In Silico Biology 3 (2003) 235–240
IOS Press

The CpG Island Searcher: A New WWW Resource

Daiya Takai* and Peter A. Jones

Department of Biochemistry and Molecular Biology, USC/Norris Comprehensive Cancer Center, Keck School of Medicine of the University of Southern California, Los Angeles, CA, USA

Edited by H. Michael; received 12 November 2002; revised and accepted 27 January 2003; published 4 February 2003

ABSTRACT: Clusters of CpG dinucleotides in GC rich regions of the genome called "CpG islands" frequently occur in the 5' ends of genes. Methylation of CpG islands plays a role in transcriptional silencing in higher organisms in certain situations. We have established a CpG-island-extraction algorithm, which we previously developed [Takai and Jones, 2002], on a web site which has a simple user interface to identify CpG islands from submitted sequences of up to 50kb. The web site determines the locations of CpG islands using parameters (lower limit of %GC, ObsCpG/ExpCpG, length) set by the user, to display the value of parameters on each CpG island, and provides a graphical map of CpG dinucleotide distribution and borders of CpG islands. A command-line version of the CpG islands searcher has also been developed for larger sequences. The CpG Island Searcher was applied to the latest sequence and mapping information of human chromosomes 20, 21 and 22, and a total of 2345 CpG islands were extracted and 534 (23%) of them contained first coding exons and 650 (28%) contained other exons. The CpG Island Searcher is available on the World Wide Web at http://www.cpgislands.com or http://www.uscnorris.com/cpgislands/cpg.cgi.

KEYWORDS: CpG island, DNA methylation, gene prediction, Internet

INTRODUCTION

Dinucleotide clusters of CpGs in GC-rich regions of genomes or CpG islands are present in the promoters [Gardiner-Garden and Frommer, 1987] and exonic regions of approximately 40% of mammalian genes [Larsen *et al.*, 1992]. By contrast, other regions of the mammalian genome contain few CpG dinucleotides and these are mostly methylated. The decreased occurrence of CpGs is best explained by the fact that methylated cytosines are mutational hotspots leading to CpG depletion during evolution [Coulondre *et al.*, 1978]. A large number of experiments have shown that methylation of promoter CpG islands plays an important role in gene silencing [Bird, A., 2002], genomic imprinting [Feil and Khosla, 1999], X-chromosome inactivation [Panning and Jaenisch, 1998], the silencing of intragenomic parasites [Yoder *et al.*, 1997] and carcinogenesis [Jones and Baylin, 2002]. Bird first deduced that vertebrate genes are associated with unmethylated CpG islands [Bird, A. P., 1986]. The first large-scale computational analysis of CpG islands using vertebrate gene sequences in GenBank was performed by Gardiner-Garden

*Corresponding author: Daiya Takai, Department of Respiratory Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyoku, Tokyo, Japan. Tel.: +81 3 3815 5411 ex.33126; Fax: +81 3 3815 5954; E-mail: dtakai-ind@umin.ac.jp.

Electronic publication can be found in *In Silico Biol.* **3**, 0021 <http://www.bioinfo.de/isb/2003/03/0021/>, 4 February 2003.

and Frommer in 1987. They defined a CpG island as being a 200-bp region of DNA with a high GC content (greater than 50%) and observed CpG/expected CpG ratio (ObsCpG/ExpCpG) of greater or equal to 0.6. Larsen *et al.* suggested that all housekeeping and widely expressed genes have a CpG island in the 5' region. We previously described characteristics of CpG islands on human chromosomes 21 and 22 extracted by a newly developed CpG island extraction algorithm [Takai and Jones, 2002]. We used more stringent criteria for CpG islands (using the lower limit values 500bp for length, 55% for GC content and 0.65 for ObsCpG/ExpCpG) that excluded most *Alu* repetitive elements and led to a better association between CpG islands and genes. Here we describe the availability of the CpG island extraction program on a World Wide Web site. Using newly mapped gene information, a significant association of CpG islands and promoter / exon regions has been revealed. The increased association of CpG islands and genes using this algorithm is useful for gene prediction.

METHODS

The CpG Islands Searcher is written in the PERL language, currently running on the Microsoft Windows 2000 Server platform. The algorithm to search CpG islands is described in the previous report [Takai and Jones, 2002]. Availability of command line version will be announced at http://www.uscnorris.com/cpgislands/. All of these programs were coded by D. Takai with PERL COMPILER (ActiveState, Vancouver, http://www.activestate.com/).

