Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 358 (2007) 336–341
BBRC
www.elsevier.com/locate/ybbrc

Prediction of C-to-U RNA editing sites in higher plant mitochondria using only nucleotide sequence features

Pufeng Du [1], Tao He [1], Yanda Li *

Bioinformatics Division, Tsinghua National Laboratory for Information Science and Technology (TNLIST), Department of Automation, Tsinghua University, Beijing 100084, China

Received 5 April 2007
Available online 30 April 2007

Abstract

RNA editing is a class of post-transcriptional processing which contributes to the organism complexity. C-to-U RNA editing is commonly observed in higher plant mitochondria. The *in vivo* mechanism of recognizing C-to-U RNA editing sites is still unknown. In recent years, many efforts have been made to computationally predict C-to-U RNA editing sites. But all existing methods require using knowledge other than an RNA sequence. In the present work, we propose the first method for predicting C-to-U RNA editing sites using only nucleotide sequence features. This method was developed based on the SVM algorithm combined with a triplet scoring model. Our method can achieve 84% overall accuracy which is comparable to other methods. We also computationally found that several triplets never appear upstream near an edited cytidine, indicating that these triplets may protect a cytidine from being edited. This discovery suggests the need for further experimental research and may be helpful in understanding the editing site recognition mechanism.

Keywords: C-to-U; RNA editing; Prediction; Triplets; SVM

Transcription is an important step in central dogma. DNA is transcribed into messenger RNA in this process. For most cases, mRNA has the same sequence as the DNA template. However, there are some organisms that can edit their mRNA sequences by inserting, deleting or substituting single or multi-nucleotides of the mRNA [1]. RNA editing is recognized as a class of post-transcriptional processing (like polyadenylation, 5′ capping and splicing) which increases the organism complexity [2]. The mechanism of RNA editing in some organisms is clear, but it remains largely unknown for others [3,4]. The first observed RNA editing event was C insertion in *trypanosome* mitochondria [5]. From then on, many types of RNA editing events have been discovered in various species [6–8]. C-to-U and A-to-I are two main types of substitution RNA editing. C-to-U RNA editing events are mostly reported in higher plant mitochondria and chloroplasts [9–12]. A-to-I RNA editing events are commonly found in animals [13–17].

C-to-U RNA editing is considered as a deamination process in which a cytidine at a specific site is converted to a uridine [3,18–20]. Since most of the known C-to-U RNA editing events are found within coding regions [21–23], the knowledge of C-to-U RNA editing focuses on its effect on the protein product of an edited gene. C-to-U RNA editing events prefer to appear at the second codon position and make the coded amino acid more hydrophobic [21]. The protein which is translated from edited mRNA is more conserved across species than the protein sequence predicted from genomic DNA [24–26].

The *in vivo* recognition mechanism of C-to-U RNA editing sites remains largely unknown. Computationally identifying C-to-U RNA editing sites is still an open problem. Well-designed algorithms for predicting C-to-U RNA

* Corresponding author. Fax: +86 10 62794295.
E-mail address: daulyd@tsinghua.edu.cn (Y. Li).
[1] The authors wish to be known that in their opinion, the first two authors contributed equally to this paper and should be regarded as joint first authors.

0006-291X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2007.04.130

editing site not only provide powerful tools for discovering new RNA editing events, but also give helpful suggestions for experimental research to understand its site recognition mechanism. Thus, in recent years, many efforts have been made to computationally predict C-to-U RNA editing sites. Cummings and Myers proposed the first *ab initio* method to predict C-to-U RNA editing sites by a using classification tree and random forest method [27]. PREP-Mt used the homologous alignment of protein sequences [28] to identify non-synonymous editing events. REGAL introduced the second *ab initio* method based on the genetic algorithm [29,30]. To make an accurate enough prediction, however, all these methods require information other than an RNA sequence, such as the codon position of the cytidine or the homology gene name of the RNA sequence. Because C-to-U RNA editing events do exist in non-coding regions, like tRNA genes, UTRs and introns (though the number is much less than that in coding regions [21,28]), it is likely that the codon position is not involved in the *in vivo* editing site recognition.

