



An evaluation of new criteria for CpG islands in the human genome as gene markers

Yong Wang and Frederick C.C. Leung*

Department of Zoology, University of Hong Kong, Pokfulam, Hong Kong, HKSAR, China

Received on April 3, 2003; revised on November 4, 2003; accepted on November 7, 2003
Advance Access publication February 5, 2004

ABSTRACT

Motivation: Recently, more stringent criteria for CpG islands have been introduced to exclude *Alu* repeats, thereby enabling a higher proportion of CpG islands associating with genes to be identified. Using these new criteria, several types of associations between CpG islands and genes were investigated to further establish the importance of CpG islands as gene markers.

Results: The CpG islands were searched by CpGIE, a java software program developed for CpG island identification. CpGIE was advanced in identification accuracy compared with other tools. According to our results, about 70% of the identified CpG islands were associating with the human genes and over half of them are in the promoters. Furthermore, the investigation of genes in the confirmed gene model showed that 56% of them had a CpG island overlapping the transcription start sites. In comparison, the new criteria were found capable of filtering a large fraction of *Alu* repeats that was identified as CpG islands by the generally accepted criteria within the genes, but very few CpG islands associating with the promoters were affected. The genes in the predicted gene model were not obviously associated with CpG islands, suggesting that CpG islands can be used to evaluate the accuracy of gene annotation.

Availability: <http://bioinfo.hku.hk/cpgieintro>

Contact: fcleung@hkucc.hku.hk

INTRODUCTION

CpG dinucleotide is generally very deficient in mammalian genomes, but many CpG dinucleotide clusters or 'CpG islands' can be found dispersed in the genomes, particularly close to or within the genes. These CpG islands are critical in gene expression regulation and cell differentiation (Bird, 2002). Over time, the criteria for CpG island identification have evolved. The original criteria for a CpG island were a DNA sequence longer than 200 bp with a G + C content $\geq 50\%$, and a CpG obs/exp (*o/e*) ratio ≥ 0.6 (Gardiner-Garden and Frommer, 1987). Currently, the

generally accepted criteria have since become more stringent, requiring a minimum DNA sequence length of 500 bp. The importance of these criteria lies in that they are able to exclude most *Alu* repeats, which were identified as CpG islands by the old criteria. Thereafter, Takai and Jones (2002) proposed more stringent criteria, with G + C content and CpG *o/e* ratio increased to 55% and 0.65 respectively, which would be more effective in excluding *Alu* repeats (Takai and Jones, 2002).

The drawback of criteria improvement was that some CpG islands associating with genes would also be excluded under these stringent criteria. However, reports showed that the proportion of genes with CpG islands was almost the same under the old and the generally accepted criteria (Larsen *et al.*, 1992; Antequera and Bird, 1993; Ponger and Mouchiroud, 2002). Similarly, under the new criteria, only a small fraction of the CpG islands associating with 5' end of genes was excluded (Takai and Jones, 2002), because CpG islands have been thought to be markers of genes in mammalian genomes (Larsen *et al.*, 1992; Antequera and Bird, 1999; Ioshikhes and Zhang, 2000). Since, basically, only the CpG islands outside of the genes were filtered by the stringent criteria, the new defined CpG islands might serve as better gene markers in the human genome. The present study is an attempt to test the hypothesis by comparing the new criteria with the generally accepted criteria.

Generally, CpG islands associating with genes are located near the promoters (Larsen *et al.*, 1992; Ioshikhes and Zhang, 2000). These have been used to predict promoters and first exons in the human genome (Scherf *et al.*, 2000; Davuluri *et al.*, 2001; Hannenhalli and Levy, 2001; Ponger and Mouchiroud, 2002). One of the reported programs can accurately predict 86% of the first exons with 17% false positives (Davuluri *et al.*, 2001). Nevertheless, it has been noticed that many CpG islands appear within or at the 3' end of the genes (Antequera and Bird, 1993). These association types are worth studying as well. We therefore defined six association types between CpG islands and genes. The plotting of CpG islands and genes on the human contigs was applied to show how significant the association between promoters and CpG islands was.

*To whom correspondence should be addressed.

In the NCBI, the genes are labeled with six types of evidence codes, C, E, PE, I, P and ?. Of these evidence codes, C represents a confirmed gene model and PE and P label the genes predicted by the GenomeScan program. Since the evidence codes were not utilized in previous studies, we expected the difference among the results from the genes labeled with different evidence codes. The result in this study showed a high association rate between CpG islands and the genes in an evidence code of C. The genes in other evidence codes were also subjected to survey and the prediction accuracy of different annotation methods was evaluated.

METHODS

Datasets

All the contigs in human chromosomes 21 and 22, and some in chromosome 1 were extracted from the NCBI (<http://www.ncbi.nlm.nih.gov>). The available contigs in the chromosomes 21 and 22 were: NT_029490.3, NT_011512.6, NT_030187.1, NT_030188.2, NT_011515.8, NT_011516.5, NT_028395.1, NT_011519.9, NT_011520.8, NT_011521.1, NT_011522.3, NT_011523.8, NT_030872.1, NT_011525.4, NT_019197.3 and NT_011526.4. The contigs in the chromosome 1 were: NT_004321.15, NT_028054.15, NT_021937.15, NT_004873.14, NT_030584.9, NT_004610.15, NT_004391.15, NT_037485.3, NT_004852.15 and NT_004483.15. All the information about the genes in these contigs, including start and stop sites, transcription orientation and evidence code, was also obtained from the NCBI.

