Mining and Predicting CpG islands

Christopher Previti, Oscar Harari and Coral del Val

Abstract— **A DNA sequence can be described as a string composed of four symbols: A, T, C and G. Each symbol represents a chemically distinct nucleotide molecule. Combinations of two nucleotides are called dinucleotides and CpG islands represent regions of a DNA sequence, certain substrings, which are enriched in CpG dinucleotides (C followed by G). CpG islands represent a prominent and enigmatic feature of vertebrate genomes. They are associated with the promoters of more than 60% of all human genes and represent a critical target for transcriptional control. Methylation of these CpG islands leads to structural changes in the DNA that stops the expression of any associated gene (gene-silencing). The factors that provoke or impede methylation are currently all but unknown. In general, the maintenance of a particular pattern of methylated CpG dinucleotides represents a critical regulatory system during a host of normal developmental processes, but the erroneous methylation of CpG islands and the resulting gene-silencing can lead to the development of cancer.**

In this work, we present a novel unsupervised machine learning method that is capable of distinguishing biologically significant classes of CpG islands, including the separation of methylated and unmethylated CpG islands. This method represents an important novel approach that will aid in the computational prediction of methylation, which is commonly used in the pre-selection of worthwhile sequences for methylation experiments.

I. Introduction

A DNA sequence can be described as a string composed of four symbols: A, T, C and G. Each symbol represents a chemically distinct nucleotide molecule. Combinations of two nucleotides are called dinucleotides and CpG islands represent regions of a DNA sequence, certain substrings, which are enriched in CpG dinucleotides (i.e., a cytosine directly followed by a guanine). CpG dinucleotides are 4-fold underrepresented in the human genome compared to other dinucleotides since they are usually targeted for methylation and methylated CpG dinucleotides are prone to mutate irreparably [1, 2].

Christopher Previti is with the Department of Molecular Biophysics (B020) at the German Cancer Research Institute (DKFZ), D-69120 Heidelberg, Germany (email: cpreviti@gmail.com).

Coral del Val is with the Departamento de Ciencias de la Computación e Inteligencia Artificial Escuela Técnica Superior de Ingeniería Informática c/. Daniel Saucedo Aranda, s/n, 18071 Granada, Spain (phone: +34 958 240469; Fax: +34 958 243317; email: delval@decsai.ugr.es)

Oscar Harari is with the Departamento de Ciencias de la Computación e Inteligencia Artificial Escuela Técnica Superior de Ingeniería Informática c/. Daniel Saucedo Aranda, s/n, 18071 Granada, Spain (phone: +34 958 240468; Fax: +34 958 243317; email: oharari@decsai.ugr.es)

CpG islands [3] represent remarkable exceptions to this rule. Their CpG dinucleotide frequency is approximately the same as expect if all combinations of dinucleotides [4] were equally represented in the human genome and their CpG dinucleotides are often not methylated.

DNA methylation is a frequent DNA modification of vertebrate genomes [5] that is both reversible and heritable, but doesn't actually alter the sequence of nucleotides. CpG islands are often associated with the regulatory region of a gene (promoter) (Figure 1). The methylation of such a CpG island impedes the accessibility of the transcription machinery to the promoter [2, 6, 7] and stops the gene from being expressed.

The maintenance of a particular pattern of methylated CpG dinucleotides represents a critical regulatory system during a host of normal developmental processes such as cell differentiation [8], imprinting [9], X-chromosome inactivation [10] and the silencing of repetitive genomic elements [11], but also coincides frequently with cancer development and progression [12-14].

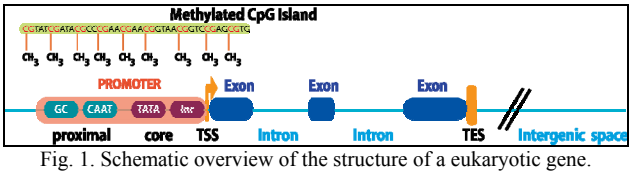

Fig. 1. Schematic overview of the structure of a eukaryotic gene. The promoter region represents the main transcriptional control center and can be separated into proximal and core promoters. CpG islands overlap with this region in about 60% of all human genes and often even extend beyond the transcriptional start site (TSS). The exons contain the protein coding sequence and the end of a gene is defined by the transcription end site (TES).

