Journal of Bioscience and Bioengineering
VOL. 107 No. 5, 569–578, 2009

www.elsevier.com/locate/jbiosc

Prediction of non-canonical polyadenylation signals in human genomic sequences based on a novel algorithm using a fuzzy membership function

Masami Kamasawa,[1,2] and Jun-ichi Horiuchi[1,*]

Department of Biotechnology and Environmental Chemistry, Kitami Institute of Technology, 165 Koen-cho, Kitami, Hokkaido 090-8507, Japan[1] *and Toyo Engineering Corporation, 2-8-1 Akanehama, Narashino, Chiba 275-0024, Japan*[2]

Received 9 November 2008; accepted 5 January 2009

Computational prediction of polyadenylation signals (PASes) is essential for analysis of alternative polyadenylation that plays crucial roles in gene regulations by generating heterogeneity of 3′-UTR of mRNAs. To date, several algorithms that are mostly based on machine learning methods have been developed to predict PASes. Accuracies of predictions by those algorithms have improved significantly for the last decade. However, they are designed primarily for prediction of the most canonical AAUAAA and its common variant AUUAAA whereas other variants have been ignored in their predictions despite recent studies indicating that non-canonical variants of AAUAAA are more important in the polyadenylation process than commonly recognized. Here we present a new algorithm "PolyF" employing fuzzy logic to confer an advance in computational PAS prediction — enable prediction of the non-canonical variants, and improve the accuracies for the canonical A(A/U)UAAA prediction. PolyF is a simple computational algorithm that is composed of membership functions defining sequence features of downstream sequence element (DSE) and upstream sequence element (USE), together with an inference engine. As a result, PolyF successfully identified the 10 single-nucleotide variants with approximately the same or higher accuracies compared to those for A(A/U)UAAA. PolyF also achieved higher accuracies for A(A/U)UAAA prediction than those by commonly known PAS finder programs, Polyadq and Erpin. Incorporating the USE into the PolyF algorithm was found to enhance prediction accuracies for all the 12 PAS hexamers compared to those using only the DSE, suggesting an important contribution of the USE in the polyadenylation process.

[**Key words:** Polyadenylation signal; Poly(A) site; Downstream sequence element; Upstream sequence element; Single-nucleotide variants; Fuzzy logic; Membership function]

The formation of a mature mRNA is completed by polyadenylation, which consists of two tightly coupled steps: the specific cleavage at the 3′ end of a pre-mature mRNA, followed by the addition of a stretch of 100 to 300 adenine residues (known as the poly(A) tail) to the 3′ terminus of the last exon (1, 2). Polyadenylation requires a hexamer motif that is called the polyadenylation signal (PAS). The PAS is generally located ca. 10 to 30 nucleotides (nt) upstream of the cleavage site (poly(A) site), serving as the binding site for the cleavage and polyadenylation specificity factor (CPSF). In vertebrate, AAUAAA and its common variant AUUAAA are known as the canonical signals among highly conserved variants, where a single nucleotide in AAUAAA is substituted (1, 3, 4, 5). Recent studies revealed that 30–50% or more of human genes have alternative polyadenylation sites (3, 4, 6, 7, 8), which produces mRNAs with various lengths of 3′-UTR from a single gene. Because the selection of a PAS is thought to be specific with cell-type, tissue-type, developmental stage or disease-type, alternative polyadenylation plays an important role in regulatory mechanisms of gene expression by affecting the stability of mRNAs, attenuating protein synthesis and controlling subcellular localization (2, 6, 9). Nevertheless, the annotations of PASes for a gene of interest are often incomplete or frequently missing from the records in major gene databases such as GenBank. Therefore, PAS prediction is essential for identifying genes or gene boundaries to confer gene annotations, and delineating heterogeneity of the 3′ ends of mRNAs generated from the alternative selection of poly(A) sites. Because a PAS consists of only six nucleotides, each PAS motif could occur quite frequently in genomic sequences at rate of approximately once every 4096 bases. Therefore, it is crucial for PAS prediction algorithms to identify true signals from randomly occurring false signals based on features of the sequences around true poly(A) sites or PAS hexamers. The region known as the downstream sequence element (DSE) is located approximately 20 to 40 nt downstream of a poly(A) site or 30 to 50 nt downstream of a PAS (10, 11). Although several different motifs have been proposed for the DSE, it is generally described as U- or G/U-rich regions (12). Alternatively, the DSE is suggested not to be a G/U-rich region rather a simple U-rich region and interestingly, the "strong" polyadenylation sites have the DSEs with higher uracil frequencies than those for the "weak" sites (5). The DSE with high uracil content possibly can compensate weak PAS signal as a strong DSE (1, 13, 14). In

[*] Corresponding author. Tel.: +81 157 26 9415; fax: +81 157 24 7719.
E-mail address: horiucju@mail.kitami-it.ac.jp (J. Horiuchi).

1389-1723/$ - see front matter © 2009, The Society for Biotechnology, Japan. All rights reserved.
doi:10.1016/j.jbiosc.2009.01.001

addition, the USE is known to exist upstream of PAS hexamers. The USE is more ambiguous compared to the DSE, except that an increased uracil frequency in the upstream regions of PAS hexamers has been proposed (12, 15). The role of the USE in polyadenylation mechanism still remains controversial. A large scale observation of upstream regions of poly(A) sites concluded that the USE is not as strongly linked to polyadenylation efficiency as the DSE (5) while several studies suggested the USE is associated with polyadenylation efficiency. For example, the USE was shown to directly stabilize the binding of the CPSF to the poly(A) signal in the polyadenylation of human lamin B2 gene (13, 16). In another case, a 53 nt sequence with high uracil frequency (42%) that is located immediately upstream of the AAUAAA was required for full activity in the polyadenylation of the complement factor *C2* gene (17).