For analyses of chromosomes 20, 21 and 22, we obtained sequence and mapping information from the GenBank Database. We used the contigs (build 28), NT_011387, NT_028391, NT_025215, NT_028392, NT_011362, NT_030871, NT_025929, NT_011333, NT_025218 (chromosome 20), NT_029490, NT_011512, NT_030187, NT_030188, NT_011515 (chromosome 21), NT_011516, NT_028395, NT_011519, NT_011520, NT_011521, NT_011522, NT_011523, NT_030872, NT_011525, NT_019197, NT_011526 (chromosome 22). Repetitive elements were detected by the Repeat Masker mail server (University of Washington Genome Center, Seattle, http://ftp.genome.washington.edu/cgi-bin/RepeatMasker).

CpG ISLAND EXTRACTION ALGORITHM

As described in our previous report, we initially applied a “sliding and jumping” 200 base pair window algorithm for the CpG Island Searcher [Takai and Jones, 2002]. The algorithm was originally designed according to the criteria of CpG islands described by Gardiner-Garden and Frommer to avoid missing any CpG islands which meet these criteria. The original algorithm merged two CpG islands when they were less than 100bp apart and the merged CpG island still met the criteria. This differs from the the approach used by another CpG island search web site, CpG plot (http://www.ebi.ac.uk/emboss/cpgplot/ #and-Newcpgseek). The latest version of the CpG Island Searcher allows the user to select the gap size between two putative CpG islands. Previous observations [Takai and Jones, 2002] showed that the use of more stringent criteria of increased lower limits for %GC, ObsCpG/ExpCpG and length led to a better association between CpG islands and genes.

The user-interface of the web site enables user-defined criteria (i.e. variable lower limit of %GC, ObsCpG/ExpCpG and length) for CpG island extraction. Previously a 200 bp sliding window was used for the initial localization of a CpG island. However, the revised algorithm employs the user-defined lower limits of parameters for the initial scanning of a submitted sequence to avoid missing any CpG islands using the criteria defined by the user. Using a larger window size potentially allows extraction of CpG islands which

could not be extracted with a smaller window size. Since our first priority for this algorithm was not to miss any sequences meeting the criteria, the initial result might differ from the perception of the user. Thus the readout of the web site includes a graphical map of CpG sites so that the user can narrow the CpG island region using a larger %GC, ObsCpG/ExpCpG and smaller length of lower limit after the initial search.

OVERVIEW OF THE CpG ISLAND SEARCHER WEB SITE

Sequences of up to 50kb can be submitted and the CpG island parameters selected for search can be selected. The web site is easy to use and the default values (%GC >55%, ObsCpG/ExpCpG >0.65, length >500bp) are the ones described in our previous report [Takai and Jones, 2002], but can be easily modified (Figure 1).