On the other hand, several short sequences which are critical for efficiently editing some cytidines have been identified experimentally in their upstream and downstream regions [31–33]. Some computational evidence shows that the inferred secondary structure of the transcript is not involved in recognizing the C-to-U RNA editing sites [18,29], while others show the opposite [27]. With the above knowledge, we could hypothesize that the flanking sequence of a cytidine is necessary and sufficient to determine whether it should be edited.

In this paper, we develop an algorithm to predict C-to-U RNA editing sites accurately based only on the flanking sequence of a cytidine. Our method can achieve 84% overall accuracy which is comparable to other methods. By analyzing the algorithm, we find that a group of triplets may be able to protect a cytidine from being edited. We hope that our method is a useful complement to those existing methods and may suggest further experimental research.

Materials and methods

Data collection. EdRNA [34], dbRES [35], and REDIdb [36] are three recently published RNA editing databases focusing on different aspects of RNA editing. In this study, we constructed our dataset from GenBank and our database dbRES which is a collection of all types of experimentally validated RNA editing events. Three complete mitochondria genome sequences were obtained from GenBank database. They are the mitochondria genomes of *Arabidopsis thaliana* (GenBank Accession No. Y08501), *Brassica napus* (GenBank Accession No. AP006444) and *Oryza sativa* (GenBank Accession No. BA000029). All GenBank annotations related with C-to-U RNA editing of these three mitochondria genomes were extracted to build a working set. The RNA editing events in this working set were then mapped to the correct strand of genome sequence. The obviously incorrect editing events annotations (like the editing site is A) were excluded. The editing events which are not shared by GenBank and dbRES were also excluded because they lack of experimental evidence. Finally, we constructed a working dataset containing 454 C-to-U RNA editing sites from *A. thaliana*, 416 sites from *B. napus* and 445 sites from *O. sativa*.

After collecting all these editing sites, we adopted the strategy which had been used by both REGAL [29] and the classification tree based method [27] to construct a null observation set as negative samples for training our predictor. Since our algorithm did not concern the codon position of a cytidine, we simply randomly selected an equal number of cytidines without any type of RNA editing annotation from the coding regions for each genome. The data statistic of our working set is shown in Table 1 with a comparison to the training sets of other methods.

Feature extraction. It has been proven that there are some short sequences which are critical for efficient editing in the upstream and downstream region of an edited cytidine. The locations of these short sequences can vary in a long range [31–33,37]. By considering that several consecutive editing sites may share such *cis*-elements [38,39], the flanking sequence of a cytidine in a long range may have effect on its efficient editing. On the other hand, a highly non-random nucleotide distribution has been observed in the immediate proximity of edited sites [21]. So we need to construct a set of sequence features which can represent the information from the long flanking sequence and the critical two or three nucleotides in the immediate proximity of the editing site.

The sequence features used in this study contained two parts. One was the triplet composition of the long range flanking sequence of a cytidine; the other was two triplets at position −3 to −1 and +1 to +3 of a cytidine (the cytidine position was 0).

The triplet composition was calculated in this way. First, the flanking sequence from $-N$ to $+N$ nt of a cytidine was extracted from genome sequence. It was easy to define $2N-1$ triplets in this $2N+1$ nt long sequence. Suppose the numbers of 64 different types of triplet were $t_1, t_2, \ldots, t_{64}$, the triplet composition can be denoted as a 64 dimension vector in Eq. (1). To get an optimized result, we chose $N = 250$ for calculating triplet composition. The reason why we chose $N = 250$ is explained in the Supplementary material.

$$\vec{V}_t = \frac{1}{2N-1}[t_1, t_2, \ldots, t_{64}] \tag{1}$$

Algorithm design. We combined two machine learning algorithms in this study. One was SVM; the other was the triplet scoring model.

SVM is a machine learning algorithm based on Statistical Learning Theory which was introduced by Vapnik [40]. It searches for an optimal separating hyper plane which maximizes the margin in feature space. SVM is designed to solve the binary classification problem. We used SVM with RBF kernel function to classify the 64 dimension triplet composition vectors in this study (SVM software is libSVM [41]).

The triplet scoring model is an extended version of the commonly used positional weighted matrix model. It was used to score two triplets which were defined in the Feature extraction section.

