Title

Rare k-mer DNA: Identification of sequence motifs and prediction of CpG Island and promoter

Author names and affiliations

Ezzeddin Kamil, Mohamed Hashim[1] and Rosni, Abdullah[1,2]

[1]School of Computer Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia.

[2]Nav6, Universiti Sains Malaysia, 11800 Penang, Malaysia.

Corresponding author

Ezzeddin Kamil: ezzeddin@usm.my

Present address

School of Health Sciences, Universiti Sains Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia.

Abstract

Empirical analysis on *k*-mer DNA has been proven as an effective tool in finding unique patterns in DNA sequences which can lead to the discovery of potential sequence motifs. In an extensive study of empirical *k*-mer DNA on hundreds of organisms, the researchers found unique multi-modal *k*-mer spectra occur in the genomes of organisms from the tetrapod clade only which includes all mammals. The multi-modality is caused by the formation of the two lowest modes where *k*-mers under them are referred as the rare *k*-mers. The suppression of the two lowest modes (or the rare *k*-mers) can be attributed to the CG dinucleotide inclusions in them. Apart from that, the rare *k*-mers are selectively distributed in certain genomic features of CpG Island (CGI), promoter, 5' UTR, and exon. We correlated the rare *k*-mers with hundreds of annotated features using several bioinformatic tools, performed further intrinsic rare *k*-mer analyses within the correlated features, and modelled the elucidated rare *k*-mer clustering feature into a classifier to predict the correlated CGI and promoter features. Our correlation results show that rare *k*-mers are highly associated with several annotated features of CGI, promoter, 5'UTR, and open chromatin regions. Our intrinsic results show that rare

k-mers have several unique topological, compositional, and clustering properties in CGI and promoter features. Finally, the performances of our RWC (rare-word clustering) method in predicting the CGI and promoter features are ranked among the top three, in eight of the CGI and promoter evaluations, among eight of the benchmarked datasets.

Keywords

Rare-word; n-mer; k-tuple; CGI; Genome; and classification.

1.0 Introduction

The human genome contains the complete information on the extremely complex biological system of humans. At the most basic level, this information is encoded in DNA sequences of Adenine, Cytosine, Guanine, and Thymine biomolecules (abbreviated as A, C, G, and T letters respectively). DNA sequences which have a biological significance (or motif) such as regulatory, genic, or structural elements are known as sequence motifs. Identifying sequence motifs within three billion letters of the human genome is not an easy task due to the complexities and flexibilities of biological motifs in term of their organizations, sizes, and interaction mechanisms (Michelson and Bulyk, 2006; Pennisi, 2012).

Previous studies have shown that empirical analysis on DNA *k*-mers can be an effective mean in identifying sequence motifs by studying their occurrence, location, and organization in certain genomic contexts. Many aspects of empirical *k*-mer properties have been associated with diverse sequence motifs such as dinucleotide compositions in genomes to define their signatures (Gentles and Karlin, 2001), using DNA hexamers to predict promoter (Chan and Kibler, 2005), conserved and/or frequent DNA words in promoter regions to identify regulatory motifs (Das and Dai, 2007), *k*-mer preferences for thousands of DNA binding proteins to build sequence motif profiles (Badis et al., 2009), abundance and rarity of *k*-mer words to index functional genomic vocabularies (Castellini et al., 2012), and enriched *k*-mers in yeast promoters are utilized to identify promoters (Hariharan et al., 2013). The findings from such studies are used to characterize the associated motifs or implemented as prediction tools to predict them such as detection of: promoter regions (Chan and Kibler, 2005; Down and Hubbard, 2002; Li and Lin, 2006; Lin and Li, 2011; Lin et al., 2014); recombination spots

(Chen et al., 2013; Liu et al., 2012); nucleosome positioning (Guo et al., 2014; Segal and Widom, 2009); and translation initiation site (Chen et al., 2014a). Other uses include sequence alignment (Kent, 2002), probe design (Fofanov et al., 2004), repeat annotation (Kurtz et al., 2008), genome assembly (Compeau et al., 2011), and drug design (Chou, 2015).

Our interest is to investigate biological motifs for rare *k*-mers due to their unique sequence properties in the human genome. We summarized their properties from the following study which is the one that coined the term rare *k*-mers. Chor et al. (2009) did extensive empirical analyses on over a hundred multi-species genomes (of Archea, Bacteria, and Eukaryota), only few species under the Tetrapod clad exhibit multi-modal spectra (which include all mammals), while the rest exhibit unimodal spectra (see Section 3.1). Only certain ranges of *k*-values give distinct multi-modal *k*-mer spectra. For the human genome, the multi-modality is apparent for *k*-values in between 7 to 11-mers and we extracted the rare *k*-mers from under the lowest two modes of such spectra (see Section 2.1). The most outstanding property of the multi-modal spectra is *k*-mers with more CGs accumulate in the leftmost side of the spectra (see Figure 4 in their paper), in order words *k*-mers with more CGs tend to have lower frequency. Even though the rare *k*-mers under the first and the second modes have very low frequencies, collectively there are many variants of them in the narrow spectrum which inflate the lower bands, even higher than the third mode of the average frequency. The low frequency 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). However, there are several unicellular species which have low CG frequency but exhibit unimodal spectra and thus, the CpG suppression is not the only factor that determines the multi-modality. From their plot of G+C content and P(CpG) of the extensive genomes, the multi-modality can be associated with G+C content in between 35-45% and P(CpG) < 0.4. Other results implicate that rare *k*-mers are surprisingly more common in exon, 5' UTR, and proximal promoters (due to their unimodal spectra) in contrast to genome, intron, distal promoter, and 3'UTR regions (due to their multi-modal spectra). Despite these unique rare *k*-mer properties, not many extensive works have been done to elucidate their biological properties, motifs, and functions. Upon searching the google scholar for the "rare *k*-mers" and DNA (or genome)

keywords, less than 20 papers were found which relate to sequence alignment or genome assembly, genome organization, immunology, and nucleosome totalling 10, 4, 3, and 1 papers respectively (refer to Appendix A).

In this current work, we begin by re-evaluating the basic properties of the rare *k*-mers. Then, we correlated rare *k*-mers with hundreds of annotated biological motifs with the help of several bioinformatic tools (i.e. EpiGRAPH and UCSC browsers) where rare *k*-mers were found to be highly correlated with CGI, promoter, 5' UTR, and exon motifs. Then we performed further empirical analyses on the rare *k*-mers within the correlated motifs to elucidate any of their unique sequence properties, and exploited the findings into a prediction method, namely RWC (Rare-word Clustering) method, to predict two of the motifs i.e. CGI and promoter regions.

2.0 Materials and Methods

We have compiled quite an extensive work for this paper. The methodologies are organized into four sub-sections of 2.1 to 2.4 to match the result sub-sections of 3.1 to 3.4 respectively.

2.1 Selection, description, and extraction of multi-modal spectra and rare *k*-mer datasets

The methods to plot the *k*-mer spectra and to extract the rare k-mers are given in Appendix B.1 which are quite simple but lengthy. For the basic sequence analyses in Section 3.1, here we elaborate the details on the selection of representative *k*-mers for the multi-modal spectra, description of the selected multi-modal *k*-mer spectra, and briefly describe the tools used for the plotting and extraction. The multi-modality of a human *k*-mer spectrum is not always looked very distinctively. When we plotted several *k*-mer distributions of *k* in between 7-to-25-mers, the multi-modality starts to appear at 7-mer and starts to fade from 11-mer which is the starting *k* where nullomers start to appear (see Sheet3 of Appendix C). As *k*-value gets larger, *k*-mer variants become quadrupled (4^{k+1} / 4^{k}) thus become more specific and lowering their frequencies into one forth (see Sheet1 of Appendix C). For higher *k*-values, most *k*-mers are nullomers and hapaxes, i.e. *k*-mers with zero and one frequency respectively (Castellini et al., 2012). Consequently, the overall *k*-mer multi-modal spectrum is turning into an exponential decay graph. Rare *k*-mers are always exists in the leftmost part of any *k*-mer

spectra but they become irrelevant in longer *k*-mer context (e.g. 13-mer onwards) since most of the longer *k*-mers are nullomers and hapaxes which are unfavourable for motif discovery study (Chan and Kibler, 2005) and the unique properties of rare *k*-mers inferred from the multi-modal spectrum are also diminished. We selected 8-, 9-, and 10-mer spectra to extract the rare *k*-mer datasets because these *k*-mer spectra give the most unique multi-modal spectra (see the list of rare 8-, 9-, and 10-mer datasets in Appendix C).

Chor et al. (2009) used 11-mer as the representative *k*-mer frequency distribution of the human genome because it is the starting *k* where nullomers start to appear. The length of *k*=8-to-10 are also in agreement with common lengths of regulatory motifs in between 5-to-15 bases (Das and Dai, 2007; Fickett and Hatzigeorgiou, 1997; Hariharan et al., 2013; Werner, 1999). From a study by Csűrös et al. (2007), we can see that the distribution of rare *k*-mers (CpG dominated *k*-mers) are concentrating in the lower tail while the distribution of frequent *k*-mers and repeat are at the opposite spectra side. Thus, there would be no significant changes to the statistical distribution of rare *k*-mers even after the repeats are removed. From our repeated runs of experiments in this work (the results are not shown due to the repetitiveness), the differences in results when using any of these three datasets are minimal because rare *k*-mers are overlapping with rare (*k+1*)-mers in genomic regions from where they were extracted and have almost the same distribution over the genomic regions as elaborated in Sheet2 of Appendix C.

The genome sequences for several organisms used to plot Figure 2 were downloaded from the NCBI RefSeq genome database (Pruitt et al., 2007). We wrote several Perl scripts to do sequence processing which are: to read the genome sequences and plot them into *k*-mer spectra, to extract the rare *k*-mer datasets, and to count the basic properties of the rare *k*-mers in certain genomic contexts. There are many automated tools exists on the web to model diverse *k*-mer DNA functional properties using various methods of sequence composition, vector feature, and pseudo component such as variations of PseKNC (Chen et al., 2015; Chen et al., 2014b; Chen et al., 2014c), Pse-in-One (Liu et al., 2015c) and repDNA (Liu et al., 2015a).

2.2 Comparative analyses using EpiGRAPH and UCSC browsers

For the correlation analyses in Section 3.2, we utilized several web-based genome analysis platforms of EpiGRAPH (Bock et al., 2005) and UCSC browsers (Rhead et al., 2010) respectively. We performed genome analyses using the EpiGRAPH with two objectives, i.e. to find possible correlations of the rare 8-mers with the EpiGRAPH built-in annotation of 896 human genomic attributes and to test if certain types of filtering done on the rare 8-mer datasets affect their correlation with the genomic attributes (see Appendix B.2 for details on the methods). We uploaded all of the rare 8-mer subsets and whole 8-mer dataset into the EpiGRAPH website and correlated each of them, one by one, against all of the built-in attributes, and collated their result in Table 3. The UCSC browsers were used to conveniently download and extract selected biological motifs, to compare the co-occurrence of the rare 8-mers with the motifs visually, and to perform Pearson pairwise correlation of the rare 8-mers with the motifs (see Appendix B.2 for details on the methods). The rare 8-mers and the selected motif datasets were uploaded into custom tracks of the UCSC genome browser and they were visually and computationally analysed to produce the results in Figure 3 and Table 4 respectively.

