupun isyarat-isyarat RW yang telah ditaabirkan boleh mengenal-pasti majoriti daripada penganjur, pretasi mereka terbantut kerana didominasi oleh isyarat CGI dalam penggalak-penggalak yang telah dikenalpasti dan terdapat isyarat yang tidak diambil kira oleh kami kerana keterbatasan kajian ini. Apabila CGI dan penggalak yang telah diramal ditanda aras terhadap tujuh data-data peramal yang lain (iaitu aturcara CpGCluster, CpG_MI, CpGProD, NCBI-CGI, dan UCSC-CGI), mereka secara konsisten tersenarai di tangga keempat teratas, berdasarkan skor F, liputan jujukan, atau nilai CC.

CLASSIFICATION OF CPG ISLAND AND PROMOTER REGIONS USING RARE K-MER MOTIFS

ABSTRACT

Empirical analysis on DNA *k*-mers is proven to be an effective means to discover functional elements in the human genomes. Among the empirical works, “rare *k*-mer" (RKM) is a very interesting subject to be studied due to their unique sequence properties. RKMs were referred as DNA *k*-mers (of *k*=7 to 11) that have a low frequency in *k*-mer frequency distributions of mammalian genomes, yet there are large variations of their mass at the lower *k*-mer spectra. Our first objective is to discover RKM motifs in the human genome; to identify potential RKM computational applications in biology; and to infer representations for the identified applications. In short, the first goal was achieved by using comparative strategy and several bioinformatic tools (of UCSC browsers, EpiGRAPH, and Galaxy) which correlated RKMs with several genomic features of CpG Islands (CGIs), promoter, 5’ Un-Translated Regions (5’UTRs), and open chromatin regions; and by using intrinsic string mining approaches which identified several unique RKM topological, compositional, and clustering properties in the correlated genomic features. These findings were summarized as the first contribution (of biology) and were analysed together to infer for three rare-word (RW) signals of clustering (RWC), maximum likelihood (RWML), and central distribution (RWCD) to represent the CGI and promoter (i.e. the identified computational applications). The second objective is to model the CGI using the inferred RW signals and incorporated them into three RW (of RWC, RWML, and RWCD) methods. Next, these RW methods are optimised using a standard Particle Swarm Optimization (PSO) algorithm, trained on several chromosomes, tested using 5-fold cross-validations, and final testing at the human genome scale. The CGI evaluations were done using four validation datasets (of Illingworth, Weber, PhastCon elements, and Alu repeats) and corresponding Hackenberg’s evaluation protocols. The third objective is to model the promoter using the same procedures as the CGI modelling. The differences are its validation datasets (of Carninci and RefSeq),

corresponding Abeel's evaluation protocols, and lower prediction accuracies than the CGI predictions. Although the inferred RW signals can identify majority of promoters, their limited performances are due to the dominance of CGI signal in the identified promoters and other signals were not considered by us due to this study limitation. When the predicted CGI and promoters were benchmarked against seven other prediction datasets (by CpGCluster, CpG_MI, CpGProD, NCBI-CGI, and UCSC-CGI programs), they consistently rank among the top four, in terms of F-score, sequence coverage, or CC-value.

CHAPTER 1 - INTRODUCTION

1.1 Introduction

Enormous biological information (to develop and to sustain life) of an organism is encoded in its genome. For a eukaryote organism, this encoding is done at three different levels of Deoxy-ribo-Nucleic Acid (DNA), Ribo-Nucleic Acid (RNA), and protein. In general, this concept is known as the central dogma of molecular biology. The most basic encoding is DNA (or dinucleotide) which is composed from four bio-molecule structures of Adenine (abbreviated as A), Cytosine (C), Guanine (G), and Thymine (T). The human genome contains more than three billion letters of DNA which provides enormous combination of DNA sequences for potentially useful information (Parida, 2007). Nevertheless, only ~5% are presumed to be functional and important which consist of protein coding genes, non-coding RNA genes, and conserved elements (Strachan and Read, 2010c). The remaining consists of repeat elements (~51% of the human genome) and other non-repetitive regions such as pseudo-gene and intron (~44%).

The functional sequence elements are also known as genomic features at a larger scale and sequence motifs at a subunit scale. Genomic feature is defined as any genomic regions which are aggregation of smaller subunits and are annotated with a common biological function such as gene, messenger RNA (mRNA), protein, CGI, promoter, and origin of replication (COGEPEDIA, 2013). Whereas sequence motif is defined as conserved or recurring DNA patterns that can be implicated with a certain biological motif such as transcription factor (TF), various core promoter elements, and structural motif (Das and Dai, 2007). Computationally, both of the functional elements are commonly been identified using broad sequence analysis techniques where the former usually requires sequence homology (comparative) and prediction (classification) based methods and the latter usually requires sequence motif discovery (intrinsic) based methods (Abouelhoda and Ghanem, 2010).

Identifying functional elements within three billion letters of the human genome is not an easy task due to the complexities and flexibilities of biological features and motifs in term of their organizations, sizes, and interaction mechanisms (Michelson and Bulyk, 2006, Pennisi, 2012). Previous studies have shown that empirical analyses on DNA *k*-mers can be an effective means in identifying various sequence motifs, in term of their location, function, and organization (Gentles and Karlin, 2001, Chan and Kibler, 2005, Das and Dai, 2007, Badis et al., 2009, Chor et al., 2009, Castellini et al., 2012, Hariharan et al., 2013). The findings from such studies are used to characterize the associated genomic features or implemented as prediction tools to predict them, as described in the aforementioned papers. Other uses include sequence alignment (Kent, 2002), probe design (Fofanov et al., 2004), repeat annotation (Kurtz et al., 2008), and genome assembly (Compeau et al., 2011).

Among the empirical works, “rare *k*-mer" is a very interesting subject to be studied due to their unique sequence properties in the human genome. The low frequency property of these rare *k*-mers might be attributed to the well-known phenomenon of CG dinucleotide suppression in vertebrate genomes (Cooper and Gerber-Huber, 1985) due to the longer *k*-mers containing the CG(s) are also under-represented in mammalian genomes (Levy, 2008). The term rare *k*-mers was conceptualized by Chor et al. (2009) as DNA *k*-mers of selected length (of *k* in between 7 to 11-mers) that fall under low-frequency modes of genomic multi-modal *k*-mer spectra that only happen in a few species under the Tetrapod clad which includes all mammals. Their results also shows that the inclusion of CG(s) causes the rare *k*-mer frequencies to be lower; the multi-modality can be determined by certain percentages of G+C content and ρ_{CG} values; and the rare *k*-mers are surprisingly more common in exon, 5’ UTR, and proximal promoters (implicated from the unimodal *k*-mer spectra in them) in contrast to genome, intron, distal promoter, and 3’UTR regions (due to their multi-modal *k*-mer spectra). Despite these unique rare *k*-mer sequence properties, not many extensive works have been done to elucidate their biological properties, motifs, and functions.

