

Large-scale human promoter mapping using CpG islands

Ilya P. Ioshikhes^{1,2} & Michael Q. Zhang¹

Vertebrate genomic DNA is generally CpG depleted^{1,2}, possibly because methylation of cytosines at 80% of CpG dinucleotides results in their frequent mutation to thymine, and thus CpG to TpG dinucleotides³. There are, however, genomic regions of high G+C content (CpG islands), where the occurrence of CpGs is significantly higher, close to the expected frequency, whereas the methylation concentration is significantly lower than the overall genome⁴. CpG islands⁵ are longer than 200 bp and have over 50% of G+C content and CpG frequency, at least 0.6 of that statistically expected. Approximately 50% of mammalian gene promoters are associated with one or more CpG islands⁶. Although biologists often intuitively use CpG islands for 5' gene identification^{7,8}, this has not been rigorously quantified⁹. We have determined the features that discriminate the promoter-associated and non-associated CpG islands. This led to an effective algorithm for large-scale promoter mapping (with 2-kb resolution) with a concentration of false-positive predictions of promoters much lower than previously obtained. Using this algorithm, we correctly discriminated approximately 85% of the CpG islands within an interval (-500 to +1500) around a transcriptional start site (TSS) from those that lie further away from TSSs. We also correctly mapped approximately 93% of the promoters containing CpG islands.

To define features that discriminate between TSS-associated and non-associated CpG islands, we initially divided a training set of CpG islands into four classes, based on the relationship of a CpG island to TSS. TSS-containing CpG islands (class 1) had a greater average length (-620 bp), and higher G+C content (69–70%) and CpG ratio (observed/statistically expected, 0.87–0.89) than CpG islands of the three other classes (~400 bp, 64–66% and 0.80–0.84, respectively; Table 1).

This effect allowed us to build an effective algorithm for localizing promoters in large-scale genomic analysis. We used the three CpG island feature variables as the discriminant variables for quadratic discriminant analysis¹⁰ (QDA). The QDA output parameters were used to classify promoter-related CpG islands on a test sequence set (Table 2). For purposes of our study, we defined promoter-related CpG islands as actual positives, and non-promoter-related CpG islands as actual negatives. If we correctly classify a CpG island as promoter-related or non-related, our prediction is true positive (TP) or negative,

correspondingly. If we classify a real promoter-related CpG island as promoter non-related or vice versa, we correspondingly get a false negative (FN) or positive (FP) prediction. Thus, for the set 1 test sequences, we obtained 8 TP predictions of TSS-containing CpG islands, along with 5 FN and 5 FP. Sensitivity $SN=TP/(TP+FN)$, the proportion of TP predictions out of the total number of actual positives, and specificity $SP=TP/(TP+FP)$, the proportion of TPs out of the total number of predicted positives, were both equal to 0.62.

To improve the prediction results, we optimized the initial classification scheme. We obtained the best results for a 2-class scheme based on whether a CpG island was inside or outside of a specific interval, within a target resolution of 2 kb. Its application to the training set itself, for the interval (-500 to +1500) around the TSS, brought $SN=0.85$, $SP=0.66$. For the test set 1 and interval (-821 to +1083) the results were $SN=0.75$, $SP=0.80$. Alternatively, for the interval (-500 to +1500) $SN=0.85$, $SP=0.73$. Other intervals (Table 2) produced intermediate values of SN and SP.

For further validation we applied this classification approach to a larger test set 2 of genomic sequences, for which statistics should be more robust than for test set 1. We obtained for it $SN=0.75$, $SP=0.47$ in the model interval (-500 to +1083). For the arbitrary interval (-500 to +1500) sensitivity was higher ($SN=0.85$) and specificity lower ($SP=0.42$). As for test set 1, other model intervals gave intermediate results. For subset 2-1, which had the most reliable TSS annotation, specificity of the predictions was better than for the entire set 2, with comparable sensitivity. Thus we obtained $SN=0.81$, $SP=0.52$ for the interval (-595 to +1083) and $SN=0.84$, $SP=0.49$ for (-500 to +1500). Although a certain drop of specificity for the test sequences versus the training set is understandable, the similar level of sensitivity is unexpected, but demonstrates the quality of the prediction.

