33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

An identification and prediction methods for feature-subsets of CpG islands methylation based on human peripheral blood leukocytes of chromosome 21q

Isse Ali*, *Member, IEEE and Hussein Sheikh ali Mohamoud*

Abstract— **The pace of technology has allowed classification of feature-subset of methylated and unmethylated of CpG islands of DNA sequence properties. As methylation of CpG islands is involved in various biological phenomena and function of the DNA methylation is correlated to various human diseases such as cancer, analysis of the CpG islands has become important and useful in characterizing and modelling biological phenomena and understanding mechanism of such diseases. However, analysis of the data associated with the CpG islands is a quite new and challenging subject in bioinformatics, systems biology and epigenetics.**

In this paper, the problems associated with prediction of methylated and unmethylated CpG islands on human chromosome 21q are addressed. In order to carry out the prediction, a data set of 132 samples of the CpG islands from human peripheral blood leukocytes of chromosomes 21q and 4 different feature sub-sets totalling 44 attributes that characterise the methylated and unmethylated groups is extracted for each sample. Due to the nature of this unbalanced data set, in order to avoid disadvantages of traditional leave-one-out (LOO) and m-fold cross validation methods, the LOO method is modified by incorporating the m-fold cross validation approach. In addition, K-nearest neighbour classifier is then adapted for the prediction.

The results gained through 440 different comprehensive analyses shows that the methylated CpG islands can be distinguished from the unmethylated CpG islands by a predictive accuracy of between 75% and 80%. More importantly, the modified LOO identifies more clearly and reliably when the feature sub-sets are combined. Another interesting observation is that the modified-LOO-based analysis reveals that the CpGI-specific feature-set achieve the highest predictive accuracy when combined with the other feature sets, which is not the case in the traditional LOO. This also further supports the robustness of the modified-LOO cross validation approach as CpGI-specific feature-set is one of the most important and effective attributes shown in other studies.

I. Introduction

DNA methylation is a biochemical modification of eukaryotic DNA, which generally occurs at the fifth (C5) position of cytosines residue in a 5'-CG-3' called CpG dinucleotides [1, 2, 3,]. In vertebrates, cytosines residue methylation in CpG nucleotides is an epigenetic marker that is necessary for physiological cell differentiation [1,3]. It is shown that more than 60% of human genes' promoter consists of unmethylated CpG islands [4].

Prediction of DNA methylation is one of the most complex and challenging problems in bioinformatics because DNA sequence features that characterize methylation, in particular CpG islands, are dispersed throughout the human genome. However, the advances in high-throughput technology for computational genomics and epigenomics has helped analyse a large variable data obtained from methylated and unmethylated DNA of CpG islands. Methylation of CpG islands are mainly involved in various biological processes such as gene silencing, structural chromosomal stability, parental imprinting and suppressing the mobility of retrotransposons [1]. The function of DNA methylation has also been linked to various human diseases such as cancer [1, 3, 5, 6]. It should be noted that, despite all the advances, analysis of DNA methylation, particularly for human genome, is just beginning.

DNA methylation profiling of the Human Major Histocompatibility Complex, which has been shown to be the most gene-dense region in the human genome and contains genes with a diversity functions (immune system) on chromosome 6 (6p21.3), was one of the first studies in Human Genome Project [7]. In addition to that Yamada [8] profiled DNA methylation data derived from human chromosomes 21q extracted from human peripheral blood leukocytes of four healthy individuals. There are also other researchers who have tried to predict DNA methylation of CpG islands [9, 10, 11, 12]. However, their studies were limited as they looked at only nucleotide sequence (CpGIs) and transcription factor binding site (TFBS) which provides only an incomplete view of the human DNA methylation. Bock et al [13] has recently extended Yamada's data by extracting DNA-sequence features associated with CpG islands and analysed the data using statistical methods. However, detailed and consistent analysis of the features was not carried out. In addition, statistical approaches used for the analysis was found to be insufficient, which could yield misleading outcome due to the nature of such complex data.

The aim of the study is therefore to develop a statistical strategy and carry out detail and comprehensive analysis of the features for a more accurate and reliable prediction of unmethylated and methylated CpG islands classes. The rest of the paper is organised as follows: section two gives an explanation the data and the method used through the study, results are presented and discussed with recommendation for

Isse Ali is with the Bio-Health Informatics Research Group within the Centre for Computational Intelligence, De Montfort University, Leicester, LE1 9BH, The United Kingdom and Hussein Sheikh ali Mohamoud is with South West Thames Regional Genetics Service, St Georges Hospital, Cranmer Terrace, London, SW17 0RE e-mails: iali@dmu.ac.uk and otomaarso@homail.com $^*CorrespondingAuthor$: $IsseAli(iali@dmu.ac.uk)$

further research in section three, and finally the paper is concluded in section four.

