JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 25, Number 2, 2018

Pp. 158–169
DOI: 10.1089/cmb.2016.0178

A CpGCluster-Teaching–Learning-Based Optimization for Prediction of CpG Islands in the Human Genome

CHENG-HONG YANG,[1,2] YI-CHENG CHIANG,[1] LI-YEH CHUANG,[3] and YU-DA LIN[1]

ABSTRACT

Many CpG island detection methods have been proposed based on sliding window and clustering technology, but the accuracy of these methods is proportional to the time required. Therefore, an accurate and rapid method for identifying CpG islands remains an important challenge in the complete human genome. We propose a hybrid method CpGTLBO to detect the CpG islands in the human genome. The method uses the clustering approach and the teaching–learning-based optimization (TLBO) algorithm. The clustering approach is used to detect CpG island candidates, and it can effectively reduce the huge volume of unnecessary DNA fragments. TLBO was used to accurately predict CpG islands among promising CpG island candidates. A comparison based on six contig data sets and a whole human genome analysis showed that the identifying stability of CpGTLBO outperformed eight existing methods in terms of sensitivity (*SN*), specificity (*SP*), accuracy (*ACC*), performance coefficient (*PC*), and correlation coefficient (*CC*) and processing time. Results indicated that ClusterTLBO can effectively overcome the drawbacks and maintain the advantages in both the CpGcluster and TLBO.

Keywords: CpG island detection, clustering technology, sliding window method, teaching–learning-based optimization.

1. INTRODUCTION

CpG dinucleotides (CpGs) are a cytosine (C) directly combined with a guanine (G) and are randomly and unevenly located in the human genome. CpG sites are the regions of DNA sequence in which the cytosine (C) base is followed by a guanine (G) base in the linear sequence of bases along $5' \rightarrow 3'$ direction. CpG is shorthand for 5′-C-phosphate-G-3′, that is, C and G separated by only one phosphate; phosphate links any two bases together in DNA sequence. CpG islands are the regions of DNA sequence with a high frequency of CpG sites, on average, 1000 base pairs (bp) long and show an elevated G + C base composition, little CpG depletion, and frequent absence of DNA methylation. These shared properties have allowed CpG islands to be isolated as a relatively homogeneous fraction of the genome, despite the heterogeneity of their

[1]Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan.
[2]Graduate Institute of Clinical Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan.
[3]Department of Chemical Engineering, Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung, Taiwan.

individual nucleotide sequences (Keles et al., 2006; Illingworth et al., 2010; Deaton and Bird, 2011; Kuo et al., 2011).

Gardiner-Garden and Frommer (GGF) noted that, ''Stretches of DNA with a high G+C content, and a frequency of CpG dinucleotides close to the expected value, appear as CpG clusters within the CpG-depleted bulk DNA, and are now generally known as CpG islands'' (Gardiner-Garden and Frommer, 1987). GGF defines a CpG island as fulfilling three conditions: (i) GC content (i.e., GC% or the percentage of nitrogenous bases on a DNA molecule that are either guanine or cytosine) exceeds 50%, (ii) observed-to-expected ratio (i.e., O/E ratio or the ratio of observed vs. expected number of CpGs) exceeds 0.6, and (iii) length exceeds 200 bp. Takai and Jones (2002) also define a CpG island as fulfilling three conditions: (i) O/E ratio exceeds 0.65, (ii) GC% exceeds 55%, and (iii) length is at least 500 bp.

Most CpG island detection methods are based on sliding window [including CpGplot (Olson, 2002), CpGProD (Ponger and Mouchiroud, 2002), CpGIS (Takai and Jones, 2002)] and particle swarm optimization (PSO) based (Chuang et al., 2011) methods. These methods scan DNA sequences to identify potential CpG islands that conform to the GGF or Takai and Jones definitions (Gardiner-Garden and Frommer, 1987). Sliding window is similar to brute force search in that it requires considerable lengths of time, and the window size is an important parameter to determine the quality of CpG island prediction results. The CpGcluster method directly uses statistical properties (p-value) to detect CpG clusters without considering CpG island definitions (Hackenberg et al., 2006), and it is likely to feature a relatively short length, high O/E ratio, and high GC% due to the strict p-value, leading to regions failing to comply with CpG island definitions (Su et al., 2010).

This study proposes a hybrid method CpGTLBO that uses the clustering approach and the teaching–learning-based optimization (TLBO) algorithm. The GGF definition was used to allow for comparisons with other methods. Results showed that CpGTLBO provides reduced search time, higher sensitivity, higher accuracy, and a higher correlation coefficient in the human genome.

2. METHODS

2.1. TLBO algorithm

TLBO was introduced by Rao et al. (2011). TLBO is based on the influence of a teacher on learner output. The main terminate processes of TLBO can be divided into two phases, that is, ''teacher phase'' and ''learner phase.'' The candidate solutions of the teacher phase are randomly distributed over the search space, and the best solution is selected from among these candidates. In the learner phase, the solutions seek to pass their own information through mutual interaction between learners.

2.2. CpGTLBO

CpGTLBO is based on the CpGcluster algorithm, which observes that the distance between CpGs in a CpG island is significantly shorter than within the bulk DNA sequences (Hackenberg et al., 2006), and CpG islands feature a high degree of local clustering of CpGs. CpGTLBO uses the physical distance between neighboring CpGs to directly detect more intensive CpG clusters in the genome according to CpGcluster. All CpG clusters are calculated by their statistical p values, and each statistically significant CpG cluster is a CpG island candidate instead of to the sliding window method, thus effectively reducing the search range in the DNA sequence. All CpG island candidates are extended both forward and backward by 200 bp for the second step to detect the CpG islands and, thus, meet the GGF criterion. These CpG island candidates are then accurately predicted by the TLBO algorithm. Figure 1 shows the CpGTLBO flowchart, and the CpGTLBO procedure pseudo-code is shown in Supplementary Figure S1.