Algorithm

The program used for CpG island searching was CpGIE. The algorithm in this program basically followed that which was developed by Takai and Jones (2002). Four major steps are as follows:

A. *Edit input.* Process FASTA or raw DNA sequence(s).

B. *Collect primary CpG islands.*

- B1. Obtain the user-defined criteria.
- B2. Move a window in a size of the minimum length along a DNA sequence by steps of 1 nt. A moving window must accommodate at least $A * B/16$ CpG dinucleotides (A representing CpG *o/e* ratio; B representing minimum length) before judging whether the criteria have been met. This can exclude mathematical CpG islands resulting from a strong compositional bias of G over C, or the reverse (Takai and Jones, 2002).
- B3. When the region in a window is identified to be a CpG island, the start site of this window is recorded. The window then moves ahead by steps of 10 nt until the region in a window is not a CpG island. The stop site of the region in the last window is recorded.

- B4. If the sequence between the recorded start and stop site meets the criteria, it is collected to be a primary CpG island; If not, both ends of the sequence will be trimmed 1 nt at the same time until the criteria are met.
- B5. From the stop site above, the window continues moving in steps of 1 nt. All the primary CpG islands are collected in their location order by repeating steps B2, B3 and B4.

C. *Combine the primary CpG islands*

- C1. Because of the small moving steps in step B3, the collected primary CpG islands mostly overlap one another. Compare the end site of a primary CpG island and the start site of the neighboring backward CpG island, and collect the overlapping and close-spaced (<100 nt in distance) primary CpG islands in a group. These primary CpG islands in the same group will be used to generate a final CpG island.
- C2. The region that they occupy in the DNA sequence (from the start site of the first primary CpG island to the stop site of the last primary CpG island in a group) is judged with the criteria. If the criteria are not met, the trimming steps in B4 are performed until a final CpG island is found in this region.
- C3. In some extreme cases, no final CpG island is found after the trimming of the sequence in the region, and thus one of the primary CpG islands in the group, generally in the middle of the region, is chosen to represent the final CpG island in this group.

Comparison with other tools

Three public tools for CpG island searching were used to compare with CpGIE in prediction accuracy. The CpG islands searcher (v1.30) written in Perl language was downloaded from <http://ccnt.hsc.usc.edu/cpgislands/> (Takai and Jones, 2003); CpGProD was downloaded from <http://pbil.univ-lyon1.fr/software/cpgprod.html> (Ponger and Mouchiroud, 2002); CpGPlot could be used directly in the following website: <http://www.ebi.ac.uk/emboss/cpgplot/> (Rice *et al.*, 2000). A contig (accession number: NT_000874.1) on human chromosome 19 was subjected to CpG island searching. The generally accepted criteria (length ≥ 500 bp, G + C content $\geq 50\%$ and CpG *o/e* ratio ≥ 0.60) were used in this comparison.

Association between CpG islands and genes

In chromosomes 21 and 22, the CpG islands were searched by CpGIE using the new criteria: length ≥ 500 bp, G + C content $\geq 55\%$ and CpG *o/e* ratio ≥ 0.65 . The CpG islands were then plotted on the contigs according to their start site and stop site locations, as were the genes in the contigs. The association types were defined into seven types: NA (non-association) and types A0–A5 (Fig. 1). The genes in the contigs were checked individually to determine whether they were associated with

