Recognition of DNase I hypersensitive sites in multiple cell lines

Wei Chen, Liaofu Luo* and Lirong Zhang

Laboratory of Theoretical Biophysics,
School of Physical Science and Technology,
Inner Mongolia University,
Hohhot 010021, China
E-mail: chenwei_imu@yahoo.com.cn
E-mail: lolfcm@mail.imu.edu.cn
E-mail: pyzlr@mail.imu.edu.cn
*Corresponding author

Hao Lin

School of Life Science and Technology,
University of Electronic Science and Technology of China,
Chengdu 610054, China
E-mail: hlin@uestc.edu.cn

Abstract: DNase I hypersensitive sites (DHSs) associate with a wide variety of functional genomic elements. Successful prediction of DHSs in computational models would dramatically accelerate the annotation of the human genome. In this study, a method of Increment of Diversity with Quadratic Discriminant analysis (IDQD) is presented for DHSs prediction in K562, $CD4^+$ T, Hela and GM06990 cell lines. The average accuracies of 10-fold cross-validation test are 98.52%, 96.50%, 99.25% and 97.58%, respectively, and the mean areas under ROC curves (auROC) are all greater than 0.90. The prediction results indicate that the IDQD method is an effective tool for DHSs recognition.

Keywords: DHSs; DNase I hypersensitive sites; IDQD; increment of diversity quadratic discriminant analysis.

Reference to this paper should be made as follows: Chen, W., Luo, L.F., Zhang, L.R. and Lin, H. (2009) 'Recognition of DNase I hypersensitive sites in multiple cell lines', *Int. J. Bioinformatics Research and Applications*, Vol. 5, No. 4, pp.378–384.

Biographical notes: Wei Chen received his BS Degree in Theoretical Physics from Tang Shan Normal University in 2005 and his MS Degree in Theoretical Biophysics from Inner Mongolia University in 2007. He is now working for his PhD at Inner Mongolia University under the supervision of Professor Liaofu Luo. His research interests include the theoretical ground of computational methods, machine learning and computational epigenetics.

Liaofu Luo is a Professor of Theoretic Biology at the School of Physical Science and Technology in Inner Mongolia University. His research interests in bioinformatics include the theoretical ground of computational methods, laws on gene splicing, discovery of promoters, protein structural prediction and computational epigenetics. He is the author of over 80 peer-reviewed papers.

Lirong Zhang is an Associate Professor at the School of Physical Science and Technology in Inner Mongolia University. Her research interests include alternative splicing and its correlation with disease.

Hao Lin received his PhD in Biophysics from Inner Mongolia University. He is an instructor at the School of Life Science and Technology and Centre of Bioinformatics, University of Electronic Science and Technology of China. His research interests include protein prediction and machine learning in computational biology.

1 Introduction

A major challenge of modern biology is the comprehensive delineation of functional non-coding sequences that regulate transcription and other chromosomal processes in the human genome. Several genome-wide strategies have been developed to identify the location of gene regulatory elements, such as sequence conservation, chromatin immunoprecipitation followed by microarray hybridisation (ChIP-chip) and computational analyses. Among these methods, one classical experimental method, mapping of DNase I Hypersensitive Sites (DHSs), has gradually made its way into the genomics domain for identifying the location of regulatory elements since 1980s. Regions of the genome that are hypersensitive to digestion by deoxyribonuclease I are generally nucleosome-free and associate with a wide variety of functional genomic elements, including promoters, enhancers, insulators, silencers, Locus Control Regions (LCRs) and suppressors (Gross and Garrard, 1988; Felsenfeld, 1992; Li et al., 2002; Felsenfeld and Groudine, 2003). Thus, mapping of DHSs by Southern blotting has become a gold-standard approach for discovering functional non-coding sequences involved in gene regulation.

