A Hybrid Genetic Algorithm for Prediction of CpG Islands in Human Genome

Li-Yeh Chuang [a], Yu-Jung Chen [c], Cheng-Hong Yang [b,c*]
[a] Dep. of Chemical Engineering, I-Shou University, Kaohsiung, Taiwan
[b] Dep. of Network Systems, Toko University, Chiayi, Taiwan.
[c] Dep. of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences, Taiwan
[*] Correspondence authors: Cheng-Hong Yang, chyang@cc.kuas.edu.tw

Abstract

CpG islands clusters of CpG dinucleotides in GC-rich fragments are considered as useful gene makers and plays important roles for genes in human. In this study, we developed a hybrid method called GA-CGI (GA-CpG Islands) for prediction of CpG islands in human genome, called GA-CGI (GA-CpG Island).The genetic algorithm was used to identify the significance of CpG islands, and then subsequently combined with the neighbor CpG islands by reinforcement learning. The developed method was compared with three public tools for assessment of prediction performance and statistically analysis. The experimental results reveal that our method not only can identify more CpG islands but also show a higher prediction performance and precise in target sequences than those of previous methods.

Keywords: CpG islands, Genetic Algorithms, reinforcement learning.

1、Introduction

It is a decade since the CpG islands was first reported [1]. CpG islands are small fragments of DNA in the GC-rich sequences, and those small fragments usually do not occur methylation. Most of the genomes' CpG dinucleotides, roughly 60% of human and 50% of all mammalian genes, are methylated [2]. In mammalian, the known CpG islands aren't randomly dispersed in genome, but often associated with more than three-quarters of all known transcription start sites and with 88% of the active promoters identified in primary fibroblasts [3]. They play important regulatory functions and can be used as powerful gene makers. For example, they bear a fundamental role for promoter prediction [4], and use as candidate regions for aberrant DNA methylation. It has contributed significant to our understanding of the epigenetic causes of cancer [5]. CpG islands are commonly defined as regions of DNA sequence longer than 200 bp, G + C content $\geq$ 50%, and CpGs obs/exp (o/e) ratio $\geq$ 0.6. This criterion was originally established by Gardiner-Garden and Frommer (GGF's) [6]. Presently, Takai and Jones (TJ's) proposed more stringent criteria, with DNA sequence longer than 500 bp, G + C content $\geq$ 55%, and a CpGs obs/exp (o/e) ratio $\geq$ 0.65 respectively, which would be more effective in excluding Alu repeats (which are highly repetitive short interspersed elements) [7].

In our proposed method, we have presented a hybrid method for prediction of CpG islands using Genetic Algorithm (GA) with reinforcement learning in human genome, called GA-CGI. The goal of our method is that using stochastic searches for CpG island subsets, then assess the relative importance of genes by examining the fitness function of the genes in the near-optimal set. The GA-CGI provides a higher prediction performance and combination ability. Furthermore, the obtained results reveal that the prediction of CpG islands were more significant via the statistics analysis.

2、Methods

2.1 Datasets

The entire DNA of human chromosomes was downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov).

2.2 A hybrid GA for prediction of CpG islands

Since the Genetic Algorithms was proposed in the 70s, researchers have investigated various methods for enhancing the algorithm. GA is stochastic search algorithms modeled on the process of natural selection underlying biological evolution. They can be applied to many search, optimization, and machine learning problems [8]. The basic concept behind a GA is designed to simulate evolutionary processes in natural systems, specifically those that follow the principle of survival of the fittest first laid down by Charles Darwin. As such, they represent an intelligent exploitation of a random search within a defined search space to solve a problem. A GA proceeds in an iterative manner by generating new populations of strings from old ones. Every string is the encoded binary, real, etc., version of a candidate solution. An evaluation function associates a fitness measure with every string, indicating its fitness for the problem. GA have been successfully applied to a variety of problems, such as scheduling problems, machine learning problems, multiple objective problems, feature selection problems, data mining problems, and traveling salesman problems [9].

GA had five major steps is included that the encoding of the chromosome and its initialization; the fitness evaluation; the selection operation; the crossover and mutation operation; and a replacement process.

