Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 740506, 4 pages
http://dx.doi.org/10.1155/2014/740506

Research Article

Prediction of DNase I Hypersensitive Sites by Using Pseudo Nucleotide Compositions

Pengmian Feng, Ning Jiang, and Nan Liu

School of Public Health, Hebei United University, Tangshan 063000, China

Correspondence should be addressed to Pengmian Feng; fengpengmian@gmail.com

Received 11 July 2014; Accepted 3 August 2014; Published 19 August 2014

Academic Editor: Hao Lin

DNase I hypersensitive sites (DHS) associated with a wide variety of regulatory DNA elements. Knowledge about the locations of DHS is helpful for deciphering the function of noncoding genomic regions. With the acceleration of genome sequences in the postgenomic age, it is highly desired to develop cost-effective computational methods to identify DHS. In the present work, a support vector machine based model was proposed to identify DHS by using the pseudo dinucleotide composition. In the jackknife test, the proposed model obtained an accuracy of 83%, which is competitive with that of the existing method. This result suggests that the proposed model may become a useful tool for DHS identifications.

1. Introduction

DNase I hypersensitive sites (DHS) are regions of chromatin which are sensitive to cleavage by the DNase I enzyme. Since the discovery of DHSs in 1980s [1], they have been used as markers of regulatory DNA regions. In general, these specific regions are generally nucleosome-free and associate with a wide variety of genomic regulatory elements, such as promoters, enhancers, insulators, silencers, and suppressors [2–4]. Therefore, mapping of DHS has become an effective approach for discovering functional DNA elements from the noncoding sequences.

Although the traditional Southern blotting technique is a gold-standard approach for identifying DHS, obtaining information from Southern blot approach is a tricky, time-consuming, and inaccurate task [5]. Recently, the DNase-seq technique (combination of DNase I digestion and high-throughput sequencing) has been proposed [6] and this technique allows for an unprecedented increase in resolution. However, methodologies for the analysis of DNase-seq data are relatively immature [7]. Therefore, computational models will be an important complement to experimental techniques for identifying DHS.

Based on nucleotide compositions, a support vector machine model for identifying DHS in K562 cell line was proposed [8]. This method yielded quite encouraging results and did play a role in stimulating the development of this area. However, further work is needed due to the following reasons. First, the sequences in their dataset share high sequence similarities. Second, the DNA structural properties were ignored. To solve these problems, we proposed a new model for identifying DHS, which is trained on a high quality benchmark dataset. In the new model, each DNA sample is encoded by using the pseudo dinucleotide composition, into which the DNA structural properties are incorporated.

2. Materials and Methods

2.1. Benchmark Dataset. The experimentally confirmed 280 DHS and 731 non-DHS sequences were obtained from http://noble.gs.washington.edu/proj/hs/, which have been used to train DHS prediction models [8]. As elucidated in [9], a predictor, if trained and tested by a dataset containing redundant samples with high similarity, might yield misleading results with an overestimated accuracy. To get rid of the redundancy and avoid bias, the CD-HIT software [10] was utilized to remove those DNA fragments that have ≥60% pairwise sequence identity to each other.

Finally, we obtained 247 positive and 710 negative samples for the benchmark dataset $\mathbb{S}$, as can be formulated by

$$\mathbb{S} = \mathbb{S}^+ \bigcup \mathbb{S}^-, \tag{1}$$

where the subset $\mathbb{S}^+$ contains 247 DHS sequences and $\mathbb{S}^-$ contains 710 non-DHS sequences, while $\bigcup$ represents the "union" in the set theory. The detailed sequences in the benchmark dataset $\mathbb{S}$ are given in Supplementary Information S1 available online at http://dx.doi.org/10.1155/2014/740506.

2.2. DNA Sequence Representation. In order to integrate the sequence-order effects and DNA physicochemical properties together, the pseudo nucleotide composition was proposed in 2011 [11]. Since then, the concept of pseudo nucleotide composition has penetrated into many branches of computational genomics, such as predicting the recombination spots [12], predicting promoters [13], predicting nucleosome positioning sequences [14], and identifying splice sites [15]. Because of its wide and increasing usage, recently, a flexible web-server, called "pseudo K-tuple nucleotide composition (PseKNC)," was developed [16], which can be used to generate various kinds of pseudo K-tuple nucleotide compositions.