Prediction results for the both test sets together were yet more successful. One can derive combined TP, FN and FP values from summing their component values over each set, and calculate corresponding SN and SP upon them (Table 2). Thus for the combination of set 1 and set 2, we obtained $SN=0.76$, $SP=0.52$ for the interval (-500 to +1083) and $SN=0.85$, $SP=0.46$ for (-500 to +1500). For the combination of test set 1 and subset 2-1 we obtained the best results for the intervals (-595 to +1083) ($SN=0.81$, $SP=0.56$) and (-500+1500) ($SN=0.84$, $SP=0.53$).

Table 1 • Characteristic feature parameters values for different classes of CpG islands

Class	Interval	Length (bp)		C+G content (%)		CpG Obs/Exp	
		med.	av. (s.d.)	med.	av. (s.d.)	med.	av. (s.d.)
1. (192 seq-s)	TSS-contain.	622	622 (347)	69.91	68.77 (5.51)	0.89	0.87 (0.09)
2. (220 seq-s)	(-500;+1500), no TSS-cont.	379	379 (172)	66.24	65.52 (5.84)	0.84	0.84(0.11)
3. (183 seq-s)	> +1500, < 3'	399	399 (256)	66.02	64.48 (6.08)	0.83	0.83 (0.13)
4. (73 seq-s)	< -500 or > 3'	433	433 (289)	64.96	63.54 (6.38)	0.80	0.81 (0.08)

Median (med.), average (av.) values and standard deviation (s.d.) are presented for each parameter. The median value for length (coinciding with its mean value) is the total length of all sequences divided by their number; for C+G content, the total number of Cs and Gs is divided by the total number of nucleotides in the sequences; for CpG Obs/Exp, the ratio of total number of CpG dinucleotides in the sequences to those statistically expected on the basis of the total number of Cs and Gs and the total number of mono- and dinucleotides within them. Average value is the mean of values of a given parameter obtained for all sequences.

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. ²Present address: Albert Einstein College of Medicine of Yeshiva University, Bronx, New York, USA. Correspondence should be addressed to M.Q.Z. (e-mail: mzhang@cshl.org).

Table 2 • Results of prediction of classes of CpG islands for different sequence sets and intervals

Sequence set	Training/test interval	TP	FN	FP	SN	SP
test set 1 (48 CpG islands from 21 GenBank records)	TSS-containing CpG island (4 classes)	8	5	5	0.62	0.62
	(-500..+835)	18	4	7	0.82	0.72
	(-500..+1083)	20	5	6	0.80	0.77
	(-595..+1083)	21	5	7	0.81	0.75
	(-821..+1083)	24	8	6	0.75	0.80
	(-500..+1500)	22	4	8	0.85	0.73
training set (668 CpG islands from 350 GenBank records)	(-500..+835)	226	132	140	0.63	0.62
	(-500..+1083)	275	105	152	0.72	0.64
	(-595..+1083)	312	89	163	0.78	0.66
	(-821..+1083)	333	72	171	0.82	0.66
	(-500..+1500)	349	63	179	0.85	0.66
test set 2 (340 CpG islands from 54 GenBank records, 135 genes, 4,870,561bp)	(-500..+1083)	73	24	81	0.75	0.47
	(-595..+1083)	79	20	93	0.80	0.46
	(-821..+1083)	88	23	110	0.79	0.44
	(-500..+1500)	88	15	120	0.85	0.42
subset 2-1 (225 CpG islands from 39 GenBank records)	(-500..+1083)	59	18	55	0.77	0.52
	(-595..+1083)	64	15	59	0.81	0.52
	(-821..+1083)	69	19	68	0.78	0.50
	(-500..+1500)	70	13	74	0.84	0.49
combined data set (training set and test set 1 and subset 2-1)	(-500..+1083)	354	128	213	0.73	0.62
	(-595..+1083)	397	109	229	0.78	0.63
	(-821..+1083)	426	99	245	0.81	0.63
	(-500..+1500)	441	80	261	0.85	0.63