Unfortunately, whereas progress has been made in identifying DHS sequences (sequences containing DHS), the traditional Southern blotting technique is not readily scalable to study the large chromosomal regions and thus precludes its use in systematic whole-genome analyses. Recently, novel methods for large-scale mapping of DHSs have been applied (Sabo et al., 2004; Dorschner et al., 2004; Crawford et al., 2006), providing an unprecedented opportunity for the computational identification of large numbers of DHS sequences that can be utilised in systematic study of transcriptional regulation and other functional elements in a genome. On the basis of these experimental data, Noble et al. (2005) trained a Support Vector Machine (SVM) model in K562 cell line and achieved an accuracy of 0.852 for DHS sequences recognition in a 10-fold cross-validation test. Therefore, the computational prediction will be an important complement to the experimental identification of DHS sequences in different tissues.

In this paper, a new method of increment of diversity with quadratic discriminant analysis, called IDQD, is presented to DHSs identification. The Increment of Diversity (ID) was first introduced by Laxton (1978) and employed in biogeography. For the purpose of improving prediction capability, the ID method combined with Quadratic Discriminant analysis (IDQD) was proposed and successfully applied in the prediction of exon–intron splice sites for several model genomes including human (Zhang and Luo, 2003). The method has also been used in the prediction of transcription start sites (Lu and Luo, 2007) and protein classification (Lin and Li, 2007). Here, the IDQD method

is generalised to the prediction of DHS sequences in multiple cell lines. The results of 10-fold cross-validation test indicate that our approach for DHSs prediction is complementary to, and equally predictive as, the SVM model proposed by Noble et al. (2005).

2 Materials and methods

2.1 Materials

The experimentally verified 280 DHS sequences and 731 non-DHS sequences in human K562 erythroid cell line were extracted from noble.gs.washington.edu/proj/hs (Noble et al., 2005). Both types of sequences were similar in size (mean length 242.1 bp vs. 242.8 bp) and constituted the positive and negative training samples, respectively.

To test the universality of the IDQD method in DHSs recognition, DHS sequences from other three cell lines, namely $CD4^+$ T, Hela and GM06990, are downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg17/encode/database/. Considering the DHS sequences typically comprise a core site-forming domain of about 150~250 bp in size (Gross and Garrard, 1988; Lowrey et al., 1992), sequences of 200~300 bp in length were chosen as candidates for positive samples. Since there is no detailed data on non-DHS sequences in these three cell lines, four 250 bp segments were appropriately extracted as negative samples for each selected positive sequence, two from its upstream and two from its downstream 2 kb of the DHS sequence. Thus, in these three cell lines, 194, 98, 168 positive samples and 776, 392, 672 negative samples were obtained, respectively.

2.2 Increment of diversity

In the state space of s dimensions, the diversity measure for any diversity source S: $\{n_1, n_2, \ldots, n_s\}$ is defined as (Laxton, 1978; Li and Lu, 2001):

$$D(S) = D(n_1, n_2, \ldots, n_s) = N \log_2 N - \sum_{i=1}^{s} n_i \log_2 n_i \quad N = \sum_{i=1}^{s} n_i \tag{1}$$

here n_i is the occurrences of the ith character in the diversity source S, and if n_i equals zero, then $n_i \log_2 n_i = 0$.

For an arbitrary sequence X to be predicted, in the same parameter space, the increment of diversity of sequence X with standard source S is defined as follows:

$$\mathrm{ID}(X, S) = D(X + S) - D(X) - D(S) \tag{2}$$

here $D(X + S)$ denotes the diversity measure of the mixed diversity source $S + X$. $D(X)$ and $D(S)$ are the diversity measures of the diversity sources S and X, respectively. $\mathrm{ID}(X + S)$ gives the relation of sequence X with standard source S. The smallest the $\mathrm{ID}(X + S)$ is, the most intimate the relation of the inquired sequence X to standard source S is.

