International Journal of
Molecular Sciences

Article

Genome-Wide Prediction of DNA Methylation Using DNA Composition and Sequence Complexity in Human

Chengchao Wu [1], Shixin Yao [2], Xinghao Li [2], Chujia Chen [1] and Xuehai Hu [1,]*

[1] College of Informatics, Agricultural Bioinformatics Key Laboratory of Hubei Province, Huazhong Agricultural University, Wuhan 430070, China; woo.kidd@gmail.com (C.W.); ccjljm@webmail.hzau.edu.cn (C.C.)

[2] College of Science, Huazhong Agricultural University, Wuhan 430070, China; yaoshixin@webmail.hzau.edu.cn (S.Y.); lxh8930217@gmail.com (X.L.)

* Correspondence: huxuehai@mail.hzau.edu.cn; Tel.: +86-181-7128-2783; Fax: +86-27-8728-8509

Academic Editor: Mateus Webba da Silva
Received: 3 January 2017; Accepted: 8 February 2017; Published: 16 February 2017

Abstract: DNA methylation plays a significant role in transcriptional regulation by repressing activity. Change of the DNA methylation level is an important factor affecting the expression of target genes and downstream phenotypes. Because current experimental technologies can only assay a small proportion of CpG sites in the human genome, it is urgent to develop reliable computational models for predicting genome-wide DNA methylation. Here, we proposed a novel algorithm that accurately extracted sequence complexity features (seven features) and developed a support-vector-machine-based prediction model with integration of the reported DNA composition features (trinucleotide frequency and GC content, 65 features) by utilizing the methylation profiles of embryonic stem cells in human. The prediction results from 22 human chromosomes with size-varied windows showed that the 600-bp window achieved the best average accuracy of 94.7%. Moreover, comparisons with two existing methods further showed the superiority of our model, and cross-species predictions on mouse data also demonstrated that our model has certain generalization ability. Finally, a statistical test of the experimental data and the predicted data on functional regions annotated by ChromHMM found that six out of 10 regions were consistent, which implies reliable prediction of unassayed CpG sites. Accordingly, we believe that our novel model will be useful and reliable in predicting DNA methylation.

Keywords: DNA methylation; predicted model; sequence complexity

1. Introduction

Although the DNA sequence of the human genome, which carries genetic information, is almost invariant in various human cells, the epigenetic features of each cell show great differences, leading to distinguishable gene expression patterns and cell-type specificities [1]. Among these features (such as histone modifications mediating changes in chromatin conformation), DNA methylation is the best studied epigenetic modification [2,3]. In mammals, DNA methylation mainly occurs at CpG dinucleotide sites with an added methyl group to the fifth carbon of the cytosine residue [2,4–8]. In general, the patterns of DNA methylation are mediated by the main three enzymes, DNMT3 (for establishment), DNMT1 (for maintenance) and MBD4 (for demethylation) [3]. The human genome contains approximately 28 million CpG sites, 60%–80% of which are generally methylated [2], and the remaining unmethylated CpG sites are mostly located near promoter or exonic regions where GC content is much greater than 50%, which are usually called *CpG islands* (CGIs) [3,9].

The biological role of DNA methylation is to repress transcriptional activity. Different cell types show distinct methylation distributions across the genome, particularly in regulatory regions of cell-specific genes [2,3]. Notably, the average methylation levels in specific regions are consistent with other signals and modifications that are related to transcriptional regulation, such as transcription factor binding sites (TFBSs), DNase I hypersensitive sites (DHSs) and various histone modifications [1,9]. Recent studies have shown that hypomethylated regions were associated with promoters near transcriptional start sites (TSSs), whereas hypermethylated regions were considered as silenced regions. Interestingly, the low-methylated regions, in which average methylation levels are approximately 0.3, an intermediate status between hypomethylation and hypermethylation, were reported to be associated with distal regulatory regions, such as enhancers [9]. Therefore, DNA methylation is a key biomarker and plays a critical role in transcriptional regulation.

Considering the significance of DNA methylation, change of its level in specific regions is usually regarded as an important factor affecting the expressions of target genes and downstream phenotypes, such as embryonic development and tumorigenesis [10]. When compared to normal cells, aberrant average levels of DNA methylation in important regulatory regions (such as promoters) were linked with the altered expression profiles of cancer cells [11–14]. Interestingly, recent findings revealed that abnormal DNA methylation levels at distal regulatory regions (such as enhancers and super-enhancers) were closely related to gene dysregulation in cancer [14,15]. In summary, the average DNA methylation level in specific regions is an ideal biomarker of tumorigenesis, and detecting aberrant methylations in these regions is a promising approach for early diagnosis and classification of cancer [10].

Traditional methods for identifying DNA methylation mainly include whole-genome bisulfite sequencing (WGBS) [16], reduced representation bisulfite sequencing (RRBS) [17] and the methylation-microarray-based Illumina 450K BeadChip [18]. The 450K BeadChip uses microarray-based technology to assay approximately 0.4–0.6 million preselected CpG sites. In contrast, WGBS is a standard method in whole-genome sequencing at base resolution with approximately 1–18 million CpG sites captured [19]. Additionally, RRBS is an efficient technique designed for enriching the regions with a high CpG content, leading to high-efficiency and saving compared with WGBS. The total coverage of RRBS is roughly equivalent to that of WGBS, and the number of CpGs captured per sample genome by RRBS ranged from 0.5 million to 13 million [19]. However, there are 28 million CpG sites in the whole human genome, only 10%–50% of which are covered by WGBS or RRBS. Why it is difficult to assay the majority of CpGs in the whole genome even by the high-throughput techniques? Generally, there are many experimental difficulties hampering the analysis of the methylation status within this “hidden” fraction of CpG sites, such as copy-number variation bias, incomplete bisulfite conversion bias, bisulfite PCR bias, GC content bias and CpG density bias [19]. For RRBS or WGBS experiments, still two coverage limitations (incomplete bisulfite conversion bias and bisulfite PCR bias) could happen [19–21], which are the reasons for their low coverages. Therefore, more quantitative methods are urgently demanded to predict the methylation status of the remaining unassayed CpG sites of the human genome.

Computational approaches are alternative methods to identify DNA methylation status. Actually, a number of computational methods had been developed to predict DNA methylation status with the rapid developments of bioinformatics and machine learning approaches, which usually contain three vital steps—data collection, feature extraction and a classification algorithm [22]. Similarly, many efforts have been made for predicting the functional sites in proteins (such as cysteine S-nitrosylation sites [23–27], protein methylation sites [28], hydroxyproline and hydroxylysine in proteins [29], lysine ubiquitination sites [30], lysine succinylation sites [31]. For more details, please refer to two recent reviews [32,33].

Due to the restrictions of experimental data (such as microarray), some early methods limited their attentions on specific genomic regions (such as CGIs) and obtained satisfactory prediction accuracies (with accuracy >90%) [34–37]. In contrast, when turning to more common regions predicted using RRBS and WGBS data, the prediction accuracies clearly decreased to 75%–89% [38]. This is because when we focused on the prediction in CGI regions, GC contents of positive samples are high, whereas GC contents of negative samples are lower. Thus, it is easy to distinguish the methylated CpGs and unmethylated ones only by GC content feature. However, when turning to more common regions, the predictive role of GC content will not be as significant as that of CGI regions and the corresponding prediction accuracies will decrease. In these situations, more complex features are needed for better predictions.

Regarding feature extraction, various methods have been employed to formulate methylation status, including: DNA composition [34,36,39,40], pseudo trinucleotide composition (PseTNC) [41–45], predicted DNA structure [34,46], single nucleotide polymorphisms (SNPs) [34], TFBSs [34,46], histone modifications [36,46], neighboring CpG site methylation status and distance [46]. It is worth noting that a powerful web-server called "Pse-in-One" has been used to extract various features from DNA or protein sequences [47].

Particularly, Das et al. analyzed GC content and Alu elements features for 800 bp regions centered on CpG sites using DNA methylation data of human brain, and developed classifiers with accuracy of 86% [40]. Moreover, Bock et al. employed 1184 DNA attributes for discriminating between CpG islands that are prone to methylation from those that remain unmethylated using CpG island methylation data on human Chromosome 21, and they found that certain sequence patterns, specific DNA repeats and a particular DNA structure played significant roles in the prediction of DNA methylation status [34]. In the same year, Fang et al. [39] used nucleotide sequence features including GC content, CpG ratio, TpG content and Alu distribution, as well as TFBSs features to predict DNA methylation status on human brain dataset. They tested four different window sizes (200 bp, 300 bp, 400 bp, and 500 bp) and found 400 bp is a better choice for prediction, with accuracy of 85%. Recently, Liu et al. invented a novel method called "pseudo trinucleotide composition (PseTNC)" which can both capture the local or short-range sequence order effects and the global or long-range effects of DNA sequences for predicting the methylation status of DNA fragments using 41-bp window centered on CpG sites [41]. Notably, Zhang et al. [46] integrated four groups of features, including neighboring features, genomic position, DNA sequence properties and *cis*-regulatory elements, to predict DNA methylation status using the whole blood sample dataset. As a result, 400-bp window was also found to be the best size for prediction and DHSs as well as GC content were found to be the most predictive features. In addition, Wang et al. combined "PseTNC" and chromatin interaction features (Hi-C features) to predict methylation states using the methylation datasets taken from two cell lines (GM12878 and K562). A series of window sizes (500–1000) were tested to search the best one, and 600 bp was found to be the appropriate choice [48].