This research consists of three progressive parts, i.e.: 1) To correlate rare *k*-mers with a wide range of genomic features, to analyse various rare *k*-mer sequence properties in the correlated features, and to infer valid representations of the correlated features based on the most profound rare *k*-mer sequence properties in them; 2) To develop classification methods to predict the correlated features utilizing the inferred rare *k*-mer signals; and 3) To evaluate the prediction results by the newly developed methods with proper validation (benchmark) datasets and to benchmark the results with other programs' results. At the end of the first part, three rare *k*-mer signals of clustering, selective distribution, and central distribution were inferred as valid representative signals for the correlated features of CGI and promoter. In the second part, the three rare *k*-mer signals was represented into a classifier (a set of parameters and a proper function) into three different RW methods of RWC, RWML, and RWCD respectively. Each prediction run of a RW method was optimized using a generic PSO algorithm based on a certain validation dataset and an evaluation protocol. For generalization of the RW methods, we trained each of them on three selected chromosomes, performed 5-fold cross-validations to test the results, and tested each of them on the human genome scale. In total, there were 24 optimised prediction datasets (2 features x 4 protocols x 3 RW methods). For the third part, we discussed several issues of the CGI and promoter models as well as of their associated experimental (validation) datasets and evaluation protocols in order to know the models' limitations and to improve their predictors' performances. Thus, certain filtering and settings were applied for particular evaluations. Then, we benchmarked RW predictions with seven other CGI based prediction datasets which are subjected to the same evaluation settings. For the CGI evaluations, RW predictions are ranked among the top three (in term of F-score and sequence coverage) when they were evaluated against four validation datasets of un-methylated regions (UMR), hypo-methylated promoters (HMP), phylogenetic conserved (PhastCon) elements, and Alu repeats. For the promoter evaluations, RW predictions are ranked at the first and within the top four (in term of F-score and CC-value) when they were evaluated against two validation datasets of transcription start regions (TSRs) and gene start regions (GSRs) respectively.

1.2 Motivation

In a related study done by Chor et al. (2009), they extensively analysed *k*-mer frequency distributions of genomes of more than one hundred species, including archaea, bacteria, and eukaryotes. Most of them exhibit a unimodal *k*-mer distribution except for a few species, which includes all mammals, exhibit multi-modal spectra. For a normally distributed *k*-mers, a bell shape distribution is expected, yet a multi-modal *k*-mer distribution is obtained in mammalian genomes. The multi-modal *k*-mer distributions comprised of two unexpectedly high peaks in the lower spectrum (encompassing the rare *k*-mers) and a shallower peak near the spectrum centre. Intrigued by the unique multi-modal *k*-mer distributions, particularly in the human genome, we focus our analysis on the rare *k*-mers, i.e. all *k*-mers (of *k*=8-to10) that fall under the first and second modes (the anomalous parts) of the multi-modal spectrum of the human genome.

One of the unique properties of the rare *k*-mers is they contain multiple CGs as implicated by Chor et al. (2009). In one of the earliest studies on genome compositions of diverse eukaryotes by Karlin and Mrazek (1997), it was discovered that CG is the most under-represented (highly suppressed) dinucleotide in vertebrate genomes. The imbalances distribution and localization of the CGs have also been correlated with other important genomic features too such as the CGI and promoter regions. CGI are regions which enriched in CGs which are drastically differ than the broad genome which is devoid of CGs whereas promoters are upstream regions that regulate the transcriptions of genes where majority of mammalian promoters overlap with the CGIs. Another recent study by Hackenberg et al. (2012) has shown that some highly clustered rare 8-mers are correlated with functional elements of exons and TFBS. These open up possibilities to use rare *k*-mers as a mean for predicting certain genomic features as experimental cost is usually higher than computational cost.

There are some intricate relationships between rare *k*-mers, CGI, and promoter regions since all of them are associated with CG suppression. Since some rare 8-mers were proven to be functional, we were interested to further investigate correlations and functions of the rare *k*-mers in the CGI and promoter. CGIs have been implied to play many important roles in biology such as contain most of the unmethylated CGs, overlap with majority of promoters, open chromatin regions, and contain high density of regulatory elements (Antequera, 2003). CGI is also been proven to be the most dominant signals for predicting promoters in mammalian genomes (Ioshikhes and Zhang, 2000, Hannenhalli and Levy, 2001, Abeel et al., 2009). Computational prediction of promoter regions has been an important task in genome annotation projects due to rarely expressed genes are hard to detect by current experimental methods and to provide means for regulatory analysis of unknown full length transcript genes. Many approaches and methods utilizing various biological features of eukaryotic promoters have been adapted by researchers to computationally address this problem. Promoter prediction has been one of the most elusive problems with limited success despite lots of study has been done in the area (Hannenhalli and Levy, 2001, Abeel et al., 2009). Perhaps, by studying rare *k*-mers properties in the CGIs and promoters could lead to better understanding of their implied roles in biology.

1.3 Problem Statement

Previous studies have shown that empirical analysis on DNA *k*-mers can be an effective mean in identifying various genomic features and sequence motifs in genomes and various DNA *k*-mer properties have been used in diverse computational applications in biology. Rare *k*-mer DNA (i.e. *k*-mers under the anomalous modes of the human multi-modal *k*-mer distribution) is a very interesting subject to be studied due to their unique sequence properties and not many extensive studies have been done on them. Among the peculiar characteristics of rare *k*-mers are: 1) abundance of rare *k*-mers which masses at the lower spectrum of *k*-mer distributions of mammalian genomes which deviate from the unimodal *k*-mer distributions of most organisms; 2) the same thing is not happening for the frequent *k*-mers in the upper spectrum; 3) only a certain range of *k* (7 to 11) has shown a clear multi-modal *k*-mer spectrum; 4) the suppression of rare *k*-mers can be attributed to CG suppression factor; and 5) the whole human genome which has the *k*-mer multi-modal spectrum shows the opposite unimodal spectrum in much smaller regions of exon, 5' UTR, and proximal promoters, but shows the same spectrum in other larger regions of intron, distal promoter, and 3'UTR. More details on these unique properties can be found in Section 3.1.3. These peculiarities give a clear indication that rare *k*-mers are somehow functional in the human genome. Moreover, CG suppression (the main factor of the rare *k*-mers) has been implied to play many important roles in biology. Therefore, we seek to answer the following research questions:

1) What are the biological functions of the rare *k*-mers in the human genome? What are the correlations of rare *k*-mers with any known genomic features? What type of rare *k*-mers motifs that we can find in the correlated genomic features?
2) Can the correlated CGI feature be the target for the rare *k*-mer computational application in biology? Can the identified rare *k*-mer motifs in the CGIs be exploited as their classifier signals?
3) Can the correlated promoter feature be the target for the rare *k*-mer computational application in biology? Can the identified rare *k*-mer motifs in the promoters be exploited as their classifier signals?

1.4 Research Objectives

Based on the motivation and problem statement of this study, three research objectives were identified which are:

1) To correlate rare *k*-mers to a wide range of known genomic features and to infer for potential representative signal for the correlated (CGI and promoter) features based on the most profound rare *k*-mer sequence properties (motifs) in them.

2) To develop classification methods to predict the CGI feature based on the inferred rare *k*-mer signals in the first step.