2.3 Quadratic discriminant analysis

To recognise a sequence, one should introduce several increment of diversities, say, $\boldsymbol{R} = (\mathrm{ID}_1, \ldots, \mathrm{ID}_r)$ of r-dimensional vector. Thus, for each sample X, there will be an r-dimensional feature vector $R(X)$. Let the standard set (positive set) denoted by $\boldsymbol{G}_1$ and the contrast set (negative set) denoted by $\boldsymbol{G}_2$. The discriminant function that differentiates with the potential sequence X belonging to positive or negative set is defined as follows (Zhang and Luo, 2003; Lu and Luo, 2007; Luo and Lu, 2007):

$$\xi = \log_2 N_1/N_2 - \left\{(R-\mu_1)^T \sum_1^{-1}(R-\mu_1) - (R-\mu_2)^T \sum_2^{-1}(R-\mu_2)\right\} \Big/ 2 - 1/2 \log_2 \left(\left|\sum\nolimits_1\right| \Big/ \left|\sum\nolimits_2\right|\right) \tag{3}$$

here $\boldsymbol{\mu}_i$ is the average of $\boldsymbol{R}$ over all N_i samples of positive ($i = 1$) or negative set ($i = 2$) and $\sum_i$ is the corresponding $\boldsymbol{r} \times \boldsymbol{r}$ covariance matrix. The parameter ξ gives an order of sample $\boldsymbol{X}$. When $\xi > \xi_0$, $\boldsymbol{X}$ is classified into positive set and when $\xi \leq \xi_0$, $\boldsymbol{X}$ is classified into negative set. In the common use of quadratic discriminant analysis, the threshold ξ_0 is taken to be 0. However, as there is large difference in size between positive and negative sets, the optimal threshold ξ_0 may not be 0. It should be empirically determined in principle to obtain optimal prediction.

2.4 Information extraction from DHS sequences

Setting the polynucleotide composition of DHS sequence as the source of information, five k-mers ($k = 3, 4, \ldots, 7$) were selected as the feature variables for DHSs recognition. Using the occurrence frequency of k-mers in arbitrary sequence X, we can define the five diversities of sequence X as $D(X_1)$, $D(X_2), \ldots, D(X_5)$ according to equation (1), and the corresponding IDs between $D(X_i)$ ($i = 1, 2, \ldots, 5$) and standard source in positive (negative) training set as $I_1(I_2), I_3(I_4), \ldots, I_9(I_{10})$ according to equation (2). In addition, since about 60% of the DHSs are enriched within CpG islands in human genome, the G + C content in each sequence was calculated and defined as another feature variable I_{11}. Any sample is, therefore, depicted by a discrete vector $\boldsymbol{R} = (I_1, I_2, \ldots, I_{11})$.

2.5 Performance assessment

The performance of our method can be measured in terms of sensitivity (*Sn*), specificity (*Sp*) and accuracy (*Acc*). They are defined through True Positives (*TP*), True Negatives (*TN*), False Positives (*FP*) and False Negatives (*FN*) by the following equations.

$$Sn = TP/(TP + FN) \tag{4}$$

$$Sp = TN/(TN + FP) \tag{5}$$

$$Acc = (TP + TN)/(TP + TN + FP + FN). \tag{6}$$

Since the only adjustable parameter that exists in IDQD method is the threshold ξ_0, the sensitivity and specificity vary with ξ_0. The best choice of ξ_0 is obtained by adjustive optimisation of sensitivity and specificity. In the meantime, by varying ξ_0 and plotting the ROC curve that gives the relation between True Positive rate and False Positive rate at different threshold values, the ROC score was obtained, which evaluates

the performance of the method globally, irrespective of the threshold choice (Akobeng, 2007).

3 Results

To objectively evaluate the performance of IDQD method on DHS and non-DHS sequences classification, 10-fold cross-validation test was adopted. Given the threshold $\xi_0 = -2.45$ that maximised the prediction accuracy, the results on DHS sequences recognition in K562 cell line are shown in the first line of Table 1.

The average sensitivity, specificity, accuracy and the mean ROC score of 10-fold cross-validation test are listed in the third to sixth columns of the table. The prediction results on K562 cell line demonstrate that the accuracy of IDQD used in DHS recognition is comparable with that of the SVM method published in a recent literature (Noble et al., 2005).

