Published online 23 October 2009

Nucleic Acids Research, 2010, Vol. 38, No. 1 ***e6***
doi:10.1093/nar/gkp882

CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes

Jianzhong Su, Yan Zhang*, Jie Lv, Hongbo Liu, Xiaoyan Tang, Fang Wang, Yunfeng Qi, Yujia Feng and Xia Li*

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 150081, China

Received May 22, 2009; Revised September 30, 2009; Accepted October 2, 2009

ABSTRACT

CpG islands (CGIs) are CpG-rich regions compared to CpG-depleted bulk DNA of mammalian genomes and are generally regarded as the epigenetic regulatory regions in association with unmethylation, promoter activity and histone modifications. Accurate identification of CpG islands with epigenetic regulatory function in bulk genomes is of wide interest. Here, the common features of functional CGIs are identified using an average mutual information method to differentiate functional CGIs from the remaining CGIs. A new approach (CpG mutual information, CpG_MI) was further explored to identify functional CGIs based on the cumulative mutual information of physical distances between two neighboring CpGs. Compared to current approaches, CpG_MI achieved the highest prediction accuracy. This approach also identified new functional CGIs overlapping with gene promoter regions which were missed by other algorithms. Nearly all CGIs identified by CpG_MI overlapped with histone modification marks. CpG_MI could also be used to identify potential functional CGIs in other mammalian genomes, as the CpG dinucleotide contents and cumulative mutual information distributions are almost the same among six mammalian genomes in our analysis. It is a reliable quantitative tool for the identification of functional CGIs from bulk genomes and helps in understanding the relationships between genomic functional elements and epigenomic modifications.

INTRODUCTION

CpG islands (CGIs) are generally considered as the epigenetic and functional elements (1,2). CpG dinucleotides are notably depleted in mammalian genomes where the observed frequency of CpG dinucleotides is only 0.20–0.25 of the expected from the G and C base composition (3). In contrast, CpG islands are CpG enriched regions where the frequency of CpGs is exceptionally high (4). At least 50% of gene promoters are associated with CGIs, which can be considered as gene markers (5,6). DNA methylation is usually absent in CGIs, compared to about 70–80% of CpG dinucleotides methylation in human somatic cells (7). CpG island methylation plays important roles in gene expression (8), X chromosome inactivation (9) and genomic imprinting (10). Hypermethylation of promoter CGIs is associated with the silence of tumor suppressor genes in human cancer cells (11). Functional CGIs are generally associated with epigenetic regulatory function, such as the unmethylation, promoter activity and histone modifications (2,12).

Currently, two approaches have been widely used for CGIs identification: the computational method based on the DNA sequence features (3) and the experimental method based on the absence of methylated CpG dinucleotides (4,13). A common problem of computational method is that a proportion of identified CGIs are actually Alu repetitive elements. Alu repetitive elements are abundant mobile elements about 300 base pairs (bp) long in human genome with high G + C content and high CpG observed-to-expected ratio (CpG O/E). Alu repetitive elements are often falsely identified as CGIs by the original criteria of Gardiner-Garden and Frommer ($G + C$ content $\geq 50\%$, CpG $O/E \geq 0.6$ and length ≥ 200 bp) (3). Alu repetitive elements are usually highly methylated and transcriptionally silent, which is contradictory to the original definition of CGIs by Bird (14). Thus, the CGIs overlapped by Alu repetitive elements are usually considered as the CpG clusterings rather than functional CGIs. The overlapping rate between the CGIs and Alu repetitive elements has become an important criterion for assessing the false positive rate for various algorithms of CGIs identification. The original

*To whom correspondence should be addressed. Tel/Fax: +86 451 866 67543; Email: yanyou1225@yahoo.com.cn
Correspondence may also be addressed to Xia Li. Tel/Fax: +86 451 866 15922; Email: lixia@hrbmu.edu.cn

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

© The Author(s) 2009. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/article/38/1/e6/3112467 by guest on 24 April 2024

criteria for identifying CGIs exhibit high false positive rate, since about 60% of the identified CGIs are dependent on Alu repetitive elements (15). Takai and Jones developed more stringent criteria to reduce the overlapping frequency between CGIs and Alu repetitive elements (G + C content $\geq$55%, CpG O/E $\geq$ 0.65 and length $\geq$500 bp) (16). Using Takai and Jones's criteria, the algorithms of CpGProD (17) and CpGIS (18) were developed to identify CGIs from human and mouse genomes. However, a part of functional CGIs were missed because of the stringent constraints on these parameters. These functional CGIs are usually shorter in length, lower in CpG O/E or G + C content than the ad hoc thresholds. Moreover, these sliding window methods are computationally inefficient as the identification depends on the window size and step size. CpGcluster is a new algorithm based on the physical distances between neighboring CpGs. This approach is independent of the three parameters mentioned above and is computationally more efficient because of its exclusive use of integer arithmetic in scanning the linear chromosomes (19). However, it still has the drawback of low sensitivity, partly due to the criteria of short length, high CpG O/E and G + C content. CpGIF (20) extended the algorithm of CpGcluster. It detected high CpG content regions as seeds and then extended them to final CGIs iteratively by reducing the content. This algorithm increases the length of the identified CGIs and significantly improves the sensitivity. However, it has low specificity and the constraints of the original criteria of Gardiner-Garden and Frommer.

Few computational algorithms have been developed for the identification of CGIs of non-human mammalian genomes. One reason is that it is unclear whether the criteria for human CGIs can be applied to other mammalian genomes, especially when taking different genomic features among different species into consideration. Another reason is that the scarcity of epigenetic data for these mammalian species is available. Han *et al*. (21) have identified the CGIs of ten mammalian genomes using three different computational algorithms. Takai and Jones' stringent criteria were considered as the most appropriate method for the CGIs identification of mammalian species. However, the stringent criteria would result in loss of the functional CGIs in the human and other species' genomes.

In addition to computational methods, the experimental CGI library has been constructed according to the feature of frequent lack of methylated cytosine in CGI regions (4,13). Illingworth *et al*. have detected many new CGIs in human blood genomic DNA that were not identified by current computational algorithms. However, the experimental method is limited to one tissue type or one cell type, and mainly depends on the feature of absence of CpG methylation. A fraction of functional CGIs are likely to be missed, as CpG methylation is tissue specific or cell specific. To take advantages of both computational and experimental approaches, Bock *et al*. predicted the combined epigenetic scores of CGIs using the conventional sliding window method by incorporating DNA sequence features with epigenetic state, such as DNA methylation, histone modifications and chromatin accessibility (2). The 'bona fide' CGIs having higher values ($>$0.5) were distinguished from the remaining ($\leq$0.5) by their combined epigenetic scores of CGIs. This strategy provides a new prospect for identifying functional CGIs by incorporation of genomic and epigenomic information. However, all the CGIs predicted by Bock *et al*. rely on the conventional sliding window method and genome-wide experimental epigenetic data. Therefore, the CGIs predicted by this approach are confined to the conventional thresholds. It is also difficult to apply this strategy to non-human species due to paucity of epigenetic data of these species. Analyzing common features of functional CGIs in bulk genome has always been challenging. The investigation regarding CpG dinucleotides enrichment in CpG island regions prompted us to analyze mutual information of physical distances between two neighboring CpGs. We then further investigated whether these functional CGIs had common sequence features by the theory of mutual information. A modified mutual information method was implemented to quantify how well these enriched CpG dinucleotides were correlated with each other, and to provide more accurate identification of epigenetic regulatory regions. In fact, mutual information is a useful tool for identifying different segments that are correlated in the genomes. Mutual information between single nucleotides as signatures of DNA sequences has been widely used in bioinformatics research (22,23).