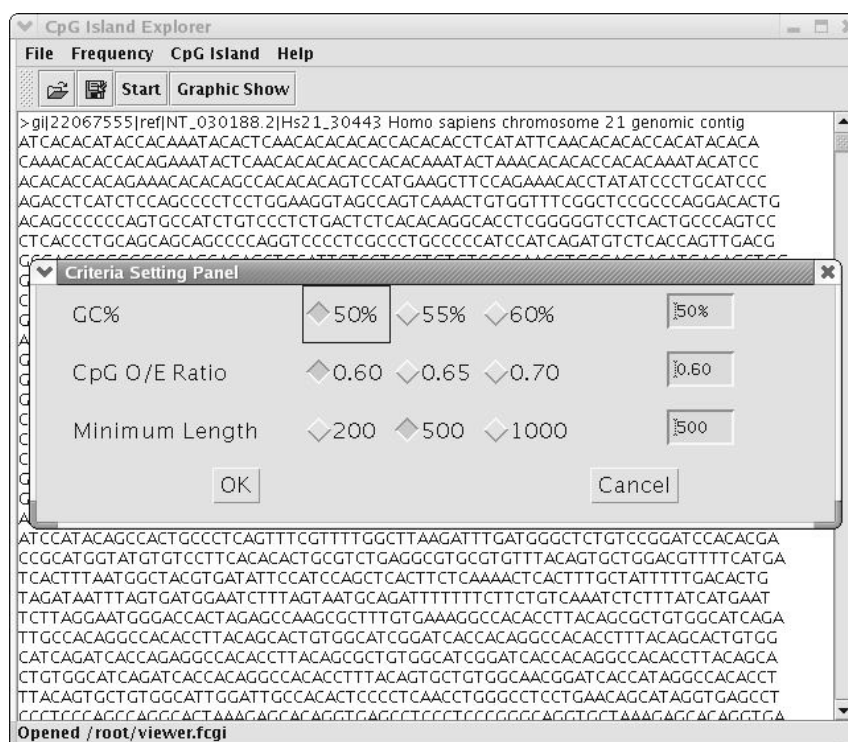


Fig. 1. User interface of CpGIE showing criteria setting panel. In the criteria setting panel, more sets of criteria can be specified by typing parameters inside the frames directly.

the CpG islands, and if so with which type. Caution was taken with the transcription orientation of the genes in this process.

The genes in the confirmed gene model on chromosome 1 were used in criteria comparison. The *Alu* repeats were identified on the contigs by using RepeatMasker, (<ftp.genome.washington.edu/cgi-bin/RepeatMasker>). On the contigs, we plotted the genes, *Alu* repeats, CpG islands under the generally accepted criteria and the new criteria. According to the association types between CpG islands and genes, we classified the associations into promoter association, within association and end association. The CpG islands containing *Alu* repeat(s) were recorded.

A complementary program was developed to display the associations graphically. CpGIE is available in the following link: <http://bioinfo.hku.hk/cpgieintro>

RESULTS AND DISCUSSION

CpGIE: a java program for CpG island exploration

CpGIE is an executable java program that can be run on the Windows, MacOS, Linux and Unix platforms. The program has a user-friendly interface, enabling users to specify their own criteria in the criteria-setting panel (Fig. 2). The program also enables identified CpG islands to be graphically demonstrated. It takes about 10 min to process NT_004873.14

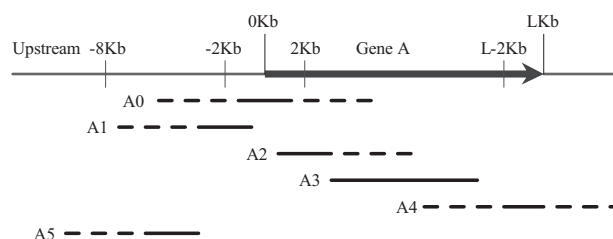


Fig. 2. Description of association types between CpG islands and gene A. L represents the length of gene A. The dashed line indicates that the line has been allowed to extend in that direction. In type A0, a CpG island must overlap the promoter of gene A. In type A1, a CpG island terminates in the upstream region between -2 and 0 kb. In type A2, a CpG island starts in the downstream region between 0 and 2 kb. In type A3, a CpG island is located within the downstream region of 2 kb to the ends. In type A4, a CpG island must overlap the $3'$ end, defined to be the last 2 kb. Type A5 terminates in the region between -8 and -2 kb.

(2.7 Mb in size) with a computer in Intel PIII (700 MHz) CPU. We suggest 10 Mb to be the maximal input file size at present. Multiple sequences in FASTA format can be processed in one operation. In a server, the size can be dramatically enlarged, depending on the increased heap size of Java virtual machine (JVM). The program also shows mononucleotide frequency and dinucleotide frequency.

CpGIE is more advanced in CpG island searching

The locations and sizes of CpG islands predicted by using different algorithms are far from identical. At present, the commonly used tools are CpGPlot and CpGReport in Emboss package (Rice *et al.*, 2000). But a more powerful program is available now and has been applied to the NCBI (Takai and Jones, 2003 and see the NCBI website). This implies that the algorithm used by Takai and Jones's program has been accepted for CpG island identification in mammalian genomes. This algorithm was further improved and used in CpGIE. In this study, the comparison between CpGIE and other programs shows that CpGIE is advanced in identification accuracy.

First, the result in Table 1 suggests that the performance of CpGPlot and CpGProD in identifying CpG islands is poorer than that of CpGi130 and CpGIE. CpGPlot found only a total of two CpG islands, and CpGProD found six CpG islands. Note that, in the result of CpGProD, 4 out of 10 CpG islands are false identifications, because the parameters of these identified 'CpG islands' do not meet the specified criteria.

Second, Table 1 also shows that CpGIE is more advanced than the other tools in prediction accuracy. CpGi130 (CpG island searcher v1.30) identified 13 CpG islands, quantitatively one CpG island more than CpGIE did. However, all the CpG islands identified by CpGi130 are covered by those identified by CpGIE. This is due to the higher capability of CpGIE in combining close-spaced CpG islands. For example, The 5th CpG island (from 15 123 to 17 190) identified by CpGIE covers the 5th (from 15 041 to 16 251) and the 6th CpG island (from 16 300 to 16 965) identified by CpGi130. The 12th CpG island (from 32 457 to 35 000) identified by CpGIE covers the 12th CpG island (from 32 503 to 33 472) and the 13th CpG island (from 33 587 to 35 040) identified by CpGi130. In the first case, CpGi130 failed to combine the two neighboring CpG islands separated in a space of 49 nt. Moreover, the 11th CpG island (from 30 742 to 31 242) in the prediction result of CpGIE is not present in that of CpGi130.