The triplet scoring model was defined as the following. The frequencies of 64 different triplets at position −3 to −1 and +1 to +3 were estimated for edited and non-edited sites respectively on the training set. Suppose the frequencies of triplets at position −3 to −1 and +1 to +3 of edited cytidines were $P_{uf}(i)$, $1 \leqslant i \leqslant 64$ and $P_{df}(i)$, $1 \leqslant i \leqslant 64$, and the frequencies of triplets at position −3 to −1 and +1 to +3 of non-edited cytidines were $P_{ub}(i)$, $1 \leqslant i \leqslant 64$ and $P_{db}(i)$, $1 \leqslant i \leqslant 64$. If a test sample had the ith triplet

Table 1
Data set distribution and comparison

Species	Our method	REGAL	Tree based	PREP-Mt
A. thaliana	454	344	444	433
B. napus	416	397	422	417
O. sativa	445	419	481	485
Over all	1315	1160	1347	1335

The number in the table shows the amount of edited cytidines of each genome.

at position −3 to −1 and the *j*th triplet at position +1 to +3, the triplet score was defined in the following equation:

$$S_{\text{tri}} = \ln\frac{P_{uf}(i)}{P_{ub}(i)} + \ln\frac{P_{df}(j)}{P_{db}(j)} \quad (2)$$

In order to combine the results of the two algorithms, we used a two hierarchy structure to fuse the prediction results of SVM and the score of triplet scoring model. The flowchart of the full algorithm is shown in Fig. 1.

The first step was to use SVM to classify the 64 dimension triplet composition. According to the prediction result made by SVM, different processing strategies were used. If the prediction result of SVM was positive and the triplet scoring model gave a score lower than the negative acceptance cutoff, the prediction result made by SVM was reversed to make it negative. Similarly, if the prediction result of SVM was negative and the triplet scoring model gave a score higher than the positive acceptance cutoff, the SVM prediction result was changed to positive. If neither of the above two conditions was true, the prediction result of SVM was considered as the final result.

After scanning the positive acceptance cutoff value from +2.0 to +1.0 with 0.1 step length and the negative acceptance cutoff value from −2.0 to −1.0 with 0.1 step length, we finally set them to +1.9 and −1.2, respectively, to get an optimized result.

Evaluation methods. We used leave-one-out cross validation to estimate the performance of our algorithm on each of the three mitochondria genomes respectively. We employed sensitivity, specificity, accuracy and positive predictive value (PPV) which were commonly used in evaluating the performance of C-to-U RNA editing predictors [27–30] to describe the performance of our method. These statistics are defined in Eqs. (3)–(6).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (3)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \quad (4)$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (5)$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (6)$$

TP, TN, FP, and FN in these formulas denote the number of true positives, true negatives, false positives, and false negatives.