2.3 Dataset preparations for intrinsic analyses

In Section 3.3, we have done various empirical analyses on several sequence properties of the rare 8-mer dataset. Here we only describe the materials i.e. the required sequences. The methods will be explained in the description of corresponding figures and tables in Section 3.3. The empirical analyses required specific sub-class sequences of genes (i.e. pseudo, non-coding, coding, reviewed, and validated), of coding genes (i.e. intron, CDS, 5' UTR, and 3' UTR), of promoters (i.e. core, proximal, distal, CGI+, and CGI- promoters), and of CGIs (i.e. relax, strict, promoter+, and promoter- CGIs). The NCBI Reference Sequence (or RefSeq) is one of the best resources for obtaining various types of high-quality sequence and annotation datasets due to its extensive curation process (Pruitt et al., 2007). Most of the required datasets are not readily available from the NCBI-RefSeq. We generated the required sequences using our own developed Perl scripts which read the relevant data in the RefSeq annotation files of "seq_gene.md" and "seq_cpg_islands.md" and extract the sequences of the

annotated features from corresponding chromosome files of the human genome. Another human polymerase II promoter dataset was downloaded from the Eukaryotic Promoter Database (EPD) website at the following address (http://www.epd.isb-sib.ch/). Refer to Appendix B.3 for details on the dataset extraction procedures and Table B.3 for the basic statistics of the datasets. Most of the time, we only show the results of intrinsic analyses based on the rare 8-mer dataset. The selection of the rare 8-to-10-mer datasets and the similarities of their results were already explained in Section 2.1.

2.4 Rare-Word Clustering method and performance evaluation

Rare-word clustering (RWC) feature of the rare 8-mers is inferred as a strong representative signal for CGI and promoter features (see the introduction in Section 3.4). This RWC feature is modelled into a classifier, namely the RWC method, to classify the CGI and promoter features. We improvise the three conventional sequence parameters of CGI (i.e. CGI length, CG suppression ratio, and G+C%) introduced by Gardiner-Garden and Frommer (1987) and one non-conventional parameter of CGI (i.e. *p*-value of CG neighbouring distance) by Hackenberg et al. (2006) into three RWC search parameters (or criteria) which are minimum rare-word occurrence count (min_rwo_count), minimum rare-word cluster size (min_rwc_size), and maximum rare-word neighboring distance (max_rwn_dist). The first criterion is used to filter spurious feature (of CGIs or promoters) while the second and the third criteria are used as direct measurements of the RWC motif. By combining signals from both conventional and non-conventional methods, we hoped that they can compensate each other to yield better result. The RWC method predicts the CGI or promoter features by searching the human genome for DNA regions that satisfy these three RWC criteria. The RWC method was implemented in a Perl script (see Appendix D for the script) and is described as follows (see Figure 1 for the pseudocode):

1. Input the paths to a RKM dataset and chromosome files and input values for the RWC parameters of *max_rwn_dist*, *min_rwc_size*, and *min_rwo_count*. Read all RW positions in the input chromosomes into an array for fast searching later;

2. In line 1, start at the first RW position, set it as a new RWC cluster start. In line 2-4, try to expend the RWC cluster by setting the next encountered RW as the RWC cluster end, repeat the process until the next RW fails the *max_rwn_dist* criteria;

3. In line 7-9, when the RWC cluster stop growing, check for the other two RWC criteria of *min_rwc_size* and *min_rwo_count*. If success, accept the current RWC cluster (marked by the current RWC cluster start and end) as a predicted CGI (or promoter) feature and set the next RW searching after the current RWC cluster end. If fail (line 10-12), set the next RW searching after the current RWC cluster start.

```
Parameters:
max_rwn_dist = maximum distance to the next RW position
min_rwo_count = minimum RW count for a RW cluster
min_rwc_size = minimum size for a RW cluster

Inputs:
Paths to the RKM dataset and chromosome files
Values for min_rwo_count, min_rwo_count, and min_rwc_size parameters

Pseudo-code for the RWC classifier:
1   Foreach next encountered RW in a selected chromosome
2     Set it as the RWC cluster start
3     while next_rw_pos < max_rwn_dist
4       Expand the RWC cluster end
5     End
6
7     If the current RWC size > min_rwc_size && curr_rw_count > min_rwo_count
8       Accept the RWC cluster as a predicted CGI or promoter
9       Set the next RW searching after the current RWC cluster end
10    Else
11      Set the next RW searching after the current RWC cluster start
12    End
13  End

Output:
RWC cluster locations in the selected chromosomes
```

Figure 1: Pseudo-code for the RWC method.

Optimization is a process of searching for model that gives the best fitting to the data (see Appendix B.4.1). Manual parameter tuning is very tedious due to the RWC parameter search space is quite large (3 dimensions, each has range between 5 to 1000 values). We utilized the Particle Swarm Optimization (PSO) algorithm introduced by Kennedy and Eberhart (1995) which was compiled into a Perl module by Jaquiery (2011), to optimize the parameter values of the RWC method. PSO is a meta-heuristic method which makes no (or few) assumptions about the problem being optimized and it can search very large search space but does not guarantee an optimal solution to be found. It is well suited for our optimization problems with a large search spaces and there are unclear dependencies

between the parameters. We applied several adjustments to increase the PSO parallelization to cover the large search space which are small neighbourhood of 10%, higher exploitation using c_0=0.9, reinitialized zero speed particles, and quite large swarm of 20-30 particles. To avoid overfitting, we used several **generalization** strategies which are: 5-fold cross validation to decide on which rare *k*-mer dataset gives the best model; training (i.e. optimizing the RWC parameters) on chromosomes 1, 12, and 21; and testing on the whole human genome (see Appendix B.4.1). These strategies are done to ensure performance consistency, strategically reduce computational time, and to increase accuracy respectively. For the validation test, three resampling tests are usually used to extrapolate the success rate of a predictor which are independent dataset, sub-sampling (or *k*-fold cross-validation), and Jackknife (Chou and Zhang, 1995). Among the three, the Jackknife is deemed as the most objective and the least arbitrary as it always produces a unique result for a given dataset. The Jackknife test is too computational expensive for our problem due to the large search space, large validation datasets, and heavy computational requirement imposed by the RWC method and PSO optimization. We utilized the 5-fold cross validation test to optimize the model of the RWC method only as done by many others (Guo et al., 2014; Lin and Li, 2011; Lin et al., 2014; Liu et al., 2012).

Several evaluations were performed to assess the performance of the RWC method in predicting CGIs and promoters. Each evaluation is done by comparing an optimized RWC prediction dataset against a validation (gold standard) dataset using a specific protocol (see Appendices B.4.2 and B.4.3 for the details on the evaluation protocols and validation datasets). **For CGI evaluation**, we followed 4 protocols by Hackenberg et al. (2006 & 2010), each protocol utilizes a different validation dataset of the Weber hypo-methylated promoter (HMP), Illingworth unmethylated region (UMR), Alu repeat, and phylogenetic conserved (PhastCons) element (Illingworth et al., 2008; Jurka et al., 2005; Siepel et al., 2005; Weber et al., 2007). For the Weber dataset, we extracted 14,182 HMPs consists of 11,260 HMPs in two cell lines and 2,922 HMPs in at least one cell line. We also extracted 20,371 UMRs, ~1.2 million Alu repeats, and ~2.3 million PhastCon elements from their corresponding validation datasets (their basic properties are summarized in Table B.5 in Appendix B.4.3). The Hackenberg's approach uses a minimum one base overlap criterion to determine if the predicted and

validated regions are intersected or not. **For the promoter evaluation**, we adapted two protocols by Abeel et al. (2009), each utilizes two same validation datasets of the Carninci Tag Cluster (CTC) and RefSeq Gene Start (RGS) datasets (Carninci et al., 2006; Pruitt et al., 2007). We extracted ~181,000 CTCs and ~20,000 RGSs from their respective datasets (see Appendix B.4.3) and we use +/-1000 bp detection regions around the CTCs and RGSs to represent the promoters and called them as Carninci and RefSeq Transcription Start Regions (TSRs). **For the performance scores**, we only focus on the sensitivity (SN), positive predictive value (PPV), F-score (harmonic mean between SN and PPV), and correlation co-efficient (CC) value. Technically, certain performance scores can be increased at the expense of other scores by tweaking the parameters of a prediction program. Bajic et al. (2004) suggested that the balance between SN and PPV scores (i.e. the F-measure) is the most unbiased setting for a prediction program. In our opinion, the F-measure is selected due to it is not dependent on the TN criterion which is usually high in most of the CGI and promoter evaluation results. **Benchmarking** was done by comparing the evaluation results of the RWC method against 7 other evaluation results from 5 other similar programs (which use CGI related signals for prediction). We selected 5 recent CGI prediction programs which are CpGcluster (Hackenberg et al., 2006), CpGMI (Su et al., 2010), CpGProD (Ponger and Mouchiroud, 2002), NCBI CGI (Maglott et al., 2007), and UCSC CGI (Rhead et al., 2010) as described in Appendix B.4.4. Based on our experience, there are many factors affecting the evaluation scores (see Appendix B.4.5). It is quite confusing to directly compare our benchmark results with other published results which used different evaluation settings that produced asymmetrical results. Thus, we re-evaluate other prediction datasets using the same evaluation setting (see Section 3.4) and compare their evaluation results with our prediction results.

3.0 Results and Discussions

3.1 Re-evaluating several basic properties of multi-modal spectrum and the rare *k*-mers

In Section 1.0, we have summarized several unique features of the rare *k*-mers from an earlier study by Chor et al. (2009). Here, we plotted 10-mer frequency distributions of a few selected organisms in logarithmic scale to review the distribution of 10-mers in those genomes. Figure 2 reconfirms results from the earlier study where organisms from the Tetrapod clade exhibit multi-modal *k*-mer spectra

(only human, chimpanzee, mouse, cow, dog, and chicken genomes are plotted in Figure 2) while other organisms (only Takifugu and Zebra fish are plotted) exhibit unimodal spectra. The multi-modal spectra have similar *k*-mer distributions which form two modes at the lower spectra and form the third mode near the genome average frequency of 10-mers (see the calculation in the description of Figure 2). Another apparent attribute is the heavy tail component of the spectra which is still visible at higher logarithmic scale. The heavy tail is not entirely caused by the repeat elements, but also frequent words in large genomes, particularly for short *k*-mers (Castellini et al., 2012; Csűrös et al., 2007).