The total length of the CpG islands is 13 637 nt in the result of CpGIE, which is in contrast to 12 725 nt in CpGi130. From the stand point of accuracy and length of the identified CpG islands, CpGIE is much better than CpGPlot and CpGProD, and slightly more advanced than CpGi130.

The difference between CpGIE and CpGi130 in prediction accuracy comes from the improvements in the algorithm of CpGIE. The major improvement in CpGIE is that the moving step span is shortened to 10 nt after the criteria are first met in step B2 (see the algorithm in Methods section), while CpGi130 keeps this step span in 500 nt. This long step span in CpGi130 will result in too many trimming steps (step B4) and finally separate one potentially long CpG island into two or more short CpG islands. Sometimes it will fail to identify small CpG islands.

Table 1. Comparison of different tools for CpG island identification

| Program | Start | End | GC% | <i>o/e</i> | Length (nt) |
|---------|---------|--------|------|------------|-------------|
| CpGIE | 2109 | 4220 | 71.9 | 0.74 | 2112 |
| | 5588 | 6095 | 56.3 | 0.6 | 508 |
| | 11 304 | 12 372 | 55.1 | 0.61 | 1069 |
| | 13 263 | 14 253 | 51.5 | 0.6 | 991 |
| | 15 123 | 17 190 | 59.8 | 0.6 | 2068 |
| | 17 802 | 18 303 | 55.2 | 0.6 | 502 |
| | 18 757 | 19 429 | 52.9 | 0.6 | 673 |
| | 19 605 | 20 128 | 50.4 | 0.6 | 524 |
| | 22 566 | 23 116 | 50.8 | 0.6 | 551 |
| | 24 294 | 25 887 | 71 | 0.65 | 1594 |
| | 30 742 | 31 242 | 50.3 | 0.6 | 501 |
| | 32 457 | 35 000 | 62.5 | 0.65 | 2544 |
| | CpGi130 | 2149 | 4247 | 72 | 0.74 |
| 5330 | | 5829 | 55.4 | 0.6 | 500 |
| 11 344 | | 12 412 | 55 | 0.6 | 1069 |
| 13 215 | | 14 191 | 51.8 | 0.6 | 977 |
| 15 041 | | 16 251 | 56.4 | 0.6 | 1211 |
| 16 300 | | 16 965 | 64.1 | 0.6 | 666 |
| 17 842 | | 18 343 | 55.1 | 0.6 | 502 |
| 18 782 | | 19 384 | 55.8 | 0.6 | 603 |
| 19 645 | | 20 167 | 50.2 | 0.61 | 523 |
| 22 605 | | 23 156 | 50.7 | 0.6 | 552 |
| 24 329 | | 25 927 | 70.8 | 0.65 | 1599 |
| 32 503 | | 33 472 | 57.4 | 0.6 | 970 |
| 33 587 | | 35 040 | 65.6 | 0.72 | 1454 |
| CpGPlot | 2326 | 3498 | 77.6 | 0.84 | 1173 |
| | 33 885 | 34 643 | 72.9 | 0.87 | 759 |
| CpGProD | 2109 | 4220 | 71.9 | 0.74 | 2112 |
| | 5290 | 6133 | 55.7 | 0.55 | 844 |
| | 11 304 | 12 372 | 55 | 0.6 | 1069 |
| | 13 171 | 14 344 | 50.1 | 0.58 | 1174 |
| | 15 001 | 17 305 | 58.3 | 0.58 | 2305 |
| | 17 802 | 18 303 | 55.2 | 0.6 | 502 |
| | 18 647 | 20 127 | 55.2 | 0.6 | 1481 |
| | 22 553 | 23 129 | 50.4 | 0.58 | 577 |
| | 24 289 | 25 887 | 70.9 | 0.65 | 1599 |
| | 32 457 | 35 000 | 62.5 | 0.65 | 2544 |

The subject sequence is gi|5867303|ref|NT_000874.1|Hs19_1479| *Homo sapiens* 19p13.3. The criteria used in searching are length ≥ 500 bp, G + C content $\geq 50\%$ and CpG *o/e* ratio ≥ 0.60 .

The reason is that, during the trimming steps (step B3), CpGi130 allows much more nucleotides to be cut off, especially when CpG dinucleotides distribute mostly in one or two ends of the last window (Takai and Jones, 2002). For example, if the trimming steps start after the first 500 nt moving step (i.e. a region of 1000 nt is under the trimming steps), the presence of a CpG dinucleotide disperse in the middle region of the second 500 nt window will cause a lot of cutoffs from 5' end of the first 500 nt window. As a result, probably no CpG island can be identified in this region finally. If more than one 500 nt moving steps have been performed, a large amount of cutoffs at both ends will shorten the length of CpG island. In contrast, CpGIE takes at the most 10 trimming steps in the same process. This is perhaps the reason why the total