Combining all three data sets brought further refinement in prediction results. In particular, for a combination of the training set, test set 1 and subset 2-1, we obtained SN=0.85, SP=0.63 at the interval (-500 to +1500). As we see, careful optimization eliminates certain controversy of the initial classification of the training set, where likely promoter-related class 2 CpG islands seem to be more similar to non-promoter related class 3 and 4 than to TSS-containing class 1 CpG islands. The controversy was caused by the space constraint of class 2: many long CpG islands could not fit the interval provided without overlapping the TSS, so class 2 was biased by short islands. The optimal classification scheme combines classes 1 and 2 as the positives and classes 3 and 4 as the negatives.

To assess the stability of our algorithm, we performed 10 cross-validation tests, standard in discrimination analysis¹¹. We merged training set, test set 1 and test subset 2-1, all with reliable TSS positioning, into one data set. In each test, we randomly chose 30% of the data as the test set and the remaining 70% as the training set. The idea is to use the training set to determine the classification surface, and then to do the prediction on the disjoint test set. The prediction results for the 10 tests were quite robust, with sensitivities varying around 0.72 and specificities around 0.66 in average (Table 3).

Table 3 • Cross-validation results on 10 partitions of the training set, test set 1 and subset 2-1

N	TP	FN	FP	SN	SP
0.	115	53	59	0.68	0.66
1.	116	34	52	0.77	0.69
2.	123	43	58	0.74	0.68
3.	105	54	49	0.66	0.68
4.	121	48	57	0.72	0.68
5.	118	42	66	0.74	0.64
6.	112	47	50	0.70	0.69
7.	110	48	52	0.70	0.68
8.	120	47	74	0.72	0.62
9.	124	38	73	0.77	0.62
Av.				0.72	0.66
s.d.				0.04	0.03

Overall, the results indicated that our method correctly identified approximately 85% of CpG islands centred in the interval (-500 to +1500) around the TSS, and thus detected the position of the TSS with this precision, whereas the proportion of correct predictions was more than 60% of their total number. The sequences of test set 2 were entirely different from those of

the training set, yet all the CpG islands have been mapped in a consistent way, which lends credibility to our results. The main restriction of our method is its applicability to identifying 5' ends only of genes that have CpG island in close proximity to their promoter, which was about half of the total (68/135) genes, consistent with a previous estimation⁶. We correctly predicted the TSS of 93% of these genes (63/68) with a resolution of 2 kb. This means that we correctly localized promoters of about half (~47%, 63/135 in our set) of all genes (63 TPs). Promoters of the other 72 genes were not correctly identified and therefore we obtained 72 FN predictions. (In this paragraph, the terms 'true positive', 'false negative' and 'false positive' refer to the prediction of gene promoters, unlike in the previous paragraphs where

they referred to classification of CpG islands.) Still, we obtained 120 FP predictions (Table 2). Sensitivity of promoter prediction by this method was therefore 0.47, and specificity 0.34. The relatively modest value of SN reflects the main restriction of our method, its applicability is only for detection of promoters associated with CpG islands, only about 50% of all genes. Sensitivity for this group of genes alone gives a figure of SN=0.93. The total length of the GenBank sequences of test set 2 is 4.9 Mb. Containing 135 genes, they average 1 gene per 36 kb, which is consistent with existing estimates⁶. Our approach gives 1 positive (either true or false) promoter prediction per 26.5 kb, far closer to the correct figure than was obtained by other algorithms of promoter recognition⁹. In particular, we compared our results with the corresponding ones obtained by program PromoterScan¹² (<http://www-bimas.cit.nih.gov/molbio/proscan/index.html>) on the sequences of test set 2. PromoterScan gave us 60 TPs, 75 FNs and 966 FPs, with sensitivity SN=0.44 and specificity SP=0.06 only, with one positive per 4.7 kb. Our algorithm also correctly identified 5' ends of the three prototype genes with CpG islands around their promoters¹, with no false-positive predictions. Our results provide an advance towards building an efficient algorithm for promoter localization in genomic BAC sequences. Because aberrant methylation of CpG islands is one mechanism of inactivating tumour suppresser genes in neoplasia and altered cytosine methylation is important in cancer development^{13,14}, identification of promoter-associated CpG islands will have an impact on cancer research and gene silencing. For example, cytosine methylation of promoter-associated CpG islands is involved in the allele-specific inactivation of imprinted genes¹⁵ and genes on the inactive X chromosome¹⁶.

