Interdiscip Sci Comput Life Sci (2014) 6: 241–249
DOI: 10.1007/s12539-013-0195-4

Exon Skipping Event Prediction based on Histone Modifications

Wei Chen[1*], Hao Lin[2*], Pengmian Feng[3], Jinpeng Wang[1]
[1](Center for Genomics and Computational Biology, School of Sciences, Hebei United University, Tangshan 063000, China)
[2](Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China)
[3](School of Public Health, Hebei United University, Tangshan 063000, China)

Received 3 December 2013 / Revised 30 December 2013/ Accepted 7 February 2014

Abstract: Alternative splicing is a tissue and developmental stage specific process and greatly increases the biodiversity of proteins. Besides the trans- and cis-factors on the genome level, the process of RNA splicing is also regulated by epigenetic factors. In the present work, we proposed a new method to predict exon skipping events by using the histone methylation and acetylation information. The maximum relevance minimum redundancy method followed by incremental feature selection was performed to select the optimal feature set. Based on the optimized features, our method obtained an overall accuracy of 68.5% in a 10-fold cross validation test for exon skipping event prediction. It is anticipated that our method may become a useful tool for alternative splicing events prediction and the selected optimal features will provide insights into the regulatory mechanisms of epigenetic factors in alternative splicing.
Key words: alternative splicing, histone methylation, histone acetylation, feature selection, quadratic discriminant function.

1 Introduction

Alternative splicing is a widespread event in eukaryotic species and is responsible for much of the complexity of the proteome (Black 2003). Data from next generation sequencing indicated that more than 90% of the human genes undergo alternative splicing and produce multiple transcript variants from a single gene locus (Pan *et al.*, 2008; Wang *et al.*, 2008a). These variants exhibit different molecular functions and structural properties (Kelemen *et al.*, 2013). A large number of genetic diseases are also closely associated with the defects of RNA splicing (David *et al.*, 2010; Garcia-Blanco *et al.*, 2004; Lai and Greenberg 2013).

Elimination of specific introns and exons is the key process of RNA splicing, which is performed by a large macromolecule known as spliceosome. Studies over the past decades have demonstrated that it can be regulated on two different levels, the genome level and the epigenome level (Luco *et al.*, 2011).

On the genome level, there are lots of splicing motifs (i.e., splicing enhancers, splicing silencers, etc.) that can be recognized and bound by snRNAs and proteins to form the spliceosome (Hoskins and Moore 2012). Based on the information encoded in the genome level, a lot of alternative splicing prediction models were proposed (Zhang and Luo, 2003; Wang and Burge 2008; Zhang *et al.*, 2010). However, the number of the tissue and developmental stage specific splicing events that can be clearly explained by these models is very limited (Barash *et al.*, 2010).

In fact, the eukaryotic genome is packaged in the form of nucleosomes. The histone components of the nucleosome undergo multiple post-translational covalent modifications including acetylation and methylation (Bernstein *et al.*, 2007; Kouzarides 2007). A series of studies have reported that the information from the epigenome level also contributes to the regulation of RNA splicing (Luco *et al.*, 2011). Schor and his colleagues demonstrated that the exon skipping correlates with H3K9ac and H3K36me (Schor *et al.*, 2009). Luco *et al.*, also found that histone methylation could medicate the splicing site selection in fibroblast growth factor receptor 2 gene (FGFR2) (Luco *et al.*, 2010). Through analyzing the relationship between histone modification and RNA splicing in human H1 and IMR90 cells, Shindo *et al.*, revealed that H3K36me3 and H3K79me1 correlate with the inclusion and exclusion of alternative exons (Shindo *et al.*, 2013). Recently, based on histone methylation and acetylation information, Enroth *et al.*, proposed a model to predict

*Corresponding authors.
E-mail: chenweiimu@gmail.com;
greatchen@heuu.edu.cn (Wei Chen);
hlin@uestc.edu.cn (Hao Lin)

the most common alternative splicing mode, exon skipping events (Fig. 1), and obtained an average accuracy of 72% for 27% of the exons (Enroth *et al.*, 2012). All these results indicate that the information encoded in the epigenome level play important roles in RNA splicing regulation.

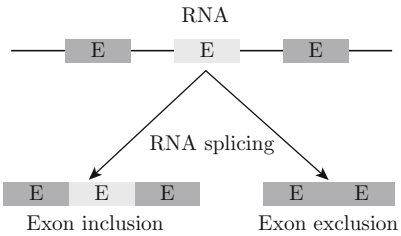

Fig. 1 Example of exon skipping. Exons (E) are indicated by rectangles and introns are shown by lines. In exon skipping event, the middle exon might be reserved in the transcript, namely, exon inclusion (left bottom panel), while in some other tissues or developmental stages it will be skipped, namely, exon exclusion (right bottom panel).

