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].

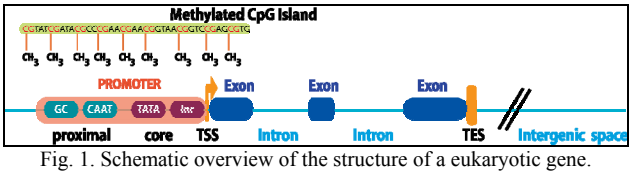

Fig. 1. Schematic overview of the structure of a eukaryotic gene. The promoter region represents the main transcriptional control center and can be separated into proximal and core promoters. CpG islands overlap with this region in about 60% of all human genes and often even extend beyond the transcriptional 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.

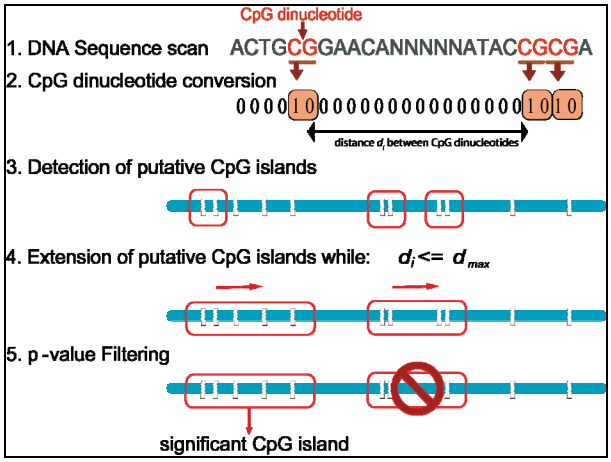

Fig. 2. Schematic overview of CpG island prediction.
1. *CpGcluster* scans the DNA for CpG-dinucleotides and 2. records the positions occupied by the Cytosine ('C'): x_1, x_2... x_N, N being the total number of CpG dinucleotides in the sequence; 3. The distance separating two neighboring CpG dinucleotides 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

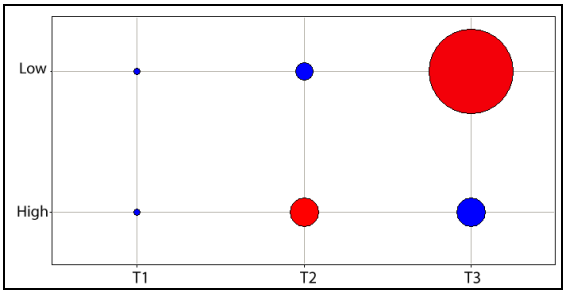

Fig. 3. Coincidence with methylation classes.
The *T2*- and *T3-CpG islands* are differentially distributed between the two methylation classes. The size and intenseness of the coloring are indicators for the number of elements and significance of the cluster, respectively.

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.

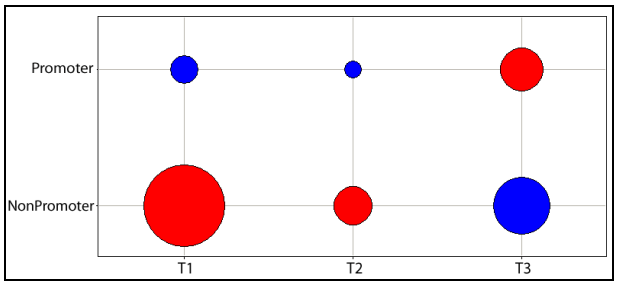

Fig. 4. Coincidence with promoter classes.
Clusters *T1* and *T2* are most representative for the class of *nonpromoter-CpG islands*, while *T3* defines a very significant subset of *promoter-CpG islands* but also a less significant, larger number of *nonpromoter-CpG islands*.

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

sequences whose methylation status was determined experimentally. Our method could only be compared indirectly with these two programs, since neither was available for large scale use at this time. We therefore had to rely on published accuracy values for the HEP-data [20] from chromosome 6. Both the overall accuracy (*Acc*) (equation 4), representing the percentage of CpG islands whose methylation status was predicted correctly, and the Correlation Coefficient (*CC*) (equation 5), a single scalar value that summarizes both sensitivity as well as specificity [31], were computed to this end.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}} \tag{5}$$

TP (true positives) represents the number of times a sequence was *correctly* predicted as being *methylated*, *TN (true negatives)* represents the number of times a sequence was *correctly* predicted to be *unmethylated*, *FP (false positives)* was the number of times a sequence was *incorrectly* predicted to be methylated and *FN (false negatives)* was the number of times a sequence was *incorrectly* predicted to be *unmethylated*.

The following accuracy data was taken from Fang et al. [15] and represents a comparison of these accuracy values for the algorithms *MethCGI* and the method implemented by *Bock et al.* [17]. The accuracy of our algorithm (*FCM*) was calculated over our entire methylation dataset. Due to the differential enrichment of methylated and unmethylated CpG islands in clusters T2 and T3, respectively, methylated CpG islands were counted as *TP* when they were a member of cluster T2 and unmethylated CpG islands were counted as *TN* when they were a member of cluster T3. All methylated CpG islands that were part of cluster T3 were counted as *FP* and all unmethylated CpG islands that were part of cluster T2 were counted as *FN*.