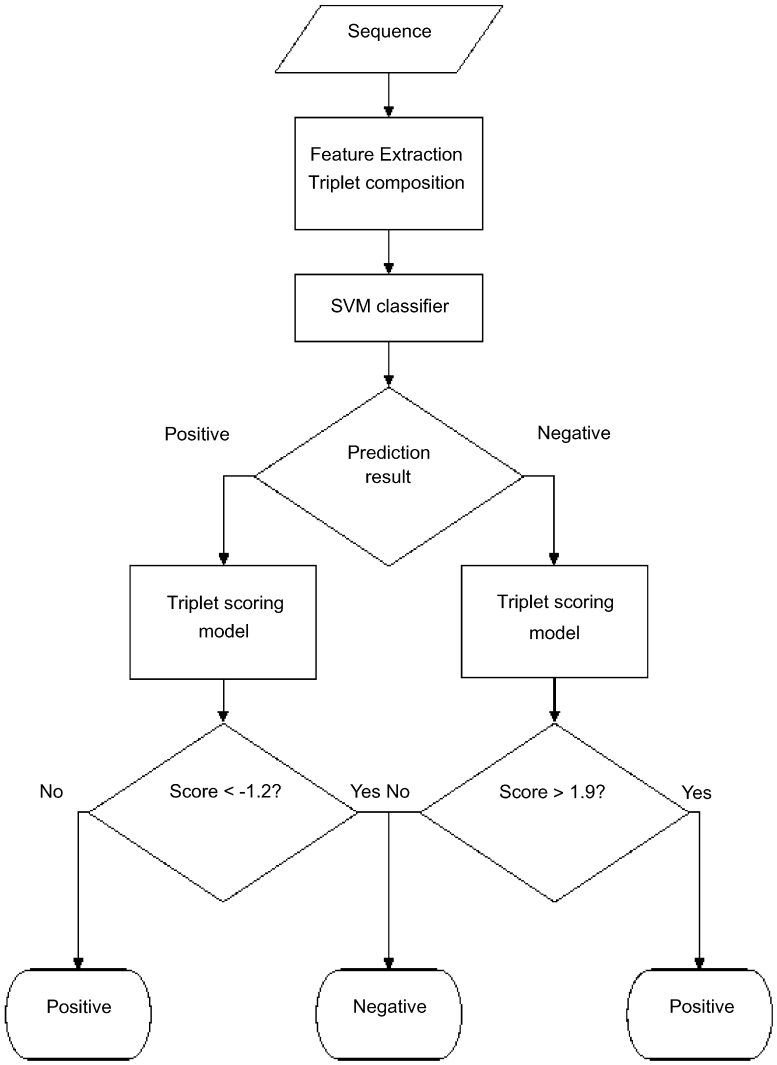

Fig. 1. Flowchart of the full algorithm. The first step is to use SVM to classify the 64 dimension triplet composition vectors. And then the prediction result made by SVM will be adjusted according to the output of triplet scoring model. The input of the triplet scoring model is two triplets in the immediate proximity of a cytidine.

Results

Algorithm performance

The performance detail is shown in Table 2. We can see that the prediction accuracy of the algorithm is about 85%, which shows it to be effective. Because of the leave-one-out cross validation method we used, the possibility of over-fitting can be eliminated.

Comparison with other methods

We compared the performance of our algorithm with all the other three methods. First, we compared our method with the classification tree method and random forest method. The comparison result is shown in Table 3.

As shown in Table 3, the accuracy of our method is much higher than either of the other two methods. The sensitivity of our method is also much higher than either of the other two methods, but the specificity is lower than classification tree method.

The comparison with REGAL method was based on its overall performance [30]. The comparison result is shown in Table 4. The performance of our method is higher than REGAL except the sensitivity estimated on *B. napus* mitochondria genome.

The comparison between PREP-Mt and our method cannot be carried out directly, because the training set of PREP-Mt was not a balanced training set [29]. We employ the comparison method and performance data which were reported while REGAL was compared to PREP-Mt [29]. The comparison result is shown in Table 5. The difference of the performance between our method and PREP-Mt is not significant. The sensitivity of our method is a little higher than PREP-Mt, while the positive predictive value is a bit lower than PREP-Mt. The estimated over all accuracy is the same.

These results show that our method, which is based on nucleotide sequence features, achieves better performance than classification tree based methods and REGAL. And the performance of our method is comparable to PREP-Mt method which is based on protein sequence homologous information.

Table 2
Estimated performance of the algorithm in this paper

Species	Sensitivity	Specificity	Accuracy	PPV
A. thaliana	0.86	0.85	0.85	0.85
B. napus	0.81	0.89	0.85	0.88
O. sativa	0.82	0.85	0.83	0.85
Over all	0.83	0.86	0.84	0.86

The performance in this table is estimated using leave-one-out cross validation.

Table 4
Comparison with REGAL

Species	REGAL			Our method		
	Sn.	Sp.	Acc.	Sn.	Sp.	Acc.
A. thaliana	0.81	0.80	0.81	0.86	0.85	0.85
B. napus	0.83	0.72	0.77	0.81	0.89	0.85
O. sativa	0.79	0.71	0.75	0.82	0.85	0.83
Over all	0.81	0.74	0.77	0.83	0.86	0.84

Sn., stands for sensitivity; Sp., stands for specificity; Acc., stands for accuracy.

Table 5
Comparison with PREP-Mt

Species	PREP-Mt			Our method		
	Sn.	PPV	Acc.	Sn.	PPV	Acc.
A. thaliana	0.79	0.86	0.82	0.86	0.85	0.85
B. napus	0.87	0.87	0.87	0.81	0.88	0.85
O. sativa	0.81	0.85	0.83	0.82	0.85	0.83
Over all	0.82	0.86	0.84	0.83	0.86	0.84

Sn., stands for sensitivity; PPV, stands for positive predictive value; Acc., stands for accuracy.