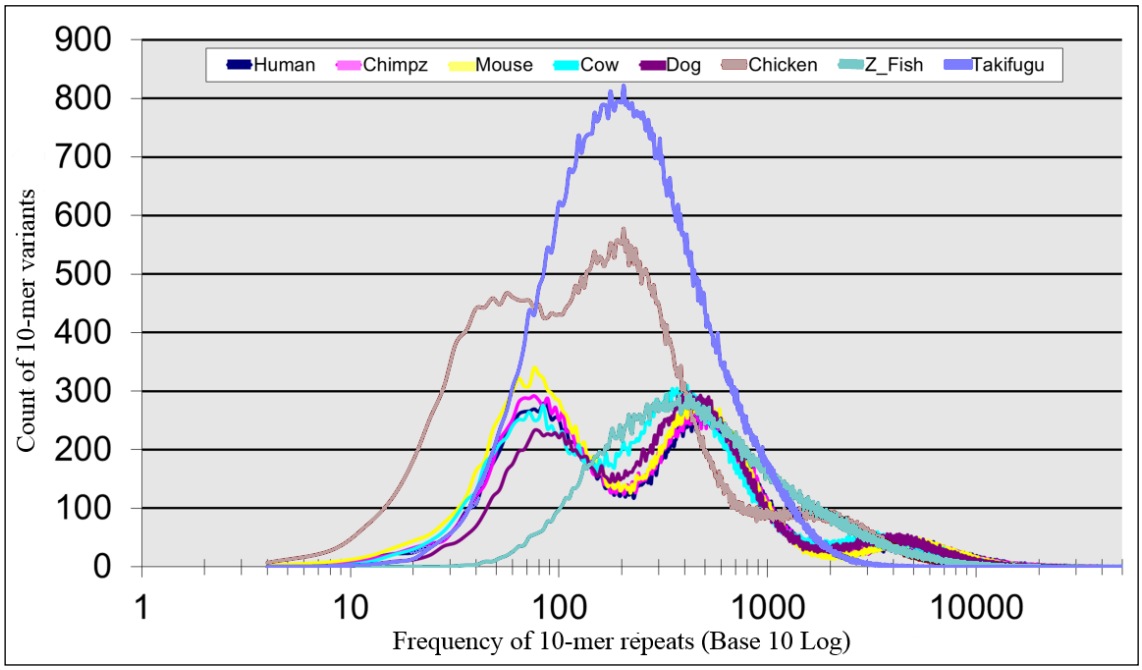

Figure 2: The 10-mer spectra of several organisms including mammals, a bird, and two fish.
The average 10-mer frequency in the human genome is equal to the total bases of the human genome divided by the total 10-mer variant (~3 billion bases / 4^10) which is ≈3000 occurrences for one 10-mer variant.

Several attempts have been done to model the statistical background for the multi-modal spectra using various distributions and approaches such as Bernoulli, copy/insert model, Pareto log normal (PLN), and Markov chain (Chor et al., 2009; Csűrös et al., 2007; Reinert et al., 2000). The Markov chain can model the multi-modal spectra of the human genome generally well although have inherent limitations to fully model the heavy tail and heterogeneity in the genome. Even at the first (low) order, the models can fit the multi-modal spectra well and thus dinucleotide frequencies are deemed sufficient to define such models (Chor et al., 2009).

To assess the impact of the sixteen dinucleotides in the 8-mer subsets, we tabulated each of the dinucleotide relative abundance within the R8MD1, R8MD2, and O8MD3 subsets and all 8-mer dataset. Table 1 shows the distribution of dinucleotide abundances in the all 8-mer dataset almost reflect the genomic signature of the human genome as reported by Burge et al. (1992) and Gentles and Karlin (2001) where CA-TG and CG are the highest and the lowest occurred dinucleotides respectively. Our calculated values are CC-GG and CG which is more accurate due to back then the human genome was less complete than now. The R8MD1 dataset is constrained by the over-represented of CG, AC-GT, and GA-TC dinucleotides while the R8MD2 dataset is constrained by the lesser over-represented of CG. Lastly, the O8MD3 dataset has very little 8-mers containing CG. The rest of the dinucleotides more or less are within the normal ranges which do not affect the dataset biasness.

Table 1: List of relative dinucleotide abundances in several 8-mer subsets and the whole 8-mer dataset. The measures of dinucleotide relative abundance are categorized by 8-mer subsets of R8MD1, R8MD2, O8MD3, and all 8-mer dataset. The measure is given by the following formula $\rho^*_{XY} = f^*_{XY}/f^*_X f^*_Y = 2(f_{XY} + f_{(XY)'})/(f_X + f_{X'})(f_Y + f_{Y'})$ where the ρ, *, and f denote measure of dinucleotide bias, concatenation with its complement, and fraction of nucleotide or dinucleotide (refer to Burge et al. (1992) for more details). Based on empirical studies, Gentles and Karlin (2001) use $\rho^*_{XY}<=0.78$ and $\rho^*_{XY}>=1.23$ to indicate under and over representations of dinucleotides. However we use stricter criteria <<0.78 (underlined) and >>1.23 (bold) to highlight the under and over represented dinucleotides in this table. The raw frequency data is kept in Sheet4 of Appendix C.

8-mers vs. nt	CG	GC	TA	AT	CC-GG	TT-AA	TG-CA	AG-CT	AC-GT	GA-TC
R8MD1	**2.3788**	1.0314	1.1058	0.9676	0.4095	1.0613	0.3234	0.3625	**1.5170**	**1.4997**
R8MD2	**1.5573**	0.9481	0.9434	0.9716	0.8506	1.1309	0.7720	0.7745	1.1274	1.1105
O8MD3	<u>0.0580</u>	1.0142	0.7425	0.8772	1.2805	1.1138	1.2373	1.1903	0.8252	0.9839
All 8-mers	<u>0.2358</u>	1.0199	0.7519	0.8849	1.2459	1.1211	1.2014	1.1574	0.8340	0.9821

The study by Chor et al. (2009) also showed that *k*-mers containing more CGs appear further to the left of the human 8-mer spectrum. We tabulated the CG counts in the first (abbreviated as Dist_1), the second (Dist_2), and the third (Dist_3) modes of the human 8-mer spectrum (see Table 2). This data clearly shows that the inclusion of CG(s) correlate with lower *k*-mer frequencies where 8-mers containing 2-CG, 1-CG, and 0-CG are dominating (made up more than 90% of) the Dist_1, Dist_2, and Dist_3 subsets respectively, which reaffirms the finding in the study.

Table 2: Distribution of 8-mers categorized by CG counts and modal groups.
Each mode of the human 8-mer spectrum is extracted based on particular frequency bands i.e. i) <2000; ii) 2000-18000; iii) >18000, and iv) all 8-mers. The extracted subsets are labelled as Dist_1, Dist_2, Dist_3, and Dist_all respectively.

CG Occurrence VS 8-mer freq. bands	Dist_1 (<2000)	Dist_2 (2k-18k)	Dist_3 (>18000)	All distributions
0 CG	0	396	40149	40545
1 CG	102	20501	865	21468
2 CG	2146	1168	52	3366
3 CG	104	50	2	156
4 CG	0	1	0	1
TOTAL	2352	22116	41068	65536

3.2 Correlations of rare *k*-mers to annotated motifs using Bioinformatics tools

In this experiment, we utilized the EpiGRAPH, an online bioinformatic tool, to correlate the rare 8-mer dataset with a wide range of genomic attributes and to test if particular rare 8-mer subsets have different degrees of correlation with the attributes. We partitioned the primary 8-mer dataset into nine different subsets using several discriminator types (and values) i.e. unimodal 8-mer datasets (of Dist_1 and Dist_2 modes), datasets of rare 8-mers with specific CG counts (of 1, 2, and 3 CGs), and datasets of rare 8-mers with specific frequency bands (of 0-500, 501-1000, 1001-1500, and 1501-2000 bands). All of the datasets were uploaded as the inputs to the EpiGRAPH, they were run one by one through the EpiGRAPH pipeline, and their performance summary table results were collated into Table 3. To explain the correlation of a particular rare 8-mer subset to a correlated attribute group, we looked-up the ranked attribute table associated with the rare 8-mer subset for individual attributes which have high significant *p*-values (refer to Appendix E on how to look up for the top individual attributes). Those significant individual attributes come with an attribute group label which can be used to associate it as a biological function to the correlated attribute group of a particular rare 8-mer subset.

Most of the rare 8-mer subsets have high correlations to the attribute groups of DNA sequence (expected due to CG enrichment in them), regulatory regions and epi-genome and chromatin structure groups. Among the top ranking item attributes (based on the *p*-values from corresponding ranked attribute tables) for the rare 8-mer subsets are: CG, CC, AG, and CA dinucleotides (from the sequence attribute group); bona-fide CGI (from the regulatory attribute group); and open chromatin

regions of Polymerase II overlapped and by several histone modifications of H3K4me2, H3K4me3, H3K9me1, H4K20me1, and H2A_Z (from the epi-genome and chromatin structure attribute group). All of the aforementioned histone modifications are related to active transcription as reported by Su et al. (2010). However, rare 8-mers have imperfect correlations to the aforementioned regions (see Appendices F.1 and F.2). In addition, we did three follow-up experiments to analyse rare *k*-mer epigenetic properties (see Appendix F.3). Although rare *k*-mers are lack of methylated CGs which are similar to CGI regions, their total methylated CGs are almost double than CGIs which shown lesser correlations of rare *k*-mers with epigenetic attribute than CGI. Based on the aforementioned highly correlated attribute groups and corresponding top ranked item attributes, in general, rare *k*-mers are likely to be functional motifs related to active transcription site and regulatory regions (promoter).

Table 3: Mean correlation coefficients between rare 8-mer subsets and attribute groups of EpiGRAPH.
Mean of correlation coefficients (CC) of several clustered rare *k*-mer datasets with the 9 attribute groups of the EpiGRAPH. CC takes a value between -1 to +1 where +1, 0, -1 represent perfect prediction, random, and total disagreement respectively. The average of CC of the 10-fold cross validations is taken as the MCC. The MCC indicates how well a genomic attribute group discriminate between positive and negative inputs using ML methods and 10-fold cross validations (Bock et al., 2009). MCC > 0.3 (*italic & bold*) can be considered as relevant and MCC > 0.6 (bold) indicates a strong correlation.

Clustering methods vs. Genomic Attribute Groups	Modal filtering on all 8-mers dataset		CG count filtering on 8-mer Dist_1 subset			Frequency band filtering on 8-mer Dist_1 subset			
	Dist_1	Dist_2	1-CG	2-CGs	3-CGs	0-0.5k	0.5-1k	1-1.5k	1.5-2k
DNA Structure	0.29	0.28	0.24	0.17	0.54	0.26	0.26	0.30	0.27
Repetitive DNA	0.11	0.05	0.10	0.12	0.09	0.09	0.07	0.07	0.02
Chromosome Organization	0.05	0.13	-0.02	-0.01	0.02	0.10	0.05	0.10	0.08
Evolutionary History	0.03	0.05	0.01	0.00	0.24	0.01	0.06	0.06	0.02
Population Variation	0.09	0.10	0.08	0.08	0.01	0.06	0.08	0.09	0.05
Genes	0.13	0.11	0.01	0.01	0.16	0.09	0.11	0.07	0.01
Regulatory Regions	0.31	0.29	0.11	0.10	0.50	0.31	0.33	0.31	0.27
Transcriptome	0.10	0.09	0.01	-0.05	0.27	0.09	0.08	0.11	0.05
Epigenome and Chromatin structure	0.72	0.72	0.80	0.83	0.64	0.78	0.75	0.68	0.65

Next, we utilized the UCSC browsers to visually and computationally analyze the spatial distribution of rare 8-mers in annotated genomic features of the human chromosome 21 (refer to Appendix B.2 for details on the applied steps). Figure 3 shows the rare 8-mers mostly occur at higher density within the strict UCSC-CGI and RefSeq promoter regions and occur at significantly lower

density in the rest of other visible regions. This visual observation reaffirms the previous finding where the rare 8-mers have high associations with strict CGI and promoter regions.

