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.

ISSN 0973-2063
Bioinformation 2(7): 316-321 (2008)

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 $4\times4^5 = 4096$. 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 $16\times4^4 = 4096$.

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 $R = I_1, I_2, \ldots, I_{28}$. $X \in G_i$ (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

ISSN 0973-2063
Bioinformation 2(7): 316-321 (2008)