Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Prediction for human transcription start site using diversity measure with quadratic discriminant

Jun Lu[1, 2,*] and Liaofu Luo[1, 3]

[1]Laboratory of Theoretical Biophysics, Faculty of Science and Technology, Inner Mongolia University, Hohhot 010021, P.R.China; [2]Inner Mongolia University of Technology, Hohhot, 010051, P.R.China; [3]Center for Theoretical Biology, Peking University, Beijing 100871, P.R.China; Jun Lu* - E-mail: lujun@imut.edu.cn; Phone: 81 471 6552521; * Corresponding author

received January 28, 2008; revised 17 March, 2008; accepted April 15, 2008; published April 28, 2008

Abstract:
The accurate identification of promoter regions and transcription start sites is a challenge to the construction of human transcription regulation networks. Thus, an efficient prediction method based on theoretical formulation is necessary for this purpose. We used the method of increment diversity with quadratic discriminant analysis (IDQD) to predict transcription start sites (TSS). The method produced sensitivity and positive predictive value of more than 65% with positives to negatives ratio of 1:58. The performance evaluation using Receiver Operator Characteristics (ROC) showed an auROC (area under ROC) of greater than 96%. The evaluation by Precision Recall Curves (PRC) showed an auPRC (area under PRC) of about 26% for positives to negatives ratio of 1:679 and about 64% for positives to negatives ratio of 1:113. The results documented in this approach are either better or comparable to other known methods.

Key words: promoter; transcription start site; increment of diversity; quadratic discriminant analysis

Background:
The accurate prediction of promoter sequence and transcription start site (TSS) is important in the construction of human transcription-regulation networks. Transcription initiation is a difficult problem which is dependent of DNA sequence and chromatin remodeling **[1]**. Lack of complete understanding on the distribution of chromatin remodeling in relation to known DNA sequences is a challenge in promoter and TSS prediction **[2]**. There are number of promoter prediction tools known till date **[3-6]**. Bajic and colleagues reviewed eight promoter prediction programs and reported that all of them showed a positive predictive value of less than 65% **[7]**. Thus, recognition of transcription starts site is a difficult task and not much progress has been made in its predictions.

Sonnenburg and colleagues used support vector machine with advanced sequence kernels for the accurate recognition of transcription start sites in human sequences with high prediction accuracy **[8]**. They evaluated four leading TSS prediction tools using performance evaluation parameters, area under the Receiver Operator Characteristics (ROC) and area under the Precision Recall Curves (PRC). The area under ROC was 90% at a chunk size resolution of 50 or 500 bases. However, the area under PRC was not significantly high in the analysis. This is attributed to variability of regulation element and complexity of regulation network in eukaryotic genomes. Regulation elements are short and are both easily erased and generated in evolution. The on and off of these elements during short evolutionary time can result in diversity of regulation elements and in their genomic arrangements.

The discovery of a large amount of alternative promoters in the human genome and our current knowledge on differentially regulated alternative TSSs in different tissues and families of genes increases the complexity of TSS recognition and prediction **[2, 9]**. Therefore, it is important to develop an efficient tool utilizing comprehensive sequence information for recognizing a variety of TSSs in a simple and unified formalism with high sensitivity and positive predictive value. We used increment diversity with quadratic discriminant analysis (IDQD) as a prediction method for splice junction identification. The method predicted exon/intron boundaries successfully for model genomes **[10]**. Here, we describe IDQD as a method to predict TSS with good accuracy.

Methodology:
Dataset
We used the database dbTSS version 4.0 **[11]** to create a dataset of 4254 genes to generate TSS data for training. This set is similar to the genes studied elsewhere **[8]**. We extracted a window of size 2000 (labeled by -1000,…, -1, +1, …, +1000) around the TSS (at site +1) for each gene. This set constituted the training positive set. We drew 10 negatives of 2000 bp at random from locations between 100bp downstream of the TSS and the end of the gene for each gene

as described elsewhere **[7]**. This set constituted the training positive set. We created a test set as described elsewhere **[8]** and used the set of genes 1024 that are new in dbTSS version 5.2 **[2]**. We then extracted sequences of size 2000 containing TSS for each gene. They consist of the test set (positives). We then identified classical TSSs as described by Bajic and colleagues **[7]**. The negatives are drawn from 100bp downstream of the TSS to the end of the gene in a shifted window of 2000 bp with step 50, 500, or 1000 bp. They consist of the negative test set. The ratio of the size of negative set to positive is 679 (for step 50) 113 (for step 500) and 58 (for step 1000), respectively. The negatives/positives ratios are comparable with those used in chunk method described elsewhere **[8]**. The performance evaluation of any prediction method is strongly dependent on the ratio of the size of negatives to positives. To make a fair comparison of our method with other programs we changed the step of shifted window to obtain different sizes of negative set and therefore the different ratios of negatives to positives.