2.2.1. Cluster the CpGs by distance-based algorithm. This step determines that the CpGs are set in a cluster and whether this cluster is an CpG island candidate. All steps are described in detail as follows:

1. All CpG positions are scanned from 5′ to 3′ in a DNA sequence, and the CpG positions are collected into a set $C=\{c_1, c_2, \ldots, c_n\}$, where n is the total number of CpGs.
2. Calculate the physical distance of all adjacent CpGs, which is computed by $d_i=c_{i+1}-c_i-1$, where $i=1$ to n–1, and collect all physical distances into a set C_s. In sequence, the shortest distance between adjacent CpGs (i.e., CGCG) is 1.

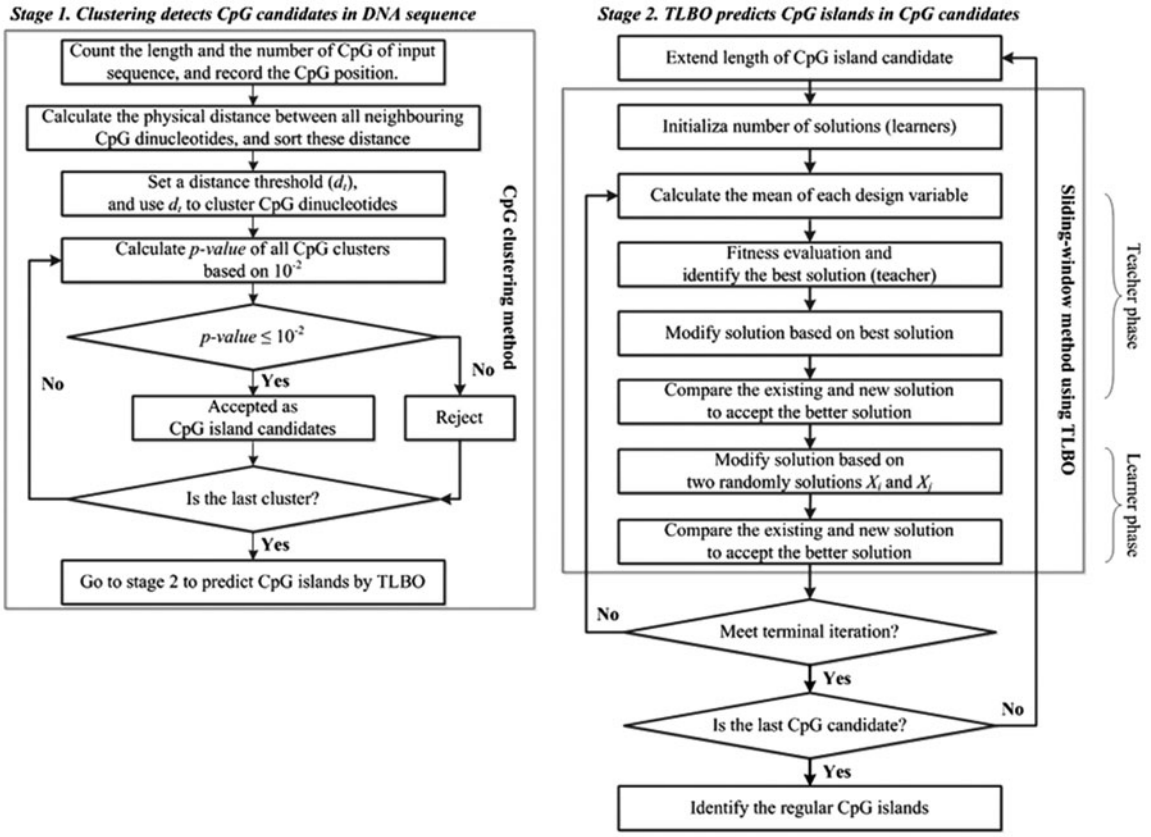

FIG. 1. CpGTLBO flowchart.

3. Set a distance threshold (d_t), which is used to determine whether the adjacent CpGs belong to a same cluster or not.
4. The clustering process begins from the first CpG of set C and extends downstream ($5' \rightarrow 3'$) to the last CpG. All CpGs are divided into clusters based on the distance threshold. When the adjacent distance between CpGs is smaller than d_t, the two adjacent CpGs are regarded as belonging to the same cluster; otherwise, the position of the downstream C nucleotide of the adjacent CpGs is defined as the start position of a new CpG cluster.

2.2.2. Detect statistically significant CpG island candidates by p*-value criterion.* The clustering technology detects all CpG clusters, and the p-value of a CpG cluster is used to calculate the probability of a CpG cluster appearing in a random sequence. When the p-value of a cluster is smaller than 0.01 (Hackenberg et al., 2006), the cluster is retained; otherwise, the cluster is discarded. Each statistically significant CpG cluster is a CpG island candidate. The probability mentioned earlier is calculated by the cumulative density function at point n_f of the CpG cluster, and it is taken as the p-value:

$$P_{N,p}^{cum}(x <= n_f) = \sum_{x=0}^{n_f} \binom{x-(N+1)-1}{(N-1)-1} \times p^{N-1} \times (1-p)^x, \tag{1}$$

$$n_f = L - 2 \times N, \tag{2}$$

$$p = \frac{N_s}{N_{is}}, \tag{3}$$

where N is the number of CpGs in the cluster, n_f is the number of independent non-CpGs in the cluster, L is the number of nucleotides in the cluster, and p is the probability of successfully finding a CpG. N_s and N_{is} are, respectively, the number of CpGs and the number of independent dinucleotides in the DNA sequence.