These methods were mainly developed based on some common machine-learning algorithms: support vector machine (SVM) [34–36,38,40,49], random forest (RF) [37,46], naive Bayes (NB) and stacked denoising autoencoders (SDA) [48]. One can refer to the textbook written by James et al. [50] for more details of these algorithms and their implementations in R. The majority of these studies used SVM as the classifier due to its powerful classification ability and universality for various data types. Notably, some research showed the superiority of RF compared with SVM [46], and the SDA method from deep learning field emerged for prediction of DNA methylation [48].

The methods above achieved remarkable results, but there were still some defects among them: (a) most of them could not perform predictions at the whole genome-wide level because the data they used were specific fragments of the genome (such as the 450K BeadChip); (b) although some of them developed their classifiers based on RRBS and WGBS and achieved satisfactory results [46], obtaining important features for prediction (such as neighboring CpG site methylation status and DHSs) is difficult and would require comprehensive and expensive epigenetic experiments; and (c) most of them did not test the validities of their predicting models on unassayed CpG sites.

In this work, we introduced a novel computational algorithm called "sequence complexity", together with DNA composition (72 features in total) to predict the DNA methylation status of CpG sites in the whole human genome. Our prediction method has the following advantages compared with current classifiers: (a) a group of novel features called “sequence complexity” were developed, and subsequent analysis confirmed that these new features played significant roles for predictions; (b) by integrating the fundamental features (DNA composition), the prediction model achieved satisfactory results; (c) all the features we used were only extracted from the primary DNA sequence of the human genome without additional experiments, and comparisons with previous works showed the superiority of our method; and (d) a statistical test of the experimental data and the predicted data on functional regions annotated by ChromHMM found that six out of 10 regions were consistent, which implies reliable prediction of unassayed CpG sites. Thus, we believe that our novel model will be useful and reliable in predicting DNA methylation.

As illustrated by many recently published papers [51–57], to establish a powerful predictor for a biological problem, one should obey the following five steps: (a) build or choose a benchmark dataset to train and test the predictor; (b) transform the raw biological sequences into mathematical feature vectors that can truly extract their intrinsic features from the target to be predicted; (c) employ or create a powerful algorithm to operate the prediction; (d) accurately use cross-validation tests to impersonally evaluate the predicting ability of the predictor; and (e) build a user-friendly web-server to make their dataset and predictor publicly available. Next, the current study will be organized following these steps one-by-one.

2. Results

2.1. DNA Methylation Dataset and Data Preprocessing

In this work, we developed a SVM-based model to predict the DNA methylation status of CpG sites in the human genome. In this work, Homo sapiens embryonic stem cell methylation profiles that were measured by RRBS [58] were downloaded from the NCBI Gene Expression Omnibus (GEO, GSE49828). After downloading, eight important cell stages (MII oocytes; zygotes; 2-cell, 4-cell, and 8-cell embryos; morulae; ICM of blastocysts; and post-implantation embryos) out of 12 were selected for analysis. There are 27,762,346 methylation sites, 4,476,329 of which are CpG sites. Embryonic stem cells are suitable for comprehensively studying DNA methylation status for two reasons. One is the large amount of number of assayed CpG sites (more than ten million CpG sites in this research) [58], because our machine-learning system requires larger samples for reducing the false positive rate. The other is the strategy we used for selecting positive and negative samples, which are stably methylated or unmethylated during all eight stages of embryonic stem cells. Whereas the adult cells display a cell-specificity patterns of DNA methylation.

Specifically, given a CpG site in a fixed stage, its methylation level is quantified by a variable β, which represents the number of methylated reads divided by the sum of both methylated and unmethylated reads at the same positions of the reference genome. Therefore, the term β of each CpG site ranges from zero (unmethylated) to one (fully methylated). Here, we adopted a similar strategy to that of Wang et al. [48] for choosing positive and negative samples: the intersection of CpG dinucleotide sites with methylation levels $\beta \geq 0.7$ during all eight stages were labeled positive, whereas the intersection of CpG dinucleotide sites with methylation levels $\beta \leq 0.01$ during all 8 stages were labeled negative. Through this procedure, 139,931 methylated CpG sites (139,931/4,476,329 = 3.13%) and 536,183 (536,183/4,476,329 = 11.98%) unmethylated CpG sites were obtained for further work.

To investigate the appropriate window size for prediction, for each CpG site, a series of DNA sequence window sizes (100 to 1000 bp) with the CpG site as its center, were generated using the HG19 reference genome. For comparable computations and so that different window sizes have the same number of positive and negative samples, we focused on the DNA sequence window size of 1000 bp for all CpG sites and the other window sizes were just the subsets of them. Through this procedure, a positive dataset consisting of 139,931 1000-bp windows was obtained, and a negative dataset consisting of 536,183 1000-bp windows was also obtained in the same manner, which constitutes our original dataset.

To reduce the homology bias of prediction, a redundancy reduction procedure was performed on the above original dataset using the CD-HIT program [59] (http://www.bioinformatics.org/cd-hit/), and a cutoff threshold of 0.8 was imposed to exclude those DNA sequences that have 80% or greater sequence identity to any other in a same subset. Based on this pre-processing procedure, the positive and negative datasets were determined (104,876 + 159,090), and are shown in Table 1.

Table 1. Detailed information on the four datasets used in this work. The cutoff threshold of CD-HIT was 0.8.

Species or Cell Lines	Number of Positive Samples	Number of Positive Samples after CD-HIT	Number of Negative Samples	Number of Negative Samples after CD-HIT	Data Resource (GEO Number)
Homo sapiens	139,931	104,876	536,183	159,090	GSE49828
Mouse	190,000	135,733	190,000	138,207	GSE56697
GM12878	263,698	166,116	563,702	252,188	GSM683906
K562	292,791	159,310	540,465	251,948	GSM683856

2.2. DNA Methylation Pattern

To better understand the DNA methylation profiles, the distribution of DNA methylation level β in the whole human genome is shown in Figure 1A. For simplicity, we chose three representative chromosomes (chr1, chr11, and chr21) by size and gene number for exhibition. As in previous studies [2,3,46,48], the DNA methylation level of each chromosome all showed a bimodal landscape, i.e., most CpG sites either have fully methylated status ($\beta \geq 0.9$) or have unmethylated status ($\beta \leq 0.1$), and few have intermediate methylated status ($0.1 < \beta < 0.9$). Notably, the three representative chromosomes generally showed bimodal distributions, but their distributions still showed some differences (Figure 1A). For example, chromosome 1 has more unmethylated CpG sites, whereas chromosome 21 has more fully methylated CpG sites. This explains why different chromosomes present different prediction accuracies in the following analysis because changing in distribution will result in changing of prediction result by the basic principle of statistical learning theory.

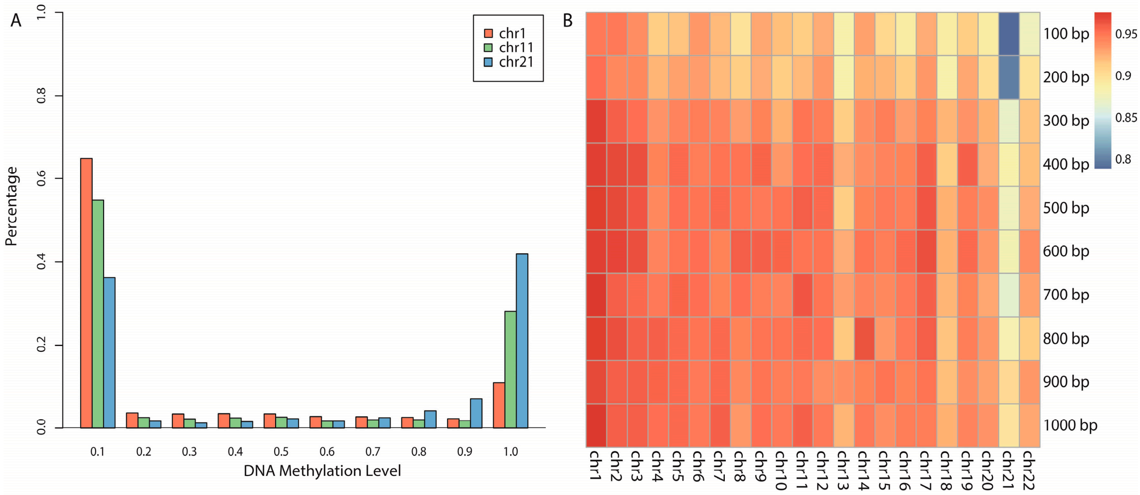

Figure 1. DNA methylation patterns and the heat map of overall prediction accuracies: (**A**) bimodal distributions of DNA methylation patterns; and (**B**) the heat map of prediction results for combinations of different chromosomes and different window sizes.

2.3. Overview of Binary Methylation Status Prediction

Here, we focused on binary methylation status prediction, which asserted that the methylation status of each CpG site was encoded as a binary variable (1 for fully methylated sites, 0 for unmethylated sites). Several previous studies have all found that DNA methylation status was strongly correlated with local DNA sequence [34–41,46,48]. To investigate which windows are most predictive for methylation, we tested our prediction approach with ten different window sizes (100–1000 bp).

Concerning feature extraction, 72 features were used to formulate DNA fragment sequences with different lengths, which included two groups of features (for details, see Materials and Methods):

1. DNA composition (DC, trinucleotide frequency and GC content, 65 features); and
2. Sequence complexity (SC, 7 features).

For the classifier, SVM was chosen by comparisons between the other three common classifiers (Figure 2A). For the evaluation method, 10-fold cross-validation (see Materials and Methods) was chosen for testing the prediction performance of our method.