IDQD algorithm

The characters of a sample, a sequence or a group of sequences, are described by a set of numbers is the assumption in the model. The i-th character is expressed by number n_i. n_i describes the number of certain base in a given site of a sequence or a group of sequences. We call n_i the character number or informational parameter of the sample (i=1,…,s). Consider the sequence X to be classified and define the diversity of sequence X as given in equation 1 (see supplementary material).

To give a classification of sequence X we should compare it with some standard samples (called standard diversity source S). Let the i-th character in standard source expressed by number m_i (i=1,…,s) where m_i is the sum of the i-th character number over all standard samples. The same definition is applied to diversity of standard source S. Likewise, the total diversity of the system X+S, D(X+S), can be defined in the same manner. The increment of diversity is defined by equation 2 in supplementary material. ID gives the relation of sequence X with standard source S. The smallest ID has the most intimate relation of X to S.

When there are r set of sequence characters, we have r feature variables ID_1 to ID_r and we need to integrate them by quadratic discriminant analysis. Given a problem (or a test) of classification, we average the increment of diversity (ID_j, j=1, …, r) over positive group or negative group in training set (denoted by μ_1 or μ_2 respectively) and thus deduce the corresponding covariance (denoted by $\mathbf{r}\times\mathbf{r}$ matrix $\sum_1$ or $\sum_2$ respectively). The increment of diversity is denoted by $\boldsymbol{R}$ for sequences to be classified. The discriminant function that differentiates with X belonging to positive group or negative group is given in equation 3 (see supplementary material). The sample X is classified into positive group as $\xi > \xi_0$ or negative group as $\xi \leq \xi_0$. In quadratic discriminant analysis, the threshold ξ_0 is taken as 0. However, due to the limited size of positive and negative group and the large difference between them, the optimal threshold ξ_0 is not 0 and it should be empirically determined. The ROC and PRC curves through varying the parameter ξ_0 are plotted for single-number performance evaluation. We initially calculated 28 ID parameters (for definitions see supplementary material) from promoter sequences.

Discussion:

We used two groups of performance measures to assess the accuracy on TSS prediction. The first group includes sensitivity (Sn), specificity (Sp), false positive rate (FPR), positive predictive value (PPV) and correlation coefficient (CC). Please see supplementary material for definitions for these performance parameters. The second group named "single-number" performance measure include auROC (area under the curve receiver operator characteristics) and auPRC (area under the curve precision recall curves).

The results of TSS prediction are summarized in Table 1 and Table 2 (see supplementary material for Tables). Table 1 (see supplementary material) shows the results depending on threshold ξ_0. The sensitivity decreases with increasing ξ_0 while positive predictive value and correlation coefficient increase with ξ_0. In test set of window step 1000bp, both sensitivity and PPV can achieve a higher value, higher than 65% under threshold ξ_0=3, which seems better than eight promoter prediction programs analyzed elsewhere **[7]**. We notice here that the distance between two negatives equal 1000 bp (the false positive rate 0.007 per 1000 bases) and the ratio of the number of positives to negatives is 1:58. This is lower than or comparable with the corresponding value in eight programs analyzed elsewhere **[7]**. In **[7]** a window of [-2000, +2000] was set relative to the TSS location. If we assume the distance between two negatives is 2000bp or more, the IDQD program will give higher prediction accuracy.

A detailed comparison of IDQD with other programs is made using auROC and auPRC. In IDQD we obtained auROC >96% and auPRC of about 26% for window step 50, auPRC of about 64% for step 500 and auPRC of about 76 for step 1000. We find these results are comparable with the performance of ARTS **[8]** and higher than other methods with the same positive to negative ratio (see Table 2 in supplementary material). The ROC and PRC curves for test set (corresponding to Table 2 under supplementary material) of window steps 500 are plotted in Figure 1.

The above prediction on typical TSSs of 1024 genes shows the effectiveness of IDQD algorithm. Results on TSS prediction were improved using algorithm with the same ID parameters for whole genome searching on classical TSSs (collected in database dbTSS2006) in chromosomes 4, 21 and 22. With window size 2000 bp and step 1000 bp , we obtained an auROC of 97% and auPRC of 65%. Both Sn and PPV exceeding 65% for optimal ξ_0 = 10.