In this study, a new mutual information algorithm was developed to identify functional CGIs based on the physical distances between two neighboring CpGs. Common features of 'bona fide' CGIs have been explored by calculating the average mutual information (AMI) of the physical distances between two neighboring CpGs. The AMIs of functional CGIs were remarkably different from that of the remaining CGIs and random segments (Figure 1). CpG_MI, a more precise and efficient tool, was developed to differentiate functional CGIs from the bulk genome using cumulative mutual information (CMI). This approach achieved the highest prediction accuracy among current approaches. CpG_MI has successfully identified 40 926 CGIs from human genome. It achieved the highest coverage rate with experimental unmethylated CGIs identified by Illingworth *et al*. Moreover, almost all CGIs identified by this approach are associated with histone modification marks. H3K4me3 is strongly correlated with gene transcription activation (24,25). About 81.6% of the CGIs identified by CpG_MI overlapped with H3K4me3 marks. The CMIs of random segments of bulk genome from human and other five mammals were also investigated. The CpG contents in the six mammalian species are almost identical, and the CMI distributions of random segments from six mammalian genomes follow nearly the same exponential distribution. As the model of mutual information for identification of CGIs mainly relies on the CpG content and CMI distribution of the bulk genome, CpG_MI could also be used for identifying functional CGIs in other mammalian genomes. Based on CpG_MI, a web service was developed for CGIs identification, which is available at http://bioinfo.hrbmu.edu.cn/cpgmi/.

Downloaded from https://academic.oup.com/nar/article/38/1/e6/3112467 by guest on 24 April 2024

MATERIALS AND METHODS

Genome datasets

The sequences and corresponding genome information of six mammals and four fish species were downloaded from the UCSC Genome Browser (26). These mammalian genomes are from human (hg18), chimpanzee (panTro2), mouse (mm9), rat (rn4), cow (bosTau4) and dog (canFam2). Four fish genomes are from zebrafish (danRer5), medaka (oryLat2), stickleback (gasAcu1) and tetraodon (tetNig1). Random genomic segments of these mammalian genomes were retrieved using the UCSC Galaxy tool (27) and Alu repetitive elements were downloaded from the UCSC Table Browser (28).

Histone modification datasets

The histone modification data were obtained from Barski *et al.* (25) and Wang *et al.* (29), where ChIP-seq experiments were used to sequence histone methylation and acetylation modifications in human CD4+T cells. It is the most comprehensive human genome-scale high-resolution profiles including 20 hisonte methylations and 18 histone acetylations.

CGIs library

The library of 14 022 experimental CGIs was obtained from Illingworth *et al.* (4), where nonmethylated CpG affinity chromatography technique was used. Human CGIs with their combined epigenetic scores predicted by Bock *et al.* based on the criteria of Gardiner-Garden and Frommer (G + C content $\geq$ 50%, CpG O/E $\geq$ 0.6, and length $\geq$ 200 bp) were downloaded from http://neighborhood.bioinf.mpiinf.mpg.de/CpG_islands_revisited (2). The four sets of CGIs were defined as: B1 (0–0.33), B2 (0.33–0.50), B3 (0.50–0.67) and B4 (0.67–1) according to the range of combined epigenetic scores of CGIs classified by Bock *et al.*

Distribution regions of genome

The bulk genome assemblies were divided into three parts as classification in CGIs proposed by Ioshikhes and Zhang (5): promoter, intragenic and intergenic regions. Promoter region is defined as the region between 2k bp upstream of transcriptional start site (TSS) and the end of first exon. Intragenic region is defined as the region from the end of first exon to 2k bp downstream of transcriptional end site (TES). Intergenic regions are defined as the remaining regions in a bulk genome except the promoter and intragenic regions.

Test set for evaluation

To assess the prediction accuracy of CpG_MI, we compared it with other algorithms by evaluating procedure including the sensitivity (SN), specificity (SP), accuracy (ACC) and correlation coefficient (CC). SN represents the proportion of experimental CGIs that have been correctly identified as CGIs. SP represents the proportion of random segments that have been correctly identified as random segments. ACC and CC are two important global accuracy scalars that are used to balance SN and SP. The 14 022 experimental CGIs (4) were set as the positive CGIs of the test set. Correspondingly, 14 022 segments were randomly selected as the negative CGIs of the test set from the human bulk genome. These negative CGIs do not overlap with the experimental CGIs of the test set and have the same length distribution as the experimental CGIs. In order to evaluate various algorithms, 1000 sequence segments were randomly chosen respectively from the positive CGIs and negative CGIs of the test set. This sampling process was repeated ten times. The calculation formulas for SN, SP, ACC and CC are defined as follows:

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},$$

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}},$$

$$\mathrm{CC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}},$$

where TP, TN, FP and FN represent true-positives, true-negatives, false-positives and false-negatives respectively.

Average mutual information and cumulative mutual information of DNA sequence segments

The average mutual information of two neighboring CpGs was developed to distinguish the mutual information distributions of distances between two neighboring CpGs among the sets of CGIs and random segments. The location of CpG dinucleotide is treated as a variable in DNA sequence segments consisting of CGIs and random segments. The number of nucleotides between two neighboring CpGs is defined as the physical distance k between two neighboring CpGs, e.g. '$CG\underbrace{ACTC\cdots AA}_{k}CG$'. Suppose that S is a set of n sequence segments, the formula of average mutual information of two neighboring CpGs with distance k of the set S is defined as:

$$AMI(k) = \frac{1}{n_k}\sum_{i=1}^{n} p^{(i)}_{(cg,cg)}(k)\log\frac{p^{(i)}_{(cg,cg)}(k)}{p_{cg}\times p_{cg}}, k = 1,2\cdots 100. \quad \mathbf{1}$$

where p_{cg} is the probability of occurrence of CpG dinucleotides in a bulk genomic sequence and n_k represents the number of sequence segments ($p^{(i)}_{(cg,cg)}(k) \neq 0, i = 1,2,\cdots,n$). In order to distinguish effectively the different sets of CGIs and random segments using $AMI(k)$, the sequence segments ($p^{(i)}_{(cg,cg)}(k) = 0, i = 1,2,\cdots,n$) were filtered out in the process of computing $AMI(k)$ in the set S. $p^{(i)}_{(cg,cg)}(k)$ represents the probability of occurrence of two neighboring CpGs with distance k of the i-th measured sequence segment in set S and is defined as $p^{(i)}_{(cg,cg)}(k) = \frac{f^{(i)}_k}{N^{(i)}}, i = 1,2,\cdots,n$, where $N^{(i)}$and $f^{(i)}_k$ represent respectively the length and frequency of occurrence of two neighboring CpGs with distance k of the i-th measured sequence segment.