Methods

We used the EMBL CPGISLE database of human CpG islands¹⁷ (<ftp://ftp.no.embnet.org/cpgisle/>). We used the CpGPlot program (<http://www.sanger.ac.uk/Software/EMBOSS/>) in the extended GCG (EGCG) package (see <http://www.sanger.ac.uk/Users/pmr/egcg.html> and references therein) for analysis of genes, based on the formal definition of CpG islands⁵.

Availability of both the database and the corresponding mapping program was crucial for a choice of sequence training set and software for our study. For instance, a much larger set of CpG islands was collected based on experimental rather than computational analysis^{18,19} (<http://www.sanger.ac.uk/HGP/cgi.shtml>). Though computer analysis of these sequences might retrieve some important sequence features, efficiency of their usage for promoter mapping on new genomic sequences would be quite questionable. On the other hand, there are other existing programs for showing CpG islands, based on essentially the same formal definition (for example, to WWWCPG program by Milanese *et al.* at <http://www.itba.mi.cnr.it/web-gene/> and references therein). These latter programs, however, yield somewhat different results (data not shown) when applied to the CPGISLE sequences and do not provide equivalent publicly available data sets.

In analysing the CPGISLE database, we attempted to obtain discriminative features between TSS-containing and/or TSS-adjacent CpG islands and other CpG islands. We used in the training set only those CpG islands for which we could explicitly establish their position relative to TSS according to GenBank annotations.

We initially divided the CpG islands into four separate classes: (i) CpG islands containing the TSS (192 sequences); (ii) CpG islands in area -500 to +1500 but not directly overlapping TSS (220 sequences); (iii) CpG islands inside the gene (183 sequences); and (iv) other CpG islands.

We calculated mono-, di- and trinucleotide contents of the sequences by the GCG 'composition' program. We normalized the values by calculation of proportional oligonucleotide content, where total of oligonucleotides of a given length was defined as 100%.

We then carried out a comparative analysis of the results. We used three CpG island parameters, length, C+G mononucleotide content and ratio of observed to expected CpG content (the ratio of the CpG frequency to the product of the C and G frequencies), as the training data set for QDA, a standard multivariate statistical pattern recognition method¹¹ that has been previously applied to sequence discriminant analysis^{20,21}. The QDA output describes parameters of a quadratic function separating classes of a training set in optimal way, where every data point of the training set is represented as a point in three-dimensional space, according to the corresponding values of the three parameters of the CpG islands. We used these parameters to distinguish TSS-containing CpG islands in a test set of sequences (test set 1). The test sequences (48 CpG islands from 21 GenBank records) all were different from those of the training set, and con-

sisted of additional sequences from either the CPGISLE database or GenBank sequences. These latter sequences were longer than 30 kb, and arbitrarily chosen from GenBank (Gb_pr:*). We mapped CpG islands for these sequences by the same CPGPLOT program (with default parameters), which was used originally to construct the database of CpG islands.