To date, several algorithms have been developed to predict PASes. Those algorithms are based on machine learning methods, such as neural network (18), linear discriminant function (LDF) (19), quadratic discriminant function (QDF) (20), Hidden Markov Model (HMM) (21, 22), weight matrix only (5), and Support Vector Machine (SVM) (23, 24). Among those algorithms, only PolyA_SVM is designed to predict poly(A) sites in a query sequence while others are designed to predict PASes. Overall, current algorithms have achieved moderate sensitivity and specificity, but may yet be improved. Moreover, all the algorithms are designed to predict only two PASes, the most canonical AAUAAA and its common variant AUUAAA, while many of the other single-nucleotide variants were experimentally verified to be functional as PASes in vertebrate cells including human. For example, AGUAAA is reported to be located 20 nt upstream of the poly(A) site of human plasma glutathione peroxidase gene (25). AAGAAA produces the 2.8 kb transcript from human PGHS-1, which is the isoform of prostaglandin endoperoxide H synthase (PGHS), as a result of alternative polyadenylation whereas 4.5 kb transcript is derived by the canonical AAUAAA (26). Other single-nucleotide variants, UAUAAA (27, 28), AAUAUA (29, 30, 31), AAUACA (32, 33, 34), CAUAAA (35, 36), GAUAAA (37, 38), AAUGAA (27, 39), ACUAAA (40) and AAUAGA (41) are also reported or suggested to be functional in mammalian systems. Therefore, a computational tool that realizes prediction of non-canonical variant PASes as well as improves prediction accuracy must be developed. Genes are thought to include variability or noise attributed to such factors as insertion, deletion and duplication of sequence elements over the course of evolution. In PAS prediction, for example, the sequence features that define DSE and the USE could vary depending on each terminal sequence (i.e., a sequence that contains a poly(A) site and at least one corresponding PAS upstream of the poly(A) site) due to the noises or variability of the sequence feature such as the DSE locations to 3′ of PASes or the uracil contents of the DSEs. This ambiguousness included in genomic sequences makes it difficult for existing PAS prediction algorithms to identify authentic PASes accurately from randomly occurring hexamer motifs. To achieve higher accuracy in PAS prediction, a method that is able to cope with the uncertainties and imprecisions involved in identifying terminal sequences is necessary.

Here, we present a new algorithm "PolyF" using fuzzy membership functions based on fuzzy logic for extensive PAS prediction. Fuzzy logic was derived from fuzzy set theory, which was first introduced by Zadeh (42, 43), to deal with qualitative or linguistic information such as human knowledge to use in applications, and a system for which mathematical modeling is difficult due to its complexity or ambiguity. Reasoning process by fuzzy logic is known to be robust because noise that is possibly contained in the input data does not affect the output drastically by use of the membership function. Fuzzy logic has been used for versatile applications, including biotechnology fields such as bioprocess control (44, 45, 46, 47). Moreover, recent studies have revealed that fuzzy theory could be applied to Bioinformatics such as prediction of peptides that bind to the major histocompatibility complex (48, 49), data clustering (50, 51, 52), prognostication of cancers (53, 54, 55), and prediction of genetic network (56, 57). PolyF was developed by making use of fuzzy membership functions to improve prediction accuracies for A(A/U)UAAA, and to enable prediction of non-canonical polyadenylation signals in human genomic sequences. We have tested two types of algorithms for PolyF in this study. The first algorithm "PolyFd" ("d" for DSE) uses only the sequence features of the DSE for the PAS prediction. The second algorithm "PolyFud" ("ud" for USE and DSE) incorporated the sequence feature of the USE into its reasoning process together with the DSE to examine whether the USE could contribute to enhancing the prediction accuracies. The performances of these two algorithms were tested in comparison with common tools of PAS prediction, Erpin (5) and Polyadq (20), and PolyFud was found to be superior in its predictive value. In addition, poly(A) sites were grouped into "strong" and "weak" sites for genes with multiple poly(A) sites. PolyFud was applied to detect those poly(A) sites in comparison with PolyA_SVM. Frequencies of PASes that appeared upstream of strong and weak sites were also examined.

MATERIALS AND METHODS

Dataset

For membership function We have previously identified a total of 20,347 terminal sequences from human genomic sequences (UniGene Build 195) by modified EST clustering (8). Each sequence contains one of the 12 PAS hexamers (A(A/U)UAAA plus its 10 single-nucleotide variants) and −150/+150 nt region (total length: 306 nt) surrounding the PAS. Membership functions that define "the DSE location" were determined based on the DSE distributions (See Fig. 6 in Kamasawa and Horiuchi (8)). The uracil contents of upstream regions of PASes were measured for the 20,347 terminal sequences to construct membership functions defining "the USE maximum uracil content".