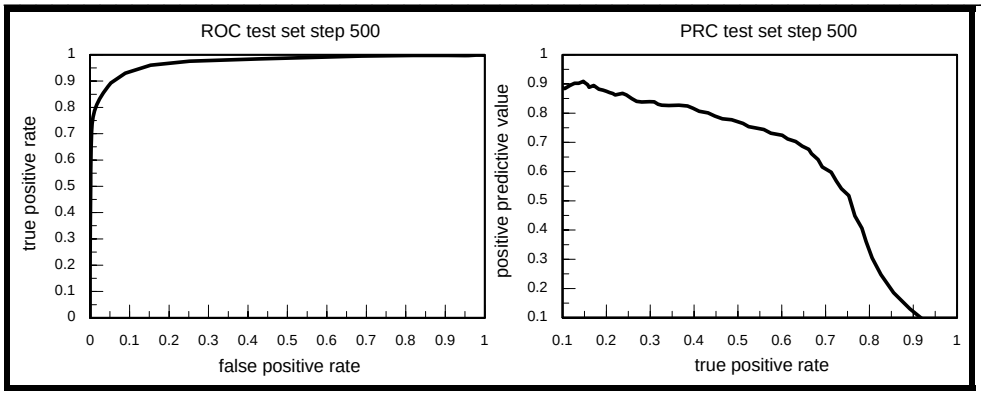

Figure 1: The ROC and PRC curves for test set of step 500.

IDQD algorithm and the choice of ID parameters

In a classification problem when the feature of the sample are denoted by a frequency distribution and compared with a standard source, the system can be described by diversity measure and classification is achieved by ID method. The program contains the calculation of a given character projected onto high dimensional space. When the size of standard sample set is large, making influences of fluctuation negligible, the diversity of standard samples includes accurate information about the frequency distribution of the character.

Thus, we use ID to evaluate the detailed difference between any sample and the standard set to find the optimal hyperplane for the classification of samples in multi-dimensional space. The estimation error is reduced in this approach since the difference among all real samples has been carefully considered in training set. Moreover, one can always define several different diversity functions as feature variables to describe sequence characteristics. The method is feasible enough for solving classification problem. Although the definition of diversity measure is similar to Shannon information in some respects, the ID method is far more different from mutual information and like information theoretical approaches in its application.

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 nonlinear discriminant function ξ through quadratic discriminant analysis. The only adjustable parameter that existed in IDQD algorithm is the threshold of ξ, namely ξ_0. Thus, the algorithm is easy to evaluate. The parameters are empirically determined in principle to obtain optimal evaluation. However, in ROC and PRC analysis the threshold ξ_0 is terated as a variable to plot curves for performance evaluation.

In IDQD, the ID selection is an important step. It is known that hexamer score and pentamer score were used as discriminant functions in promoter recognition **[12, 13]**. Many regulatory motifs in human promoters are composed of 6- to 8-nucleotide fragments **[14]**. The sequence information on base frequency and base correlation in TATA region and initiator region is also represented by hexamer distribution. The TSS signals exist mainly in a sequence of 2000 bp. In the present study, we use 6-mer frequencies in four 500 bp segments as the most important information (namely, diversities X_1 to X_4, see supplementary material) for the reorganization of promoter and TSS. A segment length of 500bp has been taken under consideration of the variation of 6-mer frequency in different regions of promoters. To emphasize the local information on base frequency and DNA structure in the vicinity of initiator, we introduced diversities on 5-mer and 4- mer frequencies in several consecutive 25 bp sequences. After comparing with other segment lengths we found that 25bp is a better choice for describing the local peculiarities of promoter sequences. However, both diversities on 5-mer and 4-mer frequencies (X_5 to X_9 and X_{10} to X_{12}, see supplementary material) are important for TSS recognition.

The use of a single diversity will lower prediction efficiency. The pentamer information is essentially related to initiator sequence while the tetramer information is responsible for the local deviation of DNA helix structure and di-nucleotide stacking energy **[15]**. Besides, G+C content (X_{13}, see supplementary material) and CpG DIMER content (X_{14}, see supplementary material) in 2000bp long sequence are also useful, since the base frequency distribution in promoters is different from non-promoters and the CpG content is important for recognizing a specific class of promoters.