2.2.3. Predict CpG islands by TLBO algorithm. If stage 1 obtains CpG island candidates, the CpG island candidates may be shorter than 200 bp. The lengths of these CpG island candidates must be extended to apply the TLBO algorithm for predicting the CpG island region. All CpG island candidates are extended forward and backward by 200 bp, after TLBO is implemented in each specific CpG island candidate to predict the CpG island. TLBO is able to find a better CpG island region in a specific CpG island candidate and confirms the CpG island to comply with the CpG island definition. The process of the TLBO algorithm involves five steps (Fig. 2), including initialization, fitness evaluation, teacher phase, learner phase, and termination criterion. The initialization step generates the population (learners) for searching the rational CpG island region. The fitness evaluation step analyzes the learner value according to the CpG island properties. The teacher phase step generates new learners by simulating while learners are learning from the teacher. The learner phase step simulates that learners increase their knowledge via interaction among themselves. If the learner gives a better function value, the learner is accepted. The termination criterion step determines whether TLBO iteration is capable of reaching the terminate criteria. The learner with the highest fitness value is predicted as the CpG island for the specific CpG island candidates. All steps are explained in detail as follows:

1. Initialization

In TLBO, the learner is defined as a possible CpG island region. Equation 4 is the learner encoding formula.

$$X_i = (xs_i, xl_i), \tag{4}$$

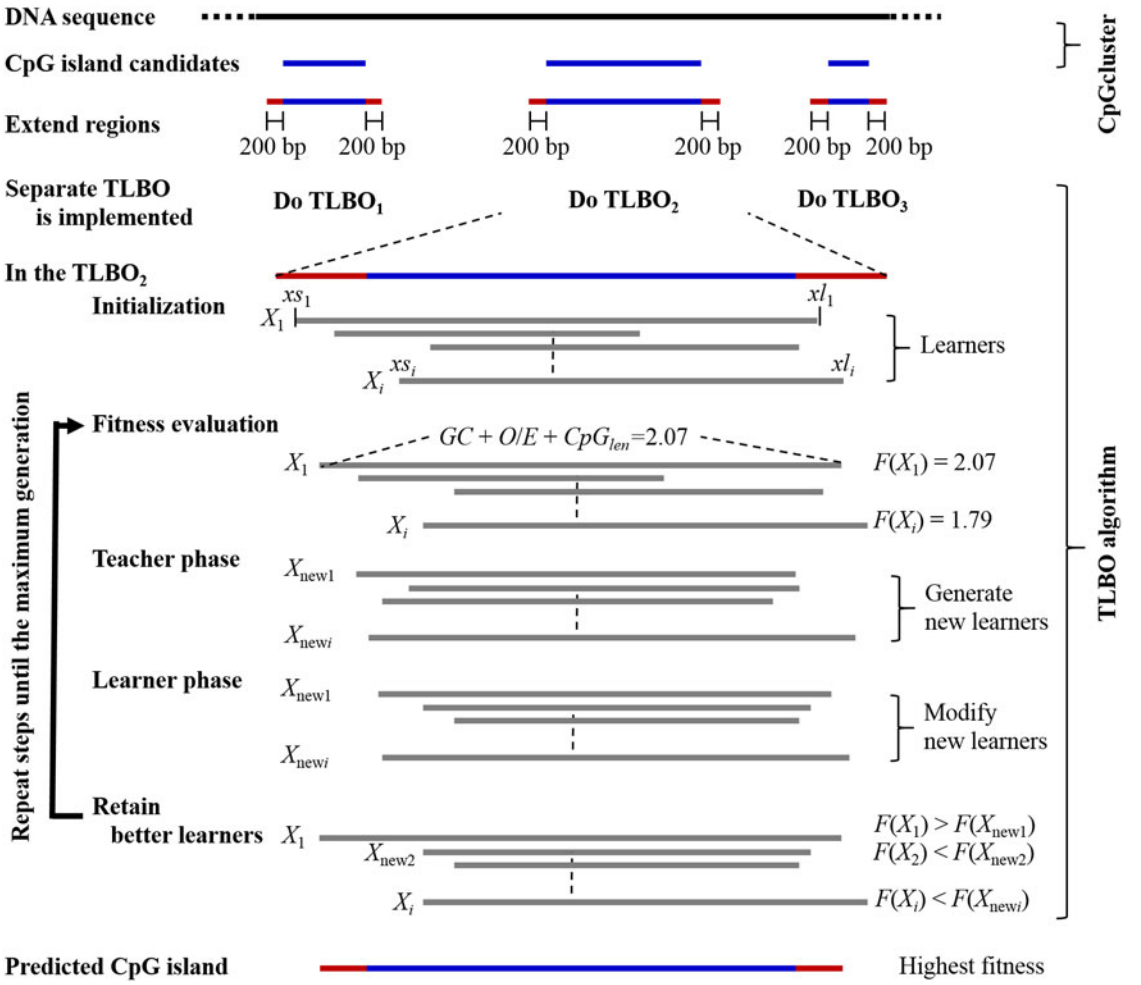

FIG. 2. The diagram of TLBO algorithm predicts CpG islands. TLBO, teaching–learning-based optimization.

where X_i is the ith learner, xs_i and xl_i are, respectively, the start position and length of a predicted CpG island in a CpG island candidate. At initialization, each learner X_i is randomly generated by the parameters xs_i and xl_i.

2. Fitness evaluation

According to the GGF CpG island definition (Gardiner-Garden and Frommer, 1987), we use properties of length ≥200 bp, GC% ≥50%, and O/E ratio ≥0.6 to design a fitness function for evaluating the learners and to determine whether the learner range includes the CpG island or not. Equation 5 is the normalization function for calculating the predicted CpG island length. Equations 6 and 7 are, respectively, the GC% function and the O/E ratio function. Equation 8 is the designed fitness function that sums the values of three properties, with a higher fitness value indicating that the learner range has a better CpG island prediction result.

$$CpG_{len}(X_i) = \frac{CpG_{length}}{X_{\max} - X_{\min}}, \tag{5}$$

$$GC(X_i) = \frac{\#C + \#G}{CpG_{length}}, \tag{6}$$

$$\mathrm{Obs}_{\mathrm{CpG}}/\mathrm{Exp}_{\mathrm{CpG}}(X_i) = \frac{\#CpG}{CpG_{length}} / \left(\frac{\#C}{CpG_{length}} \times \frac{\#G}{CpG_{length}}\right), \tag{7}$$

$$Fitness(X_i) = GC(X_i) + Obs_{CpG}/Exp_{CpG}(X_i) + CpG_{len}(X_i), \tag{8}$$

where CpG_{length} represents the number of nucleotides in the learner X_i. $X_{\min}$ is the starting position of the CpG island candidate minus 200, and $X_{\max}$ is the ending position of the CpG island candidate plus 200. *#C* and *#G* are, respectively, the numbers of *C* and *G* nucleotides in the learner X_i. *#CpG* is the number of CpGs in the learner X_i.