For program tuning and prediction test To evaluate accuracy of a PAS prediction algorithm, it is crucial to eliminate false signals from datasets used for the program tuning and the prediction test. False sites in PAS identification are mostly generated in the extracting step of EST/mRNAs that contain putative poly(A) sites, because those sequences are often tailed with a poly(A) stretch that is the pseudo poly(A) tail generated by internal priming. Hence, we imposed more stringent conditions on the extraction step of EST/mRNAs. The number of consecutive As at 3′ terminus (Ts for 5′ terminus) of the sequences was increased from 10 to 18 as the extracting condition for EST/mRNAs. Also the number of supporting EST was increased from two to three, which means that at least three EST/mRNAs should end at the same position for the PAS to be identified as authentic when EST/mRNAs were aligned to the corresponding gene. For a PAS that was identified as authentic, the −150/+150 nt region (total length: 306 nt) around the PAS was extracted for the dataset of terminal sequences. As a result, a total of 4723 terminal sequences were extracted from the original 20,347 records. For A(A/U)UAAA, each dataset was divided in half. One half was used for program tuning and the other was used for prediction test. For 10 single-nucleotide variants, one-third was used for the program tuning and the rest of the data was used for the prediction test since the numbers of terminal sequences that contained those variant PASes were much smaller than those with A(A/U)UAAA. As negative sequences for the program tuning, 8,397 sequences surrounding putative false PAS hexamers (−150 to +150 nt region around a hexamer motif), which were found in the coding sequences (CDS) on human chromosome 1, were used. Those CDSes were downloaded from Ensembl genome

TABLE 1. Number of datasets of terminal sequences used for tuning and prediction test

PAS motif	Positive data		Negative data			
	Tuning data	Test data	#1 CDS	#2 CDS	#4 CDS	#6 CDS
AAUAAA	1242	1260	666	724	456	378
AUUAAA	378	390	425	574	335	356
UAUAAA	91	195	420	442	311	267
AGUAAA	51	105	478	491	324	301
AAGAAA	66	145	2608	2269	1507	1541
AAUAUA	60	140	438	389	257	229
AAUACA	32	80	546	618	352	390
CAUAAA	29	60	367	354	239	219
GAUAAA	29	60	529	544	359	414
AAUGAA	50	105	1164	1104	760	704
ACUAAA	27	60	373	358	238	205
AAUAGA	18	50	383	360	236	255

#1 CDS, coding sequence from human chromosome 1 (for tuning data); #2, #4, #6 CDS, coding sequence from human chromosome 2, 4, and 6 (for test data).

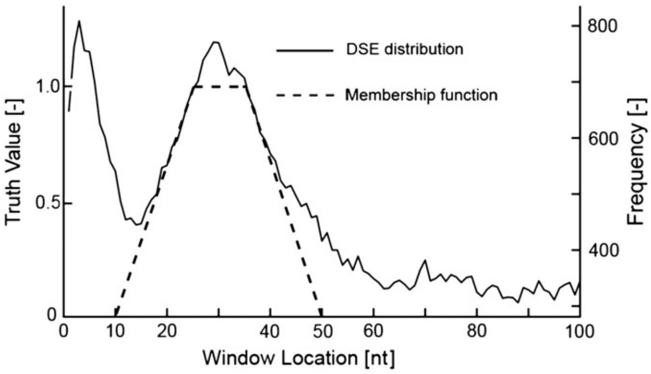

FIG. 1. Determining a membership function that defines the DSE location. The membership function (broken line) was determined based on the shape of the DSE distribution (solid line). Location is the first nucleotide of the DSE relative to the 3′ end of a PAS hexamer (position 0).

database (version 41) using MartView (http://www.biomart.org/biomart/martview). As negative sequences for the prediction test, 18,860 sequences with putative false PAS hexamers were collected from the CDSes on human chromosome 2, 4, and 6 in the same manner. Table 1 summarized the number of datasets of the terminal sequences used for the program tuning and the prediction test.

Membership function

Membership function of DSE location The DSE is considered to be located ca. 35 to 50 nt downstream of a PAS (5, 24, 58). In a previous study, we simply defined the DSE as the region with more than 60% uracil content measured in a 9 nt sliding window, which advanced by one nucleotide toward the 3′ end and stepped over the 100 nt downstream region of a PAS. According to this criterion, the distributions of DSEs for each PAS hexamer were examined (8). If a PAS motif found in a sequence is authentic, the DSE (i.e., a 9 nt window sequence with more than 60% uracil content) should be located in a particular range of distance from 3′ of the PAS. In this study, we generated the membership functions, which define the window locations to 3′ of each PAS, to evaluate the possibility that the window is a true DSE with regard to location (see Fig. 1 for an example). This membership function is trapezoidal and maps a window location to a membership degree (truth value), which can vary between 0 and 1. In fuzzy set theory, the truth value 1 can be interpreted as that the window location "completely meets the location for true DSE". The truth value 0 can be interpreted that the window location "does not meet the condition for true DSE at all". The truth values between 0 and 1 can be interpreted that the window location "meets the condition for true DSE" corresponding to the truth value. For example, the truth value will be 0.67 if the window location is 20 nt downstream of a PAS (see upper left in Fig. 2). The truth value 0.67 implies that the window is a true DSE with moderate to high possibility, with regard to the location to the PAS. Because the DSE distribution varies considerably between the PAS hexamers, the membership functions that define the window locations for the DSEs were constructed for each PAS hexamer. For the PASes that showed a broad distribution in the DSE, we limited the distance from 3′ of each PAS to the window location in the range between 10 and 70 nt based on common knowledge, as follows: The PAS is located approximately 10 to 30 nt upstream of a poly(A) site and the DSE is located 20 to 40 nt downstream of a poly(A) site (1, 9).