Table 2. The frequency of association types found by genes and CpG islands plotting

| Evidence code | Gene number | NA (%) | A0 (%) | A1 (%) | A2 (%) | A3 (%) | A4 (%) | A5 (%) |
|---------------|-------------|----------|----------|--------|--------|----------|---------|--------|
| C | 198 | 53 (27) | 111 (56) | 7 (4) | 2 (1) | 50 (25) | 22 (11) | 9 (5) |
| E | 65 | 27 (42) | 25 (38) | 4 (6) | 1 (2) | 5 (8) | 3 (5) | 6 (9) |
| PE | 340 | 197 (58) | 32 (9) | 13 (4) | 13 (4) | 68 (20) | 31 (9) | 8 (2) |
| I | 66 | 33 (50) | 24 (36) | 2 (3) | 1 (2) | 2 (3) | 2 (3) | 5 (8) |
| P | 38 | 28 (74) | 2 (5) | 2 (5) | 2 (5) | 3 (8) | 0 (0) | 1 (3) |
| ? | 383 | 68 (18) | 258 (67) | 11 (3) | 6 (2) | 122 (32) | 35 (9) | 8 (2) |

Note: in some genes, more than one association type was discovered.

According to the annotation in the NCBI, the meanings of the evidence codes are as follows: C: confirmed gene model based on alignment of mRNA, or mRNAs plus ESTs, to the genomic sequence; E: the model based on EST evidence only; PE: the model predicted by GenomeScan and EST evidence; P: the model predicted by GenomeScan only; ?: conflict model; and I: the model based alignment of mRNAs, or mRNAs plus ESTs, to the genome, in which the aligning transcripts could not be unambiguously assigned to a preexisting LocusID. Association type of NA means no association; Association types of A0–A5 are described in Figure 2. The percentage of the genes in that association type is shown in brackets.

length of CpG islands in the result of CpGi130 is smaller than that of CpGIE. Due to the large amount of cutoffs in CpGi130, the space between two neighboring CpG islands is occasionally so large that some original close-spaced CpG islands are unable to be combined. The small moving step of 10 nt in CpGIE takes more complexities into the algorithm of CpGIE, for instance, the additional steps in primary CpG islands collection. Nevertheless, the accuracy of CpG island identification is notably improved.

Evaluation on the new criteria

A total of 1266 CpG islands were identified by CpGIE in human chromosomes 21 and 22. Since *Alu* repeats were not eliminated in these contigs, the true number of CpG islands will be smaller.

A total of 1090 genes in human chromosomes 21 and 22 were classified according to the evidence codes labeled by the NCBI. Their association types with the CpG islands surveyed in all the groups and the number of the seven association types are shown in Table 2. The graphical demonstration results showed that 896 CpG islands were associating with the genes in the association type A0–A5 (Table 2). In other words, 71% of the identified CpG islands were located near or within the genes. Frequently, more than one association type was found in the long genes. By comparison, when the original criteria (length 200 bp; G + C content 50%; CpG *o/e* ratio 0.60) were used, 14 062 CpG islands were identified in human chromosomes 21 and 22 (Takai and Jones, 2002), and thus a much smaller fraction of these CpG islands was found to be associating with the genes. The new criteria are therefore clearly superior to the previous criteria, principally because of the far higher probability that an identified CpG island is associating with a gene.

In order to find the dominant association type, two χ^2 tests were performed on each of the evidence code groups. Two important findings were made. First, the number of type A0 gene was significantly higher than the combined number of

type A1 and type A2 genes, and the number of type A3, A4, A5 genes individually (χ^2 test, $P < 0.001$) in the evidence group of C, E, I and ?. Type A0 is therefore the dominant type in these groups. Because the boundary of type A0 can extend in two directions, the sum of type A1 and type A2 was applied in the test. Second, the sum of type A0, A1 and A2 (5' end association) genes was significantly higher than that of type A3, A4 and A5 genes in the above groups (χ^2 test, $P < 0.001$). The results indicate that the CpG islands associating with the genes are not only more likely to occur in the 5' end, but also that most of them overlap the transcription start sites of the genes. However, a dominant association type could not be found in the evidence code groups of P and PE.

Since the evidence code of C stands for a confirmed gene model, the results from this evidence code group are theoretically most reliable. About 56% of these genes have promoters that are overlapped by CpG islands, and 73% have a CpG island. In two previous works, only 57% (Larsen *et al.*, 1992) and 55.9% (Antequera and Bird, 1993) of the genes surveyed were found associated with CpG islands. The possible explanation for the higher percentage of associated genes revealed in this study is that the earlier studies included some unascertained or incomplete gene models. A high frequency of CpG islands has already been suggested (see above) as a potential marker for genes in mammalian genomes. The results demonstrate that the use of the new criteria enables CpG islands to be distinguished with much greater accuracy, thereby enhancing their usefulness as predictive markers. It is therefore strongly recommended that greater use be made of the potential offered by the new criteria.