3) To develop classification methods to predict the promoter feature based on the inferred rare *k*-mer signals in the first step.

1.5 Research Framework

Figure 1.1 gives a comprehensive overview of all research undertakings in this thesis. The framework divides this research into three progressive parts of a preliminary study on rare *k*-mer motifs in the human genome (Chapter 4), development of three RW methods to predict the CGI and promoter features (Chapter 5), and evaluations and benchmarks of the three RW methods (Chapter 6). Chapter 4 is purposely referred as a preliminary study for the next two chapters so that readers will focus their attention on the inferences (the final conclusions) of this chapter, not on the details of the applied experiment methods to achieve the inferences. The first part is more related to the Life Sciences domain where most of its theory and axioms are often fuzzy and incomplete, there is usually no definitive answer, and its finding is always hypothetical (or empirical) by connecting different pieces of evidences. Once the inferential works in Chapter 4 were done, the newly discovered knowledge on the rare *k*-mer applications and their related representative signals were utilized in Chapter 5 to develop the three RW methods to predict the CGI and promoter motifs. Three research objectives are already given in Section 1.4 where the first objective corresponds with Chapter 4 but both of the second and third objectives correspond with Chapter 5. Chapter 6 elaborates on the evaluations and benchmarks of the three RW methods and discusses their results.

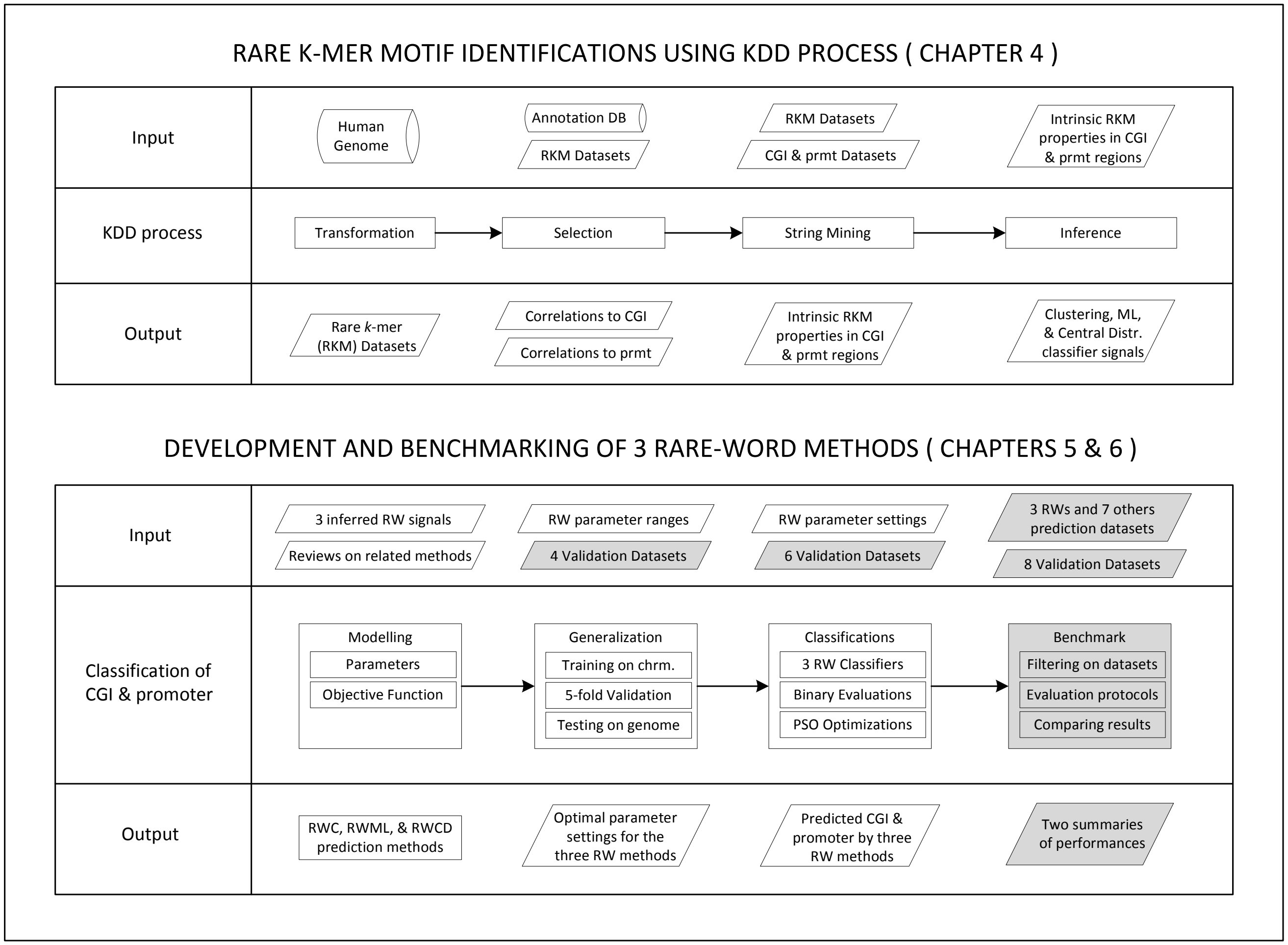

RARE K-MER MOTIF IDENTIFICATIONS USING KDD PROCESS (CHAPTER 4)
Input
Human Genome
Annotation DB
RKM Datasets
RKM Datasets
CGI & prmt Datasets
Intrinsic RKM properties in CGI & prmt regions
KDD process
Transformation
Selection
String Mining
Inference
Output
Rare k-mer (RKM) Datasets
Correlations to CGI
Correlations to prmt
Intrinsic RKM properties in CGI & prmt regions
Clustering, ML, & Central Distr. classifier signals
DEVELOPMENT AND BENCHMARKING OF 3 RARE-WORD METHODS (CHAPTERS 5 & 6)
Input
3 inferred RW signals
Reviews on related methods
RW parameter ranges
4 Validation Datasets
RW parameter settings
6 Validation Datasets
3 RWs and 7 others prediction datasets
8 Validation Datasets
Classification of CGI & promoter
Modelling
Parameters
Objective Function
Generalization
Training on chrm.
5-fold Validation
Testing on genome
Classifications
3 RW Classifiers
Binary Evaluations
PSO Optimizations
Benchmark
Filtering on datasets
Evaluation protocols
Comparing results
Output
RWC, RWML, & RWCD prediction methods
Optimal parameter settings for the three RW methods
Predicted CGI & promoter by three RW methods
Two summaries of performances

Figure 1.1: The complete research framework of this study. The shaded objects are covered in more details in Chapter 6.