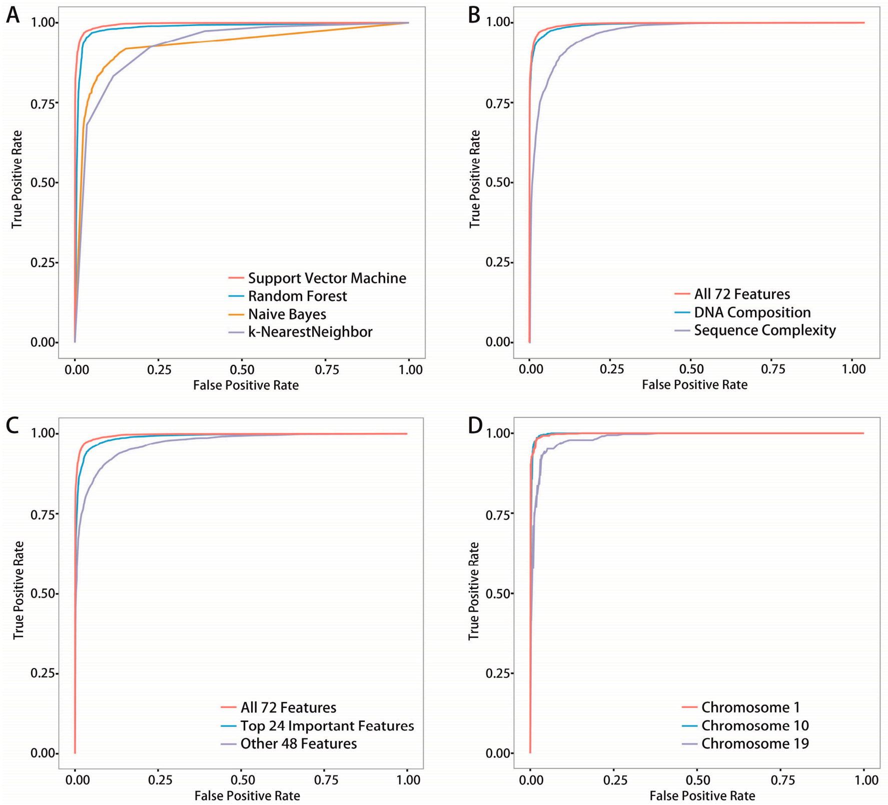

Figure 2. Receive Operating Characteristic (ROC) curves of different comparisons: (**A**) ROC curves of comparisons between two groups of features using 10-fold cross-validation; (**B**) ROC curves of comparisons between the top 24 important features and remaining 48 features using independent testing; (**C**) ROC curves of comparisons between four common classifiers using independent testing; and (**D**) ROC curves of three mouse chromosome predictions.

An overview of prediction results was exhibited in the form of a heat map (Figure 1B), in which each grid represented a corresponding chromosome and window size used for prediction (for detailed values, see Table S1). The closer to red, the better the prediction result was in that grid, whereas, the closer to blue, the worse the result was.

Firstly, we observed that all chromosomes with all window sizes achieved prediction accuracies ≥0.8. Secondly, if we focused on each column (representing each chromosome), the best results are all above 0.9 even in chromosome 21 (chromosome 21 with a 900-bp window, 0.9022, see Table S1). These results imply that our approach is a powerful predictor for DNA methylation. Notably, if we focus on each row, the 600-bp window shows consistent predictive power, and 40% of chromosomes (9 out of 22) achieve the best accuracies with 600-bp window. Even in the remaining chromosomes, the prediction differences between 600-bp window and the corresponding window in which the best accuracy was achieved were not significant (T-test, p value = 0.3521). This leads us to confirm that 600 bp is the most appropriate window for DNA methylation prediction. Hereafter, we only focused on 600-bp windows with the CpG site as its center in the rest of paper.

2.4. Comparison with Different Classifiers

To test which classifier would achieve the best prediction result and to explain why we chose SVM as the classifier in this work, a detailed comparison with other classifiers was performed. Here, three frequently used classifiers, Random Forest (RF), Naive Bayes (NB) and K-Nearest Neighbors (KNN), were chosen to compare with SVM (see Table 2) and their respective ROC (Receiver Operating Characteristic) curves in Figure 2A.

Table 2. Comparison of four commonly used classifiers using independent testing on IT-set. MCC is the abbreviation of Matthew's Coefficient Of Correlation; Sens is the abbreviation of sensitivity; Spec is the abbreviation of specificity.

Classifiers	Dimension	ACC	AUC	MCC	Sens	Spec
Support Vector Machine	72	0.971	0.996	0.940	0.972	0.969
Random Forest	72	0.959	0.984	0.917	0.958	0.959
NaiveBayes	72	0.881	0.937	0.766	0.927	0.844
K-Nearest Neighbor	72	0.849	0.941	0.706	0.803	0.911

A smaller subset of the whole dataset was constructed for subsequent in-depth analyses due to the computation requirements for large samples. This subset contains a training sample-set (T-set) with 5000 positive samples and 5000 negative samples (randomly selected from all chromosomes) and an independent testing sample set (IT-set), which has no overlaps with the T-set with another 5000 positive samples and 5000 negative samples. As a result, SVM achieved the highest AUC (Area Under Curve) of 0.996% and ACC (accuracy) of 97.1%, and RF also performed well with the similar prediction with AUC of 0.984% and ACC of 95.9%. The other two classifiers could not achieve satisfactory prediction results compared with SVM.

2.5. Feature Importance

In Figure 1B, we found that 72 features performed well for our prediction. However, each feature may not contribute equally to prediction. Subsequently, we used the T-set and IT-set with SVM to discover the importance of each feature.

We examined the contributions of two groups by plotting the ROC (Receiver Operating Characteristic) curves of each feature group (see Figure 2B and Table 3) and listing their evaluation indexes (see Table 3). As a result, it was found that fusional method with all 72 features achieved the best prediction accuracy of 0.971, and DC, SC achieved accuracies of 0.946 and 0.910 respectively. Notably, although DC achieved ACC of 0.946 which is greater than 0.910 achieved by SC, average contribution of each feature of SC is 0.910/7 = 0.13 which is nearly tenfold greater than that of DC (0.946/65 = 0.015). These results reveal that DC is the fundamental group of features and SC is an important complementary group of features for predicting DNA methylation.

Table 3. Comparison of two groups of features using independent testing on IT-set.

Feature Set	Dimension	ACC	AUC	MCC	Sens	Spec
All 72 Features	72	0.971	0.996	0.940	0.972	0.969
DNA Composition	65	0.946	0.990	0.892	0.953	0.939
Sequence complexity	7	0.910	0.968	0.819	0.911	0.909

Furthermore, we also evaluated the contribution of each feature by its feature importance in the SVM classifier, which was computed as the normalized regression coefficients of each feature in the linear kernel SVM. The feature importance map of the top 24 features according to their ranks is shown in Figure 3, in which GC content was shown to be the most important and five SC features (for more details of SC features, please refer to Material and Methods) were found in the top 24 features. Among them, SC-1 feature (the first feature of SC features, that is the fourth point of complexity function for the 600-bp window) achieved the second rank and SC-2 (the second feature of SC features, that is the fifth point of complexity function for the 600-bp window) was ranked in 9th, which was consistent with a previous work [60] that found the preceding points of entropy points (SC-2 here for 600 bp) were very important for describing complexity information of DNA sequence. In this work, we further found that the subsequent points of entropy points (SC-3 and SC-4, ranked in 11th and 7th, respectively) also provided additional information for prediction. Finally, traditional DNA motif features reported in previous works [40,48], such as TTT, AAA, CGG, CCG, etc., were also proven to be of high importance.

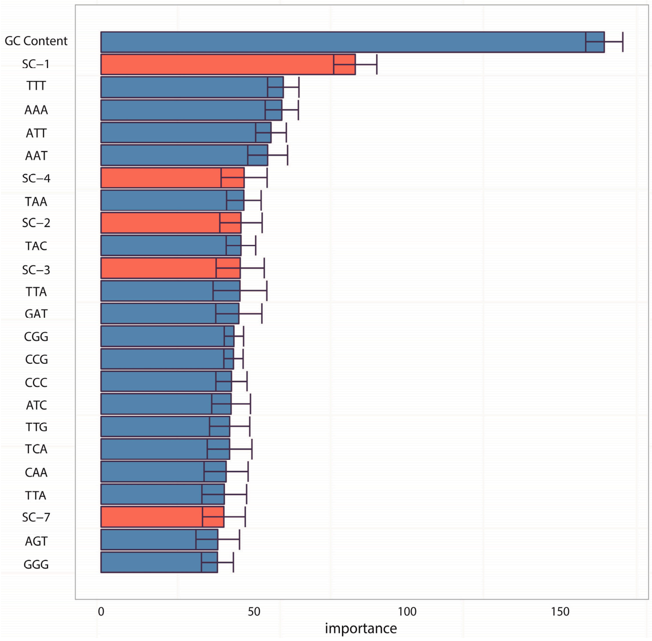

Figure 3. Top 24 important features by normalized regression coefficients in linear kernel SVM. The importance of the features was obtained by resampling statistics and the corresponding error bars of the top 24 features are represented. The colors represent different groups of features: DNA composition is blue, and sequence complexity is red.

To further show that the top 24 important features are significant, the remaining 48 (72 − 24 = 48) features were used to perform prediction and the corresponding predicting accuracy was only 90.7%, whereas the top 24 important features achieved a 95.8% prediction accuracy with only half the features (Figure 2C and Table 4).

Table 4. Comparison of top 24 important features and remaining 48 features using independent testing on IT-set.