A. Encoding chromosome and Initialization

To large-scale computational analysis of CpG islands in human DNA sequences. It was divided into some windows Pv to analysis the CpG islands in bulk DNA. (Pv is indicated that the individuals window size), Pv = (F_s, F_l) are randomly generated as initial population without duplicate. The window size was set at three thousand. Fs was randomly generated start sites of fragments within 2500. F_l was randomly generated length of CpG islands.

B. Fitness evaluation

Estimating at CpG islands in GA was depended on the fitness function; hence design a fitness function was very important. The fitness function of a chromosome was evaluated under the TJ’s criteria for CpG islands.

The CpG islands had three main factors included length, GC% and $\mathrm{CpG}\%_{e}{}_{\mathrm{ratio}}$. Aim at the length of CpG islands, which it’s from 200 bp to 2 kbp, even longer. If we add the length of islands on it fitness directly, it will result in an unfair fitness. So, reducing the length value was necessary. The CpG islands length may reach ~2 kb. To solve this challenge, we adopted the log function as each length of predicting CpG islands. It will be reduced the value of islands length and lead to the values within a small range. (i.e., the value of islands was from 2.301 to 3.301 after the function of length) The function of length value presented on Eq. 1.

The characteristic of CpG islands were highly GC-rich, so the CpG islands relation between GC%, supposing that the GC density higher as this CpG islands maybe more significant. To predict precisely CpG islands, determined the density of small regions directly. The GC% and $\mathrm{CpG}\%_{e}{}_{\mathrm{ratio}}$ function described by Eq. 2.and Eq. 3, respectively. The fitness value is provided by the following Eq. 4.

$$\mathrm{CpG}_{\mathrm{length\_value}} = \log_{10}(\mathrm{CpG\_length}) \quad (1)$$

$$\mathrm{GC\%} = \frac{\text{num of GC}}{\text{num of A\_T\_C\_G}} * 100\% \quad (2)$$

$$\mathrm{CpG}\%_{e}{}_{\mathrm{ratio}} = \frac{\mathrm{CpG\_observed}}{\mathrm{CpG\_expected}} \quad (3)$$

$$\begin{aligned}\mathrm{Fitness(Pv)} &= \mathrm{GC\%(Pv)} + \mathrm{CpG}\%_{e}{}_{\mathrm{ratio}}\mathrm{(Pv)} \\ &+ \mathrm{CpG}_{\mathrm{length\_value}}\mathrm{(Pv)}\end{aligned} \quad (4)$$

C. Selection operation with Tournament

The GAs, the selection operation viability of individuals, contained tournament selection and roulette wheel selection are two current approaches. Since the selected CpG islands, the selection operation should ensure that the CpG islands had higher fitness. Therefore, our design adopts the rank-based tournament selection scheme. The chromosomes in the population were stored by ranking, so we selected the best two chromosomes from population to performance of crossover operation and mutation operation.

D. Crossover and Mutation operation

We used the standard crossover and mutation operation. Using the tournament selected to pick up the two parents and then producing two offsprings by exchanging information. As shown in Fig. 1. However, suppose that the length of CpG islands was over the window size then a mechanism be adjusted the length. The F_s was randomly created within 2500bp, which is avoided to against the length criteria. We adjusted the length to final sites (see Fig.2). Due to the fitness values same for a given number of iterations. It's possibly trapped into local optimal. The primary purpose of mutation operation was to randomly introduce new genetic and contributive to the fitness escape from a local optimum.

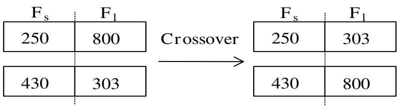

Figure 1. Description of the crossover operation to generate offsprings

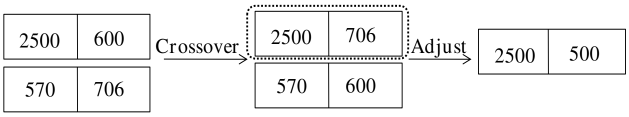

Figure 2. The dashed line indicates the length over the window size. So that mechanism of Ad_F_l was adjusted. (Ad_F_l = window size–F_s.)

E. Replacement operation

For the steady state GA, the crossover generates offspring of two parents, and the mutation operation slightly perturbs the offspring. If the offspring is superior to both parents, it replaces the similar parent; if it is in between the two parents, it replaces the inferior parent; otherwise, the most inferior chromosome in the population is replaced.