3. Teacher phase

In a population, learner improvement depends largely on the presence of a highly learned person (*teacher*) and the mean level of the population (*M*), which are defined as follows:

$$teacher_g = X_g\_best, \tag{9}$$

$$M_g = \frac{1}{n}\sum_{i=1}^{n} X_{g,\ i}, \tag{10}$$

where g represents the number of generations, n is the number of learners, and i is the target learner. In addition, the amount of learning of each learner ($Difference\_Mean_i$) is produced according to Equation 11; hence the amount of learning for different learners varies depending on the random value of the following equation:

$$Difference\_Mean_i = r_i(teacher_g - T_f M_g), \tag{11}$$

$$T_f = round\ [1 + rand(0,\ 1)], \tag{12}$$

where the range of random number r_i is between 0 and 1. T_f is a teaching factor that decides which value of the mean needs to be changed. The T_f is randomly generated as either 1 or 2, which is again a heuristic step and is decided randomly with equal probability as in Equation 12. The learners are updated by Equation 13. The difference just mentioned modifies the existing learner according to Equation 13.

$$X_{new,\ i} = X_i + Difference\_Mean_i, \tag{13}$$

4. Learner phase

Learners increase their knowledge by two different means. In this step, a learner randomly interacts with other learners (i.e., randomly selecting another learner X_j, such that $i \neq j$). Thus, fitness can be improved if the other learner has relatively more knowledge. The learner modification step is expressed as follows:

$$X_{new,\ i} = \begin{cases} X_{old,\ i} + r_i(X_i - X_j), f(X_i) > f(X_j) \\ X_{old,\ i} + r_i(X_j - X_i), f(X_i) \leq f(X_j) \end{cases}. \tag{14}$$

5. Confirm whether the stop criterion is met or not

This step repeats the subsections 1 to 4 until the maximum generation is reached. When all potential CpG island candidates are predicted by TLBO, all identified CpG islands are the CpG island detection results. In the clustering stage, the parameters *p*-value and distance threshold are, respectively, set to 0.01 and 65th. In the TLBO stage, the population size is 300, and the maximum generation is 100 (Gudise and Venayagamoorthy, 2003).

2.3. Example of CpGTLBO

An example is provided in the Supporting Information to illustrate how the CpGTLBO works.

3. RESULTS

3.1. Availability of supporting data

All contig sequences, entire chromosomes, and verified CpG islands were obtained from GenBank Database in NCBI (www.ncbi.nlm.nih.gov), along with the entire human genome (NCBI.36). We selected six regular contig sequences in chromosomes 21 and 22. Chromosome 21 contains the NT_113952.1 (184,355 bp), NT_113953.1 (131,056 bp), NT_113954.1 (129,889 bp), NT_113955.2 (281,920 bp), and NT_113958.2 (209,483 bp), whereas NT_028395.3 (647,850 bp) belongs to chromosome 22. Known CpG islands published in the NCBI are used as the gold standard. These CpG islands in contig sequences were verified based on sequence analysis and bisulfite sequencing (BS-seq). CpG islands were extracted from these contigs and entire chromosomes with the following detection algorithm. The CpGTLBO program and datasets can be downloaded from the following link https://goo.gl/0jtp0R.

3.2. Comparison of the six contig sequences for CpG island detection methods

Contig sequences are often used to test the CpG island detection methods in terms of detecting the CpG islands (Chuang et al., 2011). Table 1 shows the detection results of six contig sequences by using nine methods taken from the relevant literature. CpGPlot showed excellent specificity (*SP*) results in all contig sequences. The performance measurement was introduced in Supplementary Material. CPSORL showed better sensitivity (*SN*), performance coefficient (*PC*), and correlation coefficient (*CC*) values than CpGPlot, CpGcluster, CpGProD, CpGIS, PSO, PSORL, and CPSO. However, CpGTLBO outperforms CPSORL for *SN*, accuracy (*ACC*), *PC*, and *CC* in the six contig sequences. Supplementary Table S1 shows the detailed results for CPSORL and CpGTLBO in the six contig sequences, including GC% (average), O/E ratio (average), and number of detection CpG islands. CpGTLBO had a higher GC% (average) and O/E ratio (average) than CPSORL, indicating that CpGTLBO tends to search shorter but CpG-richer CpG island regions. We used the *p*-value of the Wilcoxon Signed-Rank test for pairs of result groups to determine whether the difference between the two methods is significant or not. In the six contig sequence results, all *p* values of CpGTLBO compared with the other eight methods were $p < 0.001$, indicating that the CpGTLBO was, indeed, effective in identifying CpG islands in the DNA sequences.

3.3. Comparison of the whole human genome for various CpG island detection methods

A comparison of CpG island detection methods can facilitate the evaluation of three conditions for detected CpG islands in the whole human genome. In Table 2, the results of the CpG island detection methods are obtained from the published literature, including CpGplot (Olson, 2002), CpGProD (Ponger and Mouchiroud, 2002), CpGIS (Takai and Jones, 2002), and PSO based (Chuang et al., 2011). CpGcluster detected the minimum CpG island length (8 bp). In chromosome 21, the total length of the true CpG islands is 1,719,555 bp and the coverage of true CpG islands is 3.66%. CpGTLBO detected a total length of 1,733,292 bp, and the coverage was 3.69%. The CpG island length and coverage values of CpGTLBO approximate the real values. In chromosome 22, CpGTLBO, respectively, detected the total length of true CpG islands and detected CpG islands as 3,114,716 and 2,998,371 bp, with respective coverage of 6.27% and 6.17%. The results indicate that CpGTLBO outperformed the other methods for detecting CpG islands in chromosomes 21 and 22.