To test the universal applicability of the IDQD method in DHSs recognition, the method was generalised to other three cell lines: $CD4^+$ T, Hela and GM06990. For comparison, the predictive results calculated by use of Nobel's SVM method were also listed in the table.

Table 1 The prediction results of IDQD method for DHSs recognition in different cell lines

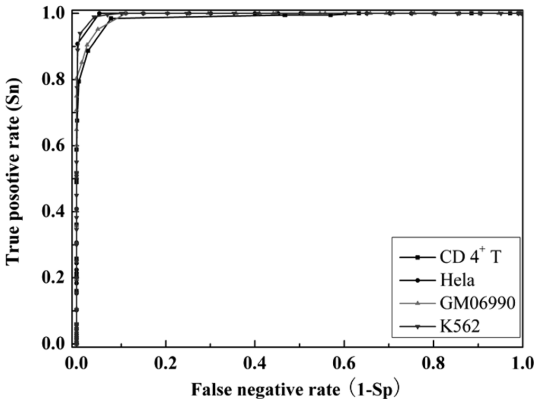
Cell line	*Method*	*Sn* (%)	*Sp* (%)	*Acc* (%)	*auROC*
K562	SVM[a]	73.91	88.61	85.29	0.842
	IDQD ($\xi_0 = -2.45$)	*95.83*	*98.06*	*98.52*	*0.996*
$CD4^+T$	SVM[b]	76.23	88.45	86.46	0.865
	IDQD ($\xi_0 = 0.21$)	*92.34*	*98.23*	*96.50*	*0.990*
Hela	SVM[b]	75.12	89.31	84.13	0.854
	IDQD ($\xi_0 = 1.20$)	*98.31*	*99.53*	*99.25*	*0.999*
GM06990	SVM[b]	79.59	90.47	88.47	0.895
	IDQD ($\xi_0 = -0.77$)	*92.24*	*98.53*	*97.58*	*0.994*

[a]Results are from Noble et al. (2005).

[b]The results are calculated by use of Nobel's method.

The high prediction accuracies (Figure 1) from IDQD indicate that our method is applicable to other cell lines for DHSs annotation. It also illuminates that the *k*-mer ($k = 3, 4, \ldots, 7$) frequency distributions are important feature parameters for classifying DHS and non-DHS sequences. In literature (Noble et al., 2005), *k*-mers with $k = 2$ to 6 were taken into account. But, following our experience, the 7-mer frequency is an important factor for the correct prediction of DHSs. The heptamer may reflect the main range of most deoxyribonuclease interactions. The IDQD method with *k*-mer frequencies as the input is easily manipulated and can be used in large-scale genome DHSs annotation.

Figure 1 The ROC curves for DHSs recognition in different cell lines



4 Discussion

On the basis of the different characteristics between DHS and non-DHS sequences, the IDQD method was successfully applied to the recognition of DHSs for four cell lines: K562, $CD4^+$ T, Hela and GM06990. These results indicate that IDQD, complementary to other existing methods, is a promising approach in future genome-wide DHSs annotation.

The ID defined in this paper is essentially a measure of entropy increase as a sample merged to a standard source. As the size of standard source is large enough, the influences of fluctuation can be negligible and the diversity of standard source will include accurate information about the frequency distribution of selected characters. Synthetically using several IDs by quadratic discriminant analysis, we are able to evaluate the detailed difference between any potential sample and the standard source.

The efficient extraction of sequence information by use of diversity measure in high-dimensional space and the synthesis of different types of sequence information into one discriminant function are two important factors for the success of IDQD algorithm. The different IDs are integrated into one non-linear discriminant function ξ through quadratic discriminant analysis. The only adjustable parameter existed in IDQD algorithm is the threshold of ξ, namely ξ_0. So, the algorithm is easily evaluated. The parameter should be empirically determined in principle to obtain optimal evaluation. However, in ROC analysis, the threshold ξ_0 can be looked exactly as a variable to plotting curves for performance evaluation.

Because DHSs are important genetic markers of cis-regulatory sequences, the application of the IDQD method for DHSs recognition will be helpful in the delineation of the functional elements in the human genome.