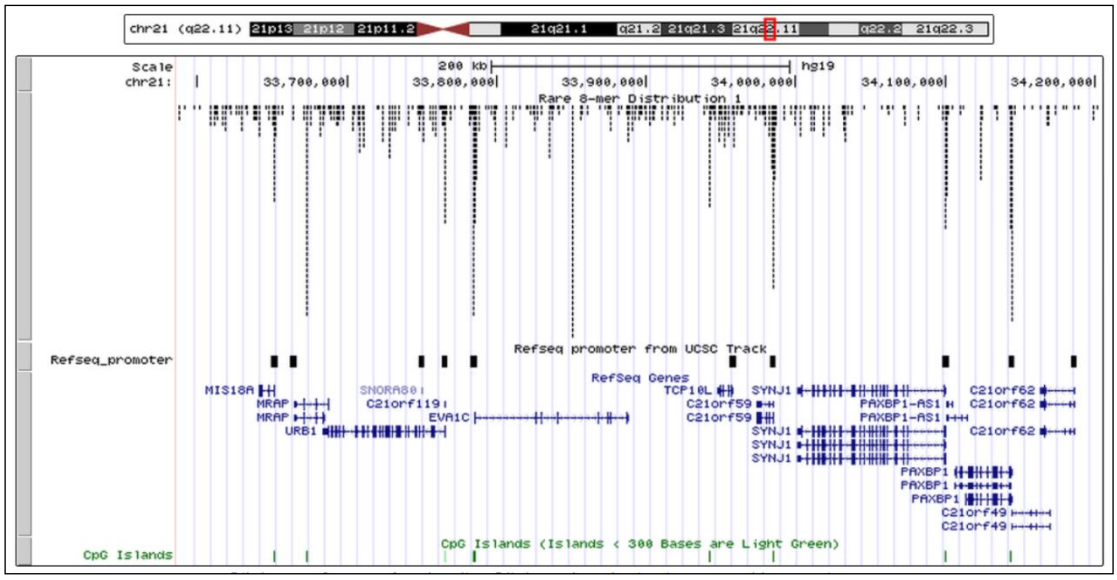

Figure 3: Distribution of the rare 8-mers in a gene rich section of the human chromosome 21.
The rare 8-mer frequency track is shown in parallel to RefSeq promoter, RefSeq gene, and UCSC-CGI tracks using the UCSC genome browser. The RefSeq gene track was been utilized to extract +/- 2000 base regions extending from the RefSeq gene start (RGS) locations and named it as the RefSeq promoter track.

However, the associations of the high density rare 8-mer regions with the strict CGI and promoter regions does not occur every time (refer to Figure 3). To get a more accurate measure for the associations, we utilized a correlation tool in the UCSC table browser (refer to Appendix B.2 for details on the applied steps) to calculate the standard Pearson pairwise correlation coefficient (i.e. the *r*-score) between the rare 8-mer, strict CGI, and promoter datasets. The *r*-score represents how well parallel pairwise items in two datasets intersected together. Table 4 shows the top 3 *r*-scores of 0.4981, 0.2940, and 0.2268 (bolded) between paired datasets of R8MD1 <=> strict-CGI, strict-CGI <=> promoter, and R8MD1 <=> promoter respectively. This result also indicates that R8MD1 has higher specificities to the strict-CGI and promoter, when compared to the R8MD2. Therefore, we focus our experiments on the R8MD1 dataset for the rest of this study. We also correlated the R8MD1 dataset to other built-in UCSC tracks but we only found moderate correlations (slightly >0.1) between the former to several histone modification and DNaseI hyper sensitive site tracks (results not shown).

Table 4: Shows Pearson correlation coefficients between pairwise datasets of several genic features in the chromosome 21 of the human genome hg18.
The listed coefficients are computed using the Pearson correlation function in the UCSC table browser. The top three scores are bolded. A score from 0 to 1 indicates a positive correlation between pairwise dataset where 1 represents the highest correlation score.

R8MD2	0.1921							
Strict CGI	**0.4981**	0.1678						
Genes	0.1137	-0.0272	0.0553					
Promoters	**0.2268**	0.0766	**0.2940**	0.1225				
5‘ UTR	0.1393	0.0459	0.1744	0.0907	0.1858			
CDS	0.1569	0.0671	0.1097	0.1658	0.0844	0.1331		
Introns	0.0790	-0.0413	0.0238	**~~0.9767~~**	0.1015	0.0250	0.0390	
3’ UTR	0.0575	0.0171	0.0307	0.1388	0.0416	0.0546	0.0798	0.0050
r_{xy}-score	R8MD1	R8MD2	s-CGI	Genes	Prmtrs.	5’UTR	CDS	Introns

3.3 Intrinsic analyses on the correlated motifs of the rare 8-mers

Empirical analysis provides an effective mean to find unique sequence patterns within a certain genomic context (or feature) which can lead to the discovery of novel sequence motifs. Based on the high correlations of the rare 8-mers with gene, promoter, and strict CGI features in the previous section, we performed more detailed intrinsic analyses on the subclasses of these features. A recent study by Hackenberg et al. (2012) has shown that the word clustering (given by enrichment ratio) can be used to indicate associations of *k*-mer words to biological functions (i.e. the associated features). Calculation for the enrichment ratio, as defined by the paper, is provided in the description of Table 5. The results show that the rare 8-mers are significantly enriched ($r >> 2$) in several genomic features of strict CGIs, 5’ UTRs, relaxed CGIs, promoters, and CDS. Apart from the enrichment ratio, we also need to consider the sequence coverage (see Column 2 in Table 5) because our next target is to implement this clustering property in a general prediction method which should cover as much items in the feature dataset. Considering both of the enrichment score and the feature coverage, only the CGI and promoter features (see the bold scores) are suitable candidates. Although the rare 8-mers have high enrichment ratios in the 5’ UTR and CDS features, but they only have a moderate sequence coverage which is less than 50%. This could mean that some of the 5’ UTR and CDS features might overlap with promoter regions or there might be different subclasses of the aforementioned features due to the high density of the rare 8-mers in subsets of them.

Table 5: Shows enrichment ratios of the rare 8-mers in selected genomic features.
When associating the rare 8-mer dataset to a biological sequence dataset, every sequence in the dataset was processed one by one. Each sequence is analysed using 1-nt step of an 8-mer sliding window, to search for any rare 8-mer occurrences in the sequence. For Column 2, one sequence is counted if it contains at least one rare 8-mer word within it. Column 3 gives the total occurrence of rare 8-mer words in a particular dataset. For Column 5, we calculate the enrichment ratio of each dataset by dividing the density of the rare 8-mer words inside a dataset over its density outside the dataset. The rare 8-mer density inside a dataset is given by the total rare 8-mer frequency inside a dataset divided by the dataset size. The rare 8-mer density outside a dataset is given by the total rare 8-mer frequency in the human genome (2643351 rare-words) excluding the total rare 8-mer frequency inside the dataset, divided by the human genome size (2861327216 nt. w/o Ns) minus the dataset size. Bold scores represent significantly enriched ratios and italic scores represent significantly enriched ratios, but limited feature coverage (Column 2).

Genomic Feature Datasets	Count of sequences containing rare 8-mers (fraction to total seq.)	Total rare 8-mer frequency	Size of feature dataset (bp)	Enrich. Ratio
RefSeq human ncRNA genes v37	1072/1275 (0.841)	44761	37026286	1.31
RefSeq human pseudo genes v37	927/1598 (0.580)	15396	11604949	1.44
RefSeq human mRNA genes v37	16220/16811 (0.965)	1215004	1043613703	1.48
- subset of 3' UTRs only	6370/18869 (0.338)	32556	22684295	1.56
- subset of 5' UTRs only	12266/26898 (0.456)	85100	4720645	*20.13*
- subset of intronic regions only	80995/170415 (0.475)	866589	986231212	0.93
- subset of CDS regions only	47812/176797 (0.270)	224207	29977551	*8.75*
- subset of promoter regions only	15861/16811 (0.943)	516857	67260811	**10.10**
RefSeq human strict CGIs v37	23750/25218 (0.942)	622813	27442917	**31.83**
RefSeq human relaxed CGIs v37	151500/315801 (0.480)	1165792	123748620	*17.45*
EPD human Pol II promoters v107	8271/8513 (0.972)	284828	34060513	**10.02**

Table 6 shows rare 8-mer enrichments in more specific subclasses of promoter and CGI features, i.e.: CpG+ and CpG- promoter regions (i.e. CpG-rich and CpG-poor promoter regions); several sub-regions of the +/- 2000-base promoter regions (also known as core, proximal, and distal promoters); and promoter+ and promoter- CGI regions (i.e. full-length CGI regions with and without overlapping to any promoter regions). This follow up investigation is important to further clarify the rare 8-mer associations in subclasses of them as mammalian promoters are commonly divided into CpG-rich and CpG-poor promoters (Saxonov et al., 2006). If we observe the top three (bold) scores in Table 6, the CGI+ core promoter regions have the highest score, followed by the promoter+ strict CGIs, and then the CGI+ proximal promoters. Concluding, this result shows that the rare 8-mer density is a very strong signal to predict strict CGIs as well as promoters.

Table 6: Shows enrichment ratios of the rare 8-mers in sub-regions of CGI and promoter regions.
Refer to the description in Table 5 for more details on the scores. Bold scores represent significantly enriched ratios. Underline percentages in Column 2 represent high sequence coverage.