Encouraged by the success of introducing pseudo nucleotide composition to computational genomics, in the current study, the pseudo dinucleotide composition was used to represent DNA sequences in the benchmark dataset, which can be expressed as [12, 16]

$$\mathbf{D} = \left[d_1 \;\; d_2 \;\; \cdots \;\; d_{16} \;\; d_{16+1} \;\; \cdots \;\; d_{16+\lambda}\right]^{\mathbf{T}}, \tag{2}$$

where

$$d_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j}, & (1 \le u \le 16)\,, \\ \dfrac{w\theta_{u-16}}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j}, & (16 < u \le 16 + \lambda)\,. \end{cases} \tag{3}$$

In (3), f_u $(u = 1, 2, \ldots, 16)$ is the normalized occurrence frequency of the dinucleotides in the DNA sequence. λ is the number of the total counted ranks (or tiers) of the correlations along a DNA sequence, and w is the weight factor. The concrete values for λ and w as well as k will be further discussed in Section 3.1, while the correlation factor θ_j represents the j-tier structural correlation factor between all the jth most contiguous dinucleotide R_iR_{i+1} at position i.

2.3. Support Vector Machine (SVM). SVM is a supervised learning algorithm and has been widely used in computational genomics and proteomics [17–23]. The basic principle of SVM is to transform the input vector into a high dimension space and then seek a separating hyperplane with the maximal margin in this space by using the decision function

$$f\left(\vec{X}\right) = \operatorname{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \cdot K\left(\vec{X}, \vec{X}_i\right) + b\right), \tag{4}$$

where α_i is the Lagrange multipliers, b is the offset, $\vec{X}_i$ is the ith training vector, and y_i represents the type of the ith training vector. $K(\vec{X}, \vec{X}_i)$ is a kernel function which defines an inner product in a high dimensional feature space, and sgn is the sign function. Due to its effectiveness and speed in nonlinear classification process, the radial basis kernel function (RBF) $K(\vec{X}_i, \vec{X}_j) = \exp(-\gamma \|\vec{X}_i, \vec{X}_j\|^2)$ was used in the current study.

The Libsvm 2.84 package [24] was used to perform the SVM, which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/. The regularization parameter C and the kernel width parameter γ were optimized via an optimization procedure using a grid search. The search spaces for C and γ are $[2^{15}, 2^{-5}]$ and $[2^{-5}, 2^{-15}]$ with steps of 2^{-1} and 2, respectively.

2.4. Performance Evaluation. Three cross-validation methods, that is, independent dataset test, subsampling (or K-fold cross-validation) test, and jackknife test, are often used to evaluate the anticipated success rate of a predictor. Among the three methods, the jackknife test is deemed the least arbitrary and most objective one [9, 25] and, hence, has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors [26–30]. Accordingly, the jackknife test was used to examine the performance of the model proposed in the current study. In the jackknife test, each sequence in the training dataset is in turn singled out as an independent test sample and all the rule-parameters are calculated without including the one being identified.

A set of parameters, namely, sensitivity (Sn), specificity (Sp), Matthew's correlation coefficient (MCC), and accuracy (Acc), are used to evaluate the performance of the proposed model and they are defined as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{5}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{6}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FP})}, \tag{7}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}, \tag{8}$$

where TP, TN, FP, and FN represent the number of the correctly recognized DHS, the number of the correctly recognized non-DHS, the number of non-DHS recognized as DHS, and the number of DHS recognized as non-DHS, respectively.

3. Results and Discussions

3.1. Parameter Optimization. By analyzing the dinucleotide composition of DHS and non-DHS sequences, we found that the frequency of CC, CG, GC, and GG is higher in DHS sequences, while the frequency of the remaining dinucleotides is higher in non-DHS (Figure 1). This is self-evident as to why the pseudo dinucleotide composition was used for the current case.

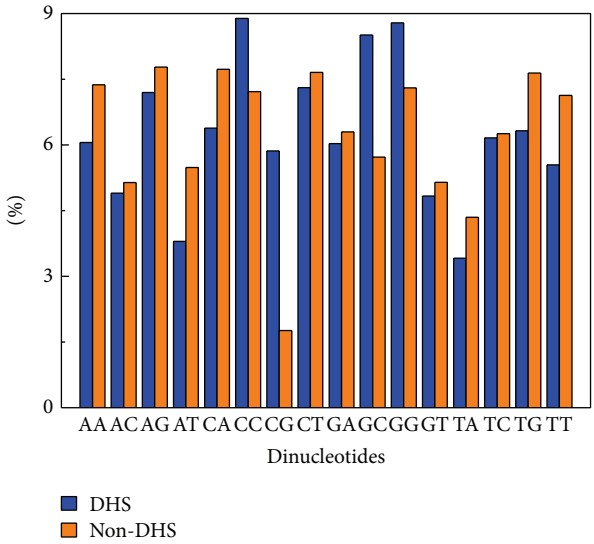

Figure 1: Comparative frequencies of 16 dinucleotides in DHS and non-DHS sequences.