II. Materials and method

1) CpG islands Data: The data set was extracted from [13]. After the data was filtered, the data used throughout this study now contains 132 CpG islands(CpGI) in human chromosome 21q, 103 of which are unmethylated samples whereas 29 samples are methylated. In order to characterize the DNA sequences, a set of features is extracted over DNA sequences which are summarised in Table I. These data were extracted from 132 samples of CpG islands which driven from peripheral blood leukocytes or placenta of four human healthy individuals. These were averaged methylation changes between CpG pairs of identical samples; in order to minimise a bias produced by the length differences of sequence windows.

As seen in Table I, there are 44 features extracted, which are further divided into four subsets. They can be described as follows: The sub-set 1 (f1: CpGI specific DNA methylation) contains 8 attributes and is averaged sequences values which calculated by using CpGcluster algorithms [14]. These are CGI-specific attributes (CG contents, CG%, number of CpG island, observed/expected ratio, CpGI distance, and CpGcluster-pvalue). The sub-set 2 (f2: Evolutionary and conservation) contains 4 attributes of conserved elements contents which is calculated by a number of CpGI overlapping with conserved elements per CpCI by using a log-odds conservation score of 100 or more without repeat masking. The sub-set 3 (f3: CG distribution) contains 16 attributes and represent score of 16 possible combinations of its observed /expected ratio. The sub-set 4 (f4: structural and physiochemical properties) contains 16 attributes and includes predicted elements such as rise, roll, tilt, twist and solvent accessible surface area as well as bending, curvature, stacking energy, turns, degree of twist, DNA cleavage, base per turn and six helical force constant. The calculations of the features were done using DNAlive algorithms [15].

TABLE I

DETAILS OF THE CPG ISLANDS EXTRACTED FEATURE-SETS

Extracted features	Abbreviations	No of features
CpGI-specific DNA methylation	f1	8
Evolutionary and conservation	f2	4
sequence distribution (Dinucleotide)	f3	16
DNA structure and properties	f4	16
All the feature sets listed above	f-all	44

2) Modified Leave-One-Out cross validation: Various cross validation methods are proposed for assessment of predictive models. As far as small data set, which is the case in this study, is concerned, leave-one-out has been widely used. M-fold cross validation method is also found to be satisfactory for various sizes of data. However, when the data set is quite unbalanced, which is the case in this study, these two methods are found to be biased towards a class with the highest number of samples and could yield misleading outcome [16, 17]. Therefore, in this study, leave-one-out method is modified by incorporating with M-fold cross validation technique. To clarify, small samples (29 samples in the methylated group) is kept constant whereas the unmethylated data is randomly divided into 10 folds of equal size of 29 samples, and then 10 different models and predictive accuracies are obtained. Therefore, these 10 divisions were analysed in a single and also all possible combinations by using modified LOO cross validation with KNN classifier and k = 1 to 11 were used.

3) Predictive method: K-nearest neighbor classifier (K-NN): K-nearest neighbor (KNN) classifier is one of the most popular non-parametric classifiers and successfully applied to various problems in bioinformatics [19, 20, 21]. It assigns to the point that the majority label among its nearest k in the training data point to x and predicts the class-label of x based on label of that k points. Increasing the k value shown reduced bias and decision boundaries becomes rather smooth and less sensitive to the outliers [19, 20]. It has been reported in some studies that KNN resulted in a higher predictive accuracy than that of Support Vector Machine being one of the most powerful methods [21]. However, it should be noted that in many cases, success of a predictive method is mainly based on a characteristic of a data set being analysed. For this study, due to its flexibility, effectiveness and power, K-NN is adapted together with the modified leave-one-out cross validation method, previous used successfully to imbalance feature problems and gave better predictive accuracy [17,18]

III. Results and Discussions

This result compromises all possible combinations (120) of the four biological feature sub-sets as listed in Table I. A total of 440 analyses was carried out and obtained predictive accuracies through these analyses are summarised and presented in Table II.

Single feature-sets, dinucleotide distribution (f3) and Evolutionary and conservation (f2) show the highest class performance as well as predictive class accuracy where the total of accuracy 77.41% and 70.34% with their standard error of 6.10 and 3.41 respectively. Evolutionary and conservation (f2) gave better predictive class performance and this confirms our previous studied chromosomes 6 and 22 [18],whereas two other features (f1 and f4) are shown with fluctuations of class performance. This may contain some noisy features.