Downloaded from https://academic.oup.com/nar/article/38/1/e6/3112467 by guest on 24 April 2024

The AMI was developed to reveal the distinction of mutual information distributions of distances between two neighboring CpG dinucleotides among the different CGI sets and the random segment set. To quantitatively identify functional CGIs with high mutual information from human bulk genome, the cumulative mutual information (CMI) was developed to measure the accumulation of mutual information of neighboring CpG dinucleotides with different distances in DNA sequence segments. The formula of cumulative mutual information of the i-th measured sequence segment in set S is defined by

$$CMI(i) = \sum_{k=0}^{M} p^{(i)}_{(cg,cg)}(k) \log \frac{p^{(i)}_{(cg,cg)}(k)}{p_{cg} \times p_{cg}}, \quad i = 1, 2 \ldots, n. \qquad \mathbf{2}$$

Where M is the maximum value of cumulative distance k between two neighboring CpGs of measured DNA sequence segments.

RESULTS

Characteristics of functional CGIs and the algorithm of CpG_MI

To find the characteristics of functional CGIs in human genome, the AMI was used to analyze the relationship among the four sets of CGIs divided by Bock *et al.* (2) and Random (random segments set). The combined epigenetic scores of the above four sets of CGIs were grouped as B1 (0–0.33), B2 (0.33–0.50), B3 (0.50–0.67) and B4 (0.67–1). The CGIs with combined epigenetic scores >0.5 are the 'bona fide' CGIs, which was determined by Bock *et al.* (2). Because the 'bona fide' CGIs are strongly associated with epigenetic regulatory function, these CGIs are considered as functional CGIs.

The AMI profiles show that B1 and B2 have the similar AMI distribution along the physical distances from 0 to 100 bp. In analogy, B3 and B4 also have similar AMI distribution. However, B3 and B4 can be distinguished from B1 and B2 by their AMI distributions. The AMI distributions of B3 and B4 are significantly higher than that of B1 and B2 when the distances are between 0 and 10 bp. When the distances are greater than 10 bp, a reversed trend was observed (Figure 1). It indicates that the AMI profile can be used to differentiate the 'bona fide' CGIs (in B3 and B4) from the remaining CGIs (in B1 and B2). Moreover, the AMIs of 100 000 random segments from human genome were computed. We found there was no significant difference of the AMIs of random segments when physical distance k varied from 0 to 10 bp (Figure 1). Therefore, the AMI distribution demonstrates that these functional CGIs have higher mutual information of physical distances between two neighboring CpGs than the remaining CGIs and the random segments following physical distance k ranging from 0 to 10 bp. In order to confirm this observation, all 'bona fide' CGIs were scanned. The results show that 83.4% of 'bona fide' CGIs (in B3 and B4) contain one or more hot spot regions (HSR) consisting of at least five CpG dinucleotides, where HSR is defined as CpG loci as the start and end boundary, and all neighboring distances of successive CpGs are ≤10 bp. As for the CGIs in B4, there are 94.5% of CGIs containing one or more HSRs. When 'bona fide' CGIs are equally divided into three parts by use of tripartite method, most of these hot spot regions locate in the middle regions of the 'bona fide' CGIs as shown in Supplementary Figure S1, for example. It can be concluded that functional CGIs have two important features: (i) each CGI contains at least one hot spot region and (ii) hot spot regions usually locate in the middle regions of the functional CGIs.

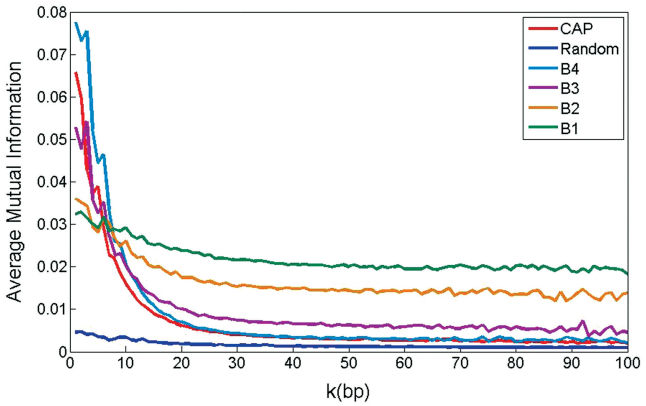

Figure 1. The distribution of average mutual information. The Figure shows that the distribution of the average mutual information of the distances between the two neighboring CpGs from 0 bp to 100 bp for each set of CGIs. B1 (0–0.33), B2 (0.33–0.50), B3 (0.50–0.67) and B4 (0.67–1) represent different sets of CGIs, which fall into the different intervals in terms of these combined epigenetic scores predicted by Bock *et al.* (2). CAP denotes the set consisting of experimental CGIs identified by Bird (1) using nonmethylated CpG affinity chromatography technique, and random represents the set of 100 000 random segments downloaded from UCSC (hg18) in the length range from 50 bp to 5000 bp. 'k' represents the distance between two neighboring CpGs in a CGI or a random segment.

The CMIs of CGIs in five different sets: B1, B2, B3, B4 and Random were calculated (Formula 2) to quantitatively evaluate the impact of mutual information of different distances between two neighboring CpGs on identifying functional CGIs from human genome. Taking the significant impact of the distances (≤10 bp) into consideration, the CMIs of CGIs in the above five sets were first computed having M set to 10 (Figure 2). The results clearly demonstrate that the CMIs of functional CGIs (in B3 and B4) are distinctly different from that of nonfunctional CGIs (in B1 and B2) as well as the random segments. Whether mutual information of the longer distances between two neighboring CpGs impacted on the effectiveness of classification between the functional CGIs and the nonfunctional CGIs was further evaluated. We found that the effect on classification was similar when M was set to 30 and 50 (Figure 2). In addition, the AMIs of identified functional CGIs had no significant difference from the random segments when the distance k is larger than 50 bp from Figure 1. Therefore, $M = 50$ should be used as the maximum of cumulative distance. Notably, the AMIs of sets B1 and

Downloaded from https://academic.oup.com/nar/article/38/1/e6/3112467 by guest on 24 April 2024

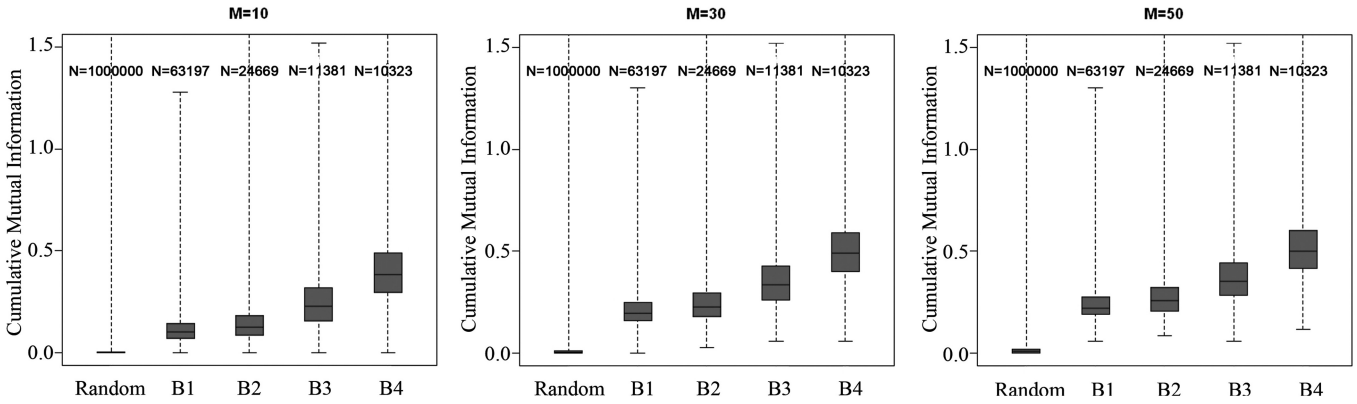

Figure 2. Box plots comparing the cumulative mutual information among five sets of CpG islands. These Figures show box plots of the cumulative mutual information of five sets of CpG islands (Random, B1, B2, B3 and B4 as mentioned in Figure 1). The three figures represent respectively their cumulative mutual information for M = 10, M = 30 and M = 50, where boxes show center quartiles, whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range from the box, and outliers are hidden.