Membership function of DSE uracil content Legendre and Gautheret (5) demonstrated that the uracil content of the DSE is between ca. 35% (for a "weak" polyadenylation site) and ca. 45% (for a "strong" polyadenylation site) in an 11 nt sliding window. Chau et al. (58) and Chen et al. (11) suggested that a pentamer, in which at least four of them (80% uracil content) are uracil residues, is required 10 to 30 nt downstream of a poly(A) site. Based on this knowledge, we generated membership functions defining uracil content to evaluate the possibility that the window is the true DSE with regard to the uracil content. The uracil content was measured in a 9 nt sliding window, as described in the previous section. This membership function maps uracil content to a truth value, which can vary between 0 and 1. Truth value 1 implies that the uracil content "completely meets the condition for true DSE". The truth value 0 implies that the uracil content dose not meet the condition for a true DSE at all. Truth values between 0 and 1 can be interpreted as showing that the uracil content "meets the condition for a true DSE" corresponding to the value. For example, if the uracil content is 50% in a given window, the membership degree will be 0.38 (see upper right in Fig. 2). The truth value 0.38 can be interpreted as "This window sequence is slightly to moderately positive for the true DSE with regard to the uracil content. We constructed 12 membership functions that define the uracil contents for the DSEs of the 12 PAS hexamers.

Membership function of USE maximum uracil content We measured the uracil contents of the upstream region of PASes in 9 nt sliding windows, which advanced by one nucleotide toward the 5′ end. Visible peaks were observed ca. 5 to 25 nt upstream of all the 12 PAS hexamers (data not shown). Based on this observation, we defined the USE as the uracil rich region that is located within 20 nt upstream of a PAS. Because the USEs were located in the very narrow range of 5 to 25 nt upstream of the PASes, we did not apply membership functions that define the window locations to 5′ of the PASes for evaluating the possibility of true USE. Instead, we used the membership functions that define the maximum uracil content for the 20 nt upstream of the PASes to evaluate whether the window is a true USE (see upper right corner in Fig. 3 for an example). The maximum uracil contents were stored for each PAS while a 9 nt sliding window advanced by one nucleotide toward the 5′ end and stepped over the 20 nt upstream region of the PAS. We assumed that the USE is not as strongly linked to polyadenylation

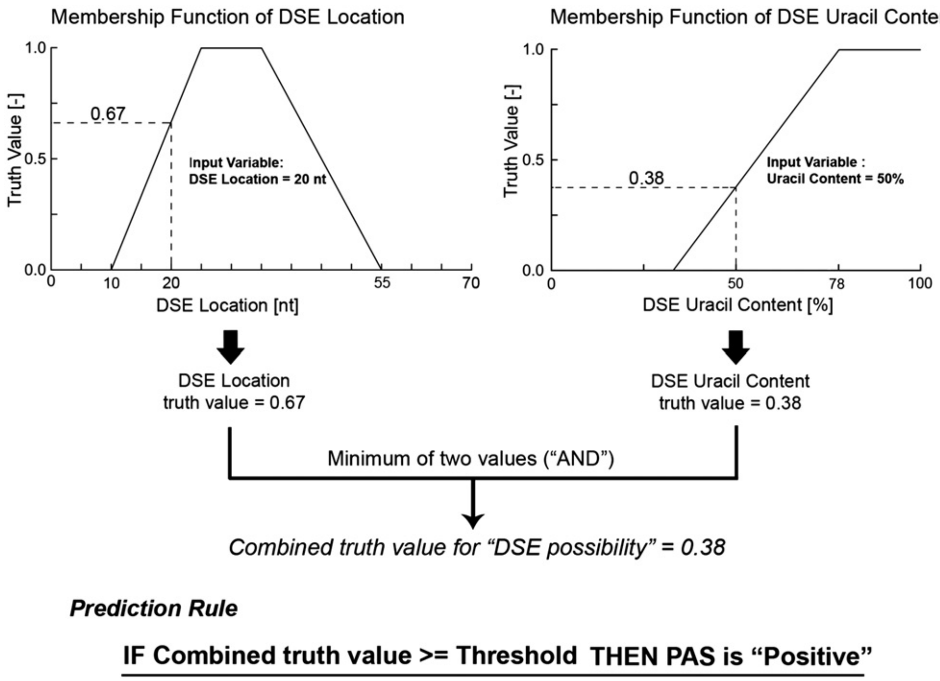

FIG. 2. PolyFd algorithms based on using the DSE for PAS prediction. Location in the membership function of the DSE location is the first nucleotide of the DSE relative to the 3′ end of a PAS hexamer (position 0).