Table IV gives an overview of the overall accuracy values *Acc* and *CC*:

TABLE IV
ACCURACY COMPARISON

Program	*Acc (%)*	*CC*
MethCGI [15]	81.48	0.42
Bock et al. [17]	74.76	0.15
Our approach	73.68	0.39

Though our method has the lowest degree of accuracy it shows the second highest *CC,* indicating that the degree of false prediction is low. The CC can be viewed as a combination of sensitivity and specificity. Therefore, a low degree of sensitivity and a high CC indicate that the lack of sensitivity is made up for by a high degree of specificity. This high degree of specificity is due to our method's lack of overtraining and provides further proof, that fuzzy methods are highly effective in avoiding the possible biases of a limited dataset.

III. CONCLUSIONS

In this work, we present an unsupervised machine learning method that is capable of separating methylated from unmethylated CpG islands and therefore represents a novel computational approach that will aid in the delineation of the methylation status of a DNA sequence, a task that is time-consuming and expensive to do experimentally.

The CpG islands that form the basis of these experiments were predicted in the human genome utilizing a novel algorithm called *CpGcluster* [19] that specifically captures the high local clustering of CpG dinucleotides in the CpG islands. Approximately 53% of all CpG islands predicted in the human genome overlap a gene contained in the *Refseq* database [21, 22] and a substantial fraction of them may still overlap putative, not yet confirmed genes.

We analyzed the biological properties of the three clusters detected in the CpG island dataset and demonstrated that they differed with regard to specific features. The viability of our approach was evaluated based on each cluster's probability of intersection with independent classes of CpG islands whose methylation status was determined experimentally.

Cluster *T1* contained most of the CpG islands that coincided with a repetitive element and therefore demonstrated the highest average overlap with these elements. It is well known, that methylation acts as defense mechanism against the transcription of repetitive elements such as *Alu* transposons [11] and repetitive elements are therefore usually methylated. However the cluster contained very few CpG islands whose methylation status was experimentally verified. This can be easily explained due to the fact that CpG islands with repetitive elements are automatically excluded from the experimental verification process resulting in a dearth of these CpG islands in the experimental data we used here. Nevertheless, our method is able to find this distinct class of CpG island.

Cluster *T2* on the other hand contained mostly CpG islands that were enriched in conserved *Phastcon* elements and were shown to be methylated. This was surprising since it was generally assumed [24] that conservation of a DNA sequence during evolution implies functional relevance. While a potential functionality of these elements can't be discarded, it is unlikely that they act in the regulation of transcription due to fact that they are constitutively methylated. Cluster *T3* contains a majority of the unmethylated CpG islands and a highly significant subset of larger, promoter-associated CpG islands, further underlining the quality of the clustering with regard to the algorithms ability to isolate potentially functional CpG islands in individual clusters. The large number of CpG islands in cluster *T3* that were classified as not promoter-overlapping

may be the result of still unknown, alternative promoters used by the genes in this dataset. It has recently been discovered that human genes have even more layers of complexity than was previously thought. The occurrence of alternative promoters seems to be a common event [32] that may result in the discovery of even more promoters co-localizing with CpG islands.

We compared our method's capacity for separating methylated from non-methylated CpG islands with existing methylation prediction algorithms. These rely on supervised learning methods such as SVMs [15, 17, 18] that require separate, predetermined dependent classes and are particularly susceptible to overtraining. Though the percentage of CpG islands whose methylation status was predicted correctly was lower for our method compared to the SVM-methods we did achieve a high degree of specificity. These values are extremely promising for an unsupervised learning method that was not specifically trained to distinguish methylated from unmethylated CpG islands and will aid us greatly in developing a more specialized methylation prediction algorithm in the future.

Our novel approach helps determine which combination of CpG island attributes influences their degree of methylation and, at the same time, characterizes the islands that overlap the promoter region, without imposing a predefined classification. Our method is therefore well suited to the challenge of finding the exact combination of features that allow CpG islands to remain unmethylated and functional within the context of the complex regulatory system of human genes and will be employed in the development of a methylation prediction tool.

ACKNOWLEDGMENT

Christopher Previti gratefully acknowledges a grant from the German Academic Exchange Service (DAAD) and would like to thank the group of J. L. Oliver [19] for their advise and the use of their facilities. Oscar Harari acknowledges the doctoral MAEC- AECI fellowship. Coral del Val was financed by the program "Retorno de Investigadores de la Junta de Andalucía".

REFERENCES

[1] F. Antequera, J. Boyes, and A. Bird, "High levels of de novo methylation and altered chromatin structure at CpG islands in cell lines," *Cell,* vol. 62, pp. 503-14, 1990.

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

[3] A. P. Bird, "CpG-rich islands and the function of DNA methylation," *Nature,* vol. 321, pp. 209-13, 1986.

[4] J. Sved and A. Bird, "The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model," *Proc Natl Acad Sci USA,* vol. 87, pp. 4692-6, 1990.