Table 1. Comparison of Nine Different Methods for CpG Island Detection

Contig		*Methods*								
		CpGPlot	*CpGcluster*	*CpGProD*	*CpGIS*	*PSO*	*PSORL*	*CPSO*	*CPSORL*	*CpGTLBO*
NT_113952.1	SN	56.43	50.46	58.07	83.98	69.22	75.58	77.43	84.88	**87.94**
	SP	**100.0**	99.95	99.50	99.05	99.61	99.02	99.58	99.05	99.64
	ACC	98.09	97.78	97.69	98.39	98.28	97.99	98.61	98.43	**99.12**
	PC	56.42	49.92	52.36	69.59	63.77	62.27	70.91	70.34	**81.48**
	CC	74.38	69.41	68.83	81.25	77.66	75.71	82.49	81.8	**89.36**
NT_113955.2	SN	47.19	67.15	68.51	85.12	54.47	59.63	77.8	87.38	**91.86**
	SP	**100.0**	99.72	99.63	99.30	99.96	99.88	99.5	99.61	99.58
	ACC	98.08	98.54	98.50	98.79	98.31	98.42	98.71	99.16	**99.30**
	PC	47.14	62.47	62.35	71.78	53.87	57.74	68.67	79.08	**82.69**
	CC	67.94	77.03	76.65	82.96	72.41	74.51	80.85	87.89	**90.17**
NT_113958.2	SN	51.29	27.16	46.41	82.13	79.27	81.65	81.08	**84.11**	80.11
	SP	**99.99**	99.94	98.93	98.26	98.13	97.90	98.17	98.34	99.22
	ACC	96.90	95.32	95.60	97.24	96.93	96.87	97.08	97.43	**98.00**
	PC	51.24	26.92	40.10	65.36	62.10	62.33	63.8	67.51	**71.82**
	CC	70.38	49.96	56.80	77.63	75.03	75.28	76.41	79.31	**82.63**
NT_113953.1	SN	22.80	57.32	29.79	74.05	60.20	64.80	70.53	75.65	**80.91**
	SP	**100.0**	99.74	99.56	98.83	99.27	99.23	99.22	99.13	99.56
	ACC	97.76	98.51	97.53	98.11	98.13	98.23	98.38	98.45	**99.02**
	PC	22.80	52.74	25.96	53.23	48.39	51.59	55.91	58.57	**70.63**
	CC	47.21	69.89	43.61	68.64	64.50	67.25	70.9	73.1	**82.31**
NT_113954.1	SN	31.24	29.86	52.01	76.31	56.92	63.58	70.54	77.68	**78.85**
	SP	**100.0**	99.46	98.72	97.62	98.40	98.13	98.34	98.23	98.49
	ACC	97.47	96.90	97.00	96.83	96.87	96.86	97.32	97.48	**97.76**
	PC	31.24	26.19	38.94	47.05	40.12	42.74	49.22	53.15	**56.56**
	CC	55.17	43.81	54.68	63.29	55.65	58.36	64.72	68.53	**71.37**
NT_028395.3	SN	27.11	44.89	54.18	76.68	68.97	72.79	72.52	77.02	**81.80**
	SP	**100.0**	99.47	99.45	98.93	99.27	98.99	99.18	98.9	99.40
	ACC	97.98	97.53	98.19	98.14	98.19	98.06	98.24	98.12	**98.78**
	PC	27.10	39.26	45.36	59.36	57.49	57.17	59.36	59.25	**70.43**
	CC	51.51	57.21	62.26	73.57	72.21	71.75	73.61	73.48	**82.02**

The bold type indicates the best value in all methods.

Table 3 shows the results of CpG island identification obtained by CpGcluster, CpGIS, CPSORL, and CpGTLBO in the whole human genome. The resulting average island length using CpGTLBO was longer than that using CpGcluster (443 bp vs. 273 bp) and shorter than CpGIS (443 bp vs. 1090 bp), but similar to CPSORL (443 bp vs. 572 bp). We examined the promoter and transcription start site (TSS) that overlaps with the CpG island region. A promoter region was defined as −1500 to +500 bp around the TSS. The TSS number for CpGTLBO was higher (below 9.65%) than that for CpGcluster, CpGIS, and CPSORL. The promoter region of CpGTLBO was higher (below 11.8%) than that of CpGcluster and CpGIS.

Supplementary Figure S2 shows the length distribution of the CpG islands. The CpG island length result of CpGTLBO ranged from 200 to 749 bp. Supplementary Figure S3 shows the distributions of GC% and the O/E ratio in the detected CpG islands using CpGTLBO. Most GC% were between 0.5 and 0.7, and the O/E ratios were between 0.6 and 1.0, indicating that the results of CpG islands detected using CpGTLBO conform to the GGF criteria. Chuang et al. (2011) proved that CPSORL was superior to CpGcluster and CpGIS. Therefore, we compared the measurement performance of CPSORL and CpGTLBO in the whole human genome. In Supplementary Table S2, CpGTLBO shows a very strong improvement in *SN*, *SP*, *ACC*, *PC*, and *CC*, with *SN* showing the most significant difference. For the Wilcoxon Signed-Rank test, the respective *p* values of *SN*, *SP*, *ACC*, *PC*, and *CC* were 2.352E-5, 6.296E-5, 1.813E-5, 1.822E-5, and 1.818E-5, indicating the strong superiority of CpGTLBO over CPSORL for CpG island detection in the whole human genome.