In the present study, we presented a discriminative computational framework to identify exon skipping events according to the histone modification information. The mRMR (minimal redundancy maximal relevance) method was used to select the optimized features. The quadratic discriminant (QD) function was proposed to perform the prediction. The predictive performance of the proposed method based on optimized features was then evaluated in a 10-fold cross validation test.

2 Materials and methods

2.1 Dataset

The benchmark dataset used in the present study was constructed by Enroth *et al.* (Enroth *et al.*, 2012). According to the following procedures, they obtained a dataset describing histone acetylation and methylation in exon and intron regions of exon skipping events (Enroth *et al.*, 2012). Firstly, based on gene expression data (Oberdoerffer *et al.*, 2008), exons were annotated as 'excluded' or 'included' according to the quote between their expression and the average gene expression. Subsequently, exons that ranked as first or last in any transcript were removed, and only exons longer than 50 bp with flanking introns longer than 360 bp and no overlap to other exons were considered.

The 20 kinds of histone acetylations (Barski *et al.*, 2007) and 18 kinds of histone methylations (Wang *et al.*, 2008b) were then considered for regions centered over, preceding as well as succeeding each exon. In this procedure, exons with no histone acetylation or methylation modification present were removed from the dataset. The histone modification information centered over, preceding and succeeding the exon was discretized as present (noted by "1") or absent (noted by "0") over the three regions (Enroth *et al.*, 2012). As a result, a benchmark dataset containing 12,692 'included' exons (left panel in Fig. 1) and 11,165 'excluded' exons (right panel in Fig. 1) with histone acetylation and methylation information was obtained and it was formulated as

$$\mathrm{E} = \mathrm{E}^{+} \cup \mathrm{E}^{-} \tag{1}$$

where $\cup$ represents the symbol for "union" in the set theory, and

$$\begin{cases} \mathrm{E}^{+} \ containing \ 12\,692 \ 'included' \ exons \\ \mathrm{E}^{-} \ containing \ 11\,165 \ 'excluded' \ exons \end{cases} \tag{2}$$

2.2 Sample representation

According to the 38 kinds of histone acetylation and methylation information centered over (or middle), preceding and succeeding exons, all the 'excluded' and 'included' exons in the dataset can be represented by a (3×38)=114-dimensional vector as,

$$\mathrm{Z} = [\varepsilon_1, \varepsilon_2, \varepsilon_3, \cdots, \varepsilon_{114}]^{\mathrm{T}} \tag{3}$$

where the components in the vector are 1 (present of histone modification) or 0 (absent of histone modification), T is the transpose operator, the first 3 components reflect the histone modification information of H3K27me3 centered over, preceding and succeeding the exons, the second 3 components reflect H3K4me2, and so forth (c.f. Supporting information S1).

2.3 Quadratic discriminant function

The quadratic discriminant (QD) function has been widely used in the realm of bioinformatics, such as out membrane protein prediction (Lin 2008), splice site prediction (Zhang and Luo 2003), nucleosome positioning prediction (Chen *et al.*, 2010; Chen *et al.*, 2012) and other DNA and protein sequence pattern recognitions as reviewed in Ref. (Lv *et al.*, 2010). The detailed deduction of QD according to Bayesian Theorem has been described in our previous work (Lv *et al.*, 2010). Therefore, we briefly described its mathematical principles as follows.

Given a sample X represented by the histone modification feature variables (ε_1 to ε_r), we obtain a r-dimensional feature vector $Z = [\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_r]$. Whether the sample X belongs to E^{+} or to E^{-} can be judged by the followings,

$$\begin{cases} X \in \mathrm{E}^{+} & if \ \xi > 0 \\ X \in \mathrm{E}^{-} & if \ \xi \le 0 \end{cases} \tag{4}$$

ξ is the quadratic discriminant function defined as (Zhang and Luo 2003),

$$\xi = \log_2 \frac{N_1}{N_2} - \frac{\delta_1 - \delta_2}{2} - \frac{1}{2} \log_2 \frac{|C_1|}{|C_2|} \tag{5}$$