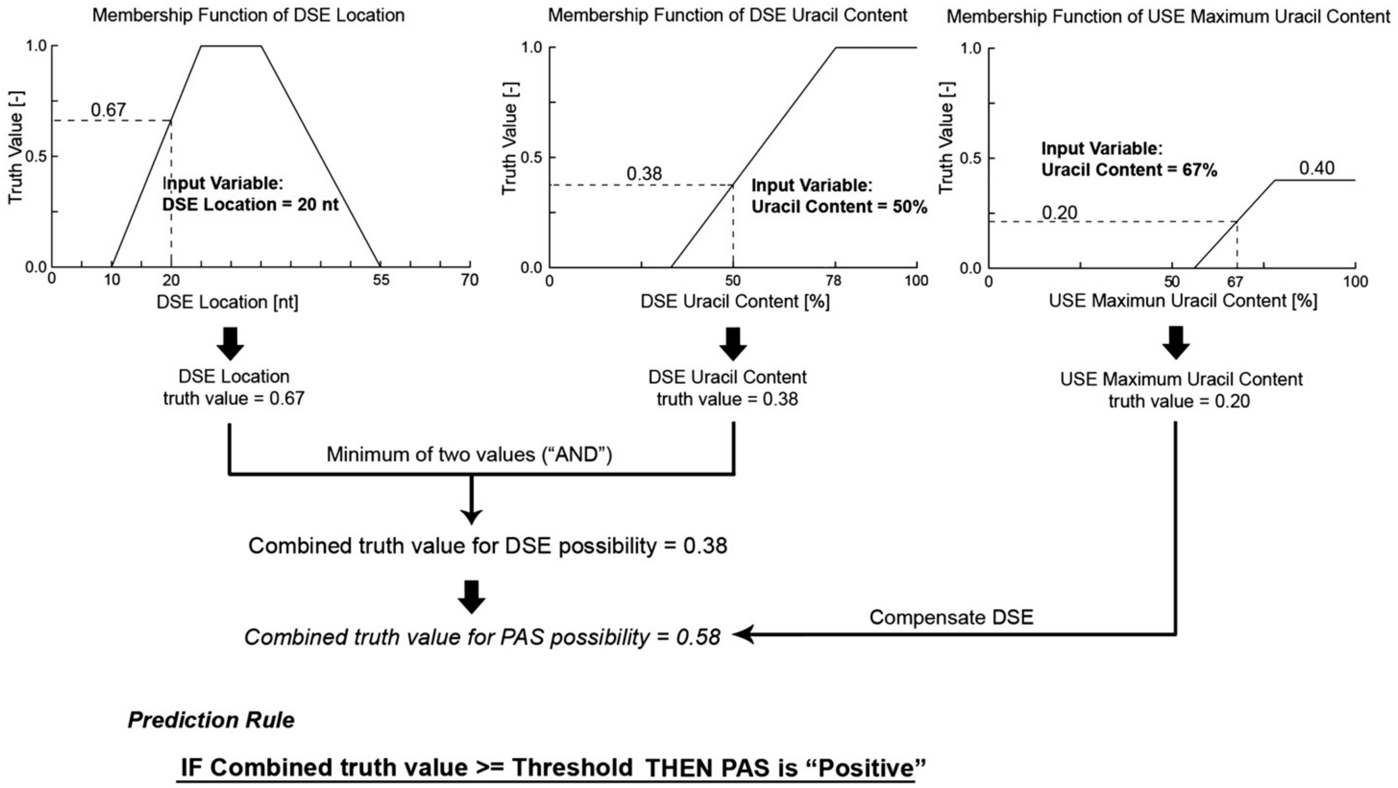

FIG. 3. PolyFud algorithms using the USE and the DSE for PAS prediction. Location in the membership function of the DSE location is the first nucleotide of the DSE relative to the 3′ end of a PAS hexamer (position 0).

efficiency as the DSE but can compensate for efficiency of the PAS or the DSE in the polyadenylation process. Based on this hypothesis, the maximum values of the truth value were set between 0.2 and 0.5 depending on each PAS, instead of 1 that is used in normal fuzzy sets. We created 12 membership functions that define the uracil contents for the USEs of the 12 PAS hexamers.

PolyF — prediction algorithm based on fuzzy logic We developed and tested two types of polyF algorithms "polyFd" and "polyFud" for PAS prediction, based on fuzzy logic. PolyFd uses only the sequence features of the DSE for PAS prediction. In addition to the DSE, PolyFud uses the USE for PAS prediction to examine whether the USE could contribute to enhance the performance of PolyF algorithm. As the programming language, Perl (Practical Extraction and Report Language: http://www.perl.org/) was used in this study. The common PAS prediction tools, Polyadq and Erpin were used as the benchmark programs to evaluate the performances of PolyF for the A(A/U) UAAA prediction. The test data were submitted to Polyadq through the web site of Zhang lab at Cold Spring Harbor Laboratory (http://rulai.cshl.org/tools/polyadq/polyadq_form.html). The default settings for cutoff scores that give the best performance were used for the prediction test. Erpin (version 4.2.5) was downloaded through the Erpin web site (http://tagc.univ-mrs.fr/erpin/) and locally installed to test the program. The score cutoff was set at 80% for AAUAAA and 90% for AUUAAA respectively to gain ca. 60% sensitivity. The default setting of Erpin was used for other parameters. For PolyA_SVM, version 2.2 was used with default settings (score cutoff: 6, HPR: 32). The test data were submitted to the PolyA_SVM web server (http://polya.umdnj.edu/polya_svm/).