Table 2. Basic Statistics for CpG Islands Identified in Human Chromosomes 21 and 22 Using Different Methods (NCBI.36)

Methods	*Real island*	*CpGPlot*	*CpGcluster*	*CpGProD*	*CpGIS*	*PSO*	*PSORL*	*CPSO*	*CPSORL*	*CpGTLBO*
Chromosome 21 (46,944,329 bp)										
Total length of CpG islands	1,719,555	347,334	639,161	1,072,192	1,280,505	1,440,953	1,564,596	1,527,114	1,607,472	1,733,292
No. of islands detected	4547	973	2703	1091	3704	2648	2648	2813	2813	4192
Island coverage (%)	3.66	0.73	1.36	2.28	2.73	3.07	3.30	3.36	3.40	3.69
Island length (bp)										
Average	378	357	237	983	346	542	591	561	571	413
Minimum	201	101	8	500	200	202	202	202	202	201
Maximum	6114	3047	3028	6732	1948	4009	4020	4032	4035	3141
GC% ± SD (%)	54.59 ± 0.05	62.17 ± 0.07	65.49 ± 0.07	54.49 ± 0.06	57.98 ± 0.04	54.63 ± 0.05	53.73 ± 0.05	54.12 ± 0.05	53.72 ± 0.05	51.97 ± 0.05
O/E ratio ± SD	0.65 ± 0.09	0.84 ± 0.1	0.87 ± 0.3	0.63 ± 0.1	0.68 ± 0.1	0.71 ± 0.14	0.64 ± 0.08	0.68 ± 0.11	0.65 ± 0.08	0.67 ± 0.1
Chromosome 22 (49,691,432 bp)										
Total length of CpG islands	3,114,716	679,803	522,748	2,067,653	2,842,255	2,772,787	2,802,675	2,873,255	2,907,983	2,998,371
No. of islands detected	8215	1642	2186	1903	6875	4571	4571	4882	4882	6356
Island coverage (%)	6.27	1.36	1.05	4.16	5.71	5.34	5.64	5.60	5.85	6.17
Island length (bp)										
Average	379	414	239	1087	413	581	613	570	596	482
Minimum	201	200	8	500	200	201	198	201	202	201
Maximum	7996	7902	7774	8363	3339	4064	4076	4064	4076	7816
GC% ± SD (%)	55.11 ± 0.05	63.70 ± 0.08	70.23 ± 0.08	55.84 ± 0.07	55.12 ± 0.06	54.91 ± 0.05	54.50 ± 0.07	55.16 ± 0.05	54.46 ± 0.07	56.66 ± 0.06
O/E ratio ± SD	0.64 ± 0.07	0.84 ± 0.1	0.95 ± 0.3	0.62 ± 0.1	0.68 ± 0.1	0.66 ± 0.08	0.63 ± 0.05	0.66 ± 0.10	0.63 ± 0.05	0.66 ± 0.09

TABLE 3. COMPARISON OF CPG ISLAND DETECTION USING FOUR METHODS IN THE WHOLE HUMAN GENOME

Methods	*CpGcluster*	*CpGIS*	*CPSORL*	*CpGTLBO*
Genome length	2.86×10^9	2.86×10^9	2.86×10^9	2.86×10^9
No. of predicted islands	198,702	37,729	208,536	272,907
Coverage (%)	1.90	1.44	4.10	4.22
Island length average	273 ± 246	1090 ± 717	572 ± 469	443 ± 392
GC% ± SD	63.78 ± 7.50	60.64 ± 5.06	53.90 ± 5.25	53.85 ± 4.92
O/E ratio ± SD	0.855 ± 0.265	0.717 ± 0.082	0.649 ± 0.087	0.671 ± 0.099
TSSs	21,741	15,106	25,477	26,333
Promoter regions	29,156	13,196	54,356	32,319

TSS, transcription start site.

4. DISCUSSION

When using the sliding window approach, CpG island prediction quality is affected by the window size. Given a large window size, regions including several short and loose CpG islands can possibly be combined into a single larger region. This large window size provides a high *SP* value, but an increased false negative (*FN*) can reduce *SN* (Lai et al., 2008; Sujuan et al., 2008; Han and Zhao, 2009; Hackenberg et al., 2010; Mabrouk, 2013). The PSO-based methods apply the optimization algorithm in each window and use the GGF to fit the CpG islands. Although this method successfully enhances the CpG island detection, PSO must be implemented in many windows that do not include CpG islands, incurring a huge computational cost; thus, PSO-based methods significantly increase computational loading. In addition, since CpGcluster replaces the CpG island definition with strict *p* values, some detected CpG islands fail to meet the definition, resulting in omissions of regions, which do not comply with the CpG island criteria. Therefore, in this study, we propose CpGTLBO to combine the CpGcluster and TLBO method. The results indicated that CpGTLBO can efficiently detect the CpG islands with GGF criteria, and the examples of performance measurement and running times are illustrated in Figures 3 and 4, respectively.

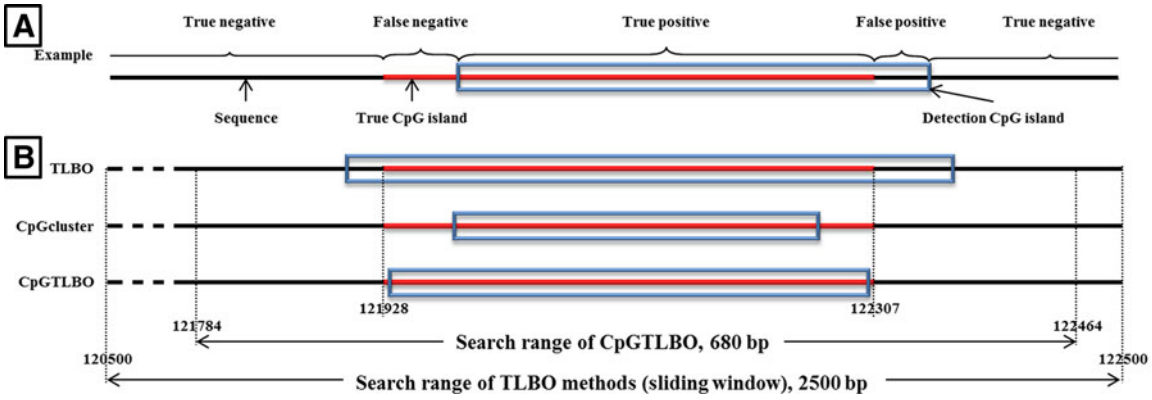

FIG. 3. Results of the three methods, showing the position of a true CpG island and the detection of CpG island positions of the TLBO, CpGcluster, and CpGTLBO methods. This figure shows the search range of the three methods in a true CpG island of NT_113955.2 contig; this island is located from 121,928 to 122,307 bp **(A)**. This figure is explained in the ''Performance measurement'' section of Supplementary Material. A true CpG island is included in the search region by CpGTLBO **(B)**. According to the GGF definition, it can accurately detect a complete true CpG island, though the detection results feature a few false positives. According to Chuang et al. (2011), they defined the window size as 2500 bp, which creates a huge number of possible solutions, making it difficult for TLBO methods to detect the CpG island position. CpGcluster detects the shortest length (region: 121,984–122,264, length: 280) to produce the highest specificity (1.0%) and lowest sensitivity (73.88%) for a sequence range between 120,500 and 122,500 bp. However, CpGTLBO extends the search region based on the relatively relaxed definition to detect the CpG islands; thus, the search range is significantly reduced to 680 bp.