Genomic Feature Datasets	Count of sequences containing rare 8-mers (fraction to the total sequences)	Total rare 8-mer frequency	Size of feature dataset (bp)	Enrich. ratio
RS human CGI+ promoter regions v37	13368/13513 (0.989)	537349	54077516	**13.25**
- core regions +/- 0-0.1 kbp from RGSs	11271/13513 (0.834)	80776	2716716	**33.17**
- prox. regions +/- 0.1-0.5 kbp from RGSs	12714/13513 (0.941)	222185	10839832	**24.13**
- distal regions +/- 0.5-2 kbp from RGSs	12848/13513 (0.951)	232913	40575032	6.72
RS human CGI- promoter regions v37	6337/7847 (0.808)	59569	31403849	2.08
- core regions +/- 0-0.1 kbp from RGSs	2135/7847 (0.272)	9386	1577649	6.46
- prox. regions +/- 0.1-0.5 kbp from RGSs	3727/7847 (0.475)	19533	6294898	3.38
- distal regions +/- 0.5-2 kbp from RGSs	5399/7847 (0.688)	30650	23562698	1.41
RefSeq human promoter+ strict CGIs v37	13097/13573 (0.965)	400011	16072443	**31.57**
RefSeq human promoter- strict CGIs v37	10653/11645 (0.915)	222802	11370474	**23.07**

The previous results in Table 6 shows the rare 8-mers have gradual enrichments in distal, then proximal, and lastly the core promoters. Here, we analysed the rate of rare 8-mer enrichments throughout smaller bins covering the whole promoter regions. Figure 4 shows that the rare 8-mer frequencies have a slightly sharp increase which peaked at the gene start bin (regions of+/-100 bp from the RGSs), followed by a gradual fall (this graph gives the general occurrences of rare 8-mers in more than 20,000 promoters). The peak looked almost symmetry with a bit skewed towards the upstream (right) promoter bins which could explain the rare 8-mer enrichments in subsets of the 5' UTR and CDS in Table 5. Nevertheless, the rare 8-mer density peaks at the core promoters. Accurate promoter prediction at the resolution of core promoter regions have become a new goal for recent promoter prediction studies and the dominance of the CGI signal have limit their performances (Zeng et al., 2009). Using the rare 8-mer high density signal (which is smaller, ~200-base) instead of the CGI signal (in average is ~1000-base) might be more effective for the purpose.

We also performed the same experiment on the CGI+ and CGI- promoter subsets. Both of them produced a similar result to Figure 4 but with a different magnitude of rare 8-mer frequencies (result not shown). The ratio of the total rare 8-mer frequencies in the CGI+ to the CGI- promoter subsets is 9:1 although their sequence count ratio is only 2:1 (the CGI+ and CGI- datasets constitute of 12,714 and 7,561 promoters respectively). The disparity distribution of the rare 8-mer frequencies

in both subsets reaffirms the result in Table 6 where the CGI+ promoters have significantly enriched rare 8-mers when compared to the CGI- counterparts.

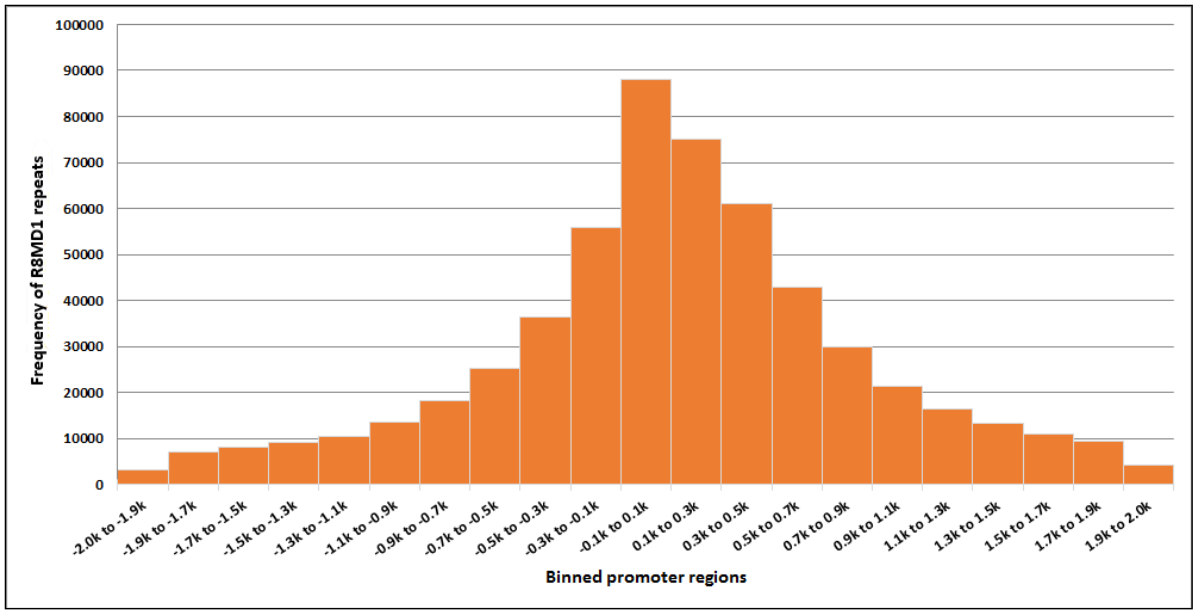

Figure 4: Shows rare 8-mer frequencies in binned-promoter regions.
Promoter regions extending +/- 2000 bases extending from the RGSs with the status of validated and reviewed were extracted (refer to Appendix B.3 for the extraction procedure). Each promoter region was binned into segments of 200 bases. The total of rare 8-mer frequency in each bin was plotted. The total R8MD1 frequency in all bins is 505,609.

Next, we searched the full promoter dataset (regions extending +/- 2000-bases from the RGSs with the RefSeq status of reviewed and validated) for certain unique topological patterns of the rare 8-mer Dist_1 (R8MD1) and Dist_2 (R8MD2) datasets to elucidate any novel sequence motifs. The core and proximal promoter regions are packed with regulatory sites which are necessary for the transcription and regulation of the associated genes (Birney et al., 2007; Werner, 1999). We discovered three unique topological configurations of the rare 8-mers in these promoters. In the first configuration (see Table 7 for some of the result), we found 301 R8MD1 variants (Column 1) which occur at a same relative position from the associated RGSs (Column 2, there are 356 of such positions) in at least five different promoters which are represented by their associated GeneIDs (Column 3, there are 645 unique GeneIDs of these promoters). For the R8MD2 dataset (result not shown), there are 4181 R8MD2 variants which occur at a same relative position (there are 26632 of such positions) within at least five different promoters (there are 14119 unique GeneIDs of these promoters).

Table 7: Examples of a R8MD1 variant occurring at a same relative position in several promoters.
Column 1 lists the R8MD1 variants that met the first configuration criteria, Column 2 lists the shared relative positions of the R8MD1 variants, and Column 3 lists the GeneIDs of the associated promoters. In the second column, we used relative positions from the RGS locations instead of the exact chromosomal position

R8MD1 variant	Relative pos. from the RGS	List of GeneIDs (represents promoters) containing R8MD1 at the same relative position.
CGCGCGGA	+15	11116, 4771, 8645, 7544, 2324, 84245, 4929
CGCGCGGA	+71	5426, 80727, 114569, 55568, 6833
GACGTCGA	-712	441324, 728430, 441326, 728753, 441327, 441314, 441328, 441315, 645651
CCGCGACG	+520	728082, 728042, 643311, 728062, 728096, 728049, 728075, 728090, 653282, 728072, 57804, 255313, 728036
ATGCGCGC	+7	79029, 79187, 54859, 286554, 25829

In the second configuration (see Table 8 for some of the result), we found 107 R8MD1 variants (Column 1) which are repeated more than five times (Column 3, list of the relative positions of the variant repeats) within the same promoters (Column 2, there are 63 unique GeneIDs for these promoters). For the R8MD2 dataset (result not shown), there are 1701 R8MD2 variants which are repeated more than five times within the same promoters (there are 940 unique GeneIDs for these promoters).

Table 8: Examples of a R8MD1 variant repeating several times within a promoter.
Column 1 lists the R8MD1 variants that met the second configuration criteria, Column 2 lists the GeneIDs of the associated promoters, and Column 3 lists the positions of the R8MD1 variant repeats. In the third column, we used relative positions from the RGS locations instead of the exact chromosomal position.

R8MD1 Variant	List of GeneIDs	List of the relative positions of the R8MD1 repeats from the RGS
AGTCCGCG	54963	-138, -107, -76, -45, -14
CGGGTCGA	729121	-1738, -1648, -1558, -1238, -1033, -738, -648, -558, -468
CGACCGAC	4161	-1626, -1622, -1618, -1614, -1610, -1593
CGTACCGA	100463500	-1723, -1705, -1687, -1669, -1651, -1633
TAGACGCG	54102	+478, +568, +598, +658, +688, +718

In the third configuration (see Table 9 for some of the result), we found 43 R8MD1 variants (Column 1) which are repeated more than three times with a same interval (Column 3, list of the relative positions of the variant repeats) within the same promoters (Column 2, there are 25 unique GeneIDs of these promoters). For the R8MD2 dataset (result not shown), there are 771 R8MD2 variants which are being repeated more than three times with a same interval within the same promoters (there are 259 unique GeneIDs of these promoters).

Table 9: Examples of a R8MD1 variant repeating several times with a same interval within the same promoters.
Column 1 lists the R8MD1 variants that met the third configuration criteria, Column 2 lists the GeneIDs of the associated promoters, Column 3 lists the R8MD1 variant repeat positions (in the bracket is the shared R8MD1 variant repeat interval that met the criteria). In the third column, we used relative positions from the RGS locations instead of the exact chromosomal position.

R8MD1 variant	List of GeneIDs	List of the relative positions of the R8MD1 repeats from the RGS [shared interval]
CGGGTCGA	729121	-1738, -1648, -1558, -1238, -1033, -738, -648, -558, -468 [90]
ACGCGTCT	4585	-319, -301, -205, -187, -169 [18]
CACGCGTC	4585	-320, -302, -206, -188, -170 [18]
CGTCGGGA	729121	-1897, -1782, -1692, -1512, -1397, -1282, -1192, -872, -782, -692, -602, -512 [90]
CGCGACAG	100499194	+37, +113, +187, +261, +335 [74]

Although the numbers of promoters (represented by the total unique GeneIDs) associated with the three topological configurations of R8MD1 (i.e. 645, 63, and 25) and of R8MD2 (i.e. 14119, 940, and 259) are relatively small in comparison to the total of 19621 promoters, such unique R8MD1 and R8MD2 configurations are unlikely to occur by chance due to the following reasons: 1) Only promoters with the RefSeq status of reviewed and validated were used; 2) A motif analysis study done by Zeng et al. (2008) on 1871 promoters from the EPD dataset discovered that only small percentages of the well-known TATA-, Inr-, and DPE-motifs (i.e. 6%, 9%, and 0.4% respectively) occur within those promoters; 3) Contiguous short repeat intervals and short oligonucleotides of the R8MD1 variants do not fit the repeat element criteria; 4) Location of a promoter motif is very crucial for its functioning (Werner, 1999). Exact location and shared interval repeats within promoter regions are not random events; 5) In most instances, the rare 8-mer variants occur within the core and proximal promoter regions which are populated with regulatory elements and are considered to be deficient of repeats and junk elements; and 6) Several small, medium, and tandem-sized repeats, within promoter and genic regions, were reported to have association with regulatory functions such as enhancer and suppressor (Cherrington and Mocarski, 1989; Eskdale and Gallagher, 1995; Tarlow et al., 1993). Thus, we deduce that the rare 8-mer configurations are indeed novel sequence motifs with (yet) unknown functions, not just a by-product of the well-known CG dinucleotide suppression in mammalian genomes (refer to Appendix G for the full lists of the three unique configurations).