F. Reinforcement learning

Reinforcement learning is an approach to intelligence. Reinforcement learning combines the fields of dynamic programming and supervised learning to yield powerful machine-learning systems. The learner is not told which actions to take, as in most forms of machine learning, but instead must discover which actions yield the most reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards. These two characteristics, trial-and-error search and delayed reward, are the two most important distinguishing features of reinforcement learning. Every reinforcement learning system learns a mapping from situations to actions by trial-and-error interactions with a dynamic environment [10].

Although prediction CpG islands were conformed to the criteria, but some CpG islands were too short than known CpG islands which its may greatly influence the sensitivity. To overcome this problem, some tools based on sliding window, however under this technique, assume target sequences so longer that result in the long of computing time. In our proposed method, the reinforcement learning severed as extend operator. The sign of extend, is called terminal state indicates whether the terminal state is a goal state (i.e., a reward, next extend length of CpG islands continue used this reward length) or a state can't extend that should be avoided (i.e., a penalty, don't use this length).

2.3 Performance evaluation

We used the Sensitivity (SN), Specificity (SP), Accuracy (ACC), Correlation coefficient (CC) and Performance coefficient (PC) to measure the prediction performance.

In this study, we calculated the five prediction performance as follows:

$$\mathrm{SN} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \quad (5)$$

$$\mathrm{SP} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) \quad (6)$$

$$\mathrm{ACC} = \mathrm{TP} + \mathrm{TN}/(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}) \quad (7)$$

$$\mathrm{PC} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN} + \mathrm{FP}) \quad (8)$$

$$\mathrm{CC} = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN}) * (\mathrm{TP} + \mathrm{FP}) * (\mathrm{TN} + \mathrm{FN}) * (\mathrm{TN} + \mathrm{FP})}} \quad (9)$$

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

3. Results and Discussion

3.1 Comparison with other tools and parameter setting

Three public tools for CpG island prediction, CpGplot/CpGReport [11], CpGProD [12] and CpGcluster [13] were compared with GA-CGI in prediction performance. In parameter setting, the population size was 30; the number of generations was 100; the crossover rate was set at 1.0; the mutation rate was set at 0.1, and the window size was set at 3000.

3.2 Statistical analysis in human genomes

Basic statistical analysis for all contigs of human chromosome 21 and 22 (Chr.21, Chr.22), we used GGF's criteria to assess at three methods, is shown in Table 1. In minimum length, CpGcluster do not meet the criteria of length in Chr.21 and Chr.22. CpGplot also do not conform the criteria of length in Chr.21. In term of the number of islands predicted, although CpGcluster showed slightly higher than GA-CGI in Chr.21 (more than 190 CpG islands). But their minimum and average islands length obtained were shorter. Here, we considered that due to the island length was shorter, it leads more islands were predicted. Moreover, some of the short CpG islands might be spurious, but some others may play a true functional role [13]. However, short CpG-rich region fails to fulfill CpG islands criteria on its own, the same region may well fulfill the criteria after small and seemingly unrelated changes of a few neighboring nucleotides. Thus, the mapping of CpG islands is inherently unstable [14]. GA-CGI can overcome above challenge; since our proposed method adopted the reinforcement learning to combine with shorter islands and leads to island of minimum length also fulfill criteria.

In Chr.21 and Chr.22, the total lengths of CpG islands were longer than other tools, this demonstrates that our search space were so wide that obtain the many islands, principally because of the GA had crossover and mutation operation, it can change the information each other and increase to search space spreading over all target sequences. These evolutionary operators can't work it in other tools, hence the total lengths surely far higher than other tools.

3.3 Assessing the CpG island prediction performance

As in the gene-finding field, the accuracy of a prediction was measured by comparing the predicted island value (island or non-island) with the true island for each nucleotide along the target sequence, and then obtains the five prediction performance. TJ's criteria have today been widely adopted for the identifications of CpG islands in genomic sequences [15]. Table 2, we used TJ's criteria to assess the CpG islands prediction performance in other methods. Generally, the larger value of SN, the more false positives are reported, and the smaller the value of SP, it is a trade-off to balance SN and SP[16]. The results reveal that GA-CGI is more reliable than the other methods tested. CpGcluster shows superb SP values, but the SN, PC and CC values are surprisingly low. This could indicate that CpGcluster depends on a composition of different input sequences. If a sequence is a non-island sequence and the length of the probability of observing a CpG island in this test sequence and thus the p-value of each CpG island would be rather low [13]. CpGProD and CpGplot using the sliding window technique, its size may greatly influence the results and get instable prediction performance (SN, PC and CC values).