Discussion

The second step is a fine adjustment

The two or three nucleotide in the immediate proximity of a cytidine is an important feature for discriminating the edited and non-edited sites. By considering the correlation between neighbored nucleotides, we use the triplet immediately near the cytidine as a correction to the result of SVM. This correction not only increases the algorithm performance, but also balances the sensitivity and specificity. Thus, we call the second step a fine adjustment.

Table 3
Comparison with classification tree/random forest method

Species	Random forest			Classification tree			Our method		
	Sn.	Sp.	Acc.	Sn.	Sp.	Acc.	Sn.	Sp.	Acc.
A. thaliana	0.70	0.81	0.74	0.65	0.89	0.71	0.86	0.85	0.85
B. napus	0.73	0.81	0.77	0.63	0.89	0.69	0.81	0.89	0.85
O. sativa	0.72	0.81	0.72	0.64	0.88	0.71	0.82	0.85	0.83
Over all	0.72	0.81	0.74	0.64	0.89	0.70	0.83	0.86	0.84

Sn., stands for sensitivity; Sp., stands for specificity; Acc., stands for accuracy.

Table 6
Usefulness of the second step of the algorithm

Species	The first step			Full algorithm		
	Sn.	Sp.	Acc.	Sn.	Sp.	Acc.
A. thaliana	0.89	0.76	0.83	0.86	0.85	0.85
B. napus	0.83	0.80	0.82	0.81	0.88	0.85
O. sativa	0.86	0.77	0.81	0.82	0.85	0.83
Over all	0.86	0.78	0.82	0.83	0.86	0.84

Sn., stands for sensitivity; Sp., stands for specificity; Acc., stands for accuracy.

To perform this fine adjustment, we define a triplet scoring model (see Feature extraction section). By the definition of triplet scoring model, if a cytidine gets a high score from triplet scoring model (higher than positive acceptance cutoff), there is strong evidence that the cytidine should be edited. Similarly, if a cytidine gets a low score (lower than negative acceptance cutoff), there is strong evidence that the cytidine should be a non-edited one. With the strong evidence provided by triplet scoring model, the prediction result of the SVM classifier should be reversed to get a more reliable and more accurate prediction.

Table 6 shows the comparison of the performance detail between the first step and the complete algorithm. From the performance data of the first step, it is obvious that the sensitivity is much higher than the specificity. This means that the SVM classifier can recognize positive samples better than negative ones. But after the second step of the algorithm, the sensitivity and specificity are much closer. The full algorithm has similar recognition ability for positives and negatives.

Further investigation of the second step

Further investigation of the second step shows some very interesting results. The results of the second step rely on the frequencies of triplet on -3 to -1 positions and $+1$ to $+3$ positions. In all three training sets, the frequencies of some kinds of triplet on -3 to -1 position of an edited cytidine are zero. But there are no zeroes observed in the frequencies of all kinds of triplets on $+3$ to $+1$ position of an edited cytosine. There are also no zeroes in the frequencies of the negatives. After hybridizing all three training set, there are still seven zeroes in the frequencies. The seven triplets are AUA, GGA, ACA, UGG, CGG, GCG, and GAG.

This is the first time to report triplets which are impossible to appear on -3 to -1 position of an edited cytidine. This discovery shows that the triplet upstream near an edited cytidine is critical for its editing. It should be able to prove that some kinds of triplets can protect a cytidine from being edited *in vivo* if further experimental validation can be carried out.

Conclusion

In this paper, we develop an algorithm which can predict C-to-U RNA editing sites directly from nucleotide sequence. The performance of our algorithm is better than other*ab initio* methods [27,29,30], and is comparable to the homology information based method [28]. The success of our algorithm suggests that the recognition of C-to-U RNA editing site may depend only on nucleotide sequence. We hope that our algorithm will provide a useful complement to those existing methods. Since the C-to-U RNA editing site recognition mechanism is still unknown, our computational discovery could suggest further experimental research.