B2 are significantly greater than zero when distance k varies from 30 to 50 bp, while the CMIs of sets B1 and B2 have no significant difference between $M = 30$ and $M = 50$. The main reason is that the CGIs ($p^{(i)}_{(cg,cg)}(k) = 0$) are filtered out in the process of computing AMIs of sets B1 and B2 and most $p^{(i)}_{(cg,cg)}(k)$ of CGIs in sets B1 and B2 are equal to 0 for distance $M = 50$ from 30 to 50 bp. Therefore, the CMI values of CGIs in the sets B1 and B2 have no significant increase when M varies from 30 to 50. We further computed the CMIs of hot spot regions located in the middle regions of the 'bona fide' CGIs and in the 100 000 random segments when M was set to 50. The results show that CMIs of the hot spot regions are larger than 0.8, and the percentage of CMIs $\geq$0.25 from the 100 000 random segments is less than 1%. Therefore, the threshold of the CMIs of functional CGIs is set to 0.25, which can significantly differentiate from the random segments from human genome.

Based on the above findings, the search algorithm of CpG_MI mainly consists of five steps:

Step 1: Scan each DNA sequence from 5′ to 3′ direction to look for all CpG dinucleotides and record their positions.
Step 2: Calculate the distance (k) of each two neighboring CpGs and then scan the distances between two neighboring CpGs from 5′ to 3′ direction to cluster the CpGs into subsequences while k is less than or equal to 10 bp.
Step 3: Compute the numbers of CpGs and the CMI of each subsequence. The subsequence is identified as hot spot region if the number of CpGs in the subsequences is greater than 5 and the CMI is larger than 0.8. Otherwise, it is filtered out.
Step 4: Extend the hot spot region to upstream and downstream iteratively by adding CpGs, and calculate the CMI of extended CpG clustering. Iteration stops when CMI of extended CpG clustering is less than 0.25. If the length of extended CpG clustering is <50 bp, it is filtered out. Return to step 2.
Step 5: Cluster two neighboring extended CpG clusterings together if the distance between them is <100 bp.

The algorithm of CpG_MI was implemented using Perl language. A web service and the executable Perl scripts are available to public at http://bioinfo.hrbmu.edu.cn/cpgmi/. CpG_MI was developed to identify the functional CGIs of different mammalian genomes. Due to differences of CpG content in different mammalian species, a species should be selected before the sequence is uploaded to CpG_MI. The output of CpG_MI includes the CGIs identified together with corresponding genomic coordinate, length, number of CpGs, G + C content and CpG O/E of the CGIs.

Evaluation of CpG_MI compared with other current approaches

It is important for CpG_MI to choose a threshold of the hot spot regions containing a minimal number of CpG dinucleotides. CpG_MI_i (i = 5, 6, 7) represents that hot spot regions are chosen as HSR_i (i = 5, 6, 7) in the algorithm CpG_MI, where HSR_i denotes the hot spot region containing at least i CpG dinucleotides. In order to assess the accuracy of different CpG_MI_ i (i = 5, 6, 7), sensitivity (SN), specificity (SP), accuracy (ACC) and correlation coefficient (CC) were calculated (see 'Materials and Methods' section). The results show that CpG_MI_6 is the most appropriate determination since it has the highest ACC and CC, and with moderate values of SN and SP (Table 1). Therefore, the HSR_6 is set as the optimal criterion for hot spot region in this algorithm. Moreover, CpG_MI was compared to other four algorithms for identifying CGIs, including two conventional sliding window programs (CpGProD and CpGIS) and the other two programs based on CpG distances (CpGcluster and CpGIF). CpG_MI performed better than the other programs with respect to sensitivity, accuracy and correlation coefficient (Table 1).

The coverage rates of the CAP experimental CGIs and the CGIs identified by five different algorithms in human chromosomes 21 and 22 were computed (Supplementary Figure S2). The data show that CpG_MI exhibits the highest coverage rate with experimental CGIs than the

Downloaded from https://academic.oup.com/nar/article/38/1/e6/3112467 by guest on 24 April 2024

Table 1. Estimation of five different predicated algorithms of CGIs

Program	SN ± SD	SP ± SD	ACC ± SD	CC ± SD
CpGProD	0.764 ± 0.027	0.811 ± 0.022	0.787 ± 0.015	0.576 ± 0.029
CpGIS	0.849 ± 0.020	0.899 ± 0.011	0.874 ± 0.013	0.749 ± 0.025
CpGcluster	0.048 ± 0.052	0.984 ± 0.001	0.516 ± 0.043	0.091 ± 0.042
CpGIF	0.950 ± 0.011	0.589 ± 0.045	0.769 ± 0.024	0.578 ± 0.041
CpG_MI_5	0.961 ± 0.008	0.782 ± 0.059	0.872 ± 0.022	0.752 ± 0.041
CpG_MI_6	0.950 ± 0.010	0.873 ± 0.027	0.914 ± 0.015	0.816 ± 0.029
CpG_MI_7	0.891 ± 0.013	0.910 ± 0.016	0.901 ± 0.011	0.800 ± 0.025

The criteria used in CpGProD and CpGIS were: length ≥ 500 bp, G + C content ≥ 55%, and CpG O/E ≥ 0.65. In CpGcluster, the 75th distance was used as the threshold of distance and the *P*-value cutoff is le-5. The criteria of CpGIF were the length ≥200 bp and other parameters with the default values. CpG_MI_*i* denoted the HSR_*i* (*i* = 5, 6, 7) was set as our scanning hot spot region in the first step of our algorithm CpG_MI. To assess above algorithms, 1000 sequence segments were randomly sampled respectively from positive CGIs and negative CGIs of the test set (see 'Materials and Methods' section). The sampling process was repeated 1000 times.