Feature Set	Dimension	ACC	AUC	MCC	Sens	Spec
All 72 Features	72	0.971	0.996	0.940	0.972	0.969
Top 24 Important Features	24	0.958	0.990	0.915	0.956	0.959
Other 48 Features	48	0.907	0.969	0.815	0.910	0.905

2.6. Comparison with Other Existing Methods

To demonstrate the superiority of our method, a detailed comparison with existing methods was performed. A number of prediction methods were developed for identifying the methylation status of CpG sites, but not all methods are suitable for comparison, because some were constructed more than ten years ago and others did not make their dataset publicly available. Here, we chose two recently published existing methods with accessible datasets for comparison, **iDNA-Methyl** [41] and **DeepMethyl** [48].

For comparison with **iDNA-Methyl**, we downloaded its dataset (787 methylated samples and 1639 unmethylated samples) from their website (http://www.jci-bioinfo.cn/iDNA-Methyl) and then computed our 72 dimensional features. The detailed comparison result in based on evaluation indexes is shown in Table 5, from which we found that the overall prediction accuracy (ACC) of our method was 78.62%, outperforming **iDNA-Methyl**, which had an ACC of 77.49%. In addition, the sensitivity and specificity of our method are 78.68% and 78.56%, respectively, which are more balanced than those of **iDNA-Methyl** (61.25% and 90.33%, respectively). This implies that the prediction results are reliable for identifying both methylated samples and unmethylated ones, whereas **iDNA-Methyl** only accurately identified 60% of methylated samples, though it can precisely predict 90% of unmethylated samples.

Table 5. Comparison with **iDNA-Methyl**.

Predictor	ACC	MCC	Sens	Spec
iDNA-Methyl	77.49	54.71	61.25	90.33
Our work	78.62	57.23	78.68	78.56

For another predictor, **DeepMethyl**, we downloaded the methylation data for cell lines GM12878 (a B-lymphocyte cell line from a normal female) and K562 (an immortalized cell line from a female patient with chronic myelogenous leukemia) from the ENCODE project (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibMethylRrbs/, the GEO accession numbers and detailed number of positive and negative samples can be found in Table 1). For the results, satisfactory prediction accuracies were achieved for the GM12878 cell line (Table S2), and all 22 chromosomes had accuracies greater than 96% with an average accuracy of 97.93%. For K562 cell line, the prediction results were also very good and the average accuracy was 97.18% (Table S2). For comparison with **DeepMethyl**, we listed sectional comparisons because only those results appearing in Table 6 can be found in the **DeepMethyl** paper [48], and all the prediction results termed as ACC are displayed in Table S2. Interestingly, for GM12878 cell line, we found that our method achieved an ACC of 98.4%, outperforming **DeepMethyl**, which had an ACC of 90.0% in larger chromosome 1, and also outperforming it 98.3% compared to 94.2% in smaller chromosome 22. Similarly, our method outperformed **DeepMethyl** for K562 cell line. In addition, another contribution of this work is to provide comprehensive prediction results for all chromosomes in both the GM12878 and K562 cell lines, whereas **DeepMethyl** provided incomplete information in their paper.

Table 6. Comparison with **DeepMethyl**.

Cell Line	Predictor	Chromosome	Window Size	ACC	MCC	Sens	Spec
GM12878	**DeepMethyl**	Chr1	500	0.900	0.800	0.905	0.894
GM12878	Our method	Chr1	600	0.984	0.969	0.985	0.984
GM12878	**DeepMethyl**	Chr21	600	0.942	0.886	0.966	0.918
GM12878	Our method	Chr21	600	0.983	0.966	0.985	0.981
K562	**DeepMethyl**	Chr1	600	0.823	0.649	0.784	0.863
K562	Our method	Chr1	600	0.976	0.952	0.975	0.978
K562	**DeepMethyl**	Chr21	800	0.876	0.753	0.904	0.848
K562	Our method	Chr21	600	0.979	0.958	0.974	0.985

2.7. Cross-Species Prediction

To show that our method has certain generalization ability, we chose the mouse genome (another mammalian animal) for cross-species prediction. Fortunately, a similar recently-published paper studied mouse embryonic stem cell methylation status for the same eight stages [61]. We downloaded the corresponding dataset from the NCBI Gene Expression Omnibus (GEO, GSE56697, see Table 1). Subsequently, each sample was processed into a 600-bp DNA fragment with a methylated CpG site (positive sample) or a unmethylated CpG site (negative sample) in its center using the mouse reference genome mm10.

DNA methylation in mouse was predicted with very high accuracy, in fact even higher that in human (Table S3). The average accuracy of all 19 chromosomes was 97.37%, and the average AUC value was 0.9955. For convenience, we listed the important evaluation indexes of each chromosome in Table S3 and generated the ROC curves of three chromosomes out of 19 chromosomes for an intuitive display (Figure 2D and Table 7).

Table 7. Prediction results of three mouse chromosomes using our method.

Chromosome	Dimension	ACC	AUC	MCC	Sens	Spec
Mouse chromosome 1	72	0.980	0.998	0.961	0.984	0.977
Mouse chromosome 10	72	0.982	0.998	0.964	0.978	0.986
Mouse chromosome 19	72	0.940	0.985	0.879	0.926	0.953

2.8. Prediction of DNA Methylation Profiles across the Whole Human Genome

Although the prediction of methylation status of CpG sites is important, accurate prediction of the DNA methylation profile of specific genomic region (the average methylation level of all CpG sites within this region) is more practical when considering their biological functions. Therefore, we applied a robust strategy to show the practicability of our model by comparing the differences between the experimental methylation profiles and predicted methylation profiles (predicted by our trained model) of some important functional regions of the human genome. ChromHMM [62] is a commonly used method that systematically segments the human genome into different functional regions according to patterns in the presence or absence of multiple chromatin marks, such as the ChIP-Seq data of various histone modifications. Significantly, many published studies have adopted ChromHMM to display their results, including the famous Epigenome Roadmap Project [1].

Here, we also applied the annotation information provided by ChromHMM to show our results. For each sample within a specific region, we chose only the suitable DNA fragments that contained more than 10 "assayed" CpG sites (=those for which DNA methylation data are available) and 10 "unassayed" CpG sites (=those for which there is no experimental data for DNA methylation), and then calculated the average methylation level of such a sample (Figure 4A). Thus, only 10 regions (active transcription start site, flanking active transcription start site, active enhancer, weak enhancer, genic enhancer, strong transcription region, weak transcription region, repressed Polycomb state, weak repressed Polycomb state and quiescent state) out of 18 were selected to test the prediction because the sample numbers of unselected regions were not large enough to perform statistical tests (less than 1000).

As a result, two semi-violin plots that clearly show the differences between the experimental data and predicted data were shown in Figure 4B. Furthermore, a Wilcox test was employed to test the significant differences between the experimental and predicted results for all 10 regions, and 6 regions including active transcription start site, flanking active transcription start site, active enhancer, weak enhancer, strong transcription region and repressed Polycomb state were confirmed to be not statistically dissimilar (p-value > 0.01). The remaining regions, though statistically significant, did not appear to be discrepant (Figure 4B). Thus, our model has the ability to accurately predict the methylation profiles of important regions across the whole human genome, which also reveals the robustness of our work.

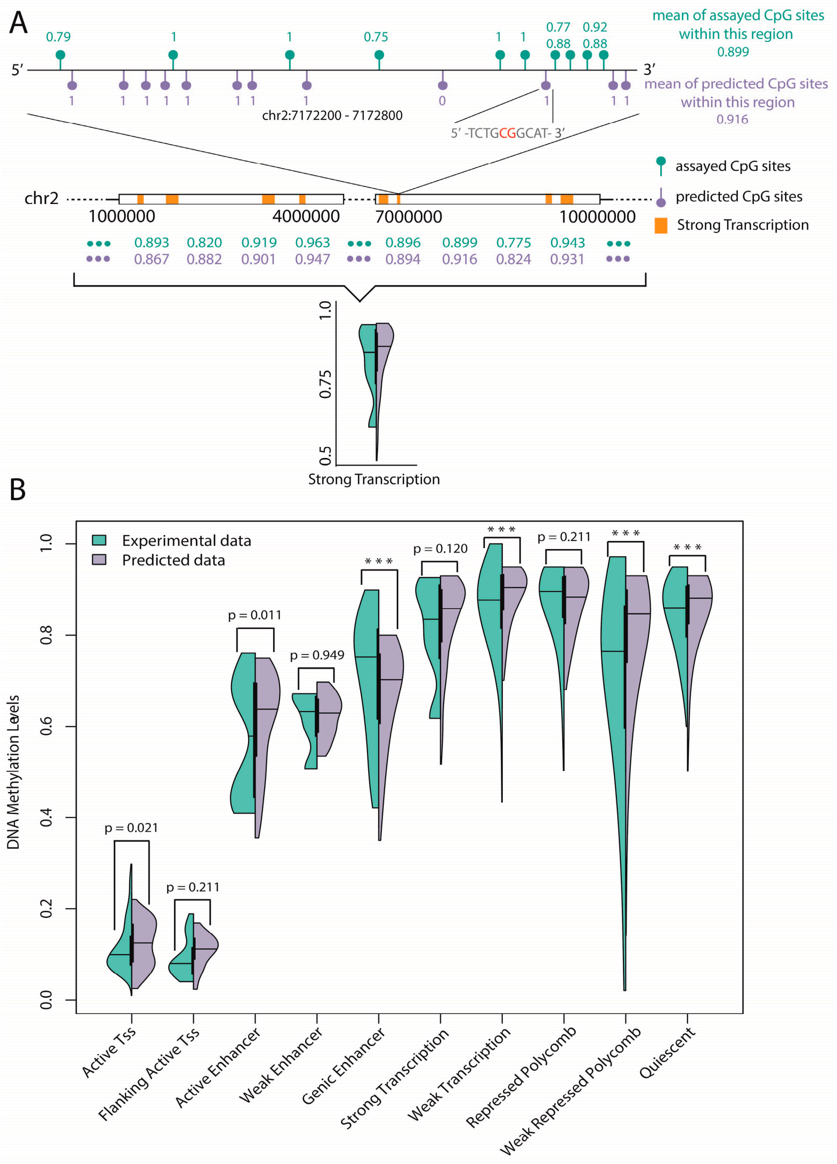

Figure 4. Statistical tests of the experimental data and the predicted data on 10 functional regions: (**A**) the computing procedure for one of 10 regions (Strong Transcription); and (**B**) the semi-violin plots show the distributions of average DNA methylation levels on 10 functional genomic regions. The *p* values of the six regions confirmed to be statistically consistent are labeled, and *** represents $p < 0.0001$.