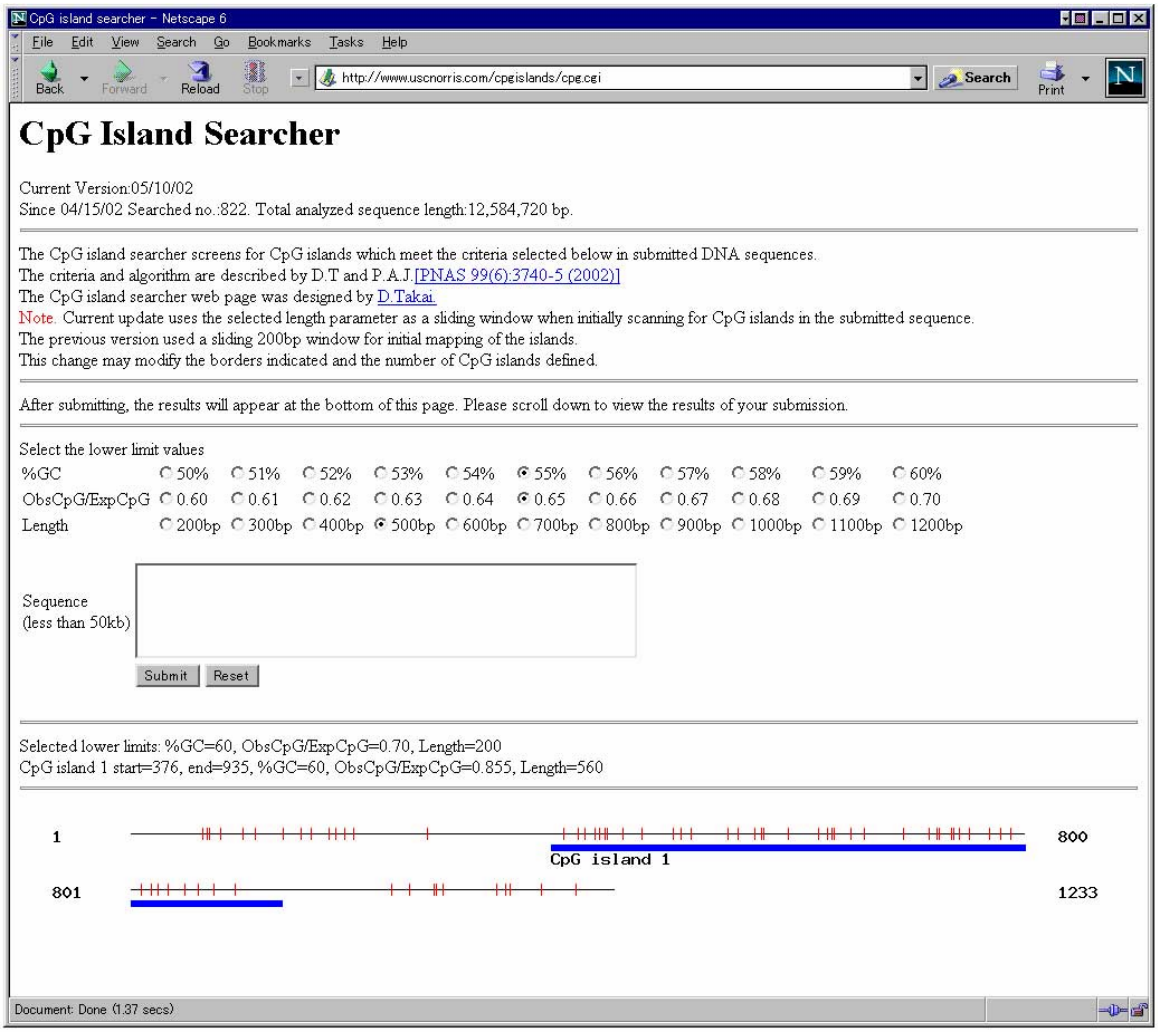

Fig. 1. An example of the analysis of a sequence containing a CpG island. The user can modify the lower limit values for search and submit sequences (up to 50kb). Selected lower limits, parameters defining CpG islands and a graphical map of CpG dinucleotide distribution are displayed.

Using the parameters (lower limit of %GC, ObsCpG/ExpCpG, length) set by the user, the web site extracts CpG islands, displays the calculated values of parameters on each CpG island, and provides a graphical map of CpG dinucleotide distribution and borders of CpG islands. Theoretically, this program can accept any sequences of any size like CpGPlot, however, we made the 50kb limitation to avoid possible data transfer errors. The limitation of 50kb is much larger than the 1kb of CpGProD (http://pbil.univ-lyon1.fr/software/cpgprod_query.html) and is probably sufficient for general use. The CpG Island Searcher is available on the World Wide Web at http://www.uscnorris.com/cpgislands/cpg.cgi. For larger sequences, a command-line version of the CpG Island Searcher with no limitation of sequence size is available on the web site (Figure 2). It also enables user-defined lower limits for CpG island extraction but does not contain a graph-generation routine.

CpG ISLANDS IN HUMAN CHROMOSOMES 20, 21 AND 22

Complete human chromosome sequencing information is currently available for chromosomes 20 [Deloukas *et al.*, 2001], 21 [Hattori *et al.*, 2000] and 22 [Dunham *et al.*, 1999]. Applying the CpG Island Searcher to the latest sequences and the default criteria and mapping information of those chromosomes, a total of 2345 CpG islands were extracted and 534 (23%) of them contained first coding exons and 650 (28%) contained other exons (Table 1).