S1 The 114 features obtained according to histone acetylation and methylation

Order	Feature	Order	Feature	Order	Feature
1	H3K27me3	39	H3K79me1_succ	77	H2AK9ac_prec
2	H3K27me3_prec	40	H3K9me2	78	H2AK9ac_succ
3	H3K27me3_succ	41	H3K9me2_prec	79	H2BK5ac
4	H3K4me2	42	H3K9me2_succ	80	H2BK5ac_prec
5	H3K4me2_prec	43	H4K20me1	81	H2BK5ac_succ
6	H3K4me2_succ	44	H4K20me1_prec	82	H3K27ac
7	H3K79me3	45	H4K20me1_succ	83	H3K27ac_prec
8	H3K79me3_prec	46	H3K27me2	84	H3K27ac_succ
9	H3K79me3_succ	47	H3K27me2_prec	85	H4K12ac
10	H3R2me1	48	H3K27me2_succ	86	H4K12ac_prec
11	H3R2me1_prec	49	H3K4me1	87	H4K12ac_succ
12	H3R2me1_succ	50	H3K4me1_prec	88	H4K91ac
13	H4R3me2	51	H3K4me1_succ	89	H4K91ac_prec
14	H4R3me2_prec	52	H3K79me2	90	H4K91ac_succ
15	H4R3me2_succ	53	H3K79me2_prec	91	H2BK120ac
16	H2BK5me1	54	H3K79me2_succ	92	H2BK120ac_prec
17	H2BK5me1_prec	55	H3K9me3	93	H2BK120ac_succ
18	H2BK5me1_succ	56	H3K9me3_prec	94	H3K14ac
19	H3K36me1	57	H3K9me3_succ	95	H3K14ac_prec
20	H3K36me1_prec	58	H4K20me3	96	H3K14ac_succ
21	H3K36me1_succ	59	H4K20me3_prec	97	H3K36ac
22	H3K4me3	60	H4K20me3_succ	98	H3K36ac_prec
23	H3K4me3_prec	61	H2AK5ac	99	H3K36ac_succ
24	H3K4me3_succ	62	H2AK5ac_prec	100	H4K16ac
25	H3K9me1	63	H2AK5ac_succ	101	H4K16ac_prec
26	H3K9me1_prec	64	H2BK20ac	102	H4K16ac_succ
27	H3K9me1_succ	65	H2BK20ac_prec	103	H2BK12ac
28	H3R2me2	66	H2BK20ac_succ	104	H2BK12ac_prec
29	H3R2me2_prec	67	H3K23ac	105	H2BK12ac_succ
30	H3R2me2_succ	68	H3K23ac_prec	106	H3K18ac
31	H3K27me1	69	H3K23ac_succ	107	H3K18ac_prec
32	H3K27me1_prec	70	H3K9ac	108	H3K18ac_succ
33	H3K27me1_succ	71	H3K9ac_prec	109	H3K4ac
34	H3K36me3	72	H3K9ac_succ	110	H3K4ac_prec
35	H3K36me3_prec	73	H4K8ac	111	H3K4ac_succ
36	H3K36me3_succ	74	H4K8ac_prec	112	H4K5ac
37	H3K79me1	75	H4K8ac_succ	113	H4K5ac_prec
38	H3K79me1_prec	76	H2AK9ac	114	H4K5ac_succ

$$\delta_i = (Z - \mu_i)^T C_i^{-1} (Z - \mu_i) \tag{6}$$

where N_1 and N_2 are the number of samples in E^+ and E^-, respectively. δ_i is the squared Mahalanobis distance between Z and μ_i. μ_i is the average of Z over all samples of E^+ ($i = 1$) or E^- ($i = 2$). $|C_i|$ and C_i^{-1} are respectively the determinant and inverse of the corresponding covariance matrix C_i denoted as

$$C_i = \begin{bmatrix} c_{11}^i & c_{12}^i & \cdots & c_{1r}^i \\ c_{21}^i & c_{22}^i & \cdots & c_{2r}^i \\ \vdots & \vdots & \vdots & \vdots \\ c_{r1}^i & c_{r2}^i & \cdots & c_{rr}^i \end{bmatrix} \tag{7}$$

The $r \times r$ elements in C_i are given by

$$c_{jk}^i = \frac{1}{N_i - 1} \sum_{u=1}^{N_i} (\varepsilon_{uj} - \overline{\varepsilon_j^i})(\varepsilon_{uk} - \overline{\varepsilon_k^i}), \quad (j, k = 1, 2, \cdots, r) \tag{8}$$

$$\overline{\varepsilon_j^i} = \frac{\sum_{u=1}^{N_i} \varepsilon_j^i}{N_i} \quad (j = 1, 2, \cdots, r) \tag{9}$$

2.4 Performance evaluation

Three cross-validation methods, namely, subsampling test, independent dataset test and jackknife test are often employed to evaluate the predictive capability of a predictor. Among the three methods, jackknife test is deemed the most objective and rigorous one that can always yield a unique outcome as demonstrated in a comprehensive review (Chou and Shen 2008), and has been widely and increasingly adopted by investigators to examine the quality of various predictors (Chen *et al.*, 2013; Feng *et al.*, 2013; Lin *et al.*, 2013). However, since the current study would involve feature selection as described below, to reduce the computational time, the 10-fold cross-validation test was adopted in the current study. In the 10-fold cross-validation test, the benchmark dataset was randomly partitioned into 10 nearly equal subsets, out of which nine subsets are used for training and the remaining one for testing. This procedure was repeated ten times and the final prediction result is the average accuracy of the ten tests.