PolyFd — algorithm using DSE for PAS prediction First, we developed and tested an algorithm "PolyFd" that exploits only the sequence feature of the DSE for the PAS prediction. PolyFd uses only two input variables in the reasoning logic — the DSE location and the DSE uracil content. PolyFd was designed to make calls all of the 12 PAS hexamers in a given query sequence. When PolyFd finds any of those 12 PAS hexamers in a query sequence, PolyFd starts measuring the uracil content in a 9 nt sliding window advancing by one nucleotide toward 3′ end, stepping over 100 nt downstream of the PAS. The truth value for the measured uracil content is calculated by mapping the uracil content into the membership function that defines DSE uracil content for the PAS. The truth value for the window location to 3′ of the PAS is also calculated by mapping the window location into the membership function that defines the DSE location for the PAS. For the PAS prediction, PolyFd relies on the common belief that a true PAS is supposed to be associated with a "strong" DSE. We assumed that a strong DSE should meet both conditions — high uracil content "AND" appropriate distance from the PAS. Therefore, the minimum (representing "AND" operator) of the two truth values is used as the combined truth value, which indicates the possibility that the PAS is associated with a strong DSE. For example, the truth value is 0.67 for the window location, and 0.38 for the uracil content as shown in Fig. 2. In this case, the combined truth value is 0.38, which was determined by taking the minimum of the two values. In the PolyFd algorithm, the PAS prediction was made as follows: PolyFd examined whether the truth value exceeded the threshold that was determined empirically by the trial prediction. If the value exceeded the threshold, the PAS was predicted as a "positive" or "true" signal. If the value did not exceed the threshold, the PAS was predicted as a "negative" or "false" signal. Note that the truth value of the DSE possibility directly represents the truth value that the PAS is authentic in the PolyFd algorithm. Each membership function as well as the threshold was tuned equally by trial reasoning using tuning datasets (both positive and negative) to gain optimum prediction accuracies. The prediction performances were evaluated using the following equations (same for all the algorithms tested):

$$\text{Sensitivity:} \quad \mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \ \text{Specificity:} \quad \mathrm{SP} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\text{Accuracy (CC):} \quad \mathrm{CC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{FP}+\mathrm{TP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

TP is the number of true positives, TN is number of true negatives, FN is number of false negatives, FP is number of false positives, and CC is correlation coefficient that represents overall accuracy in prediction. Total numbers of positive sequences and negative sequences were scaled to calculate SN, SP, and CC.

PolyFud — algorithm using USE plus DSE for PAS prediction Secondly, we developed and tested an algorithm "PolyFud" that exploits the sequence features of the USE in addition to the DSE for PAS prediction, based on the hypothesis that the USE can compensate for the efficiency of the PAS or the DSE in polyadenylation process. Hence, PolyFud uses the USE maximum uracil content as the third input variable in the reasoning process in addition to the two variables (i.e., the DSE location and the DSE uracil content) used in PolyFd. For the first step, PolyFud calculates the truth value that a given PAS has a strong DSE, in the same manner as PolyFd. For the second step, PolyFud measures the uracil content within 20 nt upstream of 5′ of the PAS in a 9 nt sliding window. The maximum uracil content is stored for the region while the sliding window advances by one nucleotide toward the 5′ end. The truth value, which represents the possibility that the PAS is associated with the true USE, is calculated by mapping the maximum uracil content into the membership function that defines the USE maximum uracil content for the PAS. For the third step, the truth value of the USE is added (simulating compensation by the USE) to the truth value of the DSE to gain the combined truth value. In Fig. 3, the truth value of the DSE is 0.38 by taking the minimum of two truth values (0.67 and 0.38 respectively), which are obtained by mapping the location and the uracil content into each membership function. In this example, the truth value of the USE is 0.2, and the combined truth value becomes 0.58 (0.38 plus 0.2). If the value exceeded the threshold, PolyFud predicts that the PAS is a positive or true signal. Otherwise, PolyFud predicts that the PAS is a negative or false signal. Tuning of the membership functions, determining the thresholds, and evaluation of the prediction accuracy were made in the same manner as PolyFd.

RESULTS AND DISCUSSION

PolyF

Parameter tuning of PolyFd and PolyFud The remarkable feature of the PolyF algorithm is that one can incorporate linguistic knowledge easily into the reasoning process by using membership functions or production rules (note that current PolyF algorithms do not use production rules in the reasoning processes). For example, common knowledge such as "The DSE is located ca. 30 to 50 nt downstream from a PAS" can be utilized to construct a membership function that defines the DSE location. A membership function represents "the degree of truth" mapped by a variable such as the DSE location to evaluate the possibility, for example, that a given 9 nt window is true DSE with regards to its location to 3′ of the PAS. As the input variables for PolyFd, the DSE location and the DSE uracil content were employed. As a result, PolyFd program was established as a simple computational algorithm that comprises two types of membership function (i.e., the DSE location and the DSE uracil content) and an inference engine based on fuzzy logic. PolyFd was tuned by adjusting two membership functions that define the DSE location and the DSE uracil content for each PAS respectively. The thresholds for the PAS prediction were also determined simultaneously using both positive and negative datasets of the terminal sequences generated in this study (Table 1). Note that extremely small datasets were used to tune for the non-canonical variant PASes. This is an advantage of reasoning methods using fuzzy logic. Tuning was performed manually by trial prediction (average of 11.3 times per PAS) for each PAS, and the membership functions were finally determined empirically to result in the best accuracies (Correlation Coefficient: CC) on the condition that both sensitivity (SN) and specificity (SP) should exceed more than 70% for A(A/U)UAAA to compare the prediction performance with Polyadq and Erpin. For PolyFud, in addition to the above parameters, the membership functions that define the USE maximum uracil content were incorporated into the reasoning process of PolyFud, and tuned against the same datasets used for PolyFd. The tuning was performed in the same manner as PolyFd (average of 3.8 times per PAS).