Thus, we used 14 diversities in our analysis. Each of the diversity is introduced with two increments of diversity (ID). One increment is relative to the diversity of positive set of standard source and another is relative to the negative set of standard source. We then calculated the prediction accuracy using double increments and this is higher than single increment. Hence, the ID selection is guided by the performance evaluation (sensitivity, specificity, accuracy, PPV, auROC and auPRC). However, if two ID selections lead to the same performance measure we use the one with lower dimension. It should be noted that the dimension of the discriminant vector is the sum of dimensions of each diversity component and the dimension of the discriminant vector $\boldsymbol{R}$ =

(I_1, I_2,…,I_{28}) presented in this analysis is of order (4096×24)+420.

Acknowledgement:
The work was partly supported by National Science Foundation of China, No. 90403010 and Scientific Research Projects of Inner Mongolia's Universities, No. NJZY07065.

References:

[01] The ENCODE Project Consortium, *Nature*, 477: 799 (2007) [PMID: 17571346]
[02] R. Yamashita *et al*., *Nucleic Acids Res*., 34: D86 (2006) [PMID: 16381981]
[03] S. Knudsen, *Bioinformatics*, 15: 356 (1999) [PMID: 10366655]
[04] R. V. Davuluri *et al*., *Nature Genet*., 29: 412 (2001) [PMID: 11726928]
[05] V. B. Bajic *et al*., *Bioinformatics*, 18: 198 (2002) [PMID: 11836231]
[06] T. A. Down *et al*., *Genome Res*., 12: 458 (2002) [PMID: 11875034]
[07] V. B. Bajic *et al*., *Nat. Biotechnol*., 22: 1467 (2004) [PMID: 15529174]
[08] S. Sonnenburg *et al*., *Bioinformatics*, 22: e472 (2006) [PMID: 16873509]
[09] N. D. Trinklein *et al.*, *Genome Res.*, 17: 720 (2007) [PMID: 17567992]
[10] L. R. Zhang & L. F. Luo, *Nucleic Acids Res*., 31: 6214 (2003) [PMID: 14576308]
[11] Y. Suzuki *et al*., *Nucleic Acids Res*., 30: 328 (2002) [PMID: 11752328]
[12] R. V. Davuluri *et al*., *Nature Genetics*, 29: 412 (2001) [PMID: 11726928]
[13] P. É. Jacques *et al*., *BMC Bioinformatics*, 7: 423 (2006) [PMID: 17014715]
[14] X. H. Xie *et al*., *Nature*, 434: 338 (2005) [PMID: 15735639]
[15] L. Tsai *et al*., *J. Biomol. Struc. and Dynamics*, 20: 127 (2002) [PMID: 12144359]

Edited by D. R. Flower
Citation: Lu and Luo, Bioinformation 2(7): 316-321 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Supplementary material

Equations:

$$D(X) = D(n_1, n_2, \cdots, n_s) = N \log_2 N - \sum_{i=1}^{s} n_i \log_2 n_i \quad \longrightarrow \quad \textbf{(1)}$$

Where $(N = \sum_i n_i)$

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

$$\xi = \log \frac{p}{q} - \frac{\delta_1 - \delta_2}{2} - \frac{1}{2} \log \frac{|\Sigma_1|}{|\Sigma_2|} \quad \longrightarrow \quad \textbf{(3)}$$

$$\delta_i = (\boldsymbol{R} - \boldsymbol{\mu_i})^T \Sigma_i^{-1} (\boldsymbol{R} - \boldsymbol{\mu_i}), \ (i=1,2)$$

where p and q denote the size of positive and negative group respectively and $|\Sigma|$ is the determinant of matrix Σ.

	ξ_0	Sn(%)	Sp(%)	PPV(%)	CC
Training set	3	76.75	97.13	72.78	0.72
	0	81.05	95.96	66.74	0.71
	-3	87.75	93.14	56.11	0.67
Test set, Step 50	3	74.98	99.27	13.09	0.31
	0	78.56	98.70	8.16	0.25
	-3	85.83	96.46	3.45	0.17
Test set, Step 500	3	75.96	99.27	47.95	0.60
	0	79.57	98.72	35.45	0.53
	-3	85.74	96.47	17.73	0.38
Test set, Step 1000	3	78.04	99.30	65.58	0.71
	0	81.46	98.69	51.50	0.64
	-3	87.97	96.52	30.26	0.50

Table 1: Prediction on typical TSSs using IDQD and evaluated by Sn, Sp, PPV and CC. Calculations are given in three thresholds, ξ_0=0, +3 and -3. When ξ_0=3 both sensitivity and positive predictive values, exceed 65% in test set of step 1000 bp where the ratio of the size of positive set to negative is 1:58.