Acknowledgments

We thank Katherine Zhang and Cathy Gibbons for helping us with the language. This work is partially supported by National Science Foundation of China (NSFC 60572086), the National Basic Research Program of China (2004CB518605), Tsinghua Basic Research Foundation (JCxx2005064) and Basic Research Foundation of School of Information Science & Technology, Tsinghua University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc. 2007.04.130.

References

[1] O. Maydanovych, P.A. Beal, Breaking the central dogma by RNA editing, Chem. Rev. 106 (2006) 3397–3411.
[2] L.P. Keegan, A. Gallo, M.A. O'Connell, The many roles of an RNA editor, Nat. Rev. Genet. 2 (2001) 869–878.
[3] V. Blanc, N.O. Davidson, C-to-U RNA editing: mechanisms leading to genetic diversity, J. Biol. Chem. 278 (2002) 1395–1398.
[4] T. Shikanai, RNA editing in plant organelles: machinery, physiological function and evolution, Cell. Mol. Life Sci. 63 (2006) 698–708.
[5] R. Benne, J.V.D. Burg, J.P.J. Brakenhoff, P. Sloof, J.H.V. Boom, M.C. Tromp, Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA, Cell 46 (1986) 819–826.
[6] R. Cattaneo, A. Schmid, D. Eschle, K. Baczko, V.t. Meulen, M.A. Billeter, Biased hypermutation and other genetic changes in defective measles viruses in human brain infections, Cell 55 (1988) 255–265.
[7] L.M. Powell, S.C. Wallis, R.J. Pease, Y.H. Edwards, T.J. Knott, J. Scott, A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine, Cell 50 (1987) 831–840.
[8] B. Sommer, M. Köhler, R. Sprengel, P.H. Seeburg, RNA editing in brain controls a determinant of ion flow in glutamate-gated channels, Cell 67 (1991) 11–19.
[9] R. Freyer, M.-C. Kiefer-Meyer, H. Kössel, Occurrence of plastid RNA editing in all major lineages of land plants, Proc. Natl. Acad. Sci. USA 94 (1997) 6285–6290.
[10] R. Hiesel, B. Combettes, A. Brennicke, Evidence for RNA editing in mitochondria of all major groups of land plants except the Bryophyta, Proc. Natl. Acad. Sci. USA 91 (1994) 629–633.