The performance of the proposed model was evaluated using sensitivity (Sn), specificity (Sp) and overall accuracy (OA), which are expressed as

$$Sn = \frac{TP}{TP + FN} \tag{10}$$

$$Sp = \frac{TN}{TN + FP} \tag{11}$$

$$OA = \frac{TP + TN}{TP + FN + TN + FP} \tag{12}$$

where TP, TN, FP and FN represent the number of the correctly recognized exon inclusion samples, the number of the correctly recognized exon exclusion samples, the number of exon exclusion samples recognized as exon inclusion samples and the number of exon inclusion samples recognized as exon exclusion samples, respectively.

2.5 Feature selection

Redundant and noisy information would cause poor prediction results and increase computational time. To improve the prediction quality and gain deeper insights into the contribution of histone modifications to RNA splicing, we evaluated the features in Z using the mRMR (minimal redundancy maximal relevance) method (Peng *et al.*, 2005). The basic idea of mRMR is to rank the features according to their relevance to the class and the redundancy among the features themselves. The ranked feature with a smaller index indicates that it has a better trade-off between the maximum relevance and minimum redundancy.

The relevance between two variables ε_i and ε_j can be defined by the mutual information I,

$$I(\varepsilon_i, \varepsilon_j) = \sum_{i,j} p(\varepsilon_i, \varepsilon_j) \log \frac{p(\varepsilon_i, \varepsilon_j)}{p(\varepsilon_i) p(\varepsilon_j)} \tag{13}$$

where $p(\varepsilon_i, \varepsilon_j)$ is the joint probabilistic density, $p(\varepsilon_i)$ and $p(\varepsilon_j)$ are the marginal probabilities, respectively.

Suppose that M is the already-selected feature set containing m features, and N the to-be-selected feature set containing n features. The relevance D between the feature ε in set N and the class c can be defined as

$$D = I(\varepsilon, c) \tag{14}$$

The redundancy R between the feature ε in N and all the features in M can be calculated by

$$R = \frac{1}{m} \sum_{\varepsilon_i \in M} I(\varepsilon, \varepsilon_i) \tag{15}$$

So the feature ε_j in the set N with the maximum relevance and minimum redundancy can be determined by

$$\max_{\varepsilon_j \in N} \left[I(\varepsilon_j, c) - \frac{1}{m} \sum_{\varepsilon_i \in M} I(\varepsilon_j, \varepsilon_i) \right], \quad (j = 1, 2, \cdots, n) \tag{16}$$

Since there are 114 features in the present work, the feature evaluation was continued 114 rounds. After these evaluations, we obtained a new feature set S,

$$\mathrm{S} = \{f_1', f_2', f_3', \cdots, f_{114}'\} \tag{17}$$

where each feature in S has a subscript index indicating at which round the feature is selected. The more importance the feature is, the smaller its subscript index is.

Based on the ranked features, the Incremental Feature Selection (IFS) (He *et al.*, 2010) was used to determine the optimal number of features. During the IFS procedure, features in the ranked feature set S were added one by one from lower to higher rank. A new feature set will be constructed when one feature had been added. Thus, the N feature sets thus formed would be composed of N ranked features. The τ-th feature set can be formulated as

$$\mathrm{S}_\tau = \{f_1, f_2, \cdots, f_\tau\} \quad (1 \le \tau \le N) \tag{18}$$

For each of the N feature sets, a QD prediction model (cf. Eqs. (5)-(6)) was constructed and examined with the 10-fold cross-validation.

By doing so, we obtained an IFS curve in a 2D Cartesian coordinate system with index τ as its abscissa (or X-coordinate) and the overall accuracy as its ordinate (or Y-coordinate). The optimal feature set is defined by

$$S_\Theta = \{f_1, f_2, \cdots, f_\Theta\} \tag{19}$$

with which the IFS curve reaches its peak. In other words, in the 2D coordinate system, when $X = \Theta$ the value of OA is the maximum. Thus, we can use the Θ features in Eq. (19) to build the final predictor.