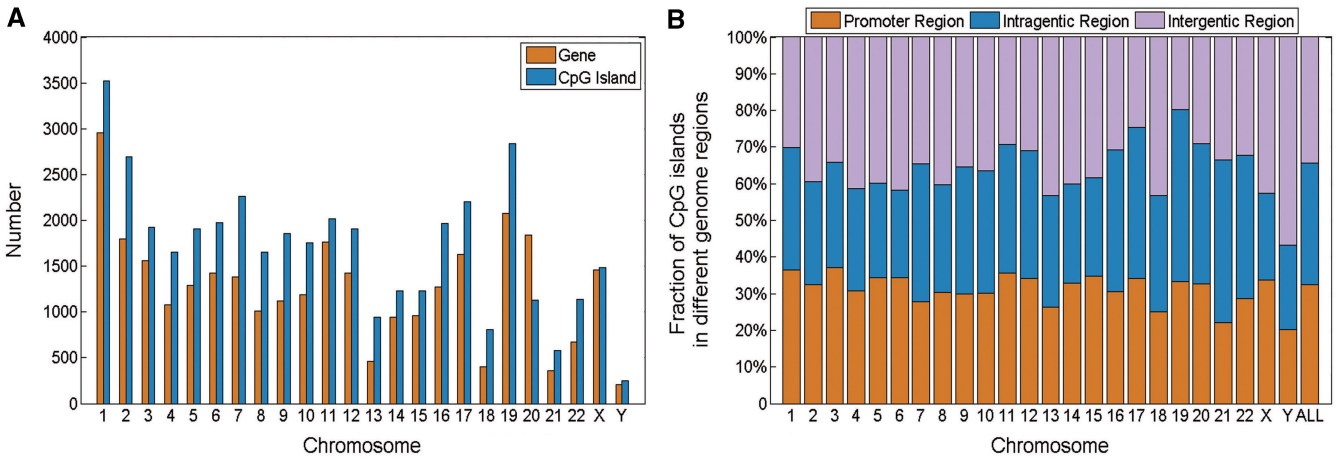

Figure 3. The distribution of CpG islands in human chromosomes. **(A)** The number distribution of CpG islands and genes in human chromosomes. **(B)** The distribution of CpG islands in different human genome regions.

other algorithms. It demonstrates that CpG_MI is a better tool for identifying the low methylated regions in human genome.

According to the repeat-dependence criteria proposed by Hutter *et al*. (15), the number of CGIs identified by CpG_MI dependent on Alu repetitive elements was computed. The result shows that the ratio (8.87%) of CGIs dependent on Alu repetitive elements identified by CpG_MI is lower than that of the conventional sliding window algorithms under the stringent criterion (≥10%) (15). The CGIs identified by CpG_MI have the lower false positive rate.

The distribution of CpG islands predicted by CpG_MI in human genome

In order to assess the regulatory function of the CGIs identified by CpG_MI, the location distributions between the CGIs and genes in each chromosome of human genome was investigated. We found the amount of CGIs was strongly positive correlated with that of genes in human chromosomes ($r = 0.89$, $P = 5.5 \times 10^{-9}$; Figure 3A). A total of 40 926 CGIs were identified by CpG_MI from human genome. About two-thirds (65.5%) of the CGIs locate within gene−environment regions, i.e. promoter regions or intragenic regions (Figure 3B). By removing parameter restriction of the length, G + C content and CpG O/E, CpG_MI identified many novel CGIs in the gene promoter regions. Out of total 29 885 annotated genes in human genome, 20 442 (68.4%) have promoter CGIs. We used CpG_MI and other four algorithms (CpGProD, CpGIS, CpGcluster and CpGIF) to identify CGIs from human genome. The results show that 63 new CGIs overlapping with gene promoter regions can be identified only by CpG_MI, where the lengths of 31 new CGIs are shorter than 200 bp, the G + C contents of 16 new CGIs are <55% and the CpG O/E values of 30 new CGIs are <0.6 (Supplementary Table S3). These promoter CGIs are missed by other algorithms due to their shorter lengths, low CpG contents or CpG O/E values. The function of new promoter CGIs identified by CpG_MI was investigated by Gene Ontology (GO) terms in Supplementary Table S3 (30,31). Gene *SYMT2*, for instance, is a coding gene of MYST histone acetyltransferase 2 and plays an important role in regulating cellular processes (32). Gene *NKX2-3* is a cancer-linked gene and aberrant hypermethylation is frequent in melanoma cell lines (33).

Downloaded from https://academic.oup.com/nar/article/38/1/e6/3112467 by guest on 24 April 2024

The relationship between histone modification and CpG islands

Histone modifications are important epigenetic regulatory elements associated with open chromatin structures and gene activity (25,34). We investigated the genome-scale profiling of 38 histone modifications derived from human CD4+ T cells in the set of CGIs identified by CpG_MI and random sets respectively. We randomly extracted 1 000 000 segments with the same length distribution as the CGIs identified by CpG_MI from human bulk genome. All random segments overlapping with CGIs were filtered out. Then, 40 926 random segments were randomly selected from the remaining random segments as the random set and the sampling process was repeated 10 000 times. The histone modification tags were mapped to the CGIs and random segments respectively. Histone modification tags are overlapped with 97.3% of the CGIs and 88.9% of the random segments.

The distribution of different histone modifications in the CGI set and the random sets was further investigated. The histone modification intensity of a CGI (or random segment) is defined as the tag number of the histone modification in the CGI (or random segment) divided by the length of CGI. The average modification intensities of the CGI set and the random sets were shown in Figure 4. The average modification intensities of H3K4me1, H3K4me2, H3K4me3, H3K9me1, H4K20me1 and H2BK5me1 are significantly high in the CGI set. The average modification intensities of these histone methylation modifications in the CGI set are more than three times greater than that of the random sets. Notably, the average modification intensity of H3K4me3 is remarkably stronger than that of the other modifications in the CGI set (Figure 4). H3K4me3 locates within 81.6% of the CGIs (Supplementary Table S4). Therefore, H3K4me3 can be regarded as an important indicator of functional CGIs. Fan *et al.* have improved the prediction accuracy of methylation status of CGIs using four histone methylation intensities (H3K4me1, H3K4me2, H3K4me3 and H3K9me1) (35). Sims *et al.* have also validated that H4K20me1 was enriched in promoter or coding regions of active genes and co-localized with H3K9me1 (36). H2BK5me1 is positively correlated with active promoters and could be regarded as a gene activation mark (25). Moreover, our data also show that 12 histone acetylation modifications are significantly enriched in the CpG set relative to the random sets: H2BK5ac, H3K27ac, H4K91ac, H2BK120ac, H3K18ac, H3K9ac, H2BK20ac, H3K4ac, H3K36ac, H4K5ac, H2BK12ac and H4K8ac. The average modification intensities of these histone acetylation modifications in the CGI set are more than three times greater than that of the random sets. It has been found that the patterns of histone acetylation modifications are associated with gene activity (29,37).

We observed significant depletion of 2 histone methylation modifications (H3K9me3 and H3K9me2) in the CGI set. However, the average modification intensities in the CGI set were only half of that in the random sets (Figure 4). The depletion of H3K9me3 and H3K9me2 in CGIs are expected, as they are considered to be marks of inactive and constitutive heterochromatin (38,39). It is indicated that the CGIs identified by CpG_MI are negatively correlated with the repressive modifications.