1.6 Research Scopes and Limitations

The research scopes and limitations for each of the research objectives are listed as follows:

1) For the identification of rare *k*-mer motifs in the human genome, the scope is to find valid representative rare *k*-mer signals to be implemented in potential computational applications in biology. Due to the vastness of available annotations which can be used for this discovery study, we utilized several bioinformatic tools (i.e. UCSC browsers, EpiGRAPH, and Galaxy) to quickly correlate rare *k*-mers to hundreds of annotated genomic features. From this step, we selected CGI and promoter as the most correlated features for the rare *k*-mers, and thus become their potential computational applications too. Then, several selected intrinsic analyses were done to the CGI and promoter features to elucidate unique rare *k*-mer sequence properties in them. Finally, three of the most profound rare *k*-mer sequence properties (i.e. RWC, RWML, and RWCD) were selected as the valid representative signals for the CGI and promoter. Another limitation is we only perform our study using the human genome only due to most of related works and existing genomic annotations are available for the human genome compared to the others.

2) In the second research objective, we want to develop three classification methods to predict the CGI. The methods were developed based on the three inferred rare *k*-mer representative signals for the CGI in the first step. We incorporated the three rare *k*-mer signals into three classifiers by improvising the parameters and functions of similar CGI classifiers from the past studies (see Section 5.1). Four proper validation datasets were used, i.e. Illingworth UMR, Weber HMP, PhastCon elements, and Alu repeats (Jurka et al., 2005, Siepel et al., 2005, Weber et al., 2007, Illingworth et al., 2008) for the CGI evaluations by and seven other CGI prediction datasets from five late CGI prediction programs were used for the CGI benchmarking, i.e. CpGcluster, CpGMI, CpGProD,

NCBI-CGI, and UCSC-CGI (Ponger and Mouchiroud, 2002, Hackenberg et al., 2006, Maglott et al., 2007, Rhead et al., 2010, Su et al., 2010).

3) For the third research objective, we limit the promoter modelling into applying the previously developed classifiers and expand our works on optimizations and evaluations of the three RW methods to predict the promoter. Since the three rare *k*-mer signals were already inferred as valid representative signals for both CGI and promoter features, naturally the classifiers using these three signals are compatible for both of the features too. However, promoters are a complex feature which constitutes lots of other motifs. Nevertheless, it is a well-known fact that CGI is the most dominant signal to predict the promoter (see Section 5.1) and presumably these three rare *k*-mer signals can also be the most dominant signals to predict promoter due to the good CGI predictive performance by the three RW methods. Two proper validation datasets were used for the promoter evaluations, i.e. Carninci TSR and RefSeq GSR (Carninci et al., 2006, Pruitt et al., 2007) and the same seven CGI prediction datasets were used for the promoter benchmarking.

1.7 Research Methodologies

Our research methodologies are divided into three parts which represent chapters 4, 5, and 6 respectively. Although validation datasets and evaluations are utilized in Chapter 5, they are not discussed in details until Chapter 6 to make way for more organized chapters of this thesis (see Figure 1.1). Our research methodologies are listed as follows:

1) The first part focuses on identifying rare *k*-mer motifs in the human genome. Therefore, we used several bioinformatic tools (such as the UCSC genome and table browsers, EpiGRAPH, and Galaxy) to quickly correlate rare *k*-mers to hundreds of annotated genomic features. Then, we performed several intrinsic analyses of string mining by using Perl scripting to elucidate unique rare *k*-mer sequence features (e.g. composition, distribution, enrichment, and topology) in the correlated features. The rare *k*-mer motif identification study can be generalized into the Knowledge Discovery in Databases

(KDD) framework (see Section 3.2). In the pre-processing step, the raw human genome data were transformed into the rare *k*-mer datasets. In the selection step, the bioinformatic tools were used to correlate rare *k*-mers to the CGI and promoter features which are then chose as the potential rare *k*-mer computational applications in biology. In the data mining step, intrinsic string mining approach was used to elucidate unique rare *k*-mer sequence properties in the CGI and promoter. In the last step, all significant evidences (results from the intrinsic analyses and related CGI and promoter knowledge from literatures) are summarized and analysed together to infer for potential rare *k*-mer representative signals for the correlated features of the CGI and promoter.

2) The second part deals with modelling of the CGI and promoter features based on the inferred rare *k*-mer signals (in the previous step) into three RW methods and optimizing and generalizing them to classify CGIs and promoters in the human genome. The development of the three RW methods can be generalized into the following steps: 1) modelling the CGI and promoter features based on the inferred rare *k*-mer signals and improvising the parameters and functions of similar prediction programs; 2) Evaluate the RW methods using proper validation datasets and evaluation protocols (see the next part); 3) Optimizing the RW classifiers using a generic PSO algorithm; and 4) Test the generalization of the prediction results by training them at several chromosome scales, performed the 5-fold cross validation test, and testing them at the human genome scale.

3) The third part concentrates on the evaluations and benchmarking of the 3 RW methods. Issues in each of the evaluation components, i.e. prediction datasets, validation datasets, and evaluation protocols, are discussed in details in order to know the limit of the CGI and promoter models and to improve on their predictors' performances by suggesting certain filtering and evaluation settings. The evaluations and benchmarking were done using the following datasets (as cited in Section 1.6) and corresponding protocols, i.e.: 1) Seven prediction datasets from five different CGI prediction programs which are CpGCluster, CpG_MI, CpGProD, NCBI-CGI, and UCSC-CGI; 2) Four CGI validation

datasets of Illingworth UMR, Weber HMP, Alu repeats, and PhastCon elements with four corresponding evaluation protocols by Hackenberg et al. (2010); and 3) Two promoter validation datasets of Carninci TSRs and RefSeq GSRs times two evaluation protocols by Abeel et al. (2009).

1.8 Research Contributions

1) New biological knowledge on rare *k*-mer properties in the correlated CGI and promoter features.

2) New CGI models based on three novel rare-word signals to classify CGI feature in the human genome by using three relevant rare-word methods.

3) New promoter models based on the above three novel rare-word signals to classify promoter feature in the human genome by using three relevant rare-word methods.

1.9 Thesis Organization

This thesis is organized into seven chapters and sixteen appendices as follows:

Chapter 1 gives comprehensive overview of this research.

Chapter 2 serves as the introduction to all of the related research areas which addresses fundamentals of cell biology, bioinformatics, and computational biology. Since this research is multi-disciplinary in nature, it is necessary to provide a concise introduction to all of the related fields so that readers from one discipline can understand the fundamental concepts and expectations of other disciplines in term of basic theories, related issues, methodologies, and results.

Chapter 3 reviews the backgrounds and methodologies of the three main objectives of this research, i.e.: 1) rare *k*-mer motif identification in the human genome (for the first objective); 2) General methodologies for genomic feature classification (for the second and the third objectives); and 3) promoter modelling (more elaborate reviews for the third objective). We purposely placed CGI modelling review in Chapter 5 for more focussed discussion.

Chapter 4 provides the details for rare *k*-mer motif identification study. We refer Chapter 4 as a preliminary study for the next two chapters so that readers will focus their attention to the main findings (or the inferences) of this chapter rather than its methodology.

Chapter 5 focuses on the development of the three novel rare-word methods based on the inferred rare *k*-mer signals in Chapter 4 to predict the CGI and promoter features. It also discusses on optimization and generalization of the three RW methods. We purposely placed the details on evaluation in Chapter 6 for more organized chapters of this thesis.