A comparison to the generally accepted criteria

In previous reports, the data sources were different, those genes not categorized in the evidence code and all mixed up. The results from these genes, to some degree, could not be referred and compared with our result. We therefore made a comparison between the generally accepted criteria and the

Table 3. Number of CpG islands under two sets of criteria in 10 contigs

| Contig | GC% | Length (Mb) | Gene number | CpG island number |
|--------------|------|-------------|-------------|-------------------|
| NT_004321.15 | 56.8 | 1.13 | 12 | 83 (160) |
| NT_028054.15 | 49.4 | 2.38 | 26 | 61 (150) |
| NT_021937.15 | 48.5 | 3.46 | 45 | 88 (196) |
| NT_004873.14 | 46.2 | 2.67 | 19 | 39 (109) |
| NT_030584.9 | 50.3 | 1.36 | 11 | 28 (54) |
| NT_004610.15 | 47.7 | 4.37 | 51 | 77 (167) |
| NT_004391.15 | 48.1 | 1.32 | 26 | 42 (89) |
| NT_037485.3 | 48 | 2.85 | 60 | 107 (191) |
| NT_004852.15 | 45.6 | 3.81 | 63 | 94 (168) |
| NT_004483.15 | 36.2 | 3.63 | 9 | 14 (40) |
| Total | | 27 | 322 | 633 (1224) |

The numbers outside the parentheses show the results from the new criteria (length ≥ 500 bp, G+C content $\geq 55\%$ and CpG *o/e* ratio ≥ 0.65) and those within the parentheses show the result from the generally accepted criteria (length ≥ 500 bp, G + C content $\geq 50\%$ and CpG *o/e* ratio ≥ 0.60). Gene number indicates the number of genes in the evidence code of C on these contigs.

Table 4. Number of CpG islands located in different gene regions

| Contig | Promoter | Within | End |
|--------------|-----------|-----------|---------|
| NT_004321.15 | 10 (10) | 53 (104) | 1 (3) |
| NT_028054.15 | 20 (21) | 29 (93) | 3 (5) |
| NT_021937.15 | 34 (34) | 26 (81) | 3 (4) |
| NT_004873.14 | 12 (13) | 9 (47) | 1 (2) |
| NT_030584.9 | 7 (7) | 5 (12) | 1 (3) |
| NT_004610.15 | 33 (34) | 14 (57) | 3 (3) |
| NT_004391.15 | 19 (21) | 6 (18) | 0 (0) |
| NT_037485.3 | 43 (43) | 17 (44) | 3 (3) |
| NT_004852.15 | 48 (50) | 24 (60) | 2 (3) |
| NT_004483.15 | 6 (6) | 0 (2) | 0 (0) |
| Total | 232 (239) | 181 (518) | 16 (26) |

The numbers outside and within the parentheses show the result from the generally accepted criteria and the new criteria (as described in Table 3) respectively. The genes in the evidence code of C were surveyed. The number of CpG islands in promoters is a combined number of the association types A0, A1 and A2. The association types between CpG island and gene are as described in Figure 2.

new criteria using the same data source from chromosome 1. *Alu* repeats were plotted on the contigs. Their association with genes in the confirmed gene model and CpG islands was investigated in order to first test the efficiency of the new criteria in filtering *Alu* repeats within the genes and second confirm that the conclusion that we made in chromosomes 21 and 22 was not influenced by the presence of *Alu* repeats.

As shown in Table 3, the number of CpG islands under the new criteria is only about half of that under the generally accepted criteria. This is in agreement with the observation of Takai and Jones's (2002) work. It is therefore interesting to know if the stringent criteria have already excluded a lot of CpG islands associating with the genes, especially those overlapping the promoters of the genes. The associations of CpG islands and the genes are shown in Table 4. The result

Table 5. Number of CpG islands associating with *Alu* repeats in three specified regions of the human genes

| Contig | Promoter | Within | End |
|--------------|----------|----------|-------|
| NT_004321.15 | 2 (2) | 3 (22) | 0 (1) |
| NT_028054.15 | 9 (10) | 8 (49) | 0 (0) |
| NT_021937.15 | 11 (18) | 15 (66) | 0 (0) |
| NT_004873.14 | 2 (3) | 2 (26) | 0 (0) |
| NT_030584.9 | 2 (2) | 3 (10) | 0 (0) |
| NT_004610.15 | 6 (8) | 9 (39) | 0 (0) |
| NT_004391.15 | 3 (11) | 4 (12) | 0 (0) |
| NT_037485.3 | 5 (13) | 10 (35) | 0 (0) |
| NT_004852.15 | 11 (18) | 13 (46) | 0 (0) |
| NT_004483.15 | 0 (0) | 0 (2) | 0 (0) |
| Total | 51 (85) | 67 (307) | 0 (1) |

For description refer to Table 4.

indicates that within the association is the main source of the difference in CpG island numbers. *t*-tests showed that the numbers of CpG islands under the generally accepted criteria were significantly larger than those under the new criteria (*t*-test, $P < 0.02$) in within associations but not significant in promoter associations (*t*-test = 0.7, $P > 0.1$). We found a total of only seven CpG islands that were excluded by the new criteria from the promoter associations. The CpG islands overlapping the promoters are therefore not affected by the new criteria. As a comparison, those at the ends of the genes are notably excluded (Table 4).