3. Discussion

We proposed a novel algorithm that could accurately extract the sequence complexity features of DNA methylation status, and developed a SVM-based prediction model by integrating DNA composition features based on human embryonic stem cell methylation profiles. Different window sizes (100–1000 bp) and different chromosomes were combined pairwise to display the overall prediction results, and 600-bp windows whose center are methylated CpG sites were found to achieve the best accuracies.

In the analyses of feature importance, the feature group of DNA composition was found to be the fundamental features and the feature group of sequence complexity was found to be the important complementary features. From the ranking of Figure 3, it is worth noting that some components of sequence complexity feature group, such as SC-1 and SC-2, are important for prediction. Let us recall the definitions of SC-1 and SC-2 features (see Materials and Methods, 600-bp window): SC-1 feature is exactly the fourth point of complexity function and SC-2 feature is exactly the fifth point of complexity function. That means that methylated samples and unmethylated samples probably have different complexities of four-nucleotide and five-nucleotide usages of DNA sequence.

For a deep discussion, we further analyzed the distribution differences between methylated samples and unmethylated samples on SC-1 and SC-2 features (600-bp window) and then performed the corresponding statistical tests (*T* test). Both results were shown in Figure 5A,B, from which we found two interesting things: (1) The *p*-value on SC-1 feature is 0.00034 and the *p*-value on SC-2 feature is 0.00093, which shows that there are statistically significant differences between methylated samples and unmethylated ones on these two features. Therefore, it is not difficult to understand why the group of sequence complexity is an important group of features; (2) The variations of methylated samples are obviously lower than those of unmethylated samples for both two features (for SC-1 feature, variation of methylated samples equals to 60.62 and variation of unmethylated samples equals to 307.94; for SC-2 feature, variation of methylated samples equals to 90.47 and variation of unmethylated samples equals to 980.82), which implies that methylated samples tend to be more conservative with having more consistent complexity of four-nucleotide and five-nucleotide usages, whereas unmethylated samples tend to use them more randomly. That leads us to conjecture that DNA sequence motifs with length four or five are influential factors for DNA methylation, which should be validated by more future studies.

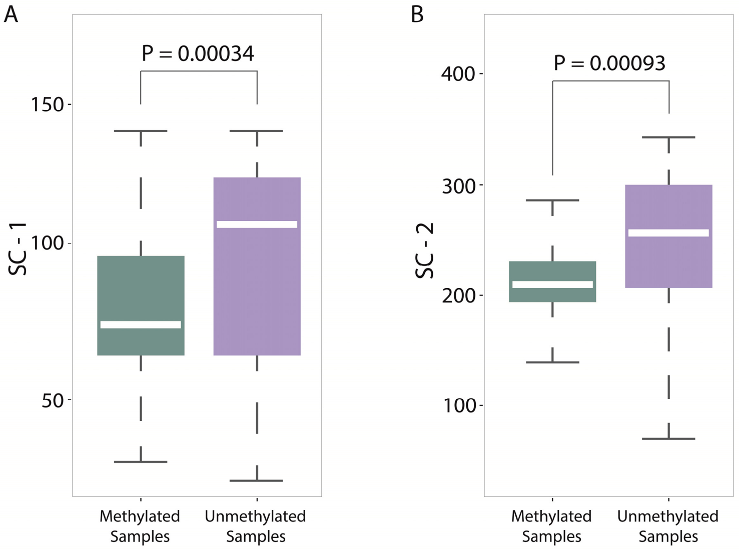

Figure 5. Distribution differences between methylated samples and unmethylated samples on two SC features: (**A**) box-plots of the distributions of methylated samples and unmethylated samples on SC-1 feature and corresponding statistical test (*p*-value = 0.00034); and (**B**) box-plots of the distributions of methylated samples and unmethylated samples on SC-2 feature and corresponding statistical test (*p*-value = 0.00093).

Moreover, we compared four common classifiers, and SVM resulted in the best records. We also compared two existing methods, **iDNA-Methyl** [41] and **DeepMethyl** [48] to show the advantages of our method. Additionally, we tested our model on methylation data of mouse genome for demonstrating that our model has certain generalization ability. Finally, a statistical test of experimental data and predicted data on 10 functional regions annotated by ChromHMM found that six regions were consistent, which implies reliable prediction of unassayed CpG sites.

Based on the above efforts, we summarize that our predictor brings some new benefits to the area of DNA methylation prediction:

1. Methodology: The biggest novelty of this work is the successful utilization of sequence complexity features for characterizing DNA methylation patterns. Earlier methods for selecting rational points when estimating topological entropy were not accurate and left out useful information. We provide a simple way to detect intrinsic features of sequence complexities, which were successfully used to predict DNA methylation status. Moreover, the feature importance analyses show that sequence complexities are the important complementary features.
2. Predicting window size: Previous works used different window sizes for predictions, such as 41 bp [41], 400 bp [46], and 100–1000 bp [48]. We found that 600 bp is the most appropriate by considering pairwise combinations of 22 chromosomes and different window sizes (from 100 to 1000 bp) based on a large dataset.
3. Prediction of unassayed CpG sites: A statistical test of the average methylation level of experimental CpG sites and predicted CpG sites (unassayed CpG sites, predicted by our trained model) on 10 functional regions annotated by ChromHMM found six regions were consistent. Based on this, we believe that the average methylation level of specific functional regulatory regions (such as promoters and enhancers) can be reliably predicted by our model.

The above in-depth analyses demonstrated the advantages of our computational model from different perspectives, however, there are still some limitations. For example, our predicting model only depends on the primary DNA sequence, which results in the same features extracted from different cells or cell lines. This might lead to the predicting deviation when predicting the DNA methylation level of different cells which show great changes in DNA methylation, especially between normal cells and cancer cells.

Moreover, important future work includes the investigation of the regulatory roles of DNA methylation, including the relationships between DNA methylation and transcription factor binding, especially in important regions of DNA regulatory elements, such as promoters and enhancers. Another future work is to establish a web-server to make our predictor publicly available just as shown in many recent publications (see, e.g., [52,53,55,57,63–69]) and to significantly enhance the influence of our predictor.

4. Materials and Methods

4.1. Features for Prediction

For our computational approach, each DNA fragment was represented as a numerical vector for inputting into the SVM for classification. In this work, the following two groups of features were used for formulating DNA fragments:

4.1.1. DNA Composition (DC, 65 Features)

1. Trinucleotide frequency (TriFre, 64 features)

Trinucleotide frequency is the simplest way to formulate DNA sequences. Precisely, a DNA sequence ω with L bases is denoted as:

$$\omega = R_1R_2R_3R_4R_5 \ldots R_L \tag{1}$$

Trinucleotide frequency of ω is defined as the normalized frequency of each trinucleotide in ω; i.e.,

$$TriFre = [f_1, f_2, f_{3,} \ldots, f_{64}]^T \tag{2}$$

where $f_i = \frac{n_i}{L-2}$, and n_i is to count the number of the i-th trinucleotide occurred in ω. Specifically, $f_1 = f(AAA) = \frac{n_1}{L-2}$, $f_2 = f(AAA) = \frac{n_2}{L-2}$ and n_1, n_2, are the occurrence number of AAA and AAC in ω, respectively.

2. GC content (GC, 1 features)

A number of previous studies all found that GC content was a predictive feature for methylation status [34,40,46], and here we also employ GC content as a feature. The detailed computational formula of GC content is simple:

$$GC = \frac{G + c}{L} \tag{3}$$

4.1.2. Sequence Complexity (SC, 7 Features)

The concept of sequence complexity originally came from the research area of "combinatorics on words", and it had wide applications in natural language processing, pattern matching and coding theory [70]. In general, it studies combinatorial features of sequence by investigating factor complexity, and complexity function and topological entropy are two important research topics in this area [60,71–75]. More precisely, to study DNA sequence, here we restricted attentions on a four-letter alphabet $\Omega = \{A, C, G, T\}$. Given a DNA sequence ω over Ω with finite length $|\omega|$, its complexity function is defined as:

Definition 1. *(Complexity Function, CF [70,71]). For a DNA sequence ω over Ω, the complexity function $p_\omega : N \to N$ is given by*

$$p_\omega(n) = \#\{u : u \prec \omega, |u| = n\} \tag{4}$$

where # denotes the number of elements of a set and $u \prec \omega$ represents that u is a factor (or a subword) of ω.