Chapter 6 discusses several issues of the CGI and promoter models as well as of their associated validation datasets and evaluation protocols in order to know the models' limitations and to improve their predictors' performances. At the end, their results were benchmarked against five other programs and their performances were discussed.

Chapter 7 concludes the main findings of this research and gives suggestions for future works.

Appendices A-P elaborate on supplementary methods and results that are indirectly related to the discussions in Chapters 4, 5, and 6 which might be of important for interested readers.

CHAPTER 2 - BACKGROUND

The rare *k*-mer motif discovery and the development of prediction tools for the rare *k*-mer motifs fall under the areas of Bioinformatics and Computational Biology. Both are multi-disciplinary areas which combine the concepts and methodologies from overlapping fields of Computer Science and Biology. Section 2.1 is dedicated to readers from a Computer Science background to comprehend the related concepts and terminologies in biological domains. Section 2.2 is prepared for positioning our research within related areas of bioinformatics, computational biology, and sequence analysis which include an introduction to fundamental concepts, major fields, basic data types, and common tools in those areas.

2.1 Fundamentals of Cell Biology

Bioinformatics is derived from the word biology and information. Hence, it is a research field which is closely related to biology. Understanding and interpretation of biological data such as cellular activities and regulations are very difficult due to they are operating in a multitude of levels, copious, and very complex. This section aims to give readers from a computer science background the fundamental of cell biology, which covers related topics of basic components of genetic information, organization and statistics of the human genome, organizations and statistics of the human genes, metaphor of genetic information flow in cell, and regulation of eukaryotic gene expression. Each topic is explained in Section 2.1.1 to 2.1.5 respectively. Note that words with "**bold"** suggest important words which are frequently used in this thesis.

2.1.1 Basic Components of Genetic Information

All living organisms depend on cell abilities to store, access, and decode the DNA instructions in genome to develop and sustain their being. In fact, the development of a multicellular organism starts from a single cell (i.e. a zygote cell in human). The single cell is then duplicated, coordinated, organized, and specialized into billions of cells until they reach their complete forms. The most fundamental of genetic information in cells is Deoxyribo-nucleic acid (abbreviate as **DNA**) for most organisms or Ribonucleic acid (**RNA**) for some type of viruses. DNA (or RNA) carries information either as regulation or as a template for other functional forms. DNA is transformed into RNA through a process known as **transcription** and RNA is transformed into **protein** through a process called **translation**. The structures and functionality of a cell are mostly made possible by the properties of proteins. For heads-up, genetic information in cells is organized into DNA-RNA-protein, genes, chromosomes, genomes, and cells (see Figure 2.1).

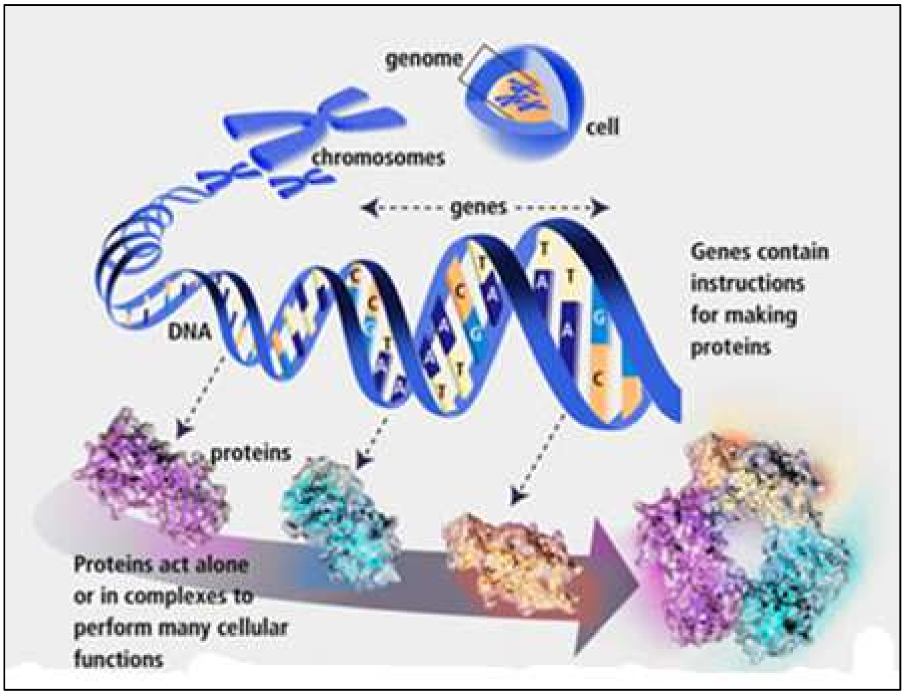

genome
cell
chromosomes
genes
Genes contain instructions for making proteins
DNA
proteins
Proteins act alone or in complexes to perform many cellular functions

Figure 2.1: Organization of genetic materials in cell and the transformation of genetic information from DNA to protein. This image was taken from (BERIS, 2008).

DNA sequence is a built-up from several consecutive units of four nucleic acids of Adenine, Thymine, Cytosine, and Guanine (abbreviated as **A**, **T**, **C**, and **G** accordingly). They are also known as **nucleotides** or **bases**. Bases that are adjacent to each other on the same **strand** are connected by phosphodiester (or covalent) bonds. When mentioning two neighbouring **dinucleotides,** it is common to insert a ‘p’ to indicate a phosphodiester bonding (e.g. **CpG** means that a Cytosine is covalently linked to a neighbouring Guanine on the same strand). A DNA strand has a unique direction denoted by a head (called the **5’ end**) and a tail (the **3’ end**). In normal condition, DNA forms antiparallel double-helical strands which are held together by weak hydrogen bonding to form a DNA duplex or double helical strands. The second (or **complementary**) strand running in the opposite direction of **3’** to **5’** than the first strand (see Figure 2.2). The hydrogen bonding between the two strands come from individual bases facing each other in both of the strands which are also known as **base-pairing**. According to the Watson-Crick rules, ’A’ specifically base-pairs with ‘T’ and ‘C’ specifically base-pairs with ‘G’ (see Figure 2.2). Collectively, individuals with weak binding of base-pairings hold both of the strands together. The DNA structure of base-pairing provides a stable mechanism for heredity (due to the redundancy) and as medium of genetic instruction (due to the capability of the transcription and translation). There also exists RNA-DNA duplex bonding structures during the transcription process (or RNA-RNA structure of viruses) where these duplexes have weaker bonds than the DNA-DNA structures of base pairing.

```
5' - ATG ACT CAC CGA GCG CGA AGC TGA - 3'
3' - TAC TGA GTG GCT CGC GCT TCG ACT - 5'
```

Figure 2.2: Shows an example of DNA double strand sequence.