A series of evidences [12, 14, 31, 32] have demonstrated that DNA local structural properties, that is, angular parameters (twist, tilt, and roll) and translational parameters (shift, slide, and rise), are effective in identifying DNA attributes. Therefore, in the present work, the six structural parameters of dinucleotides were used to calculate the pseudo dinucleotide composition by using the PseKNC web-server, which is available at http://lin.uestc.edu.cn/pseknc/default.aspx.

As we can see from (1) and (2), the present model depends on the two parameters w and λ. w is the weight factor usually within the range from 0 to 1 and λ is the global order effect. Generally speaking, the greater the λ is, the more global sequence-order information the model contains. However, if λ is too large, it would reduce the cluster-tolerant capacity so as to lower down the cross-validation accuracy due to overfitting or "high dimension disaster" problem [33]. Therefore, our searching for the optimal values of the two parameters is in the range of $w \in [0, 1]$ and $\lambda \in [1, 10]$ with the steps of 0.1 and 1, respectively.

In order to reduce the computational time, the 5-fold cross-validation approach was used to optimize the two parameters together with the parameters C and γ of the SVM. We found that when $w = 0.2$ and $\lambda = 6$ with $C = 512$ and $\gamma = 0.0078125$, a peak was observed for the Acc. Accordingly, the two numerical values were used for the two uncertain parameters in the following analysis.

3.2. Prediction Quality. The prediction quality measured by the four metrics defined in (5)–(8) for the present model in identifying DHS in the benchmark dataset $\mathbb{S}$ via the rigorous jackknife test was listed in Table 1, where, for facilitating comparison, the corresponding results obtained by the previous predictor [8] on the same benchmark data set are also given. As we can see from Table 1, the current method outperformed the existing model in all the four metrics, indicating that our proposed method may become a useful tool in identifying DHS sequences.

Table 1: Comparison of different methods for identifying DHS by the jackknife test on the same benchmark dataset.

Predictor	Sn (%)	Sp (%)	Acc (%)	MCC
Our method	72.12	86.78	83.00	0.57
Noble et al.[a]	70.43	84.23	80.12	0.52

[a] From Noble et al. [8].

4. Conclusions

Since DHS associates with a wide variety of functional elements, knowledge about the locations of DHS is helpful for deciphering the genomes. However, strong DNA sequence conservation is not observed among DHS sequences, suggesting that it is difficult to computationally identify DHS from primary DNA sequence.

A series of recent studies have demonstrated that the information coded by DNA structural properties is contributable to the identification of regulatory elements in genomes [12, 14, 31, 32]. Hence, in the present study, we proposed a SVM based model for identifying DHS by using the pseudo dinucleotide composition. In this model, we integrate dinucleotide composition with DNA structural properties. The predictive results of our model are better than existing methods. Therefore, it is anticipated that the proposed method may become a useful tool for identifying DHS sequences or, at the very least, it can play a complementary role to the existing methods in this area.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by Foundation of Science and Technology Department of Hebei Province (no. 132777133).

References

[1] C. Wu, P. M. Bingham, K. J. Livak, R. Holmgren, and S. C. R. Elgin, "The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence," *Cell*, vol. 16, no. 4, pp. 797–806, 1979.

[2] D. S. Gross and W. T. Garrard, "Nuclease hypersensitive sites in chromatin," *Annual Review of Biochemistry*, vol. 57, pp. 159–197, 1988.

[3] G. Felsenfeld and M. Groudine, "Controlling the double helix," *Nature*, vol. 421, no. 6921, pp. 448–453, 2003.

[4] G. Felsenfeld, "Chromatin as an essential part of the transcriptional mechanism," *Nature*, vol. 355, no. 6357, pp. 219–224, 1992.

[5] G. E. Crawford, I. E. Holt, J. Whittle et al., "Genome-wide mapping of DNase hypersensitive sites using massively parallel

signature sequencing (MPSS)," *Genome Research*, vol. 16, no. 1, pp. 123–131, 2006.

[6] L. Song and G. E. Crawford, "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells," *Cold Spring Harbor Protocols*, vol. 5, no. 2, Article ID pdb.prot5384, 2010.

[7] P. Madrigal and P. Krajewski, "Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data," *Frontiers in Genetics*, vol. 3, article 230, 2012.