We show below, as an example, the complexity function of a short DNA fragment CAGATGTACA:

$$p_\omega(1) = 4,\ p_\omega(2) = 8,\ p_\omega(3) = 8,\ p_\omega(4) = 7,\ p_\omega(5) = 6$$
$$p_\omega(6) = 5,\ p_\omega(7) = 4,\ p_\omega(8) = 3,\ p_\omega(9) = 2,\ p_\omega(10) = 1.$$

The complexity function of a sequence describes its combinatorial features, which imply that the more different factors the sequence uses, the larger the CF value is. Mathematically, we can find that $1 \le p_\omega(n) \le 4^n$ for an arbitrary DNA sequence, and $p_\omega(n) = 1$ represents a complete repeated sequence, whereas $p_\omega(n) = 4^n$ indicates that the sequence contains all factors with full complexity.

Definition 2. *(Topological Entropy, TE [70,71]). For an infinite sequence ω over Ω, the topological entropy H_{top} is defined by*

$$H_{top}(\omega) = \lim_{n \to \infty} \frac{\log_4 p_w(n)}{n} \tag{5}$$

By the mathematical definition of TE, it is a quantitative index for describing the exponential increasing speed of CF. That is, if $H_{top} = c$, $p_\omega(n) \approx 4^{cn}$. An early related work found that randomly generated sequences had larger TE than DNA sequences, which means that DNA sequences are not randomly evolved and have certain conserved features [72]. Subsequently, a number of previous works used TE to quantitatively study the different regions across the human genome, which included promoters, exons and introns [71,72]. They found that intron > exon > promoter when considering TE [71,72]. These remarkable progresses motivated us to use CF and TE as mathematical tools to study pattern of methylation sites.

However, the computational algorithm of TE is not easy because the mathematical definition of TE is based on an infinite sequence. For improved understanding, we generated a complexity function graph (Figure 6A) for an example, and it was observed that CF increased rapidly at the first 7 points and subsequently decreased slowly. Recall that TE represents the exponential increasing speed of CF, so we only need to focus on the first seven points. Importantly, what points should we choose to estimate TE? In 2011, Koslicki [71] provided an answer by choosing the rational point n_0 as follows:

$$4^{n_0} + n_0 - 1 < |w| \leq 4^{n_0+1} + (n_0 + 1) - 1 \tag{6}$$

For our 100-bp DNA sequence, easy computation would determine n_0 as 3, which was marked in Figure 6A. Later, Jin et al. [60] proposed that other points also provided complexity information besides n_0, which should be considered together to give a more precise estimation of TE in 2014. They concluded that TE should be computed as:

$$H_{top}(\omega) = \frac{1}{k} \sum_{i=n_0-k}^{n_0} \frac{\log_4 p_\omega(i)}{i} \tag{7}$$

which implied that k preceding points of n_0 are brought together to computation (the authors used k = 3 in their paper).

In this paper, we developed a novel method to choose rational points of CF to be predictive features for DNA methylation. This was based on the original definition of TE, which is represented as the exponentially increasing speed of CF. Thus, it is significant to figure out the exponentially increasing part of the graph of CF. For this purpose, the difference operations were employed here to determine the exponentially increasing part:

$$\Delta p_\omega(k) = p_\omega(k+1) - p_\omega(k), k = 1, \ldots, L-1 \tag{8}$$

If successive points show non-zero patterns after several difference operations, they are probably the exponentially increasing part. More precisely, after two difference operations on complexity function, most of points were zero except for consecutive seven points, i.e., the first point to the seventh point (Figure 6A). By this way, seven points (1st–7th) were determined and were considered as the exponentially increasing part of this DNA sequence. Interestingly, these seven points not only contained n_0 which was considered as the topological entropy point, but also contained the three preceding points of n_0 (1st–3rd) which were proven to be important by Jin et al. [60]. This implies that our method of choosing rational points provide more information beyond two previous works.

Next, for investigating whether other DNA sequences of 100 bp all have a similar pattern, we performed the same strategy on 100 different DNA sequences and recorded the Last Point of the Exponentially Increasing Part EIP_{LP} of each DNA sequence, and the distribution of EIP_{LP} were shown by box plot in Figure 6B with median of 7. That means, most sequences have their exponentially increasing part from the 1st to 7th points, and the sequence complexity features were computed as:

$$SC_{100} = [p_\omega(1), \ldots, p_\omega(7)]^T \tag{9}$$

We applied the above procedure to different window sizes, and the corresponding EIP_{LP} were shown by box plots in Figure 6B. For example, the sequence complexity features of 600-bp DNA sequence were computed as:

$$SC_{600} = [p_\omega(4), \ldots, p_\omega(10)]^T \tag{10}$$

For more details, the exact interval for each window size was shown in Table 8. For showing the superiority of our method, we compared three methods—entropy point (1 dim), three preceding points of entropy point (3 dim), our rational points (7 dim) by ACC index (see Table S4).

Table 8. The determinations of seven sequence complexity features of different window sizes. $|\omega|$ represents the length of window, and n_0 represents the topological entropy point in reference [35].

$\|\omega\|$	n_0	Sequence Complexity Features (SC Features)	EIP_{LP}
100 bp	3	1–7	7
200 bp	3	2–8	8
300 bp	4	2–8	8
400 bp	4	3–9	9
500 bp	4	4–10	10
600 bp	5	4–10	10
700 bp	5	4–10	10
800 bp	5	4–10	10
900 bp	5	5–11	11

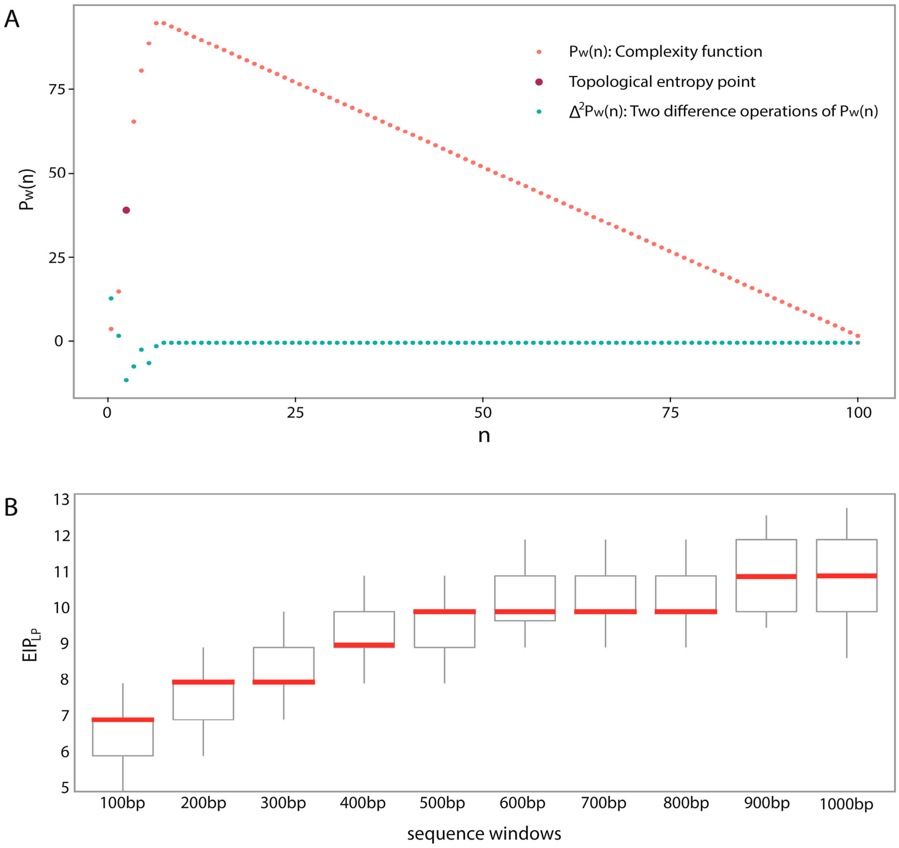

Figure 6. The determination approach of the exponentially increasing part of the complexity function graph: (**A**) The complexity function graph of 100-bp window DNA fragment with a methylated CpG site in the center. The red point represents the corresponding point of complexity function, the dark red point represents the point of the topological entropy point, and the blue point represents the corresponding point of two difference operations of complexity function; (**B**) The box plots of the distributions of EIP_{LP} for different window sizes.

4.2. Support Vector Machine

Support vector machine is a common-used machine-learning algorithm for classification or regression tasks. Because of its powerful advantages on easy implementation (only two parameters needed to determine) and small samples, it became one of the most successful machine-learning tools in the last two decades. Actually, SVM maps the input (feature vector) from Euclidean space into a higher dimensional Hilbert space by a suitable kernel function. Then, it searches the Optimal

Separating Hyper plane (OSH) to separate the positive samples and negative samples with the best accuracy by optimizing a given objective function. Comprehensive theory and its wide applications of SVM can be found in a famous monograph written by Vapnik [76]. In this research, we adopted a popular tool called "LibSVM 3.17" (open source, and can be downloaded from: http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html) to implement SVM with the linear kernel function.

4.3. Evaluating Indicator

Generally, the following three criteria are often used to evaluate a predictor for its prediction ability: independent testing, subsampling (K-fold cross-validation) test and jackknife test [77]. In this paper, we chose 10-fold cross-validation due to the new large dataset for all chromosomes and adopted independent testing for subsequent analyses. Usually, performance of a prediction method is measured by sensitivity (Sens), specificity (Spec), accuracy (ACC) and Matthew's correlation coefficient (MCC) value, calculated as:

$$\begin{cases} \text{Sens} = \frac{\text{TP}}{\text{TP+FN}} \\ \text{Spec} = \frac{\text{TN}}{\text{TN+FP}} \\ \text{ACC} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \\ \text{MCC} = \frac{\text{TP}\times\text{TN} - \text{FP}\times\text{FN}}{\sqrt{(\text{TP+FN})(\text{TP+FP})(\text{TN+FP})(\text{TN+FN})}} \end{cases} \quad (11)$$

where TP represents the number of true positives (methylated CpG sites predicted as methylated CpG sites) in one experiment, and TN represents the number of true negatives (unmethylated CpG sites predicted as unmethylated CpG sites). Similarly, FP represents the number of false positives (methylated CpG sites predicted as unmethylated CpG sites), and FN represents the number of false negatives (unmethylated CpG sites predicted as methylated CpG sites). Based on these four terms, we use the term "ACC" to represent the prediction accuracy of our model using Equation (11). For more intuitive and easier to understand these formulations, please refer to some recent publications [63–67,78,79]. Moreover, the above metrics is valid only for the single-label systems. For the multi-label systems whose existence has become more frequent in system biology and system medicine [56,80–82]. Additionally, to test the balance between true positive and false positive rates of a predictor, another important evaluating indicator is the Area Under the ROC Curve (AUC). The predictor is considered as a better predictor when the AUC value is larger.