3 Results and discussion

3.1 Prediction performance

To identify the key features for exon skipping event prediction, we used the mRMR algorithm and IFS approach as described in Materials and Methods section. After running the mRMR program, the 114 features were ranked according to their relevance to the class of samples (Supporting information S2). Such ranked features were then used in the following IFS procedure for the optimal feature set selection.

By adding the 114 ranked features one by one according to the evaluations from mRMR, we built 114 individual QD predictors for the 114 sub-feature sets. We then tested the prediction performance for each of the 114 predictors and plotted the IFS curve as shown in Fig. 2, from which we can see that, when the top ranked 71 features were used, the overall accuracy reached its peak, i.e., OA= 68.5%, with the sensitivity of 68.9% and specificity of 66.7%. As a comparison, the predictive results obtained by using all the 114 features were also listed in Table 1.

In other words, we have $\Theta = 71$ (cf. Eq. (19)) and the optimal feature set for the current biological system should be

$$S_{71} = \{f_1, f_2, \cdots, f_{71}\} \tag{20}$$

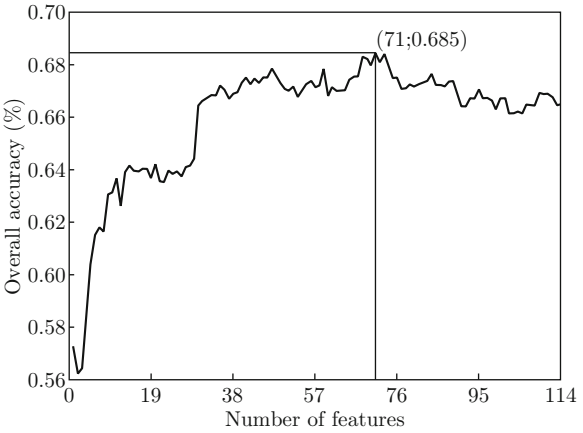

Fig. 2 A plot to show the IFS procedure. When the top 71 of the 114 features were used to perform prediction, the overall accuracy reached its peak of 0.685.

Table 1 Results of exon skipping event prediction using different parameters

Number of Features	Sn (%)	Sp (%)	OA (%)
114	65.3	67.1	66.5
71	68.9	66.7	68.5

To provide an overall view, a distribution of the 114 features and their roles for the prediction model is given in Fig. 3, where the light gray boxes indicate the features that were not contained in the optimal feature set S_{71}. The gray and black boxes indicate the features that were included in the optimal feature set S_{71}: the 34 features in gray boxes were positively correlated with exclusion event, while the other 37 features in black boxes were positively correlated with inclusion event.

3.2 Molecular mechanism analysis

As shown in Fig. 3, the 71 optimal features belong to 17 kinds of histone acetylation and 13 kinds of histone methylation. The regulatory roles of some of these histone modifications in alternative splicing have been demonstrated in a series of studies (Schor *et al.*, 2009; Kolasinska-Zwierz *et al.*, 2009; Luco *et al.*, 2011) and were also reviewed in a recent work (Zhou *et al.*, 2013).

The roles of H3K9me2 and H3K27me3 in RNA alternative splicing were found in human Fibronectin (FN1) gene (White *et al.*, 2008; Zhou *et al.*, 2013). In FN1, the heterochromatin-associated protein HP1α recognizes the H3K9me2 and H3K27me3 marks and slows down transcriptional elongation, resulting in exon inclusion event.

H3K9me3 was also found in association with alternative splicing as revealed in previous studies (Ponta *et al.*, 2003; Screaton *et al.*, 1993). In Human CD44 gene, H3K9me3 marks are recognized by HP1γ, which facilitates inclusion of the alternative exons by reducing the