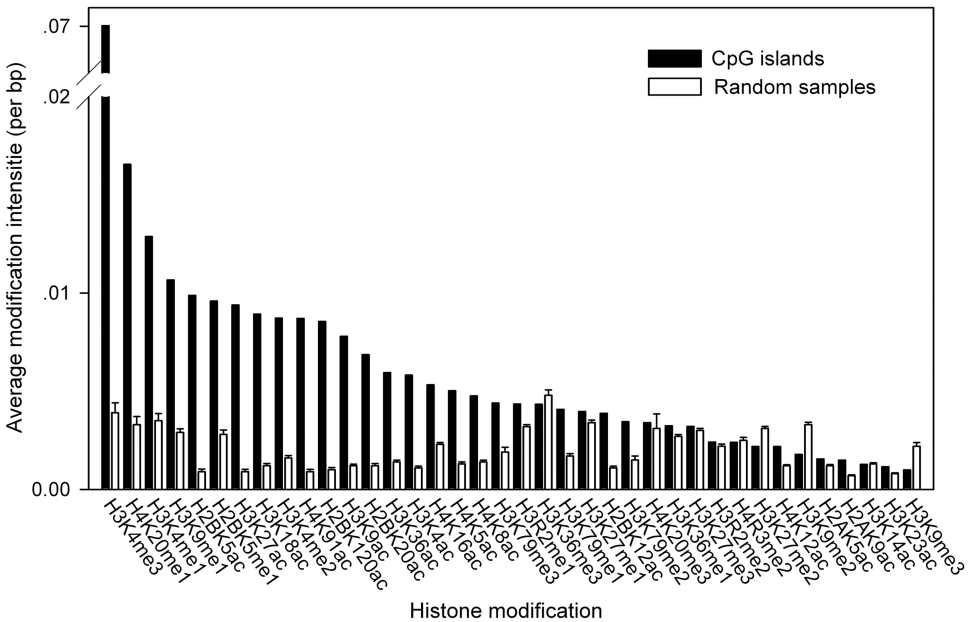

Figure 4. The average modification intensities of 38 histone modifications. The average modification intensities of 38 histone modifications in human CD4+T cells were significantly distinct between the CGI set (black bars) and random sets (white bars). Random segments (1 000 000) with the same length distribution as CGIs identified by CpG_MI were extracted from human genome. All random segments overlapping with CGIs were filtered out. Then, 40 926 random segments were randomly selected from remaining random segments as the random set and the sampling process was repeated 10 000 times. Error bars show plus and minus one standard error of the mean of 10 000 random sets.

Downloaded from https://academic.oup.com/nar/article/38/1/e6/3112467 by guest on 24 April 2024

Table 2. Basic statistics of CGIs and annotated genes containing promoter CGIs in six mammals and four fish species

Species	CpG content	Length of genome (Gb)	Average length (bp)	Number of CGIs	Number of annotated genes containing promoter CGIs (%)[a]	Number of annotated genes
Human	0.0091	2.85	1593	40 926	20 442 (68.4%)	29 885
Chimpanzee	0.0096	2.75	1318	39 797	19 331 (66.5%)	29 077
Mouse	0.0083	2.48	1105	33 408	13 353 (59.9%)	22 297
Rat	0.0089	2.48	845	42 367	8465 (54.9%)	15 431
Cow	0.0099	2.29	1213	50 642	6773 (70.4%)	9620
Dog	0.0107	2.03	1512	64 938	485 (51.4%)	944
Zebrafish	0.0166	1.52	568	92 188	7775 (27.7%)	28 058
Medaka	0.0192	0.58	428	53 721	7541 (37%)	20 399
Stickleback	0.0328	0.39	349	82 643	13 559 (57.3%)	23 654
Tetraodon	0.0335	0.19	337	37 832	7755 (49.7%)	15 613

[a]Percentage of the genes containing promoter CGIs in the total genes of the different species.

Identification of CpG islands in mammalian genomes

Owing to differences among mammalian genomes and lack of global epigenetic data of non-human genomes, it is more difficult to identify and evaluate functional CGIs in these genomes. Since the algorithm of CpG_MI is mainly dependent on the CpG content of bulk genome and the CMI distribution of random segments from the bulk genome, we performed the following two steps to test the applicability of this algorithm to other mammalian genomes. First, the CpG contents of six different mammalian genomes (human, chimpanzee, mouse, rat, cow and dog) were compared. Little variation was observed (see Table 2). Secondly, 100 000 random segments with the same length distribution as the CGIs identified by CpG_MI were selected from each of the above mammalian genomes respectively. The CMI values of these random segments were calculated using formula 2. We constructed regression models for each mammalian genome by fitting an exponential distribution model to the probability densities from the CMI distribution of 100 000 random segments. Their corresponding best-fit exponential distributions are shown in Figure 5. For each of the above six mammalian genome, Kolmogorov–Smirnov test was performed to compare the probability densities of the CMI distribution of random segments with the corresponding values from the best-fit exponential distribution. The results show the CMI distributions of random segments from these mammalian genomes follow the exponential distribution at $P < 0.01$. As shown in Figure 5, the best-fit exponential distributions of random segments from these mammalian species are similar. It indicates that the distribution of physical distances between two neighboring CpGs might be conserved during the evolution of mammalian species. Based on these observations, the criteria of identifying human CGIs by CpG_MI can be used to identify the CGIs of non-human mammalian genomes.

The CGIs from six mammal species identified by CpG_MI and the statistical results of average length, number of CGIs, annotated genes (containing promoter CGIs) are shown in Table 2. All CGIs of mammalian species mentioned in Table 2 are available for download and visualization as UCSC Genome Browser (26) tracks at http://bioinfo.hrbmu.edu.cn/cpgmi/.The results show that nearly 70% of the annotated genes in human, chimpanzee and cow genomes have promoter CGIs, and more than 54% in mouse and rat. There are only 51.4% of annotated genes having promoter CGIs in dog genome, which might be partly due to fewer genes (only 944) annotated in dog genome. Using the stringent criteria of Takai and Jones, Han *et al*. have found that the number of CGIs (58 328) in the dog genome were nearly three times more than that (19 568) in the rat genome. In this study, CpG_MI identified 64 938 CGIs in dog genomes and 42 367 CGIs in rat genome, which was not so much of disparity comparing with the result of Han *et al*. (21). The CGI average length (845 bp) of rat genome is shorter than that (≥1100 bp) of the other five mammalian genomes. It is possibly due to that some shorter functional CGIs in rat genome might not be identified by the Takai and Jones' stringent criteria. The distributions of CGIs in the three different genome regions between chimpanzee and mouse genomes were further investigated. The results were listed in Supplementary Table S5. About two-thirds of the CGIs identified by CpG_MI in chimpanzee (65.8%) and mouse (61.5%) genomes locate within the gene environment. It is consistent with the distribution of CGIs of human genome.