Supplementary Materials: Supplementary materials can be found at www.mdpi.com/1422-0067/18/2/420/s1.

Acknowledgments: Xuehai Hu was partially supported by the National Natural Science Foundation of China (NSFC) (Grant No. 11671003, 11371016) and the Fundamental Research Funds for the Central Universities Huazhong Agricultural University (Grant No. 2662015PY182). Shixin Yao, Xinghao Li and Chujia Chen were supported by the National Creative Innovation Plan of College Students of China (Grant No. 201510504113) and Students' Program of Science and Technology Innovation (Grant No. 2016BC005).

Author Contributions: Xuehai Hu and Chengchao Wu conceived this project and designed the methodology; Shixin Yao and Xinghao Li implemented the algorithm of sequence complexity using Matlab; Chengchao Wu and Chujia Chen performed the statistical analysis; and Xuehai Hu drafted the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kundaje, A.; Meuleman, W.; Ernst, J.; Bilenky, M.; Yen, A.; Heravi-Moussavi, A.; Kheradpour, P.; Zhang, Z.; Wang, J.; Ziller, M.J. Integrative analysis of 111 reference human epigenomes. *Nature* **2015**, *518*, 317–330. [CrossRef] [PubMed]
2. Smith, Z.D.; Meissner, A. DNA methylation: Roles in mammalian development. *Nat. Rev. Genet.* **2013**, *14*, 204–220. [CrossRef] [PubMed]
3. Law, J.A.; Jacobsen, S.E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **2010**, *11*, 204–220. [CrossRef] [PubMed]

4. Larsen, F.; Gundersen, G.; Lopez, R.; Prydz, H. CpG islands as gene markers in the human genome. *Genomics* **1992**, *13*, 1095–1107. [CrossRef]
5. Cedar, H.; Bergman, Y. Programming of DNA methylation patterns. *Annu. Rev. Biochem.* **2012**, *81*, 97–117. [CrossRef] [PubMed]
6. Scarano, M.I.; Strazzullo, M.; Matarazzo, M.R.; D'Esposito, M. DNA methylation 40 years later: Its role in human health and disease. *J. Cell. Physiol.* **2005**, *204*, 21–35. [CrossRef] [PubMed]
7. Tost, J. DNA methylation: An introduction to the biology and the disease-associated changes of a promising biomarker. *Mol. Biotechnol.* **2010**, *44*, 71–81. [CrossRef] [PubMed]
8. Kim, S.; Li, M.; Paik, H.H.; Nephew, K.P.; Shi, H.; Kramer, R.; Xu, D.; Huang, T.H.-M. Predicting DNA methylation susceptibility using CpG flanking sequences. *Pac. Symp. Biocomput.* **2008**, *13*, 315–326.
9. Stadler, M.B.; Murr, R.; Burger, L.; Ivanek, R.; Lienert, F.; Schöler, A.; van Nimwegen, E.; Wirbelauer, C.; Oakeley, E.J.; Gaidatzis, D. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **2011**. [CrossRef] [PubMed]
10. He, X.; Chang, S.; Zhang, J.; Zhao, Q.; Xiang, H.; Kusonmano, K.; Yang, L.; Sun, Z.S.; Yang, H.; Wang, J. Methycancer: The database of human DNA methylation and cancer. *Nucleic Acids Res.* **2008**, *36*, D836–D841. [CrossRef] [PubMed]
11. Wolffe, A.P.; Matzke, M.A. Epigenetics: Regulation through repression. *Science* **1999**, *286*, 481–486. [CrossRef] [PubMed]
12. Das, P.M.; Singal, R. DNA methylation and cancer. *J. Clin. Oncol.* **2004**, *22*, 4632–4642. [CrossRef] [PubMed]
13. Lienert, F.; Wirbelauer, C.; Som, I.; Dean, A.; Mohn, F.; Schübeler, D. Identification of genetic elements that autonomously determine DNA methylation states. *Nat. Genet.* **2011**, *43*, 1091–1097. [CrossRef] [PubMed]
14. Taher, L.; Smith, R.P.; Kim, M.J.; Ahituv, N.; Ovcharenko, I. Sequence signatures extracted from proximal promoters can be used to predict distal enhancers. *Genome Biol.* **2013**, *14*, 1. [CrossRef] [PubMed]
15. Heyn, H.; Vidal, E.; Ferreira, H.J.; Vizoso, M.; Sayols, S.; Gomez, A.; Moran, S.; Boque-Sastre, R.; Guil, S.; Martinez-Cardus, A. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol.* **2016**, *17*, 1. [CrossRef] [PubMed]
16. Lister, R.; Pelizzola, M.; Dowen, R.H.; Hawkins, R.D.; Hon, G.; Tonti-Filippini, J.; Nery, J.R.; Lee, L.; Ye, Z.; Ngo, Q.-M. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **2009**, *462*, 315–322. [CrossRef] [PubMed]
17. Meissner, A.; Gnirke, A.; Bell, G.W.; Ramsahoye, B.; Lander, E.S.; Jaenisch, R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **2005**, *33*, 5868–5877. [CrossRef] [PubMed]
18. Sandoval, J.; Heyn, H.; Moran, S.; Serra-Musach, J.; Pujana, M.A.; Bibikova, M.; Esteller, M. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **2011**, *6*, 692–702. [CrossRef] [PubMed]
19. Laird, P.W. Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* **2010**, *11*, 191–203. [CrossRef] [PubMed]
20. Dohm, J.C.; Lottaz, C.; Borodina, T.; Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **2008**, *36*, e105. [CrossRef] [PubMed]
21. Campan, M.; Weisenberger, D.J.; Trinh, B.; Laird, P.W. Methylight. In *DNA Methylation: Methods and Protocols*; Humana Press: New York, NY, USA, 2009; pp. 325–337.
22. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247. [CrossRef] [PubMed]
23. Xu, Y.; Ding, J.; Wu, L.-Y.; Chou, K.-C. Isno-pseaac: Predict cysteine s-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS ONE* **2013**, *8*, e55844. [CrossRef] [PubMed]
24. Xu, Y.; Shao, X.-J.; Wu, L.-Y.; Deng, N.-Y.; Chou, K.-C. iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ* **2013**, *1*, e171. [CrossRef] [PubMed]
25. Zhang, J.; Zhao, X.; Sun, P.; Ma, Z. PSNO: Predicting cysteine S-nitrosylation sites by incorporating various sequence-derived features into the general form of Chou's PseAAC. *Int. J. Mol. Sci.* **2014**, *15*, 11204–11219. [CrossRef] [PubMed]