We speculated that *Alu* repeats were the source of the above difference. Most *Alu* repeats could be excluded by using the new criteria, because the average G + C frequency and CpG *o/e* ratio of *Alu* repeat sequences, 53% and 0.62 respectively (Ponger *et al.*, 2001), were lower than the values specified in the new criteria. In Table 5, the data of within association show that 240 CpG islands partially or wholly associated with *Alu* repeats are deleted by using the new criteria, accounting for 71% of the difference in the numbers of identified CpG islands. Therefore, in support of the conclusion from Takai and Jones (2002), the new criteria have shown the efficiency in excluding *Alu* repeats. This will at last improve the significance of CpG islands as gene markers due to an elevated association rate between genes and CpG islands.

A second contribution of the new criteria is the shift of the dominant region where CpG islands are most often identified. The dominant region is located in the promoters for the new criteria but within the genes for the generally accepted criteria (Table 4). A χ^2 test showed that using the new criteria would much significantly change the distribution of CpG islands in the genes (χ^2 test, $P < 0.0001$). The rule of dominant association type that we conclude in the chromosomes 21 and 22 is hereby reinforced. The significance of this change is as indicated in Tables 3 and 4. About 68% of the CpG islands under the new criteria are associated with the genes in the evidence

code of C and over half of them (about 37%) were located in the promoters. The percentage is higher than the percentage of 16% discovered in a previous comparison (Takai and Jones, 2002). This is accounted to incomplete gene annotation in chromosomes 21 and 22 before.

Although a high percentage (about 64%) of CpG islands under the generally accepted criteria was found associating with the genes, the association rate between gene and CpG island was low. The rate is about 26%, in contrast to 51% in the new criteria (Table 3). These superfluous CpG islands will bring noise into the annotation process if CpG islands are referred to the gene locations. This drawback is much attenuated by using the new criteria since the predicted CpG islands have a high association rate with the genes. In conclusion, the CpG islands under the new criteria can be better gene markers than the generally accepted criteria.

In addition, under the generally accepted criteria, the CpG islands containing an *Alu* repeat within the genes are often of a short size and thus covered by *Alu* repeats for the larger part, enabling them to be easily filtered by the new criteria. But those in the promoters are generally preserved as the criteria get stringent partially due to the larger size (in most of the cases, over 1 kb) and their short overlapping with the *Alu* repeats. This indicates that with *Alu* repeats masked, the association between CpG islands and promoters can be more remarkable. Therefore, they are also useful markers of the promoters of genes.

Accurate prediction of gene promoters by CpG islands

CpG islands generally overlap the promoter, and therefore play a critical role in gene expression regulation. Larsen *et al.* (1992) showed that all housekeeping genes were associated with CpG islands in their promoters. A further study has shown that 10% of housekeeping genes are without a CpG island (Ponger *et al.*, 2001). However, about 25% of tissue-specific genes are overlapped by CpG islands in the promoter (Larsen *et al.*, 1992). Therefore, at least as far as the housekeeping genes are concerned, a CpG island is a good footprint to gene promoters.

In this study, housekeeping genes and tissue-specific genes are not particularly specified, so the results only indicate the degree (in percentage terms) to which a group of mixed genes are associating with CpG islands in their promoters. According to the results from the evidence group of C, the promoters of 56.1% of the genes in a confirmed model are overlapped by a CpG island. This percentage is even higher in the contigs from the chromosome 1, about 72% as indicated by Tables 3 and 4. Because the annotation quality is different among chromosomes, the percentages are somewhat not comparable. In a well-annotated contig, there should be a larger fraction of CpG islands covered by genes. Based on these findings, an annotated gene can be accurately placed in the genomic sequence by reference to the position of the associated CpG island.

Why is there a higher percentage of association type A0 genes in the evidence code groups of C and ? than in the other groups? The reason is probably that the genes labeled with C and ? have been identified on the basis of mRNA evidence rather than expressed sequence tag (EST) data in the NCBI. Theoretically, mRNA data in gene identification can more precisely locate the genes in full-length sequences than in EST data. The mRNA data can therefore not only confirm easily that a gene belongs to the evidence group of C, but can also identify most if not all the exons. In the case of unconfirmed gene models, CpG island can be used to find the missed exons and the transcription start sites. A high proportion of type A0 association is therefore also a hallmark of accurate gene annotation.

In the NCBI, the evidence code of ? is used to label a gene model when there is some discrepancy between mRNA evidence and the gene model, either in the alignment of the two and/or in their protein products. In this study, the high percentage of type A0 association in the evidence code group of ? indicates that a large fraction of these genes might be very promising to be in confirmed gene model. These genes probably have now been labeled with the evidence code of C.

A special type in type A0 association should also be mentioned. In the contigs from the three chromosomes, we found that 58 CpG islands were shared by the transcription start sites of the two close-spaced genes encoded in the plus strand and minus strand, respectively. Therefore, a sum of 116 associations in type A0 belongs to this special type. This feature has been noted before (Colombo *et al.*, 1992), but was considered to be exceptional (Antequera and Bird, 1993). Because the whole contigs instead of individual genes were surveyed, more cases could possibly be observed in this study. It has been suggested that this may be a mechanism of co-regulation and/or *cis*-interaction (Huxley and Fried, 1990). However, in our further investigation no evidence was found in support of the hypothetical linkage in function.