DISCUSSION

An essential goal for identifying functional CGIs is to search the epigenetic regulatory regions in genomes. The various methods face a common problem in finding a suitable minimal length of functional CGIs. In the original criterion of Gardiner-Garden and Frommer, the minimal length of CGIs was set to 200 bp, which had high false positive rate. In order to reduce the false positive rate, it was raised to 500 bp in the modified criterion by Takai and Jones. However, many functional CGIs are missed owing to this increased threshold of CGIs' length. In recent investigation, Bock *et al*. predicted the combined epigenetic scores for all CGIs identified by the criteria of Gardiner-Garden and Formmer from the

Downloaded from https://academic.oup.com/nar/article/38/1/e6/3112467 by guest on 24 April 2024

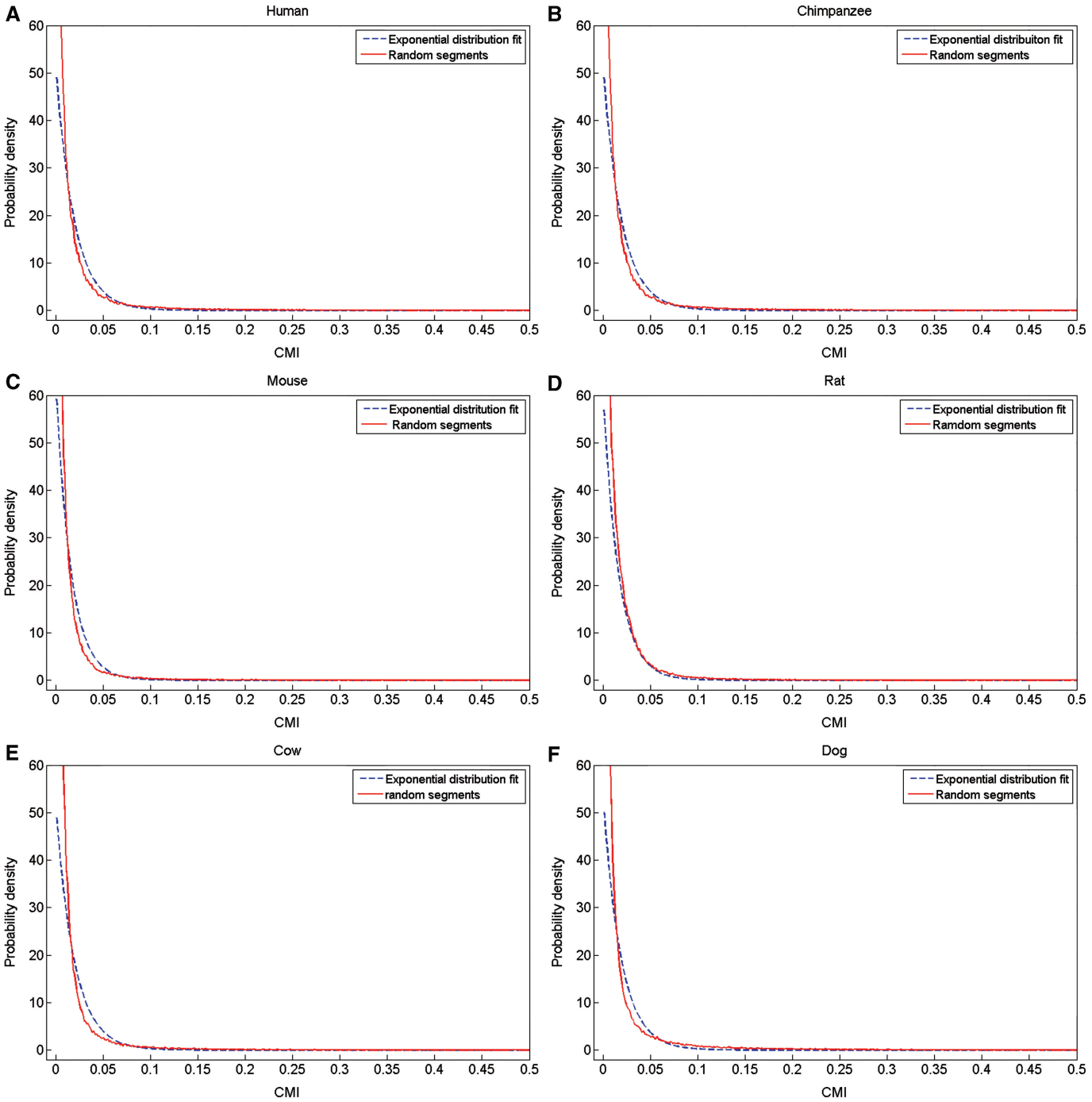

Figure 5. The distribution of the cumulative mutual information of random segments from six mammals. Random segments (100 000) are randomly selected from the human (hg18), chimpanzee (panTro2), mouse (mm9), rat (rn4), cow (bosTau4) and dog (canFam2) genomes respectively. The six figures show that the distribution curves (red) of probability densities of random segments' CMIs from six mammalian genomes and the curves (blue) of their corresponding best-fit exponential distributions.

human repeat-marked genome, and estimated the effect of three parameters (length, G + C content and CpG O/E) on the prediction accuracy of functional ('bona fide') CGIs identification respectively. Surprisingly, the results show that the length performs better than the CpG O/E and G + C content, and the G + C content is only slightly better than random. It suggests that the determination of threshold of length plays a more important role in identifying functional CGIs. Nevertheless, according to the original definition of CpG-rich regions, the CpG O/E should be the most natural sequence-based indicator of functional CGIs. The contradiction might be due to that the features of functional CGIs could not be revealed significantly by these three parameters.

Downloaded from https://academic.oup.com/nar/article/38/1/e6/3112467 by guest on 24 April 2024

The correlation between two neighboring CpGs was investigated by mutual information in this study to reveal the common feature of functional CGIs. We noted that the AMI distributions of the two sets (B3 and B4) of 'bona fide' CGIs, which resembled the set of experimental CGIs (CAP), were distinguished remarkably from other two sets (B1 and B2) and the random segments set (Figure 1). There are common mutual information features between the functional CGIs identified by computational methods and the experimental CGIs identified by CAP experiment. We found that the hot spot regions with high cumulative mutual information usually appeared in the central regions of 'bona fide' CGIs, which was an important feature of functional CGIs in our algorithm. Compared with the other algorithms, CpG_MI obtains the highest global prediction accuracy for experimental CGIs. The average length of CGIs identified by CpG_MI is 1593 bp in human genome. There are 32 shorter promoter CGIs (length $\leq$200 bp) which can be only identified by CpG_MI in human genome.

The identification of functional CGIs in other mammalian species is still difficult for various computational methods, which is handicapped by the scarcity of epigenetic data for these mammalian species. In this study, the CGIs identification in other genomes can be easily extended from human genomes through the CMI distribution of random segments. We found that there were similar CpG contents and CMI distributions of the random segments among mammalian genomes (see Table 2 and Figure 5). Therefore, the thresholds for identifying human CGIs can be directly applied to that of other mammalian species. We also calculated the CpG contents of four fishes (zebrafish, medaka, stickleback and tetraodon) whose genomes have been completely sequenced. We noted that the CpG contents between six mammals and four fish species were quite different. It is unexpected that the CpG contents of four fishes are two to four times as many as that of the mammalian species (Table 2). The possible reason may be that the mutation rates by converting CpG to TpG in these fish genomes are significantly lower than that in the mammalian genomes, which are mainly caused by deamination of 5-methylcytosine at CpG loci. The CGIs in the four fish genomes were identified by our CpG_MI respectively. The average lengths are significantly shorter than that of the mammalian species. The CGI densities of four fishes are higher than that of the six mammalian species as shown in Table 2. It indicates that the CGIs might be under significant evolution in vertebrates, which is consistent with the claim of Han *et al.* (21).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank three anonymous referees for their important suggestions, Dr Diansong Zhou and Yaoping Lei for reviewing the manuscript and Dr Illingworth R for supplying the experimental data.