Determining the methylation status of individual CpG dinucleotides can only be done experimentally and is an arduous and time-consuming task given the enormous size of the genome and the absolute number of CpG dinucleotides that would have to be analyzed. In order to accelerate this process *in silico* predicted CpG islands sequences [15, 16] are often selected as target sites for methylation analysis.

Here, we describe an unsupervised learning strategy for the detection of distinct, biologically significant classes of CpG islands. More specifically, our method is able to separate methylated from unmethylated CpG islands.

Current methylation prediction algorithms are based exclusively on supervised learning methods such as support vector machines (SVM) [15, 17, 18]. In contrast, our approach represents a completely novel strategy to the classification of CpG islands which is not overtrained due to the fact that it employs an unsupervised learning method.

1-4244-1210-2/07/$25.00 ©2007 IEEE.

II. Main Results

The purpose of this method is to identify all possible substructures, here termed clusters (i.e., groups of CpG islands sharing common biological features) that classify functional CpG islands and, in particular, define the clusters specific to CpG islands that are differentially methylated.

The common attributes of these clusters can ultimately clarify the key sequence features of CpG islands that protect certain ones from methylation while leaving others constitutively methylated and therefore transcriptionally inactive.

Since only a limited number of CpG islands with a well-defined methylation status are available (<600 CpG islands) we applied an unsupervised machine learning method where pre-existing classes are not required. This approach allows the mining of the CpG islands predicted over the entire genome (>90.000 islands), avoiding the possible biases of the limited dataset that often lead to overtraining.

These three main steps were taken in order to delineate significant clusters: *A. Database conformation; B. Cluster learning; C. Evaluation on independent classes.*

A. Database conformation

1) Instance selection

The method by which the CpG islands were predicted is described in Figure 2. The *CpGcluster* algorithm [19] was applied to the entire human genome (NCBI version 17 [20]) with a *p-value* threshold of 10^{-5}. *CpGcluster* employs a single parameter specifying the maximum permissible distance between CpG dinucleotides (d_{max}) and a statistical test that approximates the probability (*p-value*) of the same number of CpG dinucleotides appearing by chance in a random DNA sequence. Only DNA sequences with a *p-value* below a given threshold are sufficiently enriched in CpG dinucleotides to be classified as CpG islands.

This yielded a total of 197.727 CpG islands, of which 105.581 islands could be co-localized with a well-known gene from the *Refseq*-database [21] using the UCSC Table Browser data retrieval tool [22]. For a CpG island to be assigned to a specific gene, it had to overlap with the region between 2000 bp upstream of the TSS and the TES. The rest of the CpG islands were removed from the database to avoid the addition of noise to the analysis. The remaining CpG islands contained promoter-overlapping as well as non-promoter-overlapping CpG islands at a ratio of about 1 to 2.

2) Attribute selection

Multiple independent attributes were considered in the analysis of the CpG islands. Based on our previous work on the characteristics of CpG islands [19] we focused on the following 5 distinct attributes in the analysis of the CpG island database:

(i) The parameter length represents the normalized length of the CpG islands. This parameter was included in the database because promoter-overlapping CpG islands, are said to be on average longer than those located outside of the promoter region and are more likely to be functional as well as unmethylated [23]. Since both the minimum and maximum permissible length of a functional CpG island are unknown, we decided to focus on CpG islands with an "intermediate" length by removing extremely short and long islands from the dataset This was done by selecting the CpG islands whose length was within the percentiles 5 to 95 (between 70 and 868 bp in length). Of the original 105.581 islands this reduced the dataset by approx 13% to 91.687 islands.