26. Jia, C.; Lin, X.; Wang, Z. Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile bayes and Chou's pseudo amino acid composition. *Int. J. Mol. Sci.* **2014**, *15*, 10410–10423. [CrossRef] [PubMed]
27. Xu, Y.; Wen, X.; Wen, L.-S.; Wu, L.-Y.; Deng, N.-Y.; Chou, K.-C. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PLoS ONE* **2014**, *9*, e105018. [CrossRef] [PubMed]
28. Qiu, W.-R.; Xiao, X.; Lin, W.-Z.; Chou, K.-C. iMethyl-PseAAC: Identification of protein methylation sites via a pseudo amino acid composition approach. *BioMed Res. Int.* **2014**, *2014*, 947416. [CrossRef] [PubMed]
29. Xu, Y.; Wen, X.; Shao, X.-J.; Deng, N.-Y.; Chou, K.-C. iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *Int. J. Mol. Sci.* **2014**, *15*, 7594–7610. [CrossRef] [PubMed]
30. Qiu, W.-R.; Xiao, X.; Lin, W.-Z.; Chou, K.-C. iUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. *J. Biomol. Struct. Dyn.* **2015**, *33*, 1731–1742. [CrossRef] [PubMed]
31. Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K.-C. iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. *Anal. Biochem.* **2016**, *497*, 48–56. [CrossRef] [PubMed]
32. Chou, K.-C. Impacts of bioinformatics to medicinal chemistry. *Med. Chem.* **2015**, *11*, 218–234. [CrossRef] [PubMed]
33. Xu, Y.; Chou, K.-C. Recent progress in predicting posttranslational modification sites in proteins. *Curr. Top. Med. Chem.* **2016**, *16*, 591–603. [CrossRef] [PubMed]
34. Bock, C.; Paulsen, M.; Tierling, S.; Mikeska, T.; Lengauer, T.; Walter, J. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet.* **2006**, *2*, e26. [CrossRef] [PubMed]
35. Fan, S.; Zhang, M.Q.; Zhang, X. Histone methylation marks play important roles in predicting the methylation status of CpG islands. *Biochem. Biophys. Res. Commun.* **2008**, *374*, 559–564. [CrossRef] [PubMed]
36. Zheng, H.; Wu, H.; Li, J.; Jiang, S.-W. CpGIMethPred: Computational model for predicting methylation status of CpG islands in human genome. *BMC Med. Genom.* **2013**, *6*, 1. [CrossRef] [PubMed]
37. Previti, C.; Harari, O.; Zwir, I.; del Val, C. Profile analysis and prediction of tissue-specific CpG island methylation classes. *BMC Bioinform.* **2009**, *10*, 1. [CrossRef] [PubMed]
38. Ma, B.; Wilker, E.H.; Willis-Owen, S.A.G.; Byun, H.-M.; Wong, K.C.C.; Motta, V.; Baccarelli, A.A.; Schwartz, J.; Cookson, W.O.C.M.; Khabbaz, K. Predicting DNA methylation level across human tissues. *Nucleic Acids Res.* **2014**, *42*, 3515–3528. [CrossRef] [PubMed]
39. Fang, F.; Fan, S.; Zhang, X.; Zhang, M.Q. Predicting methylation status of CpG islands in the human brain. *Bioinformatics* **2006**, *22*, 2204–2209. [CrossRef] [PubMed]
40. Das, R.; Dimitrova, N.; Xuan, Z.; Rollins, R.A.; Haghighi, F.; Edwards, J.R.; Ju, J.; Bestor, T.H.; Zhang, M.Q. Computational prediction of methylation status in human genomic sequences. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 10713–10716. [CrossRef] [PubMed]
41. Liu, Z.; Xiao, X.; Qiu, W.-R.; Chou, K.-C. iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. *Anal. Biochem.* **2015**, *474*, 69–77. [CrossRef] [PubMed]
42. Chen, W.; Lei, T.-Y.; Jin, D.-C.; Lin, H.; Chou, K.-C. PseKNC: A flexible web server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.* **2014**, *456*, 53–60. [CrossRef] [PubMed]
43. Chen, W.; Zhang, X.; Brooker, J.; Lin, H.; Zhang, L.; Chou, K.-C. PseKNC-general: A cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics* **2015**, *31*, 119–120. [CrossRef] [PubMed]
44. Liu, B.; Liu, F.; Fang, L.; Wang, X.; Chou, K.-C. 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* **2015**, *31*, 1307–1309. [CrossRef] [PubMed]
45. Chen, W.; Lin, H.; Chou, K.-C. Pseudo nucleotide composition or PseKNC: An effective formulation for analyzing genomic sequences. *Mol. BioSyst.* **2015**, *11*, 2620–2634. [CrossRef] [PubMed]
46. Zhang, W.; Spector, T.D.; Deloukas, P.; Bell, J.T.; Engelhardt, B.E. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol.* **2015**, *16*, 1. [CrossRef] [PubMed]

47. Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L.; Chou, K.-C. Pse-in-one: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* **2015**, *43*, W65–W71. [CrossRef] [PubMed]
48. Wang, Y.; Liu, T.; Xu, D.; Shi, H.; Zhang, C.; Mo, Y.-Y.; Wang, Z. Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks. *Sci. Rep.* **2016**, *6*, 19598. [CrossRef] [PubMed]
49. Bhasin, M.; Zhang, H.; Reinherz, E.L.; Reche, P.A. Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett.* **2005**, *579*, 4302–4308. [CrossRef] [PubMed]
50. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2013; Volume 6.
51. Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K.-C. pSuc-Lys: Predict lysine succinylation sites in proteins with pseaac and ensemble random forest approach. *J. Theor. Biol.* **2016**, *394*, 223–230. [CrossRef] [PubMed]
52. Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K.-C. iCar-PseCP: Identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC. *Oncotarget* **2016**, *7*, 34558–34570. [CrossRef]
53. Jia, J.; Zhang, L.; Liu, Z.; Xiao, X.; Chou, K.-C. pSumo-Cd: Predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. *Bioinformatics* **2016**, *32*, 3133–3141. [CrossRef]
54. Qiu, W.R.; Sun, B.Q.; Xiao, X.; Xu, D.; Chou, K.C. iPhos-PseEvo: Identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. *Mol. Inform.* **2016**. [CrossRef]
55. Qiu, W.-R.; Sun, B.-Q.; Xiao, X.; Xu, Z.-C.; Chou, K.-C. iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. *Oncotarget* **2016**, *7*, 44310. [CrossRef] [PubMed]
56. Qiu, W.-R.; Sun, B.-Q.; Xiao, X.; Xu, Z.-C.; Chou, K.-C. iPTM-mLys: Identifying multiple lysine PTM sites and their different types. *Bioinformatics* **2016**, *32*, 3116–3123. [CrossRef] [PubMed]
57. Qiu, W.-R.; Xiao, X.; Xu, Z.-C.; Chou, K.-C. iPhos-PseEn: Identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. *Oncotarget* **2016**, *7*, 51270–51283. [CrossRef] [PubMed]
58. Guo, H.; Zhu, P.; Yan, L.; Li, R.; Hu, B.; Lian, Y.; Yan, J.; Ren, X.; Lin, S.; Li, J. The DNA methylation landscape of human early embryos. *Nature* **2014**, *511*, 606–610. [CrossRef] [PubMed]
59. Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics* **2010**, *26*, 680–682. [CrossRef] [PubMed]
60. Jin, S.; Tan, R.; Jiang, Q.; Xu, L.; Peng, J.; Wang, Y.; Wang, Y. A generalized topological entropy for analyzing the complexity of DNA sequences. *PLoS ONE* **2014**, *9*, e88519. [CrossRef] [PubMed]
61. Wang, L.; Zhang, J.; Duan, J.; Gao, X.; Zhu, W.; Lu, X.; Yang, L.; Zhang, J.; Li, G.; Ci, W. Programming and inheritance of parental DNA methylomes in mammals. *Cell* **2014**, *157*, 979–991. [CrossRef] [PubMed]
62. Ernst, J.; Kellis, M. Chromhmm: Automating chromatin-state discovery and characterization. *Nat. Methods* **2012**, *9*, 215–216. [CrossRef] [PubMed]
63. Chen, W.; Tang, H.; Ye, J.; Lin, H.; Chou, K.-C. iRNA-PseU: Identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids* **2016**, *5*, e332.
64. Chen, W.; Ding, H.; Feng, P.; Lin, H.; Chou, K.-C. Iacp: A sequence-based tool for identifying anticancer peptides. *Oncotarget* **2016**, *7*, 16895. [CrossRef] [PubMed]
65. Chen, W.; Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chou, K.-C. IRNA-Ai: Identifying the adenosine to inosine editing sites in RNA sequences. *Oncotarget* **2016**, *5*. [CrossRef] [PubMed]
66. Xiao, X.; Ye, H.-X.; Liu, Z.; Jia, J.-H.; Chou, K.-C. iROS-gPseKNC: Predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. *Oncotarget* **2016**, *7*, 34180. [CrossRef] [PubMed]
67. Zhang, C.-J.; Tang, H.; Li, W.-C.; Lin, H.; Chen, W.; Chou, K.-C. iOri-Human: Identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget* **2016**, *7*, 69783–69793. [CrossRef] [PubMed]

68. Liu, B.; Wu, H.; Zhang, D.; Wang, X.; Chou, K. Pse-analysis: A python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods. *Oncotarget* **2017**. [CrossRef] [PubMed]
69. Liu, Z.; Xiao, X.; Yu, D.-J.; Jia, J.; Qiu, W.-R.; Chou, K.-C. pRNAm-PC: Predicting N^6-methyladenosine sites in RNA sequences via physical–chemical properties. *Anal. Biochem.* **2016**, *497*, 60–67. [CrossRef] [PubMed]
70. Lothaire, M. *Applied Combinatorics on Words, Volume 105 of Encyclopedia of Mathematics and Its Applications*; Cambridge University Press: London, UK, 2005.
71. Koslicki, D. Topological entropy of DNA sequences. *Bioinformatics* **2011**, *27*, 1061–1067. [CrossRef] [PubMed]
72. Colosimo, A.; de Luca, A. Special factors in biological strings. *J. Theor. Biol.* **2000**, *204*, 29–46. [CrossRef] [PubMed]
73. Kirillova, O.V. Entropy concepts and DNA investigations. *Phys. Lett. A* **2000**, *274*, 247–253. [CrossRef]
74. Schmitt, A.O.; Herzel, H. Estimating the entropy of DNA sequences. *J. Theor. Biol.* **1997**, *188*, 369–377. [CrossRef] [PubMed]
75. Troyanskaya, O.G.; Arbell, O.; Koren, Y.; Landau, G.M.; Bolshoy, A. Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity. *Bioinformatics* **2002**, *18*, 679–688. [CrossRef] [PubMed]
76. Vapnik, V.N. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998; Volume 1.
77. Chou, K.-C.; Zhang, C.-T. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275–349. [CrossRef] [PubMed]
78. Chen, W.; Feng, P.-M.; Lin, H.; Chou, K.-C. iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* **2013**, *41*, e68. [CrossRef] [PubMed]
79. Lin, H.; Deng, E.-Z.; Ding, H.; Chen, W.; Chou, K.-C. iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.* **2014**, *42*, 12961–12972. [CrossRef] [PubMed]
80. Cheng, X.; Zhao, S.-G.; Xiao, X.; Chou, K.-C. iATC-mISF: A multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics* **2016**, *33*, 341–346. [CrossRef] [PubMed]
81. Chou, K.-C.; Wu, Z.-C.; Xiao, X. iLoc-Hum: Using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.* **2012**, *8*, 629–641. [CrossRef] [PubMed]
82. Chou, K.-C. Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.* **2013**, *9*, 1092–1100. [CrossRef] [PubMed]