Table. 1 Comparison of different methods on number of CpG islands identified in Human genomes

	Chr 21				Chr 22			
	CpGcluster	CpGProD	CpGplot	GA-CGI	CpGcluster	CpGProD	CpGplot	GA-CGI
Genome Length (bp)	46,944,329	46,944,329	46,944,329	46,944,329	49,691,432	49,691,432	49,691,432	49,691,432
Number of islands predicted	2,703	1,091	973	2,513	2,186	1,903	1,642	3,902
Total length of CpG islands	639,161	1,072,192	347,334	1,262,449	522,748	2,067,653	679,803	2,251,454
Island length (bp)								
Average	236.46	982.76	356.97	502.36	239	1,087	414	577
Minimum	8	500	101	201	8	500	200	201
Maximum	3,028	6,732	3,047	6,126	7,774	8,363	7,902	5,340
Average island GC%	65.49±6.87	54.49±6.08	62.17±7.01	54.24±4.76	70.23±7.60	55.84±6.51	63.70±7.66	55.21±5.37
Average CpG o/e ratio	0.87±0.26	0.63±0.09	0.84±0.13	0.68±0.12	0.95±0.25	0.62±0.09	0.84±0.14	0.66±0.10

Legends: CpG o/e ratio repreasent that CpG_observer / CpG_expected

Table 2. Prediction performance of other methods.

Chr.	Methods / Performance	CpG-cluster	CpGProD	CpGplot	GA-CGI
21	SN (%)	44.33	78.21	23.55	79.81
	SP (%)	99.88	97.54	100.0	98.79
	ACC (%)	97.72	96.73	96.78	97.97
	PC (%)	42.97	50.06	23.54	62.80
	CC (%)	63.50	65.82	47.73	76.13
22	SN (%)	37.47	66.61	27.17	76.33
	SP (%)	99.73	99.32	100.0	99.01
	ACC (%)	97.84	98.33	97.79	98.33
	PC (%)	34.51	54.72	27.17	57.93
	CC (%)	54.35	70.02	51.54	72.56

Legends: Chr. indicates the chromosome.

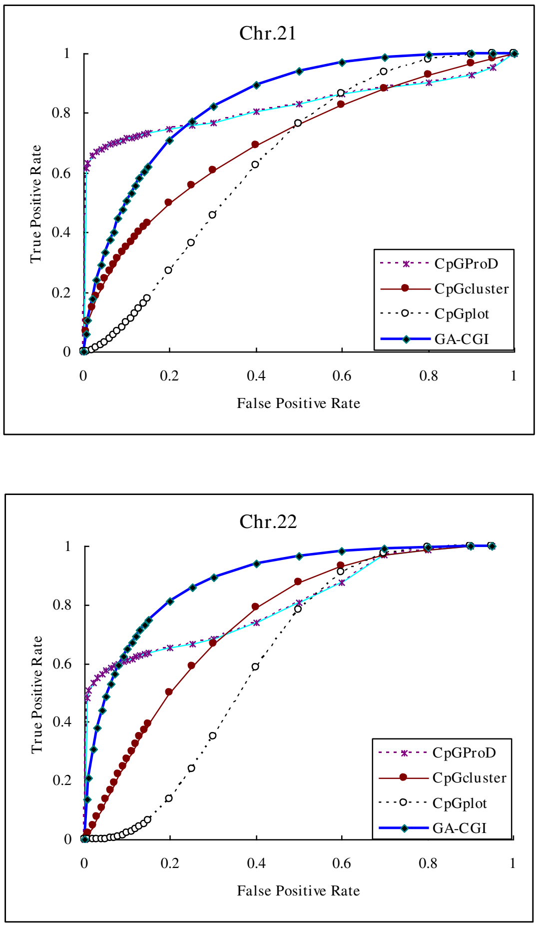

Figure 3 ROC curves comparing for other methods.