RNA has very similar chemical structures to DNA with two slight variations, i.e. RNA contains sugar Ribose components instead of sugar Deoxyribose components and Uracil (**U**) nucleotides in place of Thymine (T) nucleotides (see Figure 2.3). RNA molecules normally exist as single strand structures in cell in contrast to DNA molecules which exist as double helical strands. During the translation process, linear DNA (or nucleotide) sequence in the 3’ to 5’ strand orientation (the **template** or the **antisense strand**) is decoded into a linear RNA sequence of the 5’ to 3’ strand orientation (the **sense strand**). The translation is done using 1 to 1 base decoding fashion where the RNA sequence has similar bases as the DNA sequence except for U is replaced with T. The RNA molecules which encode for polypeptide are known as messenger RNA (**mRNA**) or coding RNA (due to they only code for protein). There are also other types of RNA available known as non-coding RNA (**ncRNA**) such as **rRNA**, **tRNA**, and **miRNA**. These ncRNAs are not translated into proteins, but they can form higher degree of secondary structures which can have very specific functions in cells.

```
5' - AUG ACU CAC CGA GCG CGA AGC UGA - 3'
```

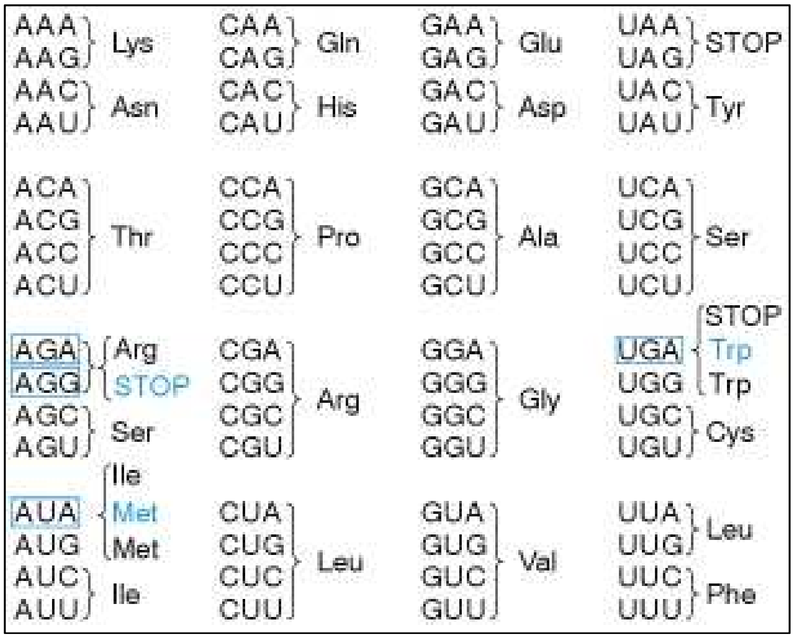
Figure 2.3: Shows the corresponding RNA sequence transcribed from the DNA sequence in Figure 2.2.

Protein is composed of one or more **polypeptide** molecules which are made up from smaller subunits known as **amino acids**. During the translation process, linear sequence of RNA molecules in the 5’ to 3’ strand is decoded in a **codon** fashion, i.e. 3-bases to 1-peptide decoding, to give a linear sequence of polypeptides (see Figure 2.4). There are 20 different types of amino acids. Each amino acid has a different side chain structure and chemical property which cause the whole amino acid chain to fold into a specific 3D structure. Collectively, the structure and property of amino acids in a protein determine the its functions in cell such as enzymes, receptors, transport, regulation, signalling, hormones, etc. (BERIS, 2008).

```
5' - Met Thr His Arg Ala Arg Ser Trp - 3'
```

Figure 2.4: Shows the corresponding protein sequence translated from the RNA sequence in Figure 2.3.

The list of three-bases to one-peptide conversions is given in the **genetic code** table in Figure 2.5. For each position in a codon, there are four possible nucleotides to choose, so there are 64 (4^3) possible variations of codon to encode for only 20 types of amino acids. Thus, the genetic code is said to be degenerate, i.e. in average there is about three different codons to encode for a single amino acid. However, the distribution of codons to code for the amino acids is not equal. Methionine or tryptophan is encoded by only a single codon while serine and leucine are encoded by six codons.



AAA	Lys	CAA	Gln	GAA	Glu	UAA	STOP
AAG		CAG		GAG		UAG	
AAC	Asn	CAC	His	GAC	Asp	UAC	Tyr
AAU		CAU		GAU		UAU	
ACA	Thr	CCA	Pro	GCA	Ala	UCA	Ser
ACG		CCG		GCG		UCG	
ACC		CCC		GCC		UCC	
ACU		CCU		GCU		UCU	
AGA	Arg STOP	CGA	Arg	GGA	Gly	UGA	STOP Trp
AGG		CGG		GGG		UGG	Trp
AGC	Ser	CGC		GGC		UGC	Cys
AGU		CGU		GGU		UGU	
AUA	Ile Met	CUA	Leu	GUA	Val	UUA	Leu
AUG	Met	CUG		GUG		UUG	
AUC	Ile	CUC		GUC		UUC	Phe
AUU		CUU		GUU		UUU	

Figure 2.5: The nuclear and mitochondrial genetic codes are almost similar except for few codons. The blue boxes highlight different interpretations of codons in the nucleus and mitochondria of mammalian cells. The different interpretation is coloured in blue for mitochondria. This figure was taken from (Strachan and Read, 1999).

Gene is a small region inside a genome, which represents a basic functional unit of heredity. Through laboratory experiments or computational methods, segments of DNA which are identified to encode for particular types of RNAs are labelled as genes. The genes remain dormant as heredity units until they become accessible to the right combinations of regulatory elements. Then the genes will serve as guides or templates for the transcription process of RNA. Each of prokaryote and eukaryote has a different organisation of a protein-coding gene. Most of gene sequences for prokaryote encode for protein while only a small amount (in average about 10%) of gene sequences for eukaryote encode for protein (see Section 2.1.3). Most parts of the latter are removed during the transcription process before the remaining are decoded into protein. The transcribed RNA of eukaryotic protein-coding genes will undergo additional post-transcriptional modifications of capping, polyadenylation, and splicing before they will become mature RNAs (i.e. mRNA) (Strachan and Read, 1999).

Chromosome is a packaging of DNA sequence. For an example, a human chromosome contains a chunk of DNA sequences with size range from 50 to 250 million base pairs (BERIS, 2008). Chromosome packaging allows a cell to keep large amount of genetic information in a neat, organized, and compact form. Chromosomes have several levels of DNA packing namely double helix, nucleosome, "beads-on-a-string" chromatin, 30-nm fibre chromatin, looping of fibre chromatin, and finally mitotic chromosome (Alberts et al., 2003). The packaging in each level is facilitated by certain protein structures (e.g. histones) and regulation (e.g. chromatin modification) which permits a certain degree of access and controls to the DNA information that is contained within (see Figure 2.6).