To refine the prediction results obtained for the test set 1, we modified the classification scheme, based on the distance between the TSS and the centre of a given CpG island. In the course of modifying the classification scheme, we also changed the classification of the sequences of the training set. At every point of the scheme optimization, we classified the training sequences according to the modified scheme, then using their parameters (length, C+G content, ratio of observed to expected CpG content) along with the modified classification for training the QDA program. We applied results of the training for classification of the sequences of the test set 1. Classification scheme for training and test sequences were always identical, and few of the tested intervals are listed in the column "Training/test interval" of Table 2. We then applied this scheme for classification of CpG islands of test set 2. This latter set contained Gb_pr:* sequences (longer than 30 kb) from GenBank, different from those of either the training set or of the test set 1. Test sets consisted of 340 CpG islands from 54 GenBank records of total length 4,870,561 bp, which contained 135 genes, 105 associated with some CpG island. We used a coding sequence (CDS) start for classification of sequences with no explicit annotation of the TSS in the GenBank records. Because the CDS start and TSS positions are generally not the same²², we gathered the sequences with unambiguous TSS annotation in a special subset of the test set 2 (Subset 2-1) and analysed them separately. An initial version of a complete software package (CpG_promoter) and implementation details of the algorithm are available (<ftp://cshl.org/pub/science/mzhanglab/ioshikhes/>).

Acknowledgements

We thank R. Bari for assistance in sequence annotation; T. Zhang for assistance in testing CpG_promoter; S.H. Cross for discussions; P. Rice and R. Lopez for consultations about the EMBOSS project and CpGPlot program; and J. Locker and S. Emmons for editing of the text. This work was supported by National Institutes of Health Grant HG01696 to M.Q.Z.

Received 22 March; accepted 10 June 2000.

- Bird, A.P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 (1980).
- Jones, P.A., Rideout, W.M. 3d, Shen, J.C., Spruck, C.H. & Tsai, Y.C. Methylation, mutation and cancer. *Bioessays* **14**, 33–36 (1992).
- Bird, A. DNA methylation de novo. *Science* **286**, 2287–2288 (1999).
- Antequera, F. & Bird, A. CpG islands. *EXS* **64**, 169–185 (1993).
- Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
- Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90**, 11995–11999 (1993).
- Cross, S.H. & Bird, A.P. CpG islands and genes. *Curr. Opin. Genet. Dev.* **5**, 309–314 (1995).
- Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
- Pedersen, A.G., Baldi, P., Chauvin, Y. & Brunak, S. The biology of eukaryotic promoter prediction—a review. *Comput. Chem.* **23**, 191–207 (1999).
- Venables, W.N. & Ripley, B.D. *Modern Applied Statistics with S-Plus* (Springer, New York, 1994).
- McLachlan, G.J. *Discriminant Analysis and Statistical Pattern Recognition* (Wiley, New York, 1992).
- Prestridge, D.S. Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* **249**, 923–932 (1995).
- Toyota, M. & Issa, J.P. CpG island methylator phenotypes in aging and cancer. *Semin. Cancer Biol.* **9**, 349–357 (1999).
- Baylin, S.B. & Herman, J.G. DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends Genet.* **16**, 168–174 (2000).
- Barlow, D.P. Gametic imprinting in mammals. *Science* **270**, 1610–1613 (1995).
- Singer-Sam, J. & Riggs, A.D. X chromosome inactivation and DNA methylation. *EXS* **64**, 358–384 (1993).
- Larsen, F., Gundersen, G., Lopez, R. & Prydz, H. CpG islands as gene markers in the human genome. *Genomics* **13**, 1095–1107 (1992).
- Cross, S.H., Charlton, J.A., Nan, X. & Bird, A.P. Purification of CpG islands using a methylated DNA binding column. *Nature Genet.* **6**, 236–244 (1994).
- Cross, S.H., Clark, V.H. & Bird, A.P. Isolation of CpG islands from large genomic clones. *Nucleic Acids Res.* **27**, 2099–2107 (1999).
- Zhang, M.Q. in *Proceedings of Pacific Symposium on Biocomputing 1998* (eds Altman, R.B. *et al.*) 240–251 (World Scientific, Singapore, 1998).
- Zhang, M.Q. Identification of protein coding regions in the human genome based on quadratic discriminant analysis. *Proc. Natl Acad. Sci. USA* **94**, 565–568 (1997).
- Zhang, M.Q. Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.* **7**, 919–932 (1998).