The increased frequency for the association of CpG islands with genes relative to our previous report [Takai and Jones, 2002] is probably due to rapid progress in the identification and mapping of genes. The total number of CpG islands on chromosomes 21 and 22 has increased from our previous observations [Takai and Jones, 2002]. A part of this increase may come from recent progress in the sequencing of both chromosomes, and another part might rely on the discovery of newly recognized CpG islands as discussed above. An increase in the number of CpG islands categorized as "*Alu*" may be caused by the same reasons. However, the fact that a better association of CpG islands with genes is found, indicates the usefulness of CpG island extraction for gene prediction.

```
Selected lower limits: %GC=55, ObsCpG/ExpCpG=0.65, Length=500
NT030871, CpG island 1, start=20905, end=21736, %GC=55, ObsCpG/ExpCpG=0.658, Length=832
NT030871, CpG island 2, start=22852, end=23352, %GC=56.4, ObsCpG/ExpCpG=0.652, Length=501
NT030871, CpG island 3, start=176468, end=177028, %GC=55.7, ObsCpG/ExpCpG=0.665, Length=561
NT030871, CpG island 4, start=197269, end=197768, %GC=55, ObsCpG/ExpCpG=0.714, Length=500
NT030871, CpG island 5, start=401243, end=401752, %GC=55, ObsCpG/ExpCpG=0.701, Length=510
NT030871, CpG island 6, start=403562, end=404510, %GC=55, ObsCpG/ExpCpG=0.796, Length=949
NT030871, CpG island 7, start=433721, end=434381, %GC=56.5, ObsCpG/ExpCpG=0.662, Length=661
NT030871, CpG island 8, start=473918, end=474468, %GC=55.1, ObsCpG/ExpCpG=0.654, Length=551
NT030871, CpG island 9, start=487979, end=488553, %GC=56.6, ObsCpG/ExpCpG=0.651, Length=575
NT030871, CpG island 10, start=504343, end=504848, %GC=63.4, ObsCpG/ExpCpG=0.65, Length=506
NT030871, CpG island 11, start=538939, end=539438, %GC=55.2, ObsCpG/ExpCpG=0.656, Length=500
NT030871, CpG island 12, start=540794, end=541315, %GC=55.7, ObsCpG/ExpCpG=0.669, Length=522
NT030871, CpG island 13, start=557674, end=558178, %GC=55, ObsCpG/ExpCpG=0.653, Length=505
NT030871, CpG island 14, start=558672, end=559303, %GC=63.6, ObsCpG/ExpCpG=0.667, Length=632
NT030871, CpG island 15, start=563409, end=564042, %GC=58.3, ObsCpG/ExpCpG=0.676, Length=634
NT030871, CpG island 16, start=567329, end=568096, %GC=55.7, ObsCpG/ExpCpG=0.65, Length=768
```

Fig. 2. A truncated example of the analysis by the command-line version of the CpG Island Searcher for the input of 1,147,210 bp of a human chromosome 20 contig (NT_030871.1).

Table 1
The distribution of CpG islands on chromosomes 20, 21 and 22

Chromosome	20	21	22	Total
Length [Mb]	62	44	47	153
Mapped Genes	1244	611	1034	2889
First coding exon	204	104	226	534
Other exon	282	127	241	650
Alu	187	76	194	457
Other repeat	267	81	126	474
Unknown	103	54	73	230
Total	1043	442	860	2345

CpG islands were categorized into five categories in this order: "First coding exon" included at least the first coding exon of a gene and might or might not include downstream introns, exons and *Alu*s. An "other exonic" CpG island did not include a first coding exon and possibly included intronic and *Alu* sequences. An "*Alu*" did not include an exonic sequence. "Other repeat" sequences did not include *Alu* and were masked by Repeat Masker (University of Washington Genome Center, Seattle) as other types of repetitive sequence. For this analysis we used the contigs (build 28), NT_011387, NT_028391, NT_025215, NT_028392, NT_011362, NT_030871, NT_025929, NT_011333, NT_025218 (chromosome 20), NT_029490, NT_011512, NT_030187, NT_030188, NT_011515 (chromosome 21), NT_011516, NT_028395, NT_011519, NT_011520, NT_011521, NT_011522, NT_011523, NT_030872, NT_011525, NT_019197, NT_011526 (chromosome 22).

ACKNOWLEDGMENTS

We thank S. Catherall for maintenance of the web server. We also thank to P. Laird and G. Coetzee for helpful discussions. Daiya Takai and Peter A. Jones are supported by National Cancer Institute Grants R01 CA82422 and R01 CA 83867.

REFERENCES

[1] Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21.
[2] Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209–213.
[3] Coulondre, C., Miller, J. H., Farabaugh, P. J. and Gilbert, W. (1978). Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274**, 775–780.
[4] Deloukas, P., *et al.* (2001). The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**, 865–871.
[5] Dunham, I. *et al.* (1999). The DNA sequence of human chromosome 22. *Nature* **402**, 489–495.
[6] Feil, R. and Khosla, S. (1999). Genomic imprinting in mammals: an interplay between chromatin and DNA methylation. *Trends Genet.* **15**, 431–435.
[7] Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282.

[8] Hattori, M. *et al.* (2000). The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* **405**, 311–319.
[9] Jones, P. A. and Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* **3**, 415–428.
[10] Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992). CpG islands as gene markers in the human genome. *Genomics* **13**, 1095–1107.
[11] Panning, B. and Jaenisch, R. (1998). RNA and the epigenetic regulation of X chromosome inactivation. *Cell* **93**, 305–308.
[12] 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.
[13] Yoder, J. A., Walsh, C. P. and Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340.