S2 The 114 ranked features according to their relevance to the class of samples

Order	Feature	Order	Feature	Order	Feature
1	H2BK5ac	39	H3K36ac	77	H2BK120ac_prec
2	H3K9ac_succ	40	H3K27me2_succ	78	H3K79me2_prec
3	H4K20me3_succ	41	H3K27me1	79	H2BK5ac_succ
4	H3K9me2	42	H4K16ac_succ	80	H2BK12ac
5	H3K14ac_prec	43	H3K4me3_prec	81	H3K27ac_prec
6	H3R2me1_prec	44	H3R2me2	82	H3K4me2
7	H3R2me2_succ	45	H3K9me2_prec	83	H3K36ac_succ
8	H2AK9ac_prec	46	H3K79me2_succ	84	H3K18ac
9	H3K36me3_succ	47	H2AK5ac_prec	85	H2BK20ac_succ
10	H3K23ac_prec	48	H4K8ac_prec	86	H3K4me1
11	H3K27me2_prec	49	H3K36me1	87	H3K9me1_prec
12	H3K36me1_succ	50	H3K18ac_succ	88	H2AK5ac
13	H3K9me3_succ	51	H4R3me2	89	H4K12ac
14	H2BK12ac_succ	52	H3K79me3_prec	90	H4K20me1_succ
15	H3R2me2_prec	53	H4K8ac_succ	91	H2BK20ac_prec
16	H4K20me3_prec	54	H2AK9ac	92	H3K4ac
17	H4R3me2_succ	55	H3K27me3_prec	93	H2BK120ac_succ
18	H3K79me2	56	H3K27ac_succ	94	H3K36me3
19	H3K9ac_prec	57	H3K9ac	95	H3K9me1_succ
20	H2AK9ac_succ	58	H3K36me3_prec	96	H3K79me1_succ
21	H3K27me3	59	H3K23ac	97	H3K27ac
22	H3K14ac_succ	60	H3K4ac_succ	98	H3K4me1_prec
23	H3K9me3_prec	61	H3K4me2_prec	99	H4K20me1
24	H4R3me2_prec	62	H4K5ac_prec	100	H4K5ac
25	H3K4me3	63	H3K14ac	101	H2BK120ac
26	H4K12ac_succ	64	H3K36ac_prec	102	H3K79me1_prec
27	H3K36me1_prec	65	H3K27me3_succ	103	H3K4me1_succ
28	H4K20me3	66	H3K4me2_succ	104	H4K8ac
29	H2AK5ac_succ	67	H2BK5ac_prec	105	H4K91ac_prec
30	H4K16ac_prec	68	H3R2me1	106	H3K9me1
31	H3K23ac_succ	69	H3K27me1_succ	107	H2BK5me1_succ
32	H3K79me3_succ	70	H3K18ac_prec	108	H2BK20ac
33	H3K4me3_succ	71	H4K16ac	109	H4K91ac_succ
34	H3K9me2_succ	72	H4K5ac_succ	110	H2BK5me1
35	H2BK12ac_prec	73	H3K27me2	111	H3K79me1
36	H3R2me1_succ	74	H3K4ac_prec	112	H2BK5me1_prec
37	H4K12ac_prec	75	H3K27me1_prec	113	H4K20me1_prec
38	H3K9me3	76	H3K79me3	114	H4K91ac

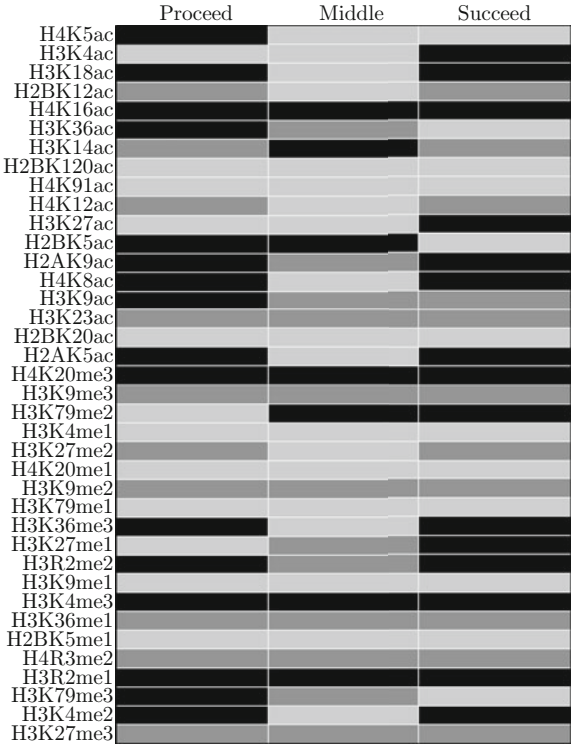

Fig. 3 A distribution overall view for the 114 features. The features that were included in the optimal feature set S_{71} are shown in the gray and black boxes: the former was positively correlated with exon exclusion event, while the latter positively correlated with exon inclusion event. Those features that were not in the optimal feature set S_{71} are shown in the light gray boxes.

local transcriptional elongation rate (Saint-Andre *et al.*, 2011; Zhou *et al.*, 2013).