Membership function — DSE location The membership functions of the DSE location were constructed for PolyFd algorithm based on the DSE distributions for each PAS, and fine-tuned by the trial prediction using the tuning datasets to result in the best CC in the PAS prediction (Table 2). Because the determination of the membership functions is arbitrary or subjective as shown in Fig. 1, the resultant membership functions could vary depending on designers. Nevertheless, a few differences in the membership functions did not affect the prediction results significantly (data not shown). This is one of the favorable features of fuzzy logic, which provides robustness in the prediction results. For PolyFud, the membership functions of the DSE location were determined in the same manner (Table 2). The incorporation of the sequence features of the USE into the reasoning process of PolyFud did not substantially affect the membership functions of the DSE location, which had been determined for PolyFd.

TABLE 2. Membership functions for DSE location

PAS motif	PolyFd				PolyFud			
	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
AAUAAA	10	25	35	55	10	25	35	55
AUUAAA	10	25	33	60	10	25	33	60
UAUAAA	10	28	40	70	10	30	40	60
AGUAAA	10	27	36	50	10	27	36	50
AAGAAA	10	27	35	70	10	27	35	70
AAUAUA	13	22	36	70	10	22	36	70
AAUACA	10	25	35	60	10	25	35	60
CAUAAA	10	27	37	55	10	27	37	55
GAUAAA	10	20	35	70	10	25	35	60
AAUGAA	10	25	35	60	10	25	40	65
ACUAAA	10	25	40	60	10	27	37	50
AAUAGA	10	30	45	70	10	30	42	60

X_1, X_2, X_3, and X_4 define a trapezoidal shape for the membership function.
The truth value will be zero if the DSE location is less than X_1 or more than X_4 nt downstream of a PAS. The truth value will be 1.0 if the DSE location is between X_2 and X_3. If the window location is either between X_1 and X_2 or between X_3 and X_4, the truth value will be the one corresponding to its window location.

TABLE 3. Membership functions for DSE uracil content

PAS motif	PolyFd		PolyFud	
	X_1	X_2	X_1	X_2
AAUAAA	0.33	0.56	0.33	0.78
AUUAAA	0.33	0.56	0.33	0.78
UAUAAA	0.33	0.56	0.33	0.78
AGUAAA	0.33	0.56	0.33	0.78
AAGAAA	0.33	0.56	0.33	0.78
AAUAUA	0.33	0.78	0.33	0.78
AAUACA	0.33	0.78	0.33	0.78
CAUAAA	0.33	0.56	0.33	0.78
GAUAAA	0.33	0.78	0.33	0.78
AAUGAA	0.33	0.56	0.33	0.78
ACUAAA	0.33	0.78	0.33	0.78
AAUAGA	0.33	0.78	0.33	0.78

X_1 and X_2 define the shape of a membership function for DSE uracil content.
The truth value will be zero if the DSE uracil content is X_1 or smaller than X_1.
The truth value will be 1.0 if the DSE uracil content is X_2 or larger than X_2. If the uracil content is between X_1 and X_2, the truth value will be the one corresponding to its uracil content.

Membership function — DSE uracil content The membership functions of the DSE uracil content were constructed for PolyFd algorithm for each PAS, and fine-tuned in the same manner as the membership functions of the DSE location (Table 3). Note that the concept of the uracil richness for the DSE could be more subjective than the DSE location because the DSE shows high sequence variability. We arbitrarily defined the DSE as the region with more than 60% uracil content in a 9 nt window in which at least six of them should be uracil residue. As in the DSE location, the membership functions of the DSE uracil content were not substantially affected when the USE was incorporated into the reasoning process of PolyFud (Table 3). These results might be evidence that the DSE and the USE independently contribute to the polyadenylation efficiency.

TABLE 4. Membership functions for USE maximum uracil content

PAS motif	PolyFud		
	X_1	X_2	Z
AAUAAA	0.66	0.78	0.4
AUUAAA	0.66	0.78	0.4
UAUAAA	0.56	0.66	0.5
AGUAAA	0.56	0.66	0.2
AAGAAA	0.56	0.66	0.4
AAUAUA	0.56	0.78	0.4
AAUACA	0.66	0.78	0.4
CAUAAA	0.66	0.78	0.2
GAUAAA	0.66	0.78	0.2
AAUGAA	0.66	0.78	0.2
ACUAAA	0.66	0.78	0.4
AAUAGA	0.56	0.66	0.4

X_1, X_2, and Z define the shape of a membership function for USE maximum uracil content.
The truth value will be zero if the DSE uracil content is X_1 or smaller than X_1.
The truth value will be Z if the DSE uracil content is X_2 or larger than X_2. If the uracil content is between X_1 and X_2, the truth value will be the one corresponding to its uracil content.

TABLE 5. Threshold used in fuzzy reasoning for PAS

PAS motif	Threshold	
	PolyFd	PolyFud
AAUAAA	0.90	0.51
AUUAAA	0.90	0.51
UAUAAA	0.95	0.74
AGUAAA	0.95	0.51
AAGAAA	0.92	0.51
AAUAUA	0.60	0.74
AAUACA	0.51	0.55
CAUAAA	0.90	0.51
GAUAAA	0.51	0.51
AAUGAA	0.90	0.51
ACUAAA	0.51	0.51
AAUAGA	0.51	0.74

If a combined truth value calculated by PolyF exceeds the threshold, the PAS is predicted as a 'positive signal'.