[5] R. Jaenisch and A. Bird, "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals," *Nat Genet,* vol. 33 Suppl, pp. 245-54, Mar 2003.

[6] F. Antequera, "Structure, function and evolution of CpG island promoters," *Cell Mol Life Sci,* vol. 60, pp. 1647-58, 2003.

[7] S. Saxonov, P. Berg, and D. L. Brutlag, "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters," *Proc Natl Acad Sci USA,* vol. 103, pp. 1412-7, 2006.

[8] K. L. Arney and A. G. Fisher, "Epigenetic aspects of differentiation," *J Cell Sci,* vol. 117, pp. 4355-63, Sep 1 2004.

[9] W. Reik, W. Dean, and J. Walter, "Epigenetic reprogramming in mammalian development," *Science,* vol. 293, pp. 1089-93, Aug 10 2001.

[10] E. Heard, "Recent advances in X-chromosome inactivation," *Current Opinion in Cell Biology,* vol. 16, pp. 247-255, 2004.

[11] J. A. Yoder, C. P. Walsh, and T. H. Bestor, "Cytosine methylation and the ecology of intragenomic parasites," *Trends in Genetics,* vol. 13, pp. 335-340, 1997.

[12] S. B. Baylin, M. Esteller, M. R. Rountree, K. E. Bachman, K. Schuebel, and J. G. Herman, "Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer," *Hum Mol Genet,* vol. 10, pp. 687-92, 2001.

[13] M. Esteller, P. G. Corn, S. B. Baylin, and J. G. Herman, "A gene hypermethylation profile of human cancer," *Cancer Res,* vol. 61, pp. 3225-9, 2001.

[14] J. P. Issa, "CpG island methylator phenotype in cancer," *Nat Rev Cancer,* vol. 4, pp. 988-93, Dec 2004.

[15] F. Fang, S. Fan, X. Zhang, and M. Q. Zhang, "Predicting methylation status of CpG islands in the human brain," *Bioinformatics,* vol. 22, pp. 2204-9, Sep 15 2006.

[16] Y. Yamada, H. Watanabe, F. Miura, H. Soejima, M. Uchiyama, T. Iwasaka, T. Mukai, Y. Sakaki, and T. Ito, "A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q," *Genome Res,* vol. 14, pp. 247-66, Feb 2004.

[17] C. Bock, M. Paulsen, S. Tierling, T. Mikeska, T. Lengauer, and J. Walter, "CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure," *PLoS Genet,* vol. 2, p. e26, Mar 2006.

[18] R. Das, N. Dimitrova, Z. Xuan, R. A. Rollins, F. Haghighi, J. R. Edwards, J. Ju, T. H. Bestor, and M. Q. Zhang, "Computational prediction of methylation status in human genomic sequences," *Proc Natl Acad Sci U S A,* vol. 103, pp. 10713-6, Jul 11 2006.

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

[20] V. K. Rakyan, T. Hildmann, K. L. Novik, J. Lewin, J. Tost, A. V. Cox, T. D. Andrews, K. L. Howe, T. Otto, A. Olek, J. Fischer, I. G. Gut, K. Berlin, and S. Beck, "DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project," *PLoS Biol,* vol. 2, p. e405, Dec 2004.

[21] D. R. Maglott, K. S. Katz, H. Sicotte, and K. D. Pruitt, "NCBI's LocusLink and RefSeq [http://www.ncbi.nih.gov/RefSeq]," *Nucleic Acids Res,* vol. 28, pp. 126-8, Jan 1 2000.

[22] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent, "The UCSC Table Browser data retrieval tool [http://genome.ucsc.edu] " *Nucleic Acids Res,* vol. 32, pp. D493-6, Jan 1 2004.

[23] L. Ponger, L. Duret, and D. Mouchiroud, "Determinants of CpG islands: expression in early embryo and isochore structure," *Genome Res,* vol. 11, pp. 1854-60, Nov 2001.

[24] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Res,* vol. 15, pp. 1034-50, 2005.

[25] J. C. Bezdek, "Pattern Analysis," in *Handbook of Fuzzy Computation*, W. Pedrycz, P. P. Bonissone, and E. H. Ruspini, Eds. Bristol: Institute of Physics, 1998, pp. F6.1.1-F6.6.20.

[26] A. P. Gasch and M. B. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biol,* vol. 3, p. RESEARCH0059, Oct 10 2002.

[27] T. M. Mitchell, *Machine learning*. New York: McGraw-Hill, 1997.

[28] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nat Genet,* vol. 22, pp. 281-5, Jul 1999.

[29] G. Saporta, "Probabilits, analyse des donnes et statistiques," *Technip,* 1996.

[30] C. L. Wilkins, "Data mining with Spotfire Pro 4.0," *Analytical Chemistry,* vol. 72, pp. 550a-550a, AUG 1 2000.

[31] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics,* vol. 16, pp. 412-24, May 2000.

[32] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume, and Y. Hayashizaki, "Genome-wide analysis of mammalian promoter architecture and evolution," *Nat Genet,* vol. 38, pp. 626-35, 2006.