Acknowledgements

We are grateful to Prof. Guojun Li for his careful reviews and valuable comments on our manuscript. We thank Dr. Jun Lu for his useful discussions. This work was supported by National Natural Science Foundation of China, No. 90403010 and No. 10447003.

References

Akobeng, A.K. (2007) 'Understanding diagnostic tests 3: receiver operating characteristic curves', *Acta Paediatr.*, Vol. 96, pp.644–647.

Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D., Davis, S., Margulies, E.H., Chen, Y.D., Bernat, J.A., Ginsburg, D., Zhou, D.X., Luo, S.J., Vasicek, T.J., Daly, M.J., Wolfsberg, T.G. and Collins, F.S. (2006) 'Genome-wide mapping of DNase Hypersensitive sites using Massively Parallel Signature Sequencing (MPSS)', *Genome Res*., Vol. 16, pp.123–131.

Dorschner, M.O., Hawrylycz, M., Humbert, R., Wallace, J.C., Shafer, A., Kawamoto, J., Mack, J., Hall, R., Goldy1, J., Sabo, P.J., Kohli, A., Li, Q.L., McArthur, M. and Stamatoyannopoulos, J.A. (2004) 'High-throughput localization of functional elements by quantitative chromatin profiling', *Nat. Methods.*, Vol. 1, pp.219–225.

Felsenfeld, G. (1992) 'Chromatin as an essential part of the transcriptional mechanism', *Nature*, Vol. 355, pp.219–224.

Felsenfeld, G. and Groudine, M. (2003) 'Controlling the double helix', *Nature*, Vol. 421, pp.448–453.

Gross, D.S. and Garrard, W.T. (1988) 'Nuclease hypersensitive sites in chromatin', *Annu. Rev. Biochem.*, Vol. 57, pp.159–197.

Laxton, R.R. (1978) 'The measure of diversity', *J. Theor. Biol.*, Vol. 70, pp.51–67.

Li, Q.L., Peterson, K.R., Fang, X.D. and Stamatoyannopoulos, G. (2002) 'Locus control regions', *Blood*, Vol. 100, pp.3077–3086.

Li, Q.Z. and Lu, Z.Q. (2001) 'The prediction of the structural class of protein: application of the measure of diversity', *J. Theor. Biol.*, Vol. 213, pp.493–502.

Lin, H. and Li, Q.Z. (2007) 'Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant', *Biochem. Biophys. Res. Commun*., Vol. 354, pp.548–551.

Lowrey, C.H., Bodine, D.M. and Nienhuis, A.W. (1992) 'Mechanism of DNase I Hypersensitive Site formation within the human globin locus control region', *Proc. Natl. Acad. Sci., USA*, Vol. 89, pp.1143–1147.

Lu, J. and Luo, L.F. (2007) 'Predicting human transcription starts by use of diversity measure with quadratic discriminant', *AIP Conf. Proc.*, Greece, Vol. 963, pp.1273–1277.

Luo, L.F. and Lu, J. (2007) 'Sequence pattern recognition in genome analysis', *AIP Conf. Proc*., Greece, Vol. 963, pp.1278–1281.

Noble, W.S., Kuehn, S., Thurman, R.Y.M. and Stamatoyannopoulos, J.A. (2005) 'Predicting the in vivo signature of human gene regulatory Sequences', *Bioinformatics*, Vol. 21, pp.i338–i343.

Sabo, P.J., Humbert, R., Hawrylycz, M., Wallace, J.C., Dorschner, M.O., McArthur, M. and Stamatoyannopoulos, J.A. (2004) 'Genome-wide identification of DNase1 hypersensitive sites using active chromatin sequence libraries', *Proc. Natl. Acad. Sci., USA*, Vol. 101, pp.4537–4542.

Zhang, L.R. and Luo, L.F. (2003) 'Splice site prediction with quadratic discriminant analysis using diversity measure', *Nucleic Acids Res.*, Vol. 31, pp.6214–6220.