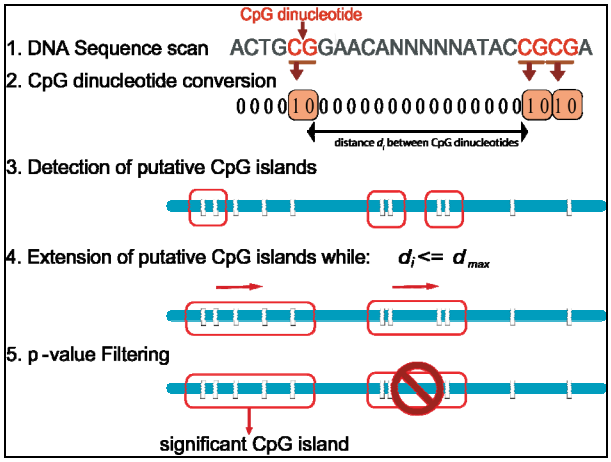

Fig. 2. Schematic overview of CpG island prediction.
1. *CpGcluster* scans the DNA for CpG-dinucleotides and 2. records the positions occupied by the Cytosine ('C'): x_1, x_2... x_N, N being the total number of CpG dinucleotides in the sequence; 3. The distance separating two neighboring CpG dinucleotides is defined as: $d_i = x_{i+1} - x_i - 1$, so that the minimal distance between two neighboring CpG dinucleotides (i.e. CGCG) is equal to 1; 4. The first pair of CpG dinucleotides whose distance falls under the chromosome-specific threshold d_{max} marks the start of a potential CpG island and is extended until d_i exceeds d_{max}. This step is iterated until all potential CpG islands are detected; 5. The *p-value* filtering is performed leaving only those CpG islands with a sufficiently high number of CpG dinucleotides.

(ii) Percentage of CpG dinucleotides and (iii) average distance (normalized) between the CpG dinucleotides. Both are good measures for capturing the overall CpG dinucleotide-enrichment of a given CpG island and were therefore included in the analysis.

(iv) Percentage of repetitive genomic elements. Since methylation of CpG dinucleotides is utilized to inactivate transcriptionally active repetitive elements such as *Alu*-repeats [11], it is reasonable to include the degree of overlap with repetitive elements as a parameter. These data were obtained using the UCSC Table Browser data retrieval tool [22].

(v) Percentage of conserved *Phastcon* elements. *Phastcons* are DNA sequences that are highly conserved in multiple vertebrate genomes [24]. Though their exact function still is unknown, CpG islands that overlap these elements should be more likely to be of functional due to their conservation across species. This measure is also based on external data obtained via the UCSC Table Browser data retrieval tool [22].

Each one of these distinct parameters is biologically significant and has implications on the methylation status as well as promoter co-localization of a CpG island.

B. Cluster Learning

We clustered the CpG islands using the Fuzzy C-means method (FCM) [25, 26] which builds models for each cluster by calculating their centroids. These models represent the prototypes, where the membership values for each CpG island is calculated by its similarity to the centroids $\overline{V_i}$:

$$\mu_{i,k} = \left[\sum_{j=1}^{c} \left(\left\| x_k - \overline{V}_i \right\|_A \Big/ \left\| x_k - \overline{V}_j \right\|_A \right)^{2/(m-1)} \right]^{-1} \qquad (1)$$

where $x_k = \{x_1, ..., x_5\}$ corresponds to the features that represent each *k* CpG island; *m* is the degree of fuzzification which is initialized as 2; $\| \|_A$ represents the Euclidean norm; and *i* indexes the prototypes.

We apply the Xie-Beni validity index [25] to estimate the optimal number of *c* clusters from the database:

$$XB(U,V) = \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{i,k}^2 \left\| x_k - \overline{V}_i \right\|^2 \Big/ n \left(\min_{i \neq j} \left\{ \left\| \overline{V}_i - \overline{V}_j \right\|^2 \right\} \right) \qquad (2)$$

This index is related historically to the FCM method, and its rationale for validating fuzzy clusters is geometric; good clusters should minimize this index (through different number of clusters *c*) by having compact representations (and therefore small numerators) and wide separators (and therefore large denominators) [25].

Fuzzy C-means Clustering algorithm (FCM) [25, 26]:

(i) Initialize $L_0 = \{\overline{V_1}, .., \overline{V_c}\}$;

(ii) while ($s<S$ and $\left\| L_s - L_{s-1} \right\| > \varepsilon$) where *S* is the maximum number of iterations. (iii) Calculate the membership of U_s in L_{s-1} as in equation (1) (iv) update L_{s-1} to L_s with U_s and $\overline{V_i} = \sum_{k=1}^{n} \mu_{ik} x_k \Big/ \sum_{k=1}^{n} \mu_{ik}$ (v) iterate