FUNDING

National High Tech Development Project of China, the 863 Program (grant number 2007AA02Z329), National Natural Science Foundation of China (grant numbers 30871394 and 30370798), the National Basic Research Program of China, the 973 Program (grant number 2008CB517302), the Natural Science Foundation of Heilongjiang Province (grant number D2007-35) and the Heilongjiang Province Department of Education Outstanding Overseas Scientist (grant number 1152hq28). Funding for open access charge: National High Tech Development Project of China, the 863 Program (grant number 2007AA02Z329).

Conflict of interest statement. None declared.

REFERENCES

1. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
2. Bock,C., Walter,J., Paulsen,M. and Lengauer,T. (2007) CpG island mapping by epigenome prediction. *PLoS Comput. Biol.*, **3**, e110.
3. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
4. Illingworth,R., Kerr,A., Desousa,D., Jorgensen,H., Ellis,P., Stalker,J., Jackson,D., Clee,C., Plumb,R., Rogers,J. *et al.* (2008) A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.*, **6**, e22.
5. Ioshikhes,I.P. and Zhang,M.Q. (2000) Large-scale human promoter mapping using CpG islands. *Nat. Genet.*, **26**, 61–63.
6. Antequera,F. and Bird,A. (1993) Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA*, **90**, 11995–11999.
7. Cooper,D.N., Taggart,M.H. and Bird,A.P. (1983) Unmethylated domains in vertebrate DNA. *Nucleic Acids Res.*, **11**, 647–658.
8. Jones,P.A. and Takai,D. (2001) The role of DNA methylation in mammalian epigenetics. *Science*, **293**, 1068–1070.
9. Plath,K., Mlynarczyk-Evans,S., Nusinow,D.A. and Panning,B. (2002) Xist RNA and the mechanism of X chromosome inactivation. *Annu. Rev. Genet.*, **36**, 233–278.
10. Reik,W. and Walter,J. (2001) Genomic imprinting: parental influence on the genome. *Nat. Rev. Genet.*, **2**, 21–32.
11. Jair,K.W., Bachman,K.E., Suzuki,H., Ting,A.H., Rhee,I., Yen,R.W., Baylin,S.B. and Schuebel,K.E. (2006) De novo CpG island methylation in human cancer cells. *Cancer Res.*, **66**, 682–692.
12. Bird,A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.
13. Heisler,L.E., Torti,D., Boutros,P.C., Watson,J., Chan,C., Winegarden,N., Takahashi,M., Yau,P., Huang,T.H., Farnham,P.J. *et al.* (2005) CpG Island microarray probe sequences derived from a physical library are representative of CpG Islands annotated on the human genome. *Nucleic Acids Res.*, **33**, 2952–2961.
14. Bird,A.P. (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.*, **8**, 1499–1504.
15. Hutter,B., Paulsen,M. and Helms,V. (2009) Identifying CpG Islands by different computational techniques. *OMICS*, **13**, 153–164.
16. Takai,D. and Jones,P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA*, **99**, 3740–3745.
17. Ponger,L. and Mouchiroud,D. (2002) CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics*, **18**, 631–633.
18. Takai,D. and Jones,P.A. (2003) The CpG island searcher: a new WWW resource. *In Silico Biol.*, **3**, 235–240.
19. Hackenberg,M., Previti,C., Luque-Escamilla,P.L., Carpena,P., Martinez-Aroza,J. and Oliver,J.L. (2006) CpGcluster: a

Downloaded from https://academic.oup.com/nar/article/38/1/e6/3112467 by guest on 24 April 2024

distance-based algorithm for CpG-island detection. *BMC Bioinformatics*, **7**, 446.
20. Sujuan,Y., Asaithambi,A. and Liu,Y. (2008) CpGIF: an algorithm for the identification of CpG islands. *Bioinformation*, **2**, 335–338.
21. Han,L., Su,B., Li,W.H. and Zhao,Z. (2008) CpG island density and its corresslations with genomic features in mammalian genomes. *Genome Biol.*, **9**, R79.
22. Aktulga,H.M., Kontoyiannis,I., Lyznik,L.A., Szpankowski,L., Grama,A.Y. and Szpankowski,W. (2007) Identifying statistical dependence in genomic sequences via mutual information estimates. *EURASIP J. Bioinform. Syst. Biol.*, **2007**, 14741.
23. Bauer,M., Schuster,S.M. and Sayood,K. (2008) The average mutual information profile as a genomic signature. *BMC Bioinformatics*, **9**, 48.
24. Berger,S.L. (2007) The complex language of chromatin regulation during transcription. *Nature*, **447**, 407–412.
25. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
26. Kuhn,R.M., Karolchik,D., Zweig,A.S., Wang,T., Smith,K.E., Rosenbloom,K.R., Rhead,B., Raney,B.J., Pohl,A., Pheasant,M. *et al.* (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.*, **37**, D755–D761.
27. Giardine,B., Riemer,C., Hardison,R.C., Burhans,R., Elnitski,L., Shah,P., Zhang,Y., Blankenberg,D., Albert,I., Taylor,J. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.
28. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–496.
29. Wang,Z., Zang,C., Rosenfeld,J.A., Schones,D.E., Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Peng,W., Zhang,M.Q. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.
30. Carbon,S., Ireland,A., Mungall,C.J., Shu,S., Marshall,B. and Lewis,S. (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.
31. Maere,S., Heymans,K. and Kuiper,M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.
32. Thomas,T. and Voss,A.K. (2007) The diverse biological roles of MYST histone acetyltransferase family proteins. *Cell Cycle*, **6**, 696–704.
33. Tellez,C.S., Shen,L., Estecio,M.R., Jelinek,J., Gershenwald,J.E. and Issa,J.P. (2009) CpG island methylation profiling in human melanoma cell lines. *Melanoma Res.*, **19**, 146–155.
34. Goll,M.G. and Bestor,T.H. (2002) Histone modification and replacement in chromatin activation. *Genes Dev.*, **16**, 1739–1742.
35. Fan,S., Zhang,M.Q. and Zhang,X. (2008) Histone methylation marks play important roles in predicting the methylation status of CpG islands. *Biochem. Biophys. Res. Commun.*, **374**, 559–564.
36. Sims,J.K., Houston,S.I., Magazinnik,T. and Rice,J.C. (2006) A trans-tail histone code defined by monomethylated H4 Lys-20 and H3 Lys-9 demarcates distinct regions of silent chromatin. *J. Biol. Chem.*, **281**, 12760–12766.
37. Kurdistani,S.K., Tavazoie,S. and Grunstein,M. (2004) Mapping global histone acetylation patterns to gene expression. *Cell*, **117**, 721–733.
38. Rosenfeld,J.A., Wang,Z., Schones,D.E., Zhao,K., DeSalle,R. and Zhang,M.Q. (2009) Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics*, **10**, 143.
39. Tachibana,M., Nozaki,M., Takeda,N. and Shinkai,Y. (2007) Functional dynamics of H3K9 methylation during meiotic prophase progression. *EMBO J.*, **26**, 3346–3359.

Downloaded from https://academic.oup.com/nar/article/38/1/e6/3112467 by guest on 24 April 2024