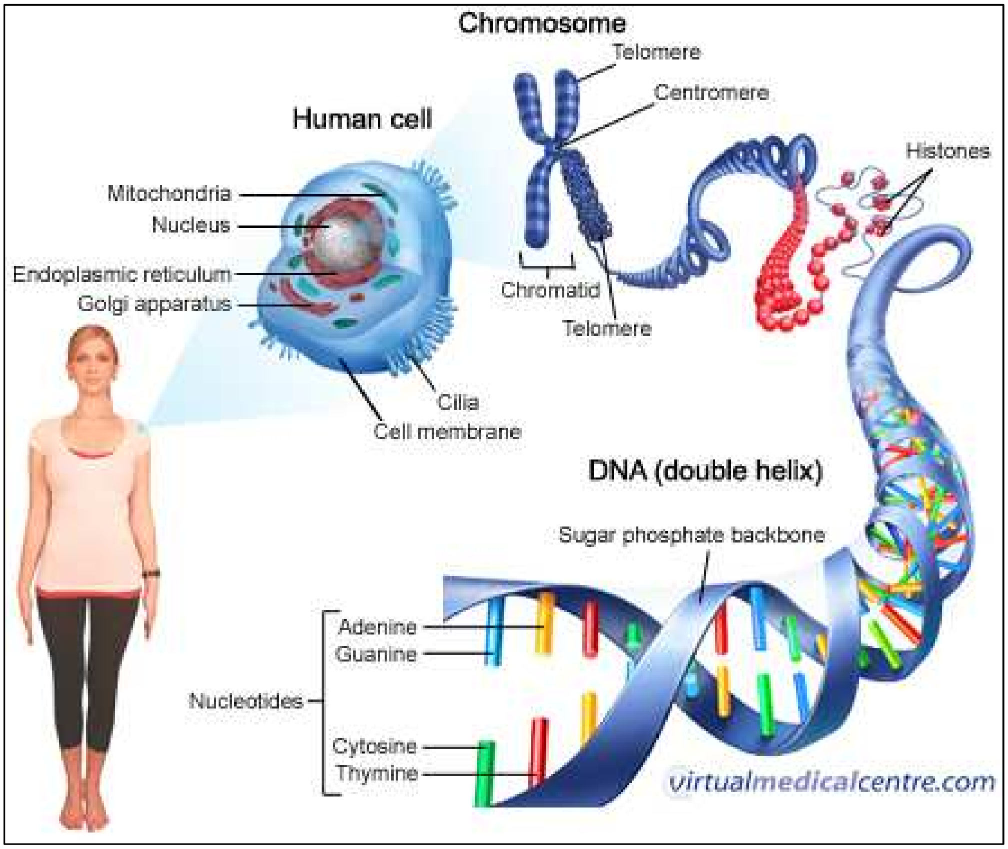


Figure 2.6: Organization and structure of DNA information in cell. This image was taken from (VMC, 2008).

Genome is the whole set of genetic material of all living organisms. Genome serves as a cellular brain to coordinate all activities which are happening inside a cell. Genome encodes for instructions on how to build, run, and maintain an organism; and as an entire set of heredity to pass life on to the next generation. Apart from the nuclear (**nucleus**) DNA, there are additional DNA information contained within other organelles of cells namely **mitochondria** (for animals) and **chloroplast** (for fungi and plants). The mitochondria and chloroplast DNA are believed to mainly encode for proteins that are used in the metabolisms of a cell. Human genome is made up from 23 pair of chromosomes of nuclear DNA and 2-10 chromosomes of mitochondrial DNA (Alberts et al., 2003).

Cell is the most basic structural and functional unit of every living organism. A cell regulates thousands of vital functions within to allow its sustenance, obtain energy, and respond to environmental stimuli. Human adult is estimated to have in between 50-100 trillion cells and each of them carries a complete genome information. The genome as hereditary information is carried on to the next generation through reproductive (or gamete) cells and to the child (or somatic) cells through cell divisions. What makes a cell different the others is the patterns of gene expressions in the cells, which define the functionalities of the cells. For examples, brain cells activate a large number of different genes, but in many other cell types, these genes are inactive. Some of the genes are essential for general cell functions which are actively transcribed across all cell types which we called them **housekeeping genes**. Other genes which are restricted to specific cell types are known as **tissue-specific genes**.

2.1.2 Organization and Statistics of the Human Genome

The human genome is comprised of nuclear and mitochondrial genomes. Each has slightly different properties as shown by their basic statistics in Table 2.1. The basic components and statistics of genome, chromosome, gene, and repeat for human organism are going to be explained as follows:

The size of the nuclear genome (~3.1 billion bases) is 186,000 times larger than the mitochondrial genome (16.6 kbps) as illustrated in Figure 2.7. Normally, the genome of higher eukaryotes such as human contains a substantial amount of non-coding sequence. Through a sequence comparison with other vertebrate genomes, only ~5% of the human nuclear genome is strongly conserved. These regions are presumed to be functional and important. Out of the 5%, only 1.1% code for protein while the other ~4% consist of non-coding RNA genes and conserved elements. For the rest of the nuclear genome, ~51% code for highly repetitive regions (such as transposon elements, satellite sequences, and heterochromatin) and ~44% code for non-repetitive regions (such as pseudo-gene, intron, and other non-repeat sequences). The nuclear genome has quite a different characteristics than the mitochondrial genome where the latter has almost none of the repeat sequences which is <2% (versus >50% in the former) and is densely populated with genes which is ~98% of its genome (versus ~5% in the former).

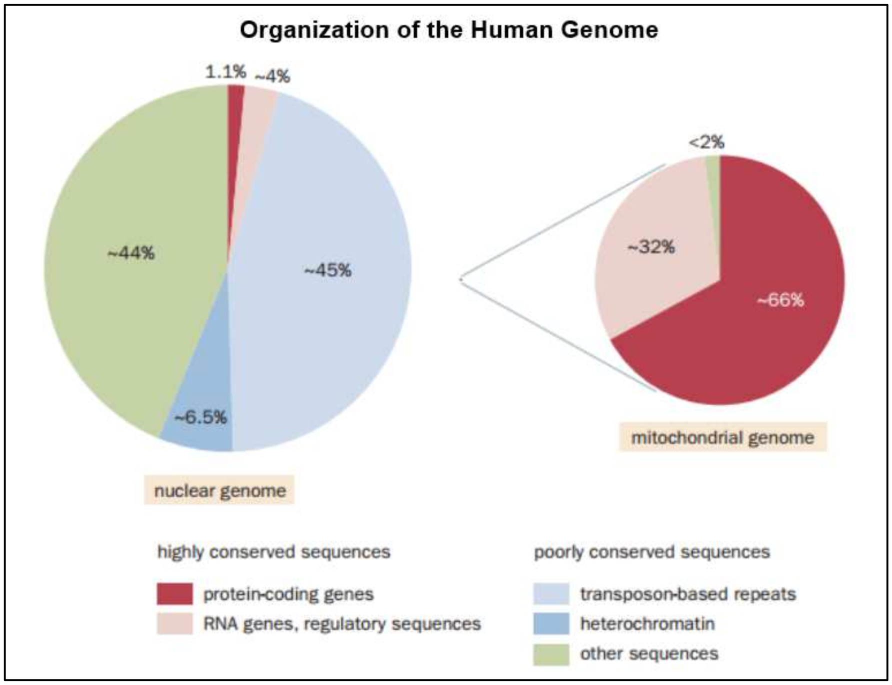


Figure 2.7: Organization of the human genome. This image was taken from (Strachan and Read, 2010c). To illustrate the different in sizes between nuclear and mitochondrial genomes, the red tiny dot at the centre of the Figure Nhows the actual size of mitochondrial genome (on the right) at the same scale as the nuclear genome on the left.