C. Evaluation of independent classes

This proposed unsupervised method does not require the specification of output classes. Consequently, the learnt clusters can be used to independently explain external classes as a process often termed labeling [27]. In order to find its classes of equivalence the method applies the hypergeometric distribution that gives the probability of intersection (PI) [28] as:

$$PI(V_i, V_j) = 1 - \sum_{q=0}^{p} \binom{h}{q} \binom{g-h}{n-q} \Big/ \binom{g}{h} \qquad (3)$$

where V_i is an alpha-cut of an internal cluster, of size h; V_j is the external class, of size *n*; *p* is the number of islands of the intersection; and *g* is the total number of candidates, such that the lower the value of *p* the better the size of the cluster association. PI is distinguished from other metrics, such as the Jaccard coefficient [29], in being an adaptive measure that is sensitive to small sets of examples, while retaining specificity with large datasets.

By applying the Xie-Beni validity index (equation 2) we estimated the optimal number of clusters at three and FCM clustering (equation 1) [25] was then used to partition the data.

The quality of the clustering of the CpG islands was evaluated using methylation and promoter co-localization classes. For this purpose experimental methylation data was acquired from two sources: the Human Epigenome Project (HEP), which currently contains information on chromosomes 22, 20 and 6 [20] and a methylation study of chromosome 21 [16].

These data indicate the average degree of methylation of individual CpG dinucleotides. These CpG dinucleotides were termed "*informative CpG dinucleotides*". We calculated the average degree of methylation of a CpG-island by averaging the degree of methylation of the individual, *informative CpG dinucleotides* over the total number of *informative CpG dinucleotides* in the CpG island. This yielded a set of 594 CpG islands with at least 70% of their CpG dinucleotides being *informative*, meaning that at least 70% of their CpG dinucleotides had information about them in the experimental methylation data. 377 CpG islands with an average degree of methylation of less than 60% were defined as being less methylated (*low*), the remaining 217 were classified as being highly methylated (*high*).

CpG islands that overlapped with the promoter, defined as the region 2000 bp upstream of the TSS to the end of the first exon, were termed *promoter-CpG islands* the rest as *nonpromoter-CpG islands*. Based on this definition the dataset contained 29184 *promoter-CpG islands* and 62503 *nonpromoter-CpG islands*. The clusters were evaluated based on their coincidence with the classes' *low*/*high* as well as *promoter-CpG islands*/*nonpromoter-CpG islands*.

The subsets of CpG islands equivalent to each cluster (*T1/T2/T3-CpG islands*) showed significant differences in their biological characteristics (Table I). The *T1-CpG islands* were enriched with repetitive elements, while at the same time containing the least number of CpG dinucleotides, the lowest degree of overlap with conserved *Phastcon* elements, the highest average distance between CpG dinucleotides and the shortest average length.

The main distinguishing characteristics between the *T2-* and *T3-CpG islands* were the average length, which was higher for the *T3-CpG islands* than for any other cluster and the amount of overlap with *Phastcon* elements which was highest for the *T2-CpG islands*.

Table I lists the absolute number of CpG islands per cluster as well as the averages and the standard deviations (SD) of the 5 parameters used to characterize the islands in the dataset for each cluster *T1-T3*.

As shown in figure 3 [30], the CpG islands in subsets *T1* through *T3* are differentially distributed with regard to the methylation classes *low* and *high.* About 75.61% of the *T3-CpG islands* belong to the *low* methylation class while approximately 66.47% of the *T2-CpG islands* are part of the

high methylation class. The *T1-CpG islands* on the other hand don't contain significant numbers of either class. This was to be expected, since only 19 (out of 594) of the CpG islands in the methylation dataset overlap with repetitive elements, while the *T1-CpG islands* contain almost exclusively CpG islands that overlap with repetitive elements.

TABLE I
AVERAGE VALUES FOR EACH BIOLOGICAL PARAMETER FOR ALL THREE CGI SUBSETS *T1-T3*

Cluster (number of CpG islands)	Length [bp] ± SD	CpG dinucleotide density ± SD	Mean distance between CpG dinucleotides [bp] ± SD	Overlap with repetitive elements [%] ± SD	Overlap with *Phastcon* elements [%] ± SD
T1 (38488)	**225.2 ± 104.2**	**0.074 ± 0.02**	**14.21 ± 3.59**	**93.36 ± 16.77**	1.08 ± 8.13
T2 (18509)	248.1 ± 153.5	0.089 ± 0.025	11.89 ± 3.64	2.64 ± 10.18	**43.87 ± 32.39**
T3 (34637)	**292.2 ± 193.3**	0.089 ± 0.026	11.84 ± 3.87	2.94 ± 9.84	14.6 ± 24.19