Next we investigated the association between at least two feature-sets when combined. This shows that the accuracy steadily increased while the class performance approximately remains the same as the single subset. The best class performance yield when CpGI-specific features (f1) and dinucleotide distribution (f3) are combined followed by f1 and f2 combination. The predictive accuracies of both classes (methylated , unmethylated and total) are 69.66%, 80.00% and 74.83% respectively. Combining feature-sets (f2 and f3, f2 and 4 or f3 and f4) shows overall higher predictive accuracy despite their predictive performance decreased. Three

TABLE II

THE HIGHEST MEAN PREDICTIVE ACCURACY (%) AND STANDARD ERROR FOR COMBINATIONS AND INDIVIDUAL OF THE FEATURE-SETS AS A RESULT OF THE M-LOO-BASED ANALYSIS OF CHROMOSOME 21.

No of features	combined feature-sets	Methylated predictive accuracy	UnMethylated predictive accuracy	total predictive accuracy	standard-error
single feature	f1	67.59	80.70	74.14	9.27
	f2*	**69.66**	**71.04**	**70.34**	**3.41**
	f3	73.79	81.03	77.41	6.10
	f4	63.4	86.55	75.00	17.80
2 features	{f1,f2}	70.00	83.10	76.55	9.27
	{f1,f3}*	**69.66**	**80.00**	**74.83**	**7.31**
	{f1,f4}	56.90	87.24	72.07	21.46
	{f2,f3}	71.72	87.93	79.83	11.46
	{f2,f4}	62.41	90.69	76.55	19.99
	{f3,f4}	66.20	86.55	76.38	14.39
3 features	**{f1,f2,f3}***	**72.07**	**88.62**	**80.34**	**11.70**
	{f1,f2,f4}	59.31	92.41	75.86	23.41
	{f1,f3,f4}	65.17	85.86	75.51	14.63
	{f2,f3,f4}	69.31	88.62	78.97	13.65
4 features	{f1,f2,f3,f4}	67.59	88.97	78.28	15.12

* bolded feature-sets/values are those which show highest class performance.

feature-sets combinations, CpGI-specific feature (f1), Evolutionary and conservation (f2) and dinucleotide distribution (f3) achieved the highest accuracy as well as good class performance compared to any other three combinations. Furthermore, combining (f4) to other two feature-sets resulted less class predictive performance but the total accuracy did not change very much. This may cause some of the physio- and chemical properties that do not complement with other feature-subsets. Four feature-sets combination have shown slight reduction of the class performance, whereas three feature-set combinations revealed a consistent class performance specially when excluded feature-set (f4). However, any combination with excluding CpGI-specific feature (f1) remains the lowest class performance. Hence, it would be expected that our prediction would be higher, if we trained on a single feature-set and compared the rest of the features, focusing only one specific attributes. However, these were no much differences among the predicted results where predictive performance for single attributes nearly the same as the average (unpublished). Moreover, M-LOO is potential candidate to predict unbalance data as well as sub feature-set that outperformed than the traditional Leave-One-Out [16]. However, combined feature-sets are complex and computational required which also needed pre-processing data. Despite this, our prediction method improves the tradition method and adds a new learning method for solving pattern recognition problems. In addition, this method is a simple yet yielded better accuracy than more complicated classifier such as Support Vector Machine. Furthermore, this method does not need dimensionality reduction as used previous study in order to predict feature-sets [13].

Further research will extend towards developing a more robust feature-set based on the DNA-patterns for a more accurate and reliable prediction of methylated and unmethylated groups and also will extend to DNA sub-set Physio- and chemical structure features to predict the best indicator of DNA methylation.

IV. CONCLUSION

In this paper, four feature-subsets of CpG islands data were studied for human peripheral blood leukocytes of chromosome 21q. The data was extracted from 132 DNA samples and a prediction was carried out to determine the features that are associated methylated from unmethylated DNA sequence patterns.