	Area under ROC (%)	Area under PRC (%)
Training set	96.69	81.28
Test set, step 50	96.63 (ARTS 92.77; Eponine 88.48; McPromoter 92.55; First EF 71.29)	26.15 (ARTS 26.18; Eponine 1.79; McPromoter 6.32; First EF 6.54)
Test set, step 500	97.28 (ARTS 93.44; Eponine 91.51; McPromoter 93.59; First EF 90.25)	64.06 (ARTS 57.19; Eponine 40.80; McPromoter 24.23; First EF 40.89)
Test set, step 1000	98.10 (ARTS 93.85; Eponine 92.07; McPromoter 93.80; First EF 92.86)	76.03 (ARTS 67.71; Eponine 52.75; McPromoter 35.43; First EF 56.00)

Table 2: Prediction on typical TSSs using IDQD and evaluated by developing auROC and auPRC is shown. The results by Sonnenburg and colleagues **[8]** are given in brackets for comparison. Our test set of step 50 (the ratio of the number of positives to negatives 1567:1063726) is comparable with chunk size 50 in the case of this reference where the ratio of the number of positives to negatives is 1588:1087664. Our test set of step 500 (the ratio of the number of positives to negatives 940:106037) is comparable with chunk size 500 where the ratio of the number of positives to negatives is 943:108783. The test set of step 1000 (the ratio of the number of positives to negatives 906:52853) is also comparable with the chunk size 1000 in this case.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net

Prediction Model

Definition:

ID definition:

6 mer:

The difference of sequence construction between promoters and non-promoters were used to choose 6-mer frequencies as the main source of information for TSS identification. We used 6-mer frequencies in a shifted window of step 1 between -1000bp and -501bp, we defined diversity D(X) of sequence X as X_1, and defined ID between X_1 and diversity of standard source in positive (negative) training set as I_1 (I_2). Similarly we used 6-mer frequencies in shifted windows between -500bp to -1bp, +1bp (TSS) to +500bp, and +501bp to +1000bp, we defined diversity of sequence X as X_2, X_3 and X_4, respectively. The corresponding IDs with positive (negative) training sets as I_3(I_4), I_5(I_6) and I_7(I_8). The dimension of each of above IDs is to the order of 4^6 = 4096.

5 mer:

The site-specific information in initiator near TSS were used to choose 5-mer frequencies in consecutive four 25 bp-long sequences in -200bp:-101bp (defining diversity X_5), -100bp:-1bp (defining diversity X_6), +1bp:+100bp (defining diversity X_7), +101bp:+200bp (defining diversity X_8) and +201bp:+300bp (defining diversity X_9) as the sources of information for recognizing TSS. The diversity X_5 is defined by 5-mer frequencies in -200bp:-176p, -175bp:-151bp, -150bp:-126bp and -125bp:-101bp, and the corresponding ID with positive (negative) training sets I_9(I_{10}) has dimension 4×4^5 = 4096. The other four diversities X_6, X_7, X_8, X_9 are defined in the same way and the corresponding IDs are denoted by I_{11} to I_{18}.

4 mer:

The structural information near the TSS were used to choose 4-mer frequencies in consecutive sixteen 25 bp-long sequences in -600bp:-201bp, -200bp:+200bp and +201bp:+600bp. The three diversities are denoted by X_{10}, X_{11}, X_{12} and the corresponding IDs by I_{19} to I_{24}. Each ID has dimension 16×4^4 = 4096.

G+C content:

The G+C contents in each of 10bp interval from -1000bp to +1000bp were used to define diversity X_{13} and the corresponding IDs I_{25} and I_{26} (with dimension 200). The CpG content was calculated in each of 200bp interval from -1000bp to +1000bp and thus was used to define diversity X_{14} and the corresponding IDs I_{27} and I_{28} (with dimension 10).

Discriminant vector:

A Discriminant vector for sequence X is defined by R = I_1, I_2, …, I_{28}. $X \in G_i$ (i = 1, positive group and i = 2, negative group) and the average of **R** over positive group or negative group is $\boldsymbol{\mu}_1$ or $\boldsymbol{\mu}_2$ and the corresponding covariance is Σ_1 or Σ_2 in Equation (3).

ROC and PRC

The ROC and PRC curves (corresponding to Table 2) for test set step 500 is developed in Figure 1.

Performance measure

$$Sn = [TP/(TP+FN)]\times 100\% \qquad PPV = [TP/(TP+FP)]\times 100\%$$

$$Sp = [TN/(TN+FP)]\times 100\% \qquad CC = \frac{(TP\times TN)-(FN\times FP)}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}}$$

$$FRP = FP/(TN+FP) = 1-Sp$$

TP = true positive; TN = true negative; FN = false negative; FP = false positive