H3K6me3 is enriched in exon regions and its function in alternative splicing has been demonstrated in FGFR2 (Warzecha *et al.*, 2012). The enriched H3K6me3 in FGFR2 gene can be recognized by the MORF-related gene 15 (MRG15) protein, which directly recruits the polypyrimidine tract-binding protein (PTB) to the intronic splicing silencer element surrounding exon IIIb to repress its inclusion in mesenchymal cells (Orr-Urtreger *et al.*, 1993; Zhou *et al.*, 2013). The H3K36me3 mark can also be recognized by thechromatin-associated protein Psip1 to regulate alternative splicing (Pradeepa *et al.*, 2012; Zhou *et al.*, 2013).

4 Conclusion

The process of RNA splicing occurs co-transcriptionally and the complicated chromatin environment can also affect RNA splicing through epigenetic factors, such as nucleosome positioning, histone methylation, histone acetylation, etc..

In the present study, based on the histone methylation and histone acetylation information, we proposed a novel method to predict the exon skipping events by using the quadratic discriminant function. By means of the feature selection algorithm, an optimal set of 71 features were selected. With the optimal features, the proposed method achieved an accuracy of 68.5% in the benchmark dataset, indicating that these features contribute significantly for the prediction of exon skipping event.

In other words, the result also demonstrates that the epigenome information is not enough to explain the complete alternative splicing patterns as well. RNA splicing is co-regulated by both genomic and epigenomic factors. There exists communications between the two levels of alternative splicing regulation by virtue of co-transcriptional deposition of protein factors (Zhou *et al.*, 2013; Shindo *et al.*, 2013; de Almeida *et al.*, 2011). Recently, the concept of pseudo dinucleotide composition (PseDNC) has been proposed to deal with DNA sequences (Chen *et al.*, 2013). Encouraged by the success of PseDNC, in the near future, we will also apply the PseDNC to identify exon skipping events by combining the information from both genome and epigenome levels and develop novel methods to precisely demonstrate the mechanism of alternative splicing.

Acknowledgements

The authors are grateful to Dr. Enroth Stefan for sharing the processed benchmark datasets. The authors also wish to thank the reviewers for their constructive comments. This work was supported by the National Nature Scientific Foundation of China (Nos. 61100092 and 61202256), the Nature Scientific Foundation of Hebei Province (No.C2013209105) and Science and Technology Department of Hebei Province (No. 132777133).

Conflict of Interest

The authors have declared that no competing interests exist.

References

[1] Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., Frey, B.J. 2010. Deciphering the splicing code. Nature 465, 53-59.

[2] Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., Zhao, K. 2007 High-resolution profiling of histone methylations in the human genome. Cell 129, 823-837.

[3] Bernstein, B.E., Meissner, A., Lander, E.S. 2007 The mammalian epigenome. Cell 128, 669-681.

[4] Black, D.L. 2003. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 72, 291-336.

[5] Chen, W., Feng, P.M., Lin, H., Chou, K.C. 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 41, e68.

[6] Chen, W., Lin, H., Feng, P.M., Ding, C., Zuo, Y.C., Chou, K.C. 2012. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. Plos One 7, e47843.

[7] Chen, W., Luo, L.F., Zhang, L.R. 2010. The organization of nucleosomes around splice sites. Nucleic Acids Res 38, 2788-2798.

[8] Chou, K.C., Shen, H.B. 2008. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3, 153-162.

[9] David, C.J., Chen, M., Assanah, M., Canoll, P., Manley, J.L. 2010. HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. Nature 463, 364-368.

[10] de Almeida, S.F., Grosso, A.R., Koch, F., Fenouil, R., Carvalho, S., Andrade, J., Levezinho, H., Gut, M., Eick, D., Gut, I., Andrau, J.C., Ferrier, P., Carmo-Fonseca, M. 2011. Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. Nat Struct Mol Biol 18, 977-983.

[11] Enroth, S., Bornelov, S., Wadelius, C., Komorowski, J. 2012. Combinations of histone modifications mark exon inclusion levels. PloS One 7, e29911.

[12] Feng, P.M., Chen, W., Lin, H., Chou, K.C. 2013 iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Anal Biochem 442, 118-125.

[13] Garcia-Blanco, M.A., Baraniak, A.P., Lasda, E.L. 2004. Alternative splicing in disease and therapy. Nat Biotechnol 22, 535-546.

[14] He, Z., Zhang, J., Shi, X.H., Hu, L.L., Kong, X., Cai, Y.D., Chou, K.C. 2010. Predicting drug-target interaction networks based on functional groups and biological features. Plos One 5, e9603.

[15] Hoskins, A.A., Moore, M.J. 2012. The spliceosome: a flexible, reversible macromolecular machine. Trends Biochem Sci 37, 179-188.

[16] Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., Stamm, S. 2013. Function of alternative splicing. Gene 514, 1-30.