GA-CGI obtained more stable and higher prediction performance than other methods tested. In this paper, we also using receiver operating characteristic (ROC) curves to aim all methods were evaluated against these two evaluation data sets. These ROC curves interpret the score of any one CpG islands as it's predicted likelihood of being CpG islands. For all possible thresholds on the CpG islands score, they describe the tradeoff between the true positive rates (i.e., the percentage of CpG islands that are predicted, also called SN) and the false positive rate (i.e., the percentage of negatives that are erroneously classified as CpG islands, which is equal to 1- SP) and thereby assess how we particular CpG islands score predicts the methods in the literature and all ROC curves shown in Fig. 3. In ROC curves, we can obtain the superior in ROC area (i.e., GA-CGI can get higher SN and SP). Since, we used the RL operation not only merged and extended CpG islands but also increased SN.

4. Conclusion

We have presented a hybrid method called GA-CGI to predict CpG islands in human genome. The proposed method was compared to three other methods, and its superiority in terms of prediction performance has been established. GA-CGI's capacity for precisely locating and predicting significant CpG islands in target sequences has been tested under TJ's criteria for CpG islands. In the future, it could be worthwhile to develop other methods, such as gene rearrangement for chromosome encoding, and apply the proposed method to identify genome-wide methylation sites.

Acknowledgement

This work is partly supported by the National Science Council in Taiwan under grant NSC97- 2622- E-151-008-CC2, NSC96-2221-E-214-050- -MY3.

References

[1] Tykocinski, M. L., Max, E. E., "CG dinucleotide clusters in MHC genes and in 5' demethylated genes," Nucleic Acids Res, vol. 12, pp. 4385-96, May 25 1984.

[2] Ioshikhes, I. P., Zhang, M. Q., "Large-scale human promoter mapping using CpG islands," Nat Genet, vol. 26, pp. 61-3, Sep 2000.

[3] Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D. Ren, B. "A high-resolution map of active promoters in the human genome," Nature, vol. 436, pp. 876-80, Aug 11 2005.

[4] Bajic, V. B., Tan, S. L., Suzuki, Y., Sugano, S., "Promoter prediction analysis on the whole human genome," Nat Biotechnol, vol. 22, pp. 1467-73, Nov 2004.

[5] Ushijima, T., "Detection and interpretation of altered methylation patterns in cancer cells," Nat Rev Cancer, vol. 5, pp. 223-31, Mar 2005.

[6] Bird, A., "DNA methylation patterns and epigenetic memory," Genes Dev, vol. 16, pp. 6-21, Jan 1 2002.

[7] Takai, D., Jones, P. A., "Comprehensive analysis of CpG islands in human chromosomes 21 and 22," Proc Natl Acad Sci U S A, vol. 99, pp. 3740-5, Mar 19 2002.

[8] Goldberg, D. E., "Genetic Algorithms in Search, Optimization and Machine Learning," Addison-Wesley, Reading, MA, 1989.

[9] Pullan, W., "Adapting the Genetic Algorithm to the Traveling Salesman Problem," in IEEE Congress on Evolutionary Computation, 2003.

[10] Harmon, M. E., Harmon, S. S., "Reinforcement learning: a tutorial," 1996.

[11] Rice, P., Longden, I., Bleasby, A. "EMBOSS: the European Molecular Biology Open Software Suite," Trends Genet, vol. 16, pp. 276-7, Jun 2000.

[12] Ponger L., Mouchiroud, D. "CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences," Bioinformatics, vol. 18, pp. 631-3, Apr 2002.

[13] Hackenberg, M., Previti, C., Luque-Escamilla, P. L., Carpena, P., Martinez-Aroza, J., Oliver, J. L., "CpGcluster: a distance-based algorithm for CpG-island detection," BMC Bioinformatics, vol. 7, p. 446, 2006.

[14] Bock, C., Walter, J., Paulsen, M., Lengauer, T., "CpG island mapping by epigenome prediction," PLoS Comput Biol, vol. 3, p. e110, Jun 2007.

[15] Hutter, B., Paulsen, M., Helms, V., "Identifying CpG Islands by Different Computational Techniques," Omics, vol. 13, Jan 23 2009.

[16] Xie, X., Wu, S., Lam, K. M., Yan, H., "PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm," Bioinformatics, vol. 22, pp. 2722-8, Nov 15 2006.