Combinations of the feature-sets, sequence features derived from DNA methylation is the most difficult one to predict because of the imbalance sample and their variables. To overcome this, we used modified M-fold cross validation which shows better performance and also higher accuracy. The combinations of some feature-sets (f1, f2 and f3) increased their class performance. However, some of the combined features (f2, f3 and f4) decreased the class performance but their total average accuracy did not change very much. This is because the number of attributes may have a negative effect for combining some of these features. When (f4) combined with other feature subsets showed a reduced class performance despite its total average accuracy remained steadily. Feature-sets, CpGI-specific (f1), Evolutionary and conservation (f2) and dinucleotide distribution (f3) showed a highest predictive accuracy and also a good class performance compared with all other three feature-sets combination, followed by two feature-set combination of CpGI-specific feature (f1) and dinucleotide distribution (f3). Comparing, the predictive accuracy of single feature-sets to the two, three and all combinations of feature-sets, shows an increase of approximately 10% of accuracy when it is combined. Our results show a significant correlation between CpGI-specifity and dinucleotide distribution as well as Evolutionary and conservation.

REFERENCES

[1] Adrian Bird. DNA methylation patterns and epigenetics memory. *Genes and dev*. vol.**16**, pp : 6–21, 2002.

[2] M. Brena Romulo. *et al*. Toward a Human Epigenome. *Nat. genet*. vol.**38**, pp : 1359–1360, 2006.

[3] Andrew P. Feinberg. Phenotypic plasticity and the epigenetics of human disease. *Nat. insight review*. vol.**447**, pp : 433–40, 2007.

[4] Francisco Antequera and Adrian Bird. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. USA*. vol.**90**, pp : 11995–11999, 1993.

[5] Ilana Keshet. *et al*. Evidence for an instructive mechanism of the novo methylation in cancer cells. *Nat. genet*. vol.**38**(2), pp : 149–153, 2006.

[6] Mill Jonathan. *et al*. Epigenetic profiling reveals DNA-methylation changes associated with Major Psychosis. *The American Journal of Human Genetics*. vol.**82**, pp : 696–711, 2008.

[7] Vardhman K. Rakyan *et al*. DNA methylation profiling of Human Major Histocompatibility Complex: A pilot study for the human epinenome project. *PLoS Biology*, vol.**2**(12): e405, 2004.

[8] Y. Yamada. *et al*. A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. *Genome Reseach*. vol.**14**, pp : 247–266, 2004.

[9] Elmar Shilling and Michael Rehli . Global, comparative analysis of tissue-specific promoter CpG methylation. *Genomics*. vol.**90**, pp : 314-323, 2007.

[10] Fang Fang *et al*. predicting methylation status of CpG islands in the human brain. *Bioinformatics*. vol.**22** (18), pp : 2204–2209, 2006.

[11] Robert J. Weeks and Ian M. Morison. Detailed Methylation Analysis of CpG Islands on Human Chromosome Region 9p21. *Genes, Chromosomes and Cancer*. vol.**45**, pp : 357–364, 2006.

[12] Srinivas Veerla. et al. Promoter Analysis of Epigenetically Controlled Genes in Bladder Cancer. *Genes, Chromosomes and Cancer*. vol.**47**, pp : 368–374, 2008.

[13] C Bock. *et al*. CpG Island Methylation in Human Lymphocytes Is Highly Correlated with DNA Sequence, Repeats, and Predicted DNA Structure, *PLoS Genetics*, vol.**2**(3): e26, 2006.

[14] Michael Hackenberg *et al*. CpGcluster: a distance-based algorithm for CpG-island detection.*BMC Bioinformatics*. vol.**7**(446), pp : 1-13, 2006.

[15] J. Ramon Goi *et al*. DNAlive: a tool for the physical analysis of DNA at the genomic scale. *Bioinformatics*. Vol. **24** (15), PP : 17311732, 2008.

[16] P. Baldi *et al*. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics review*. vol.**16**(5), pp : 412-424, 2000.

[17] I. Ali and H. Seker. Detailed methylation prediction of CpG islands on human chromosome 21. Proceedings of the 10th WSEAS International Conference on MATHEMATICS and COMPUTERS in BIOLOGY and CHEMISTRY. **ISSN: 1790-5125**, PP : 147-152, 2009.

[18] I. Ali and H. Seker. A comparative study for characterisation and prediction of tissue-specific DNA methylation of CpG islands in chromosomes 6, 20 and 22.*32nd Annual International Conference of the IEEE EMBS*. **ISSN: 1557-170X** , pp : 1832– 1835, 2010.

[19] Edward R. Dougherty *et al*. Validation of computational methods in genomics. *Current Genomics*. vol.**8**, pp : 1–19, 2007.

[20] Tao Li *et al*. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*. vol.**20** (15), pp : 2429–2437, 2004.

[21] Yvan Saeys *et al*. A review of feature selection techniques in bioinformatics. *Bioinformatics*. vol.**23** (19), pp : 2507–2517, 2007.