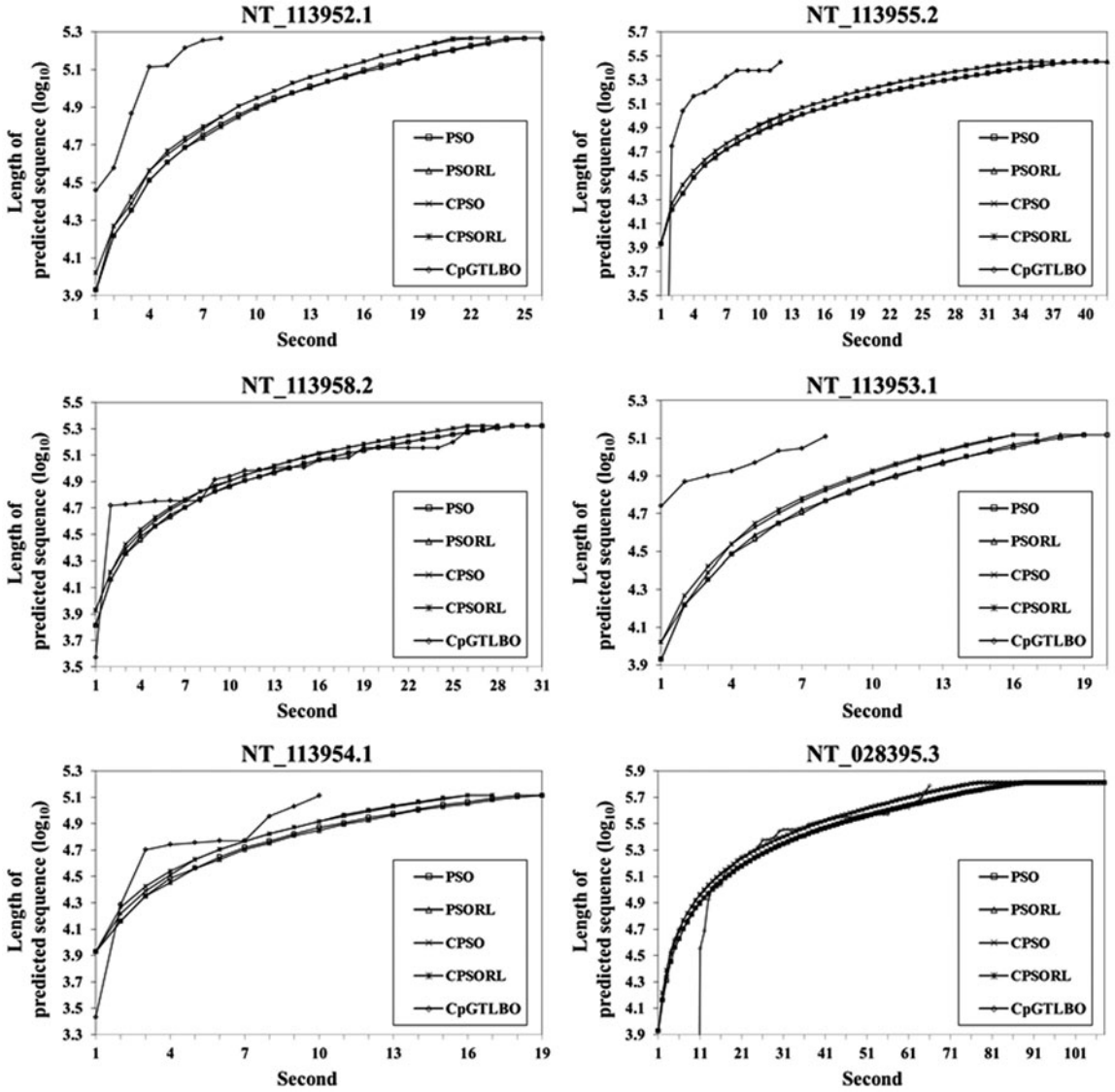

FIG. 4. Running times comparing the search efficiency among the five methods for the six contig sequences. Running times for PSO-based and CpGTLBO methods are used to assess search efficiency in the six contig sequences based on running time and the $\log_{10}$ value for the sequence's presently detected position on the x and y axes, respectively. This figure shows the running times of PSO-based methods and CpGTLBO in six contig sequences. All PSO-based methods have similar execution times and require more execution time than CpGTLBO. Although the time complexity of TLBO is greater than PSO-based methods, the first step of CpGTLBO can effectively reduce the number of unnecessary regions in DNA sequences; thus, the execution time for CpGTLBO is significantly less than that for PSO-based methods for the entire DNA sequence. Supplementary Tables S3 and S4, respectively, show the PSO-based and CpGTLBO methods with PSO and TLBO run times in six contigs and the entire human genome. The total length of the NT_113952.1 contig sequence is 184,355 bp. CpGTLBO detects the 21 CpG island candidates after the *p*-value assessment; thus, TLBO is only run 21 times, whereas the PSO-based methods need to run PSO 93 times. In the NT_113958.2 and NT_028395.3 contig sequences (long sequences), CpGTLBO requires relatively less time to be spent on preprocessing. Although the initial CpGTLBO scanning sequence may proceed more slowly than in other methods, sequence pre-treatment provides a strong time advantage for complete genome detection. PSO, particle swarm optimization.