Simple organism such as bacteria has linear or cellular DNA molecules. Higher organisms such as human have their DNA molecules packaged into several chromatin structures of chromosomes. The DNA sequences of the nuclear genome are bound by histone and non-histone proteins which caused them to compress into higher density structure of chromosomes. Within the chromosome itself, there are several levels of structural density known as chromatin. Most of nuclear genome is packaged into euchromatin (~2.9 Gb) and a minority of the genome is packaged into heterochromatin (~200 Mb). The former is a less condensed region, gene-rich, and accessible for transcriptional activity while the latter is permanently condensed structure, devoid of genes, and mostly consists of repeat elements. The nuclear genome is distributed into 23 pairs of chromosomes, i.e. 22 homologous pairs of autosomes (i.e. one from paternal and one from maternal side) and 1 non-homologous pair of sex (either X-X, both from maternal, or X-Y, each from maternal-paternal) chromosomes. Meanwhile, the mitochondrial genome is of single type only, circular in shape, and there are variations of genome copy numbers in different types of cell.

The total number of genes in the nuclear genome has been revised several times since the post genomic era. In year 2001, the Human Genome Consortium came up with estimation of >30,000 protein-coding genes. This turn-out to be over-estimated due to the lack of supporting evidence and error made in defining the genes. Almost a decade later, the estimation was stabilized around 20,000 to 21,000 for protein-coding genes, but there are still many debates on the total number of RNA genes. Identification of the RNA genes is evidently difficult because they do not have open reading frames, are short in length, their sequences are not very conserved, and their definition is still uncertain. Recent studies have shown that >85% of euchromatin human DNA are transcribed and it is still not clear about how many of them are noise or significantly functional. At least, >6000 human RNA genes were annotated by 2010 and tens of thousands of RNA gene transcript evidences were obtained, but could not be clearly defined due to their ambiguity. In total, there are >26,000 human genes were confirmed, but this figure remains provisional due to the uncertainty in the definition of RNA genes.

More than 50% of the human nuclear genome consists of repetitive noncoding DNA sequences. About 200 Mb. (or 6.5% of the genome) are large arrays of tandem-repeat DNA sequences known as satellite DNA. This repeat sequence is located in the heterochromatin parts of all chromosomes where most of them at the centromeres and small amount of them at the telomeres. The heterochromatin structure remains condensed throughout the cell cycle and devoid of genes. The remaining repeat elements are known as transposon which account for ~45% of the genome. They originated from imperfect duplication of DNA segments which resulted in pseudo- or partial genes which repeats. They are scattered throughout the nuclear genome including extra-genic (most of the time), introns, untranslated sequences, and even in coding sequences.

Table 2.1: Shows several statistics of human nuclear and mitochondrial genome. This table was adapted from (Strachan and Read, 2010c).

Genome properties:	Nuclear genome:	Mitochondrial genome:
Size	3.1 billion bases	16.6 kilo bases
Types of chromosomal DNA molecules	23 (in XX cell) or 24 (in XY cell), all are linear	One circular DNA molecule
Total num. of DNA molecules per cell	23 in haploid cells; 46 in diploid cells	Several thousand copies but varies in different cells
Association of proteins to the DNA molecules	Several classes of histone and non-histone proteins	Largely free of proteins
Number of protein-coding genes	20,000 to 21,000	13
Number of RNA genes	uncertain, but >6000.	24
Gene density	~1/120 kb, but great uncertainty	~1/0.45 kb
Repetitive DNA	>50% of genome	Very little
Transcription	Genes are often individually transcribed	Continuous transcription of multiple genes
Introns	Found in most genes	Absent
% of coding DNA	~1.1%	~66%
Codon usage	61 amino acid codons + 3 stop codons	60 amino acid codons + 4 stop codons

2.1.3 Organization and Statistics of the Human Genes

The organizations of prokaryotic and eukaryotic genes are known to be significantly different. For comparison, the organization of the former is explained first, followed by the latter. A prokaryotic coding gene consists of several elements of **coding regions** (specifically used to encode for protein sequences); specific sequences which encode for the start and stop positions of the transcription and translation process; and other regulatory elements concentrated in the upstream regions of the Transcription Start Site (TSS) of the gene (which is also known as **promoter** regions). One obvious characteristic of prokaryotic gene is its coding sequences are continuous (see Figure 2.8).

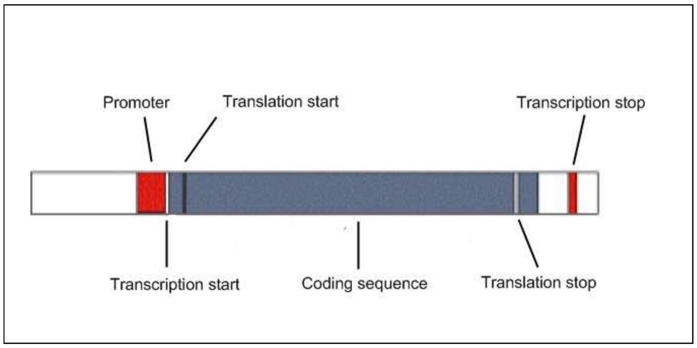


Figure 2.8: Shows elements of a bacterial gene.

The genes in higher eukaryotes are more complex. Their coding regions (or **exons**) are not linear and are alternated with non-coding regions called **introns** (see Figure 2.9). Introns are transcribed from DNA->RNA but never translated into protein (they are **spliced** out before translation). Apart from that, the region in between the transcription start site (**TSS**) and the translation start site is known as 5' untranslated region (**5' UTR**) while the region from the translation stop site to the transcription stop site is known as 3' un-translated region (**3' UTR**). The last and most important part of a eukaryotic gene is various regulatory elements and mechanisms to regulate the transcription process. Correlations between the elements and the mechanisms are very complex makes identification of the regulatory elements difficult. Usually, they are located close to the gene, especially within the immediate "**up-stream**" and "**down-stream**" regions (i.e. before and after) of the TSS.

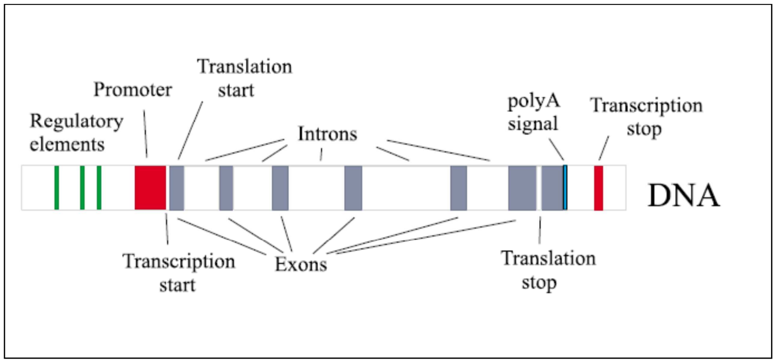


Figure 2.9: Shows elements of a eukaryote gene.