3.4 Rare-word Clustering method for prediction of CGI and promoter regions

Previously, we have correlated the rare 8-mer (of Dist_1) dataset with CGI and promoter features (see Section 3.2) and elucidated that the rare 8-mers are highly enriched and clustered in them (see Section 3.3). In relation to that, a study by Hackenberg et al. (2012) also concluded that highly clustered 8-mer words (many rare 8-mers are included the top 200 clustered words of their result) are significantly associated with human exons and TFBS. Here, we demonstrate that rare 8-mer clustering feature, which is incorporated into the RWC method, is also competent in predicting the CGI and promoter. The RWC method searches the human genome for DNA regions with high densities of the rare 8-mers using a heuristic clustering approach to identify the CGI and promoter (see Section 2.4 for the RWC pseudo code). In evaluating the CGI and promoter prediction datasets, there are various issues of the prediction datasets, and corresponding validation datasets and evaluation protocols. We only mentioned issues which are important to this section. Readers can refer to Appendix B.4 for more details. For this section, first we discuss the validation of RWC results. Then, we review CGI prediction in general and discuss our CGI evaluation results and finally followed by similar approach (of the CGI) for the promoter.

Validation tests were done based on four main validation datasets and on chromosome 1 only to determine the best RWC model for each of the validation datasets (which represent the evaluation protocols). One of the RWC method settings is the use of a certain rare *k*-mer dataset. Rare *k*-mer datasets of *k*=6-to-13-mers were inputted one by one into the RWC method, other parameters were optimized using the PSO algorithm, and the optimizations were generalized using the 5-fold cross validation test (see Section 2.4). Figure 5 gives the results of the aforementioned procedures for the validation datasets of Carninci TSR, Weber HMP, RefSeq TSR, and Illingworth UMR. Their F-score ranges are 42-48%, 51-56%, 59-64%, and 67-72% respectively. In general, higher *k*-value for the rare *k*-mer dataset translates into higher F-scores for the RWC method. There is a need to run the test on higher *k*-values than 13 but we are handicapped at memory limitation and computational time to complete the procedures. Thus, we stay with the rare 13-mers as the best model for the RWC method.

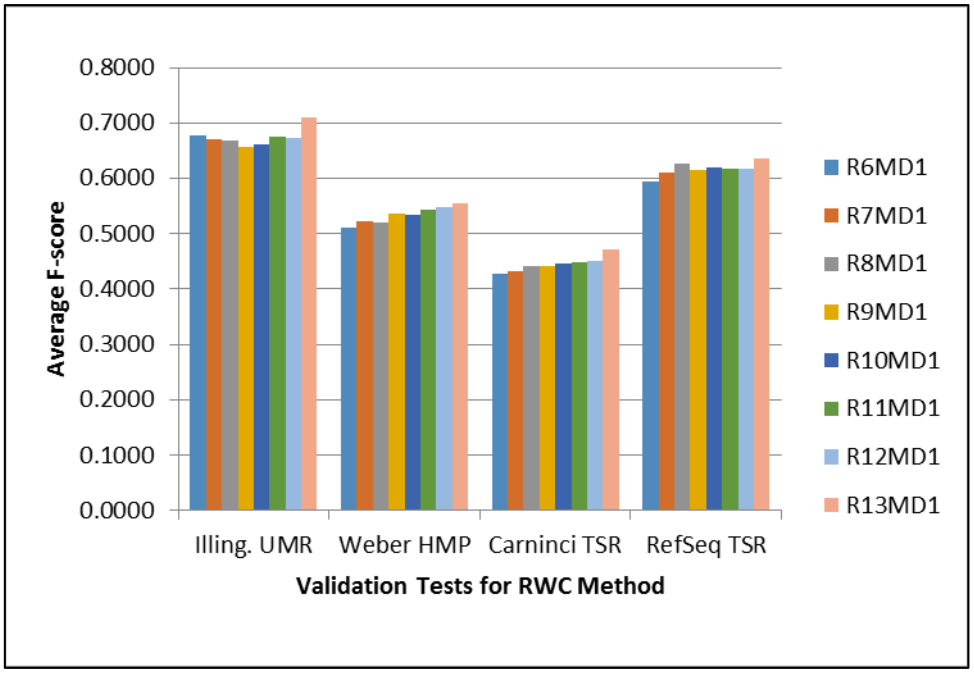

Figure 5: Histograms of average F-scores for various rare *k*-mer datasets inputted into the RWC method and evaluated using 5-fold cross validation test.

Analysis of CGI regions has become a hot topic in the post genomic era (Chae et al., 2013; Karlin, 2005; Takai and Jones, 2002; Zhao and Han, 2009). CGI plays many important roles in cells, such as gene silencing, alternative promoter, tissue specific control, and regulatory dense regions (Carninci et al., 2006; Deaton and Bird, 2011; Illingworth and Bird, 2009). CGI also has very distinct topological properties where it coincides with open chromatin regions and a majority of promoters (Hackenberg et al., 2006; Ioshikhes and Zhang, 2000; Larsen et al., 1992; Ponger and Mouchiroud, 2002). In terms of sequence characteristics, CGs inside much the smaller regions of CGIs occur at the expected frequency and are mostly un-methylated (which comprise of 20% of all CGs), whereas CGs inside the much larger regions of the human genome occur at only one fifth of the expected frequency and are mostly methylated (which is about 80% of all CGs). CGIs can be identified computationally by searching the human genome for DNA regions with a high density of CGs which is significantly differs from bulk DNA regions. Several algorithms have been developed to predict CGIs where they can be categorized into two types: 1) Algorithms dependent on the three CGI parameters, i.e. CGI length, CG observed/expected ratio, and G+C percentage, such as an algorithm by (Gardiner-Garden and Frommer, 1987), an algorithm by (Takai and Jones, 2002), and CpGProD by (Ponger and Mouchiroud, 2002); and 2) Algorithms which are not relying on the three conventional parameters

such as CpGcluster by (Hackenberg et al., 2006), CG Cluster by (Glass et al., 2007), and CpG_MI by (Su et al., 2010). Our RWC method also belongs to the second category. CGIs can also be identified using wet-lab experiments by searching for un-methylated sites. There are certain advantages and limits associated with either computational or experimental method, as mentioned in the following papers (Hackenberg et al., 2010; Han and Zhao, 2009; Illingworth et al., 2008; Illingworth and Bird, 2009).

Predicted CGI datasets were generated using all of the aforementioned prediction algorithms and were evaluated together with the RWC method for benchmarking (see Column 2 in Table 10). For the evaluations of the CGI prediction datasets, 4 validation datasets were used (see Appendix B.4.2), i.e. two experiment CGI datasets (by Illingworth and Weber) were used for the evaluations of their accuracy and two empirical CGI markers (of Alu repeat and PhastCons) were used for the evaluations of their quality (see Row 1 in Table 10). Experiment datasets are generally considered as more accurate than prediction datasets (in term of their quality and support of biological evidences) and are usually used as validation datasets to evaluate the accuracy of the latter. The Alu repeat is one of the repeat elements which are abundant in the human genome which resemble relaxed CGI regions and are considered as a major problem for conventional CGI detection methods (Takai and Jones, 2002), whereas the PhastCons elements are assumed to be functional genomic elements due to their highly conserved property (Siepel et al., 2005). Less overlapping with Alu repeats and greater overlapping with PhastCon elements indicate better qualities for computationally predicted CGIs (Hackenberg et al., 2010). Table 10 shows the performance scores and coverage percentages of the four evaluation results for eight prediction datasets. The RWC method is ranked at the third place when it was evaluated using the Illingworth validation dataset, at the first place for the Weber evaluation, at the second place for the lowest Alu coverage (the UCSC CGI dataset was excluded because repeats were pre-removed), and at the second place for the highest PhastCons coverage. Consistently, the RWC method ranked among the top three in term of accuracy and quality of the predicted CGIs among the benchmarked datasets. This indicates that the rare 8-mer density feature is a competent signal for

predicting CGIs. Detailed CGI evaluation results and basic statistics of all the datasets are provided in Appendix H.

Table 10: Evaluation and benchmark of predicted CGIs by the RWC method.
Predicted CGI regions from several algorithms were evaluated against validation datasets of Weber hypo-methylated promoters, Illingworth un-methylated regions, Alu repeats, and PhastCons elements using a search criterion of at least 1-base overlap. The SN, PPV, and F-measure scores stand for Sensitivity, Positive Predictive Value, and Harmonic Mean (between Sensitivity and PPV) respectively (refer to Appendix B.4.1 for more details on the scores). The method with higher scores (or percentages) is considered to have better performances except for Alu coverage where a lower percentage is considered as better.

	Validation Dataset (Protocol):	Illingworth UMR hg18 (overlap)			Weber HMP hg17 (1000bp dist)			Alu hg18 (coverage)	PhastCons hg18 (cvrg.)
#	Prediction Vs. Score	SN	PPV	F	SN	PPV	F	Percent	Percent
1	RWC method	0.6292	0.7091	0.6668	0.5725	0.5267	**0.5486**	1.62	16.62
2	CG-cluster (relaxed)	0.9341	0.1483	0.2559	0.7966	0.0546	0.1022	29.99	10.91
3	CG-cluster (strict)	0.7176	0.6702	**0.6931**	0.6283	0.3917	0.4826	**0.94**	15.60
4	CpGProD	0.8661	0.2068	0.3339	0.7774	0.1530	0.2556	21.96	9.42
5	CpG_MI	0.8431	0.3809	0.5247	0.7431	0.2831	0.4100	8.08	11.86
6	NCBI-CGI (relaxed)	0.8980	0.0576	0.1083	0.8188	0.0407	0.0776	40.62	7.82
7	NCBI-CGI (strict)	0.7250	0.6527	0.6870	0.6918	0.3847	0.4944	5.20	15.13
8	UCSC-CGI	0.7852	0.5321	0.6344	0.7213	0.3867	0.5035	-	**19.48**

Promoter is an integrated and an upstream part of a gene which regulate its expression. Before we can predict a promoter, we need a computer model to define what constitutes of promoters. Promoter modelling is a very difficult task due to the diversity and complexity of eukaryotic promoter elements, their organization (the promoter modules), and to what extent both of them constitute a functional promoter (Werner, 1999). There is probably unique promoter models as many as the gene number (Antequera, 2003). Many sequence features have been used to identify the location of promoters in DNA. The basic principle is promoter sequences have different properties than other DNA sequence such as: overlapping with CGI regions; contains any known core promoter elements; have higher density of any known TFBS, conserved motif, and over-represented motif; adjoining to mRNA transcripts; have a different compositional bias than non-promoter; and have sequence with structural dependency related to transcription (Bajic et al., 2004). Despite many features have been used, CGI is still the most dominant signal for predicting promoter regions of mammalian genomes (Abeel et al., 2009; Glass et al., 2007; Hannenhalli and Levy, 2001; Ioshikhes and Zhang, 2000). Due to the dependent of most promoter prediction programs on the CGI signal, their predictive

performances since the past decade are limited up to 60% detection only (Abeel et al., 2009; Bajic et al., 2004; Zeng et al., 2009).