Assessment of the gene models predicted by GenomeScan

In the NCBI, GenomeScan is applied in the process of gene annotation. As it combines hidden Markov model architecture, BLASTX, rpsBLAST and BLASTP, GenomeScan is a powerful tool for genome sequence annotation (Yeh *et al.*, 2001). During the annotation process, if the *E*-value is less than 0.0001 (indicated in the NCBI website), a gene will be labeled with P or PE. These gene models have not yet been ascertained, but represent a large percentage (35%) of all the genes investigated. So, it is therefore of great interest to assess the accuracy of the annotation on the genes, by means of the association between the gene and CpG islands. Results from the evidence group of C can be utilized as a control. In the evidence code group of P and PE, <10% of the genes have CpG islands in the association type A0, and the sum of type A0 genes is not significantly higher than that of type A1 and

type A2 genes. Instead, the percentage of type A3 genes is very high. These results suggest that the gene locations were inaccurately predicted. Moreover, some 40% of the tissue-specific genes were reported to have a CpG island (Edwards, 1990; Larsen *et al.*, 1992; Ponger *et al.*, 2001). If so, at least 50% of the genes should have a CpG island (Larsen *et al.*, 1992). However, only 26 and 42% of the genes in the evidence code groups of P and PE respectively had CpG islands. This indicates that a large number of false positive gene identifications were made in the groups of P and PE. Nevertheless, this hypothesis rests on the assumption that there was very little gene composition bias among the different evidence code groups. For example, if more housekeeping genes are identified in the evidence code group of C, the high proportion of type A0 genes is probably irrelevant to gene prediction accuracy. However, we cannot find any support for the biased composition of housekeeping genes and tissue-specific genes among the groups. Thus, it is strongly suggested that the GenomeScan program should take into consideration the presence and location of CpG islands in the algorithm when sequence similarity is low. The necessity of the involvement of CpG islands in computational prediction of promoters and first exons of genes has been strongly suggested in previous works but some related programs are still using the 'old' criteria (Scherf *et al.*, 2000; Davuluri *et al.*, 2001; Hannenhalli and Levy, 2001; Ponger and Mouchiroud, 2002). We therefore suggest the application of the new criteria in the algorithms to obtain more accurate predictions.

CONCLUSION

Our study shows that the new criteria are worth promoting in future studies, because CpG islands under these criteria are much better gene markers as well as footprints of promoters. The GenomeScan is here shown to have some drawbacks, which will make false predictions in genes model and gene location. Since all the genome data of the human genome are available at present, the genes especially those labeled with an evidence code of C in the other chromosomes should also be investigated. This will make the result more convincing. CpGIE was developed to search CpG islands. With some improvements from Takai and Jones's algorithm, the program has been shown to have advantages in accurately identifying and precisely locating the CpG islands in a target sequence.

ACKNOWLEDGEMENTS

We would like to thank Tania Cheung, who conducted partial studies that facilitated our work. We would also like to acknowledge helpful comments from Dr David Wilmshurst,

who reviewed this paper before submission. Finally, we would like to acknowledge the award of a grant from the RGC to fund this study.

REFERENCES

- Antequera, F. and Bird, A. (1993) Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci., USA*, **90**, 11995–11999.
- Antequera, F. and Bird, A. (1999) CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr. Biol.*, **9**, R661–R667.
- Bird, A.P. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
- Colombo, P., Yon, J., Garson, K. and Fried, M. (1992) Conservation of the organization of five tightly clustered genes over 600 million years of divergent evolution. *Proc. Natl Acad. Sci., USA*, **89**, 6358–6362.
- Davuluri, R.V., Grosse, I. and Zhang, M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nat. Genet.*, **29**, 412–417.
- Edwards, Y.H. (1990) CpG islands in genes showing tissue-specific expression. *Phil. Trans. R. Soc. London*, **326**, 207–215.
- Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
- Hannenhalli, S. and Levy, S. (2001) Promoter prediction in the human genome. *Bioinformatics*, **17**, S90–S96.
- Huxley, C. and Fried, M. (1990) The mouse surfeit locus contains a cluster of six genes associated with four CpG-rich islands in 32 kilobases of genomic DNA. *Mol. Cell. Biol.*, **10**, 605–614.
- Ioshikhes, I.P. and Zhang, M.Q. (2000) Large-scale human promoter mapping using CpG islands. *Nat. Genet.*, **26**, 61–63.
- Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095–1107.
- Ponger, L., Laurent, D. and Mouchiroud, D. (2001) Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res.*, **11**, 1854–1860.
- Ponger, L. and Mouchiroud, D. (2002) CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics*, **18**, 631–633.
- Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
- Scherf, M., Klingenhoff, A. and Werner, T. (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.*, **297**, 599–606.
- 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.
- Takai, D. and Jones, P.A. (2003) The CpG island searcher: a new WWW resource. *In Silico Biol.*, **3**, 0021.
- Yeh, R.F., Lim, L.P. and Burge, C.B. (2001) Computational inference of homologous gene structures in the human genome. *Genome Res.*, **11**, 803–816.