Detection stability is an important performance criterion in optimization algorithms. CpG island detection results can be affected by different random seeds. Supplementary Figure S4 shows that PSO-based methods produced similarly different results for worst and best values for *ACC*, *SN*, *SP*, *PC*, and *CC*, indicating that some random seeds may result in failed searches, thus reducing detection accuracy. However, the CpGTLBO method produces high levels of detection accuracy, even with unfavorable seed

values (see the error bars near the left boxes that indicate 10th percentiles). CpGTLBO was found to produce greater stability than the other PSO-based methods.

The GC% (average) and O/E ratio (average) results using CPSORL are similar to those of PSO-based methods for CpG island detection, but smaller than CpGcluster values. CpGcluster is usually used to identify short CpG islands that may have high O/E ratio values and high GC% levels, but low sensitivity. CpGTLBO is based on the CpGcluster approach, but we use the GGF CpG island definition. Thus, CpGTLBO can significantly reduce the O/E ratio and GC%, thus substantially improving sensitivity. Supplementary Table S2 shows the entire human genome analysis, in which CpGTLBO outperforms CPSORL in terms of detecting higher *SN*, *SP*, *ACC*, *PC*, and *CC* values. These results indicate that CpGTLBO provides accurate CpG island identification. CpGTLBO inherits the advantages of CpGcluster in terms of CpG island detection with TSSs and promoters (Table 3). Given the overlap with conserved elements or promoter regions, the CpGcluster is co-localized more specifically to TSSs and many of the small CpG islands detected by CpGcluster may be functional (Han and Zhao, 2009).

Until now, the identification of CpG islands can fetch the disease analysis. Hence, for identifying CpG islands, a CpGTLBO-based clustering and optimization has been proposed in this article. CpGTLBO has several advantages over the CpGcluster and PSO-based method alone: (i) CpGTLBO running time and search stability can be significantly improved by pre-treatment with CpGcluster technology; (ii) the lower sensitivity of CpGcluster can be improved by the use of the GGF criteria; and (iii) CpGTLBO effectively improves the CpGcluster and the PSO-based methods, and these improvements greatly enhance the accuracy for CpG island detection. The results indicate that the CpGTLBO outperforms the other methods.

ACKNOWLEDGMENTS

This work was partly supported by the National Science Council in Taiwan (under Grant nos. MOST 105-2811-E-151-002, MOST 104-2221-E-214-035-MY2, MOST 103-2221-E-151-029-MY3, and MOST 104-2320-B-037-013-MY3).

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Chuang, L.Y., Huang, H.C., Lin, M.C., et al. 2011. Particle swarm optimization with reinforcement learning for the prediction of CpG islands in the human genome. *PLoS One* 6, e21036.

Deaton, A.M., and Bird, A. 2011. CpG islands and the regulation of transcription. *Genes Dev.* 25, 1010–1022.

Gardiner-Garden, M., and Frommer, M. 1987. CpG islands in vertebrate genomes. *J. Mol. Biol.* 196, 261–282.

Gudise, V.G., and Venayagamoorthy, G.K. 2003. Evolving digital circuits using particle swarm, 468–472. In *IEEE Proceedings of the International Joint Conference on Neural Networks*. Portland, OR, USA.

Hackenberg, M., Barturen, G., Carpena, P., et al. 2010. Prediction of CpG-island function: CpG clustering vs. sliding-window methods. *BMC Genomics* 11, 327.

Hackenberg, M., Previti, C., Luque-Escamilla, P.L., et al. 2006. CpGcluster: A distance-based algorithm for CpG-island detection. *BMC Bioinformatics* 7, 446.

Han, L., and Zhao, Z. 2009. CpG islands or CpG clusters: How to identify functional GC-rich regions in a genome? *BMC Bioinformatics* 10, 65.

Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., et al. 2010. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet.* 6, e1001134.

Keles, S., Van der Laan, M.J., Dudoit, S., et al. 2006. Multiple testing methods for ChIP-chip high density oligonucleotide array data. *J. Comput. Biol.* 13, 579–613.

Kuo, H.C., Lin, P.Y., Chung, T.C., et al. 2011. DBCAT: Database of CpG islands and analytical tools for identifying comprehensive methylation profiles in cancer cells. *J. Comput. Biol.* 18, 1013–1017.

Lai, H.M., Chiang, Y.Y., Hsu, C.C., et al. 2008. A recognition machine for CpG-islands based on Boltzmann model. *J. Med. Biol. Eng.* 28, 23–30.

Mabrouk, M.S. 2013. Extracting gene markers by the identification of functional CG rich genomic regions. *J. Med. Imag. Health Inf.* 3, 38–41.

Olson, S.A. 2002. EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. *Brief Bioinformatics* 3, 87–91.

Ponger, L., and Mouchiroud, D. 2002. CpGProD: Identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* 18, 631–633.

Rao, R., Savsani, V., and Vakharia, D. 2011. Teaching–learning-based optimization: A novel method for constrained mechanical design optimization problems. *Comp. Aided Des.* 43, 303–315.

Su, J., Zhang, Y., Lv, J., et al. 2010. CpG_MI: A novel approach for identifying functional CpG islands in mammalian genomes. *Nucleic Acids Res.* 38, e6.

Sujuan, Y., Asaithambi, A., and Liu, Y. 2008. CpGIF: An algorithm for the identification of CpG islands. *Bioinformation* 2, 335.

Takai, D., and Jones, P.A. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad Sci. U. S. A.* 99, 3740–3745.

Address correspondence to:
Dr. Li-Yeh Chuang
Department of Chemical Engineering
Institute of Biotechnology and Chemical Engineering
I-Shou University
No, 1, Sec. 1, Syuecheng Road
Dashu Township, Kaohsiung 84001
Taiwan

E-mail: chuang@isu.edu.tw

Dr. Yu-Da Lin
Department of Electronic Engineering
National Kaohsiung University of Applied Sciences
No. 415, Jiangong Road
Sanmin District, Kaohsiung 80778
Taiwan

E-mail: e0955767257@yahoo.com.tw