[8] W. S. Noble, S. Kuehn, R. Thurman, M. Yu, and J. Stamatoyannopoulos, "Predicting the in vivo signature of human gene regulatory sequences," *Bioinformatics*, vol. 21, no. 1, pp. i338–i343, 2005.

[9] K. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, pp. 236–247, 2011.

[10] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[11] X. Zhou, Z. Li, Z. Dai, and X. Zou, "Predicting methylation status of human DNA sequences by pseudo-trinucleotide composition," *Talanta*, vol. 85, no. 2, pp. 1143–1147, 2011.

[12] W. Chen, P. Feng, H. Lin, and K. Chou, "IRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Research*, vol. 41, no. 6, article e68, 2013.

[13] X. Zhou, Z. Li, Z. Dai, and X. Zou, "Predicting promoters by pseudo-trinucleotide compositions based on discrete wavelets transform," *Journal of Theoretical Biology*, vol. 319, pp. 1–7, 2013.

[14] S. H. Guo, E. Z. Deng, L. Q. Xu et al., "iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition," *Bioinformatics*, vol. 30, no. 11, pp. 1522–1529, 2014.

[15] W. Chen, P. M. Feng, H. Lin, and K. C. Chou, "iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition," *BioMed Research International*, vol. 2014, Article ID 623149, 12 pages, 2014.

[16] W. Chen, T. Y. Lei, D. C. Jin, H. Lin, and K. C. Chou, "PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition," *Analytical Biochemistry*, vol. 456, pp. 53–60, 2014.

[17] W. Chen and H. Lin, "Prediction of midbody, centrosome and kinetochore proteins based on gene ontology information," *Biochemical and Biophysical Research Communications*, vol. 401, no. 3, pp. 382–384, 2010.

[18] H. Lin and H. Ding, "Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 269, no. 1, pp. 64–69, 2011.

[19] B. Liu, X. Wang, Q. Chen, Q. Dong, and X. Lan, "Using Amino Acid Physicochemical Distance Transformation for Fast Protein Remote Homology Detection," *PLoS ONE*, vol. 7, no. 9, Article ID e46633, 2012.

[20] B. Liu, X. Wang, L. Lin, B. Tang, and Q. Dong, "Prediction of protein binding sites in protein structures using hidden Markov support vector machine," *BMC Bioinformatics*, vol. 10, article 381, 2009.

[21] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, "Exploiting three kinds of interface propensities to identify protein binding sites," *Computational Biology and Chemistry*, vol. 33, no. 4, pp. 303–311, 2009.

[22] K. C. Chou and Y. D. Cai, "Using functional domain composition and support vector machines for prediction of protein subcellular location," *The Journal of Biological Chemistry*, vol. 277, no. 48, pp. 45765–45769, 2002.

[23] M. Hayat and A. Khan, "MemHyb: predicting membrane protein types by hybridizing SAAC and PSSM," *Journal of Theoretical Biology*, vol. 292, pp. 93–102, 2012.

[24] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[25] K.-C. Chou and C.-T. Zhang, "Prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 4, pp. 275–349, 1995.

[26] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203–209, 2010.

[27] C. Ding, L. F. Yuan, S. H. Guo, H. Lin, and W. Chen, "Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions," *Journal of Proteomics*, vol. 77, pp. 321–328, 2012.

[28] W. Chen and H. Lin, "Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine," *Computers in Biology and Medicine*, vol. 42, no. 4, pp. 504–507, 2012.

[29] K. Chou, Z. Wu, and X. Xiao, "iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins," *PLoS ONE*, vol. 6, no. 3, Article ID e18258, 2011.

[30] H. Mohabatkar, M. Mohammad Beigi, and A. Esmaeili, "Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine," *Journal of Theoretical Biology*, vol. 281, no. 1, pp. 18–23, 2011.

[31] Y. Zuo and Q. Li, "Identification of TATA and TATA-less promoters in plant genomes by integrating diversity measure, GC-Skew and DNA geometric flexibility," *Genomics*, vol. 97, no. 2, pp. 112–120, 2011.

[32] J. R. Goñi, A. Pérez, D. Torrents, and M. Orozco, "Determining promoter location based on DNA structure first-principles calculations," *Genome Biology*, vol. 8, no. 12, article R263, 2007.

[33] T. Wang, J. Yang, H. Shen, and K. Chou, "Predicting membrane protein types by the LLDA algorithm," *Protein and Peptide Letters*, vol. 15, no. 9, pp. 915–921, 2008.