[17] Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X.S., Ahringer, J. 2009 Differential chromatin marking of introns and expressed exons by H3K36me3. Nat Genet 41, 376-381.

[18] Kouzarides, T. 2007. Chromatin modifications and their function. Cell 128, 693-705.

[19] Lai, T.S., Greenberg, C.S. 2013. TGM2 and implications for human disease: role of alternative splicing. Front Biosci 18, 504-519.

[20] Lin, H. 2008. The modified Mahalanobis Discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 252, 350-356.

[21] Lin, H., Chen, W., Ding, H. 2013. AcalPred: A Sequence-Based Tool for Discriminating between Acidic and Alkaline Enzymes. Plos One 8, e75726.

[22] Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., Misteli, T. 2011. Epigenetics in alternative pre-mRNA splicing. Cell 144, 16-26.

[23] Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., Misteli, T. 2010. Regulation of alternative splicing by histone modifications. Science 327, 996-1000.

[24] Lv, J., Luo, L.F., Zhang, L.R., Chen, W., Zhang, Y. 2010. Increment of diversity with quadratic discriminant analysis- an efficient tool for sequence pattern recognition in bioinformatics. Open Access Bioinformatics 2, 89-96.

[25] Oberdoerffer, S., Moita, L.F., Neems, D., Freitas, R.P., Hacohen, N., Rao, A. 2008 Regulation of CD45 alternative splicing by heterogeneous ribonucleoprotein, hnRNPLL. Science 321, 686-691.

[26] Orr-Urtreger, A., Bedford, M.T., Burakova, T., Arman, E., Zimmer, Y., Yayon, A., Givol, D., Lonai, P. 1993. Developmental localization of the splicing alternatives of fibroblast growth factor receptor-2 (FGFR2). Dev Biol. 158, 475-486.

[27] Pan, Q., Shai, O., Lee, L.J., Frey, B.J., Blencowe, B.J. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40, 1413-1415.

[28] Peng, H., Long, F., Ding, C. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE transactions on Pattern Analysis and Machine Intelligence 27, 1226-1238.

[29] Ponta, H., Sherman, L., Herrlich, P.A. 2003. CD44: from adhesion molecules to signalling regulators. Nat Rev Mol Cell Biol 4, 33-45.

[30] Pradeepa, M.M., Sutherland, H.G., Ule, J., Grimes, G.R., Bickmore, W.A. 2012. Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. Plos Genet 8, e1002717.

[31] Saint-Andre, V., Batsche, E., Rachez, C., Muchardt, C. 2011. Histone H3 lysine 9 trimethylation and HP1gamma favor inclusion of alternative exons. Nat Struct Mol Biol 18, 337-344.

[32] Schor, I.E., Rascovan, N., Pelisch, F., Allo, M., Kornblihtt, A.R. 2009. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. Proc Natl Acad Sci U S A 106, 4325-4330.

[33] Screaton, G.R., Bell, M.V., Bell, J.I., Jackson, D.G. 1993. The identification of a new alternative exon with

highly restricted tissue expression in transcripts encoding the mouse Pgp-1 (CD44) homing receptor. Comparison of all 10 variable exons between mouse, human, and rat. J Biol Chem 268, 12235-12238.

[34] Shindo, Y., Nozaki, T., Saito, R., Tomita, M. 2013. Computational analysis of associations between alternative splicing and histone modifications. FEBS Lett 587, 516-521.

[35] Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., Burge, C.B. 2008a. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470-476.

[36] Wang, Z., Burge, C.B. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA 14, 802-813.

[37] Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q., Zhao, K. 2008b. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40, 897-903.

[38] Warzecha, C.C., Hovhannisyan, R., Carstens, R.P. 2012. Dynamic fluorescent and luminescent reporters for cell-based splicing screens. Methods Mol Biol 867, 273-287.

[39] White, E.S., Baralle, F.E., Muro, A.F. 2008. New insights into form and function of fibronectin splice variants. J Pathol 216, 1-14.

[40] Zhang, L.R., Luo, L.F. 2003. Splice site prediction with quadratic discriminant analysis using diversity measure. Nucleic Acids Res 31, 6214-6220.

[41] Zhou, H.L., Luo, G., Wise, J.A., Lou, H. 2013. Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. Nucleic Acids Res doi:10.1093/nar/gkt875.

[42] Zhang, Q.W., Peng, Q.K., Zhang, Q., Yan, Y.H., Li, K.K., Li, J. 2010. Splice sites prediction of Human genome using length-variable Markov model and feature selection. Expert Syst Appl 37, 2771-2782