Next we demonstrate that the rare 8-mer clustering feature is also capable to predict promoter due to its high correlation with CGI (see Section 3.2) where CGI is the best signal for predicting promoters. We used the same predicted CGI datasets in Table 10 to represent identified promoter datasets by CGI signals and they were evaluated together with the predicted promoter by the RWC method for benchmarking (see Column 2 in Table 11). Two bin-based and two distance based protocols by Abeel et al. (2009) were used for the promoter evaluations. Each protocol specifically defines how to classify predicted datasets into a correct or false promoter category based on two validation datasets of Carninci and RefSeq TSRs (see Appendix B.4.2 for more details). We used detection boundary of +/- 1000 bases extending from the validated TSRs (i.e. the accuracy level of proximal promoter) to have higher SN and PPV for comparisons. Table 11 shows the performances of the RWC method in comparison to seven other programs (that use CGI related signals) in predicting TSRs and promoters. The result shows that in terms of F-score, the predictive performances of the RWC method ranked at the first places for the evaluations using Carninci TSR and at the third place for the evaluations using RefSeq TSR. The high performances of the RWC method show that the rare 8-mers is a competent signal, comparable to the renowned CGI signal, in predicting promoter. Detailed promoter evaluation results and basic statistics of all the datasets are provided in Appendix H. We only utilized the RKM signals in our promoter prediction endeavours (which are signals rich features), we obtained similar results with the other programs. Although the accuracy is not very high, it is the state of the art of current promoter prediction programs.

Table 11: Evaluation and benchmark of predicted promoters by the RWC method.
Predicted promoters from several methods were evaluated against validation datasets of Carninci and RefSeq TSRs using detection boundary of +/- 1000 bases from the validated TSRs. The SN, PPV, and F-measure scores stand for Sensitivity, Positive Predictive Value, and Harmonic Mean (between Sensitivity and PPV) respectively (refer to Appendix B.4.1 for more details on the scores). The method with higher scores is considered to have better performances.

	Validation Dataset (Protocol):	Carninci TSRs (bin-based)			RefSeq TSRs (bin-based)			Carninci TSRs (distance-based)			RefSeq TSRs (distance-based)		
#	Predictions Vs. Scores	SN	PPV	F	SN	PPV	F	SN	PPV	F	SN	PPV	F
1	RWC method	0.1738	0.4608	**0.2524**	0.4129	0.5249	0.4622	0.3618	0.6410	**0.4625**	0.5372	0.6710	0.5967
2	CG-cluster (relaxed)	0.2443	0.1813	0.2082	0.6107	0.1473	0.2373	0.4560	0.2950	0.3583	0.7057	0.2712	0.3918
3	CG-cluster (strict)	0.1156	0.6861	0.1979	0.4118	0.5577	0.4738	0.2976	0.7425	0.4249	0.5221	0.7316	0.6093
4	CpGProD	0.1386	0.2511	0.1786	0.4487	0.2506	0.3216	0.3627	0.3165	0.3381	0.6323	0.3336	0.4368
5	CpG_MI	0.1076	0.3746	0.1672	0.3243	0.3020	0.3128	0.2822	0.5077	0.3627	0.5261	0.4494	0.4847
6	NCBI-CGI (relaxed)	0.2203	0.1011	0.1386	0.4854	0.0773	0.1334	0.4919	0.1503	0.2302	0.7083	0.1103	0.1909
7	NCBI-CGI (strict)	0.1198	0.6124	0.2003	0.4052	0.5746	**0.4752**	0.3200	0.6949	0.4382	0.5376	0.7277	**0.6183**
8	UCSC-CGI	0.1214	0.5962	0.2017	0.4344	0.4900	0.4605	0.3253	0.6796	0.4399	0.4248	0.6475	0.5130

4.0 Conclusions

Intrigued by the unique sequence properties of the rare *k*-mers, we examined the possibilities of rare *k*-mer functionalities and their applications in the human genome. Rare *k*-mers are indeed the least occurring DNA words in the human genome but they are highly correlated with several genomic features of CGIs, promoters, 5'UTR, and open chromatin regions of certain histone modification codes. These high correlations imply that the rare *k*-mers are functional but due to its short length of 8 to 10-mers and its average frequencies are in between 60-1000 (although very rare), these make the rare *k*-mers looked ubiquitous and less functional in the human genome. This dilemma is quite similar to the CGIs (which is the highest feature correlated with the rare *k*-mers) which are very prominent in terms of sequence and correlation to biological functions, but little is understood about its molecular mechanisms. Because there is no clear definition for the CGI about its structure, we are unable to come out with any positional statistics of rare *k*-mers inside them. Nevertheless, we have found three significant positional statistics of rare *k*-mers inside promoter regions (with the RefSeq status of reviewed and validated) which are: 1) several rare 8-mers have the same relative positions from the RGS locations; 2) same rare 8-mers are repeated more than five times within the same promoters, and 3) same rare 8-mer repeats have the same intervals in the same promoters. Moreover, several studies have identified that some rare 8-mers are important words and functional in the human genome (Bao

et al., 2012; Hackenberg et al., 2012; Stacey et al., 2003). Then, we extend our works by utilizing the RW clustering property (which was elucidated from the enrichment analyses of rare 8-mers in CGIs and promoters) in the RWC method to further the use of rare *k*-mers in biology. When we evaluated the predicted CGIs and promoters by the RWC method for 4 CGI and 4 promoter evaluations respectively and benchmarked both of the predicted RWC datasets with seven other datasets, the RWC method consistently achieved the top 3 F-scores for all of the evaluations. These results prove that rare *k*-mers is as good as the widely used CGI feature in most of CGI and promoter prediction programs. Another advantage is the RWC method predicts CGIs and promoters based on functional rare 8-mer words, not just by pure integer arithmetic which gives clues about novel CGI and promoter regulations. Since practical predictors are now commonly required to have a user-friendly and publicly available online version for more usability in the field of computational biology (Chou, 2011; Liu et al., 2013; Liu et al., 2015b; Liu et al., 2014; Zhang et al., 2011), we shall make efforts to provide a web-server for the RWC method in the near future.

5.0 Acknowledgements

We acknowledges and thankful for the guidance, financial supports, facilities, and technicalities provided by Malaysian Genomics Resource Centre (MGRC) Sdn. Bhd. and Synamatix Sdn. Bhd.

6.0 References

Abeel, T., Peer, Y. V. d., Saeys, Y., 2009. Toward a gold standard for promoter prediction evaluation. Bioinformatics 25, 13-20, doi:10.1093/bioinformatics/btp191.

Antequera, F., 2003. Structure, function and evolution of CpG island promoters. Cell Mol Life Sci 60, 1647-1658, doi:10.1007/s00018-003-3088-6.

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.-F., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R., Bulyk, M. L., 2009. Diversity and Complexity in DNA Recognition by Transcription Factors. Science 324, 1720-1723, doi:10.1126/science.1162327.

Bajic, V. B., Tan, S. L., Suzuki, Y., Sugano, S., 2004. Promoter prediction analysis on the whole human genome. Nat Biotechnol 22, 1467-1473, doi:10.1038/nbt1032.

Bao, T., Li, H., Zhao, X., Liu, G., 2012. Predicting nucleosome binding motif set and analyzing their distributions around functional sites of human genes. Chromosome Research 20, 685-698, doi:10.1007/s10577-012-9305-0.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., *et al.*, 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799-816, doi:10.1038/nature05874.

Bock, C., Halachev, K., Büch, J., Lengauer, T., 2009. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi) genomic data. Genome Biol 10, R14.

Bock, C., Paulsen, M., Tierling, S., Mikeska, T., Lengauer, T., Walter, J. E., 2005. CpG island methylation in human lymphocytes is highly correlated with DNA sequence patterns, repeat frequencies and predicted DNA structure. PLoS Genetics 2, e26, doi:10.1371/journal.pgen.0020026.eor.

Burge, C., Campbell, A. M., Karlin, S., 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci USA 89, 1358-1362.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engstrom, P. G., Frith, M. C., Forrest, A. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A., Hayashizaki, Y., 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38, 626-635, doi:10.1038/ng1789.

Castellini, A., Franco, G., Manca, V., 2012. A dictionary based informational genome analysis. BMC Genomics 13, 485.

Chae, H., Park, J., Lee, S.-W., Nephew, K. P., Kim, S., 2013. Comparative analysis using K-mer and K-flank patterns provides evidence for CpG island sequence evolution in mammalian genomes. Nucleic Acids Res 41, 4783-4791, doi:10.1093/nar/gkt144.

Chan, B. Y., Kibler, D., 2005. Using hexamers to predict cis-regulatory motifs in Drosophila. BMC Bioinformatics 6, 1-9, doi:10.1186/1471-2105-6-262.

Chen, W., Lin, H., Chou, K.-C., 2015. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Molecular BioSystems.

Chen, W., Feng, P.-M., Lin, H., Chou, K.-C., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic acids research, gks1450.

Chen, W., Feng, P.-M., Deng, E.-Z., Lin, H., Chou, K.-C., 2014a. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Analytical biochemistry 462, 76-83.

Chen, W., Lei, T.-Y., Jin, D.-C., Lin, H., Chou, K.-C., 2014b. PseKNC: A flexible web server for generating pseudo K-tuple nucleotide composition. Analytical Biochemistry 456, 53-60, doi:http://dx.doi.org/10.1016/j.ab.2014.04.001.

Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., Chou, K.-C., 2014c. PseKNC-General: A cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics, doi:10.1093/bioinformatics/btu602.

Cherrington, J. M., Mocarski, E. S., 1989. Human cytomegalovirus ie1 transactivates the alpha promoter-enhancer via an 18-base-pair repeat element. Journal of Virology 63, 1435-1440.

Chor, B., Horn, D., Goldman, N., Levy, Y., Massingham, T., 2009. Genomic DNA k-mer spectra: models and modalities. Genome Biol 10, R108, doi:10.1186/gb-2009-10-10-r108.

Chou, K.-C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology 273, 236-247.

Chou, K.-C., 2015. Impacts of Bioinformatics to Medicinal Chemistry. Medicinal Chemistry 11, 218-234.

Chou, K.-C., Zhang, C.-T., 1995. Prediction of Protein Structural Classes. Critical Reviews in Biochemistry and Molecular Biology 30, 275-349, doi:doi:10.3109/10409239509083488.

Compeau, P. E. C., Pevzner, P. A., Tesler, G., 2011. How to apply de Bruijn graphs to genome assembly. Nat Biotech 29, 987-991, doi:http://www.nature.com/nbt/journal/v29/n11/abs/nbt.2023.html#supplementary-information.

Cooper, D. N., Gerber-Huber, S., 1985. DNA methylation and CpG suppression. Cell Differentiation 17, 199-205, doi:http://dx.doi.org/10.1016/0045-6039(85)90488-9.