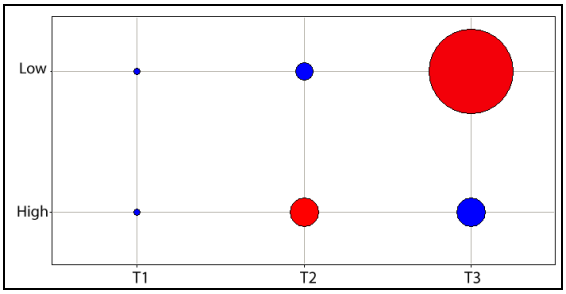

Fig. 3. Coincidence with methylation classes.
The *T2*- and *T3-CpG islands* are differentially distributed between the two methylation classes. The size and intenseness of the coloring are indicators for the number of elements and significance of the cluster, respectively.

The *PI* with regard to the methylation classes was computed using equation (3) and lends support to the significance of the clusters *T2* and *T3* with regard to the methylation classes *low* and *high*. Table II shows that the clusters *T2* and *T3* had the lowest *PI* for the methylation classes *High* and *Low* respectively.

In experimenting with various alpha-cuts, we found that 92% of the CpG islands formed a core set that was recovered while using an ample range of alpha-cuts (0.35-0.7), proving that our method is highly stable with regard to this user provided parameter and allowing us to set the alpha value at 0.40 (53 CpG islands were not recovered at this level).Though our method is highly stable, the 8% difference in the islands recovered while modifying the alpha-cut justifies the use of fuzzy clustering methods instead of crisp methods since the data do contain a certain degree of ambiguity that would cause a high degree of error in methods such as hierarchical or k-means clustering.

TABLE II
PI FOR CLUSTERS *T1-T3* FOR THE METHYLATION CLASSES *LOW* AND *HIGH*.

	PI	
Cluster	*Low*	*High*
T1	0.436	0.386
T2	1	<1.0E-20
T3	<1.0E-18	1

Figure 4 shows that the CpG island subsets *T1-T3* are also differentially distributed with regard to their promoter co-localization.

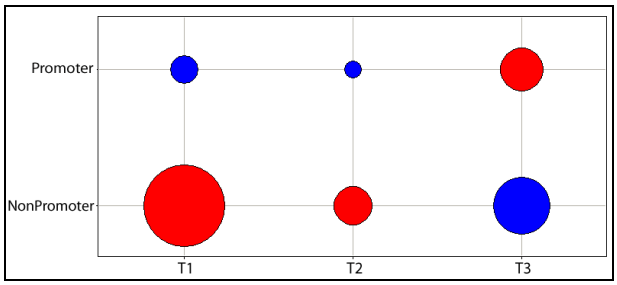

Fig. 4. Coincidence with promoter classes.
Clusters *T1* and *T2* are most representative for the class of *nonpromoter-CpG islands*, while *T3* defines a very significant subset of *promoter-CpG islands* but also a less significant, larger number of *nonpromoter-CpG islands*.

The *PI* for the promoter classes supports the significance of clusters *T1* and *T2* with regard to the *nonpromoter-CpG islands*. Despite the large number of *nonpromoter-CpG islands* in the *T3*-subset the *PI* indicates that *T3* describes the *promoter-CpG islands* with a higher level of significance than the *nonpromoter-CpG islands*.

TABLE III
PI FOR CLUSTERS T1-T3 WITH REGARD TO THE PROMOTER CO-LOCALIZATION CLASSES.

	PI	
Cluster	Non-promoter	Promoter
T1	3.12E-11	1
T2	5.72E-11	1
T3	1	< 1.0 E-9

A comparison of clusters T1/T2 versus cluster T3 showed that the principal difference between the two types of CpG islands was their average length, with the *promoter-CpG islands* being, on average, about 30% longer than the CpG islands classified as *nonpromoter-CpG islands*.

In order to evaluate our method's capacity for distinguishing between methylated and unmethylated sequences we compared it to the two most recent methylation prediction algorithms.

Both *MethCGI* [15] and a method implemented by *Bock et al.* [17] were SVM-based and trained on datasets of DNA