[11] O. Malek, K. Lättig, R. Hiesel, A. Brennicke, V. Knoop, RNA editing in bryophytes and a molecular phylogeny of land plants, EMBO J. 15 (1996) 1403–1411.
[12] S. Steinhauser, S. Beckert, I. Capesius, O. Malek, V. Knoop, Plant mitochondrial RNA editing, J. Mol. Evol. 48 (1999) 303–312.
[13] A. Athanasiadis, A. Rich, S. Maas, Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome, PLoS Biol. 2 (2004) e391.
[14] D.R. Clutterbuck, A. Leroy, M.A. O'Connell, C.A.M. Semple, A bioinformatic screen for novel A–I RNA editing sites reveals recoding editing in BC10, Bioinformatics 21 (2005) 2590–2595.
[15] D.D.Y. Kim, T.T.Y. Kim, T. Walsh, Y. Kobayashi, T.C. Matise, S. Buyske, A. Gabriel, Widespread RNA editing of embedded Alu elements in the human transcriptome, Genome Res. 14 (2004) 1719–1725.
[16] E.Y. Levanon, E. Eisenberg, R. Yelin, S. Nemzer, M. Hallegger, R. Shemesh, Z.Y. Fligelman, A. Shoshan, S.R. Pollock, D. Sztybel, M. Olshansky, G. Rechavi, M.F. Jantsch, Systematic identification of abundant A-to-I editing sites in the human transcriptome, Nat. Biotechnol. 22 (2004) 1001–1005.
[17] D.P. Morse, P.J. Aruscavage, B.L. Bass, RNA hairpins in noncoding regions of human brain and *Caenorhabditis elegans* mRNA are edited by adenosine deaminases that act on RNA, Proc. Natl. Acad. Sci. USA 99 (2002) 7906–7911.
[18] R.M. Mulligan, M.A. Williams, M.T. Shanahan, RNA editing site recognition in higher plant mitochondria, J. Hered. 90 (1999) 338–344.
[19] V.K. Rajasekhar, R.M. Mulligan, RNA editing in plant mitochondria: [alpha]-phosphate is retained during C-to-U conversion in mRNAs, Plant Cell 5 (1993) 1843–1852.
[20] W. Yu, W. Schuster, Evidence for a site-specific cytidine deamination reaction involved in C-to-U RNA editing of plant mitochondria, J. Biol. Chem. 270 (1995) 18227–18233.
[21] P. Giegé, A. Brennicke, RNA editing in arabidopsis mitochondria effects 441 C-to-U changes in ORFs, Proc. Natl. Acad. Sci. USA 96 (1999) 15324–15329.
[22] H. Handa, The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus* L.): comparative analysis of the mitochondrial genomes of rapeseed and *Arabidopsis thaliana*, Nucleic Acids Res. 31 (2003) 5907–5916.
[23] Y. Notsu, S. Masood, T. Nishikawa, N. Kubo, G. Akiduki, M. Nakazono, A. Hirai, K. Kadowaki, The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants, Mol. Genet. Genomics 268 (2002) 434–445.
[24] P.S. Covello, M.W. Gray, RNA editing in plant mitochondria, Nature 341 (1989) 662–666.
[25] J.M. Gualberto, L. Lamattina, G. Bonnard, J.-H. Weil, J.-M. Grienenberger, RNA editing in wheat mitochondria results in the conservation of protein sequences, Nature 341 (1989) 660–662.
[26] R. Hiesel, B. Wissinger, W. Schuster, A. Brennicke, RNA editing in plant mitochondria, Science 246 (1989) 1632–1634.
[27] M.P. Cummings, D.S. Myers, Simple statistical models predict C-to-U edited sites in plant mitochondrial RNA, BMC Bioinformatics 5 (2004).
[28] J.P. Mower, PREP-Mt: predictive RNA editor for plant mitochondrial genes, BMC Bioinformatics 6 (2005).
[29] J. Thompson, S. Gopal, Genetic algorithm learning as a robust approach to RNA editing site prediction, BMC Bioinformatics 7 (2006).
[30] J. Thompson, S. Gopal, Correction: genetic algorithm learning as a robust approach to RNA editing site prediction, BMC Bioinformatics 7 (2006).
[31] J.-C. Farre, G. Leon, X. Jordana, A. Araya, *cis* Recognition elements in plant mitochondrion RNA editing, Mol. Cell. Biol. 21 (2001) 6731–6737.
[32] M.L. Hayes, M.L. Reed, C.E. Hegeman, M.R. Hanson, Sequence elements critical for efficient RNA editing of a tobacco chloroplast transcript in vivo and in vitro, Nucleic Acids Res. 34 (2006) 3742–3754.
[33] M. Takenaka, J. Neuwirt, A. Brennicke, Complex *cis*-elements determine an RNA editing site in pea mitochondria, Nucleic Acids Res. 32 (2004) 4137–4144.
[34] J.-H. Hung, W.-C. Wang, H.-D. Huang, Systematic identification and repository of RNA editing site in human genome, in: International Computer Symposium, 2006, pp. 1386–1391.
[35] T. He, P. Du, Y. Li, dbRES: a web-oriented database for annotated RNA editing sites, Nucleic Acids Res. 35 (Database Issue) (2007) D141–D144.
[36] E. Picardi, T.M.R. Regina, A. Brennicke, C. Quagliariello, REDIdb: the RNA editing database, Nucleic Acids Res. 35 (Database Issue) (2007) D173–D177.
[37] D. Choury, J.-C. Farre, X. Jordana, A. Arays, Different patterns in the recognition of editing sites in plant mitochondria, Nucleic Acids Res. 32 (2004) 6397–6406.
[38] R. Bock, M. Hermann, M. Fuchs, Identification of critical nucleotide positions for plastid RNA editing site recognition, RNA 3 (1997) 1194–1200.
[39] J.A. van der Merwe, M. Takenaka, J. Neuwirt, D. Verbitskiy, A. Brennicke, RNA editing sites in plant mitochondria can share *cis*-elements, FEBS Lett. 580 (2006) 268–272.
[40] V.N. Vapnik, The Nature of Statistical Learning Theory, second ed., Tsinghua University Press, Beijing, 2000 (in Chinese).
[41] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, 2001.