Csűrös, M., Noé, L., Kucherov, G., 2007. Reconsidering the significance of genomic word frequencies. Trends in Genetics 23, 543-546, doi:10.1016/j.tig.2007.07.008.

Das, M. K., Dai, H. K., 2007. A survey of DNA motif finding algorithms. BMC Bioinformatics 8 S21, doi:10.1186/1471-2105-8-S7-S21.

Deaton, A. M., Bird, A., 2011. CpG islands and the regulation of transcription. Genes & Development 25, 1010-1022, doi:10.1101/gad.2037511.

Down, T. A., Hubbard, T. J. P., 2002. Computational Detection and Location of Transcription Start Sites in Mammalian Genomic DNA. Genome Research 12, 458-461, doi:10.1101/gr.216102.

Eskdale, J., Gallagher, G., 1995. A polymorphic dinucleotide repeat in the human IL-10 promoter. Immunogenetics 42, 444-445.

Fickett, J. W., Hatzigeorgiou, A. G., 1997. Eukaryotic Promoter Recognition. Genome Research 7, 861-878, doi:10.1101/gr.7.9.861.

Fofanov, Y., Luo, Y., Katili, C., Wang, J., Belosludtsev, Y., Powdrill, T., Belapurkar, C., Fofanov, V., Li, T. B., Chumakov, S., Pettitt, B. M., 2004. How independent are the appearances of n-mers

in different genomes? Bioinformatics 20, 2421-2428, doi:DOI 10.1093/bioinformatics/bth266.

Gardiner-Garden, M., Frommer, M., 1987. CpG islands in vertebrate genomes. Journal of Molecular Biology 196, 261-282.

Gentles, A. J., Karlin, S., 2001. Genome-scale compositional comparisons in eukaryotes. Genome Res 11, 540-546, doi:10.1101/gr.163101.

Glass, J. L., Thompson, R. F., Khulan, B., Figueroa, M. E., Olivier, E. N., Oakley, E. J., Van Zant, G., Bouhassira, E. E., Melnick, A., Golden, A., Fazzari, M. J., Greally, J. M., 2007. CG dinucleotide clustering is a species-specific property of the genome. Nucleic Acids Res 35, 6798-6807, doi:10.1093/nar/gkm489.

Guo, S.-H., Deng, E.-Z., Xu, L.-Q., Ding, H., Lin, H., Chen, W., Chou, K.-C., 2014. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics, btu083.

Hackenberg, M., Previti, C., Luque-Escamilla, P. L., Carpena, P., Martinez-Aroza, J., Oliver, J. L., 2006. CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinformatics 7, 1-13, doi:10.1186/1471-2105-7-446.

Hackenberg, M., Barturen, G., Carpena, P., Luque-Escamilla, P. L., Previti, C., Oliver, J. L., 2010. Prediction of CpG-island function: CpG clustering vs. sliding-window methods. BMC Genomics 11, 1-14, doi:10.1186/1471-2164-11-327.

Hackenberg, M., Rueda, A., Carpena, P., Bernaola-Galván, P., Barturen, G., Oliver, J. L., 2012. Clustering of DNA words and biological function: A proof of principle. Journal of Theoretical Biology 297, 127-136, doi:http://dx.doi.org/10.1016/j.jtbi.2011.12.024.

Han, L., Zhao, Z., 2009. CpG islands or CpG clusters: how to identify functional GC-rich regions in a genome? BMC Bioinformatics 10, 65, doi:10.1186/1471-2105-10-65.

Hannenhalli, S., Levy, S., 2001. Promoter prediction in the human genome. Bioinformatics 17, S90-S96, doi:10.1093/bioinformatics/17.suppl_1.S90.

Hariharan, R., Simon, R., Pillai, M. R., Taylor, T. D., 2013. Comparative Analysis of DNA Word Abundances in Four Yeast Genomes Using a Novel Statistical Background Model. PLoS ONE 8, e58038, doi:10.1371/journal.pone.0058038.

Illingworth, R., Kerr, A., DeSousa, D., Jørgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J., Humphray, S., Cox, T., Langford, C., Bird, A., 2008. A Novel CpG Island Set Identifies Tissue-Specific Methylation at Developmental Gene Loci. PLoS Biol 6, e22, doi:10.1371/journal.pbio.0060022.

Illingworth, R. S., Bird, A. P., 2009. CpG islands – 'A rough guide'. FEBS Letters 583, 1713-1720, doi:10.1016/j.febslet.2009.04.012.

Ioshikhes, I. P., Zhang, M. Q., 2000. Large-scale human promoter mapping using CpG islands. Nat Genet 26, 61-63.

Jaquiery, P., 2011. Particle Swarm Optimization Perl Module. Vol. 2015. CPAN.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research 110, 462-467.

Karlin, S., 2005. Statistical signals in bioinformatics. Proc Natl Acad Sci USA 102, 13355-13362, doi:10.1073/pnas.0501804102.

Kennedy, J., Eberhart, R., 1995. Particle swarm optimization. Neural Networks, 1995. Proceedings., IEEE International Conference on, Vol. 4, pp. 1942-1948 vol.4.

Kent, W. J., 2002. BLAT—The BLAST-Like Alignment Tool. Genome Research 12, 656-664, doi:10.1101/gr.229202.

Kurtz, S., Narechania, A., Stein, J., Ware, D., 2008. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 9, 517.

Larsen, F., Gundersen, G., Lopez, R., Prydz, H., 1992. CpG islands as gene markers in the human genome. Genomics 13, 1095-1107, doi:10.1016/0888-7543(92)90024-m.

Li, Q.-Z., Lin, H., 2006. The recognition and prediction of σ 70 promoters in Escherichia coli K-12. Journal of theoretical biology 242, 135-141.

Lin, H., Li, Q.-Z., 2011. Eukaryotic and prokaryotic promoter prediction using hybrid approach. Theory in Biosciences 130, 91-100.

Lin, H., Deng, E.-Z., Ding, H., Chen, W., Chou, K.-C., 2014. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic acids research 42, 12961-12972.

Liu, B., Wang, X., Zou, Q., Dong, Q., Chen, Q., 2013. Protein Remote Homology Detection by Combining Chou's Pseudo Amino Acid Composition and Profile-Based Protein Representation. Molecular Informatics 32, 775-782.

Liu, B., Liu, F., Fang, L., Wang, X., Chou, K.-C., 2015a. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics 31, 1307-1309.

Liu, B., Xu, J., Fan, S., Xu, R., Zhou, J., Wang, X., 2015b. PseDNA-Pro: DNA-Binding Protein Identification by Combining Chou's PseAAC and Physicochemical Distance Transformation. Molecular Informatics 34, 8-17.

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., Chou, K.-C., 2015c. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Research, doi:10.1093/nar/gkv458.

Liu, B., Zhang, D., Xu, R., Xu, J., Wang, X., Chen, Q., Dong, Q., Chou, K.-C., 2014. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics 30, 472-479.

Liu, G., Liu, J., Cui, X., Cai, L., 2012. Sequence-dependent prediction of recombination hotspots in Saccharomyces cerevisiae. Journal of theoretical biology 293, 49-54.

Maglott, D., Ostell, J., Pruitt, K. D., Tatusova, T., 2007. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 35, D26-31, doi:10.1093/nar/gkl993.

Michelson, A. M., Bulyk, M. L., 2006. Biological code breaking in the 21st century. Mol Syst Biol 2, doi:10.1038/msb4100062.

Pennisi, E., 2012. GENOMICS ENCODE Project Writes Eulogy For Junk DNA. Science 337, 1159-1161.

Ponger, L., Mouchiroud, D., 2002. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics 18, 631-633.

Pruitt, K. D., Tatusova, T., Maglott, D. R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35, D61-65.

Reinert, G., Schbath, S., Waterman, M. S., 2000. Probabilistic and statistical properties of words: an overview. J Comput Biol 7, 1-46, doi:10.1089/10665270050081360.

Rhead, B., Karolchik, D., Kuhn, R. M., Hinrichs, A. S., Zweig, A. S., Fujita, P. A., Diekhans, M., Smith, K. E., Rosenbloom, K. R., Raney, B. J., Pohl, A., Pheasant, M., Meyer, L. R., Learned, K., Hsu, F., Hillman-Jackson, J., Harte, R. A., Giardine, B., Dreszer, T. R., Clawson, H., Barber, G. P., Haussler, D., Kent, W. J., 2010. The UCSC Genome Browser database: update 2010. Nucleic Acids Res 38, D613-619, doi:10.1093/nar/gkp939.

Saxonov, S., Berg, P., Brutlag, D. L., 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci USA 103, 1412-1417, doi:10.1073/pnas.0510310103.

Segal, E., Widom, J., 2009. Poly (dA: dT) tracts: major determinants of nucleosome organization. Current opinion in structural biology 19, 65-71.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., Haussler, D., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15, 1034-1050, doi:10.1101/gr.3715005.

Stacey, K. J., Young, G. R., Clark, F., Sester, D. P., Roberts, T. L., Naik, S., Sweet, M. J., Hume, D. A., 2003. The Molecular Basis for the Lack of Immunostimulatory Activity of Vertebrate DNA. The Journal of Immunology 170, 3614-3620, doi:10.4049/jimmunol.170.7.3614.

Su, J., Zhang, Y., Lv, J., Liu, H., Tang, X., Wang, F., Qi, Y., Feng, Y., Li, X., 2010. CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes. Nucleic Acids Res 38, e6, doi:10.1093/nar/gkp882.

Takai, D., Jones, P. A., 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci USA 99, 3740-3745, doi:10.1073/pnas.052410099.

Tarlow, J. K., Blakemore, A. I., Lennard, A., Solari, R., Hughes, H. N., Steinkasserer, A., Duff, G. W., 1993. Polymorphism in human IL-1 receptor antagonist gene intron 2 is caused by variable numbers of an 86-bp tandem repeat. Human genetics 91, 403-404.

Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Paabo, S., Rebhan, M., Schubeler, D., 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 39, 457-466, doi:10.1038/ng1990.

Werner, T., 1999. Models for prediction and recognition of eukaryotic promoters. Mamm Genome 10, 168-175.

Zeng, J., Zhu, S., Yan, H., 2009. Towards accurate human promoter recognition: a review of currently used sequence features and classification methods. Briefings in Bioinformatics 10, 498-508, doi:10.1093/bib/bbp027.

Zeng, J., Zhao, X.-Y., Cao, X.-Q., Yan, H., 2008. SCS: Signal, Context, and Structure Features for Genome-Wide Human Promoter Recognition. Computational Biology and Bioinformatics, IEEE/ACM Transactions on 7, 550-562, doi:10.1109/tcbb.2008.95.

Zhang, Y., Liu, B., Dong, Q., X Jin, V., 2011. An improved profile-level domain linker propensity index for protein domain boundary prediction. Protein and peptide letters 18, 7-16.

Zhao, Z., Han, L., 2009. CpG islands: algorithms and applications in methylation studies. Biochem Biophys Res Commun 382, 643-645, doi:10.1016/j.bbrc.2009.03.076.