Membership function — USE maximum uracil content The membership functions of the USE maximum uracil content were generated to make use of the sequence features of the USE, and incorporated into the reasoning process of PolyFud. The definition of the USE is especially ambiguous; hence the membership functions defining the USE maximum content significantly depend on the subjective judgments of designers. Nevertheless, those membership functions were determined successfully by performing several (average of 3.8 times per PAS) trial predictions for each PAS in order to confer the best CC (Table 4). Incorporating the USE into the reasoning process did not substantially affect the shapes of membership functions defining the DSE characteristics, which had been determined for PolyFd. Hence, the number of trial predictions required for tuning PolyFud was much smaller than those required for polyFd.

Threshold for PAS prediction Thresholds for the prediction were determined empirically by the trial prediction for both PolyF and PolyFud (Table 5). The sensitivity of the PAS prediction can be enhanced by lowering the threshold, which results in the sacrifice of specificity of the prediction. The thresholds were determined to confer the best CC in the prediction for each PAS.

Prediction of the most canonical AAUAAA and its common variant AUUAAA We first examined the performance of prediction by PolyFd. The prediction results for A(A/U)UAAA were compared to those by Polyadq and Erpin (see Table 6 for AAUAAA and Table 7 for AUUAAA). Polyadq uses a pair of quadratic discriminant functions with three variables; a position weight matrix for the DSE, a weighted average of DSE hit position, and a downstream dimmer preference. Erpin uses a simple dinucleotide weight matrix only. Both algorithms were developed to predict the canonical AAUAAA and its common variant AUUAAA. The same datasets were used to test all the algorithms. Total numbers of positive sequences and negative sequences were scaled to calculate SN, SP, and CC. To verify whether prediction performances of PolyF depend on datasets used for tests, 3 sets of negative sequences from human chromosome 2, 4, 6 were used. The prediction performances of PolyFd and PolyFud were found to be stable across those datasets compared with Erpin and Polyadq. As a result, PolyFd identified AAUAAA with accuracy identical (−0.34% difference in CC) to that of Polyadq, but with 9.9% lower accuracy than Erpin when judged by CC. For AUUAAA prediction, the accuracy of polyFd was slightly lower (−1.9%) than Erpin and 5.0% lower than Polyadq. Overall, PolyFd achieved accuracies at approximately the same level as those two algorithms. This was a surprising achievement because the PolyFd is extremely simple algorithm that uses only two variables as inputs.

Next, we tested PolyFud in which the truth value of USE is incorporated into the reasoning process to verify whether the USE could contribute to improving the prediction accuracy. Total numbers of positive sequences and negative sequences were also scaled to calculate SN, SP, and CC. For AAUAAA, PolyFud achieved 10.1% higher accuracy than PolyFd (0.488 to 0.537), which was identical (−0.72%) to Erpin, and 4.8% higher than Polyadq respectively (Table 6). For AUUAAA, the accuracy was improved by 17.5% (0.493 to 0.579), which was 15.2% higher than Erpin, and 11.6% higher than Polyadq respectively (Table 7). PolyFud outperformed Polyadq for both AAUAAA and AUUAAA on the same datasets. PolyFud achieved accuracy identical to that of Erpin for AAUAAA prediction. Interestingly, the prediction accuracy for AUUAAA was considerably higher (15.3%) than Erpin. Overall, PolyFud outperformed Erpin on the same datasets. Higher accuracies in the A(A/U)UAAA prediction, which were gained by PolyFud, can be attributed to the following: PolyFud can incorporate the USE successfully into the reasoning logic as well as the DSE for the prediction, while neither Erpin nor Polyadq algorithms make use of the USE. Furthermore, prediction accuracies achieved by the previous methods based on machine learning including Erpin and Polyadq strongly depend on amount and quality of datasets available for learning. Generally, massive datasets are required for successful determination of weight matrices, or learning of prediction algorithms. Therefore, in case that sufficient datasets for learning process are not available, their prediction accuracy will be greatly diminished. In contrast, our approach using fuzzy logic allows us to generate membership functions of variables by using a limited number of datasets and a graphical representation, as shown in Fig. 1. In this study, the number of positive data for AUUAAA was much smaller than for canonical AAUAAA as described in Table 1. Nevertheless, PolyFud achieved higher accuracy in prediction for AUUAAA than AAUAAA, whereas Erpin detected AUUAAA with lower accuracy than AAUAAA.

Prediction of non-canonical variant PASes As described above, the previous algorithms based on machine learning methods require a large amount of datasets for a successful learning process. However, positive datasets for non-A(A/U)UAAA variant PASes are not available in sufficient amounts (Table 1), which makes previous methods extremely difficult to be applied for the prediction of those variant PASes. Furthermore, matrix-based motif discovery using weight matrices or position-specific scoring matrices was used to search A(A/U)UAAA motifs in previous methods based on machine-learning such as LDF, QDF, neural network, SVM and HMM. However, for non-A(A/U)UAAA variant PASes, it is difficult to search those motifs with single weight matrix or score matrix because those variant PASes vary significantly from A(A/U)UAAA, as shown in Table 8. This is a methodological weakness of matrix-based motif discovery, which makes it difficult for previous algorithms to predict non-canonical variant PASes. To overcome these drawbacks, we then tested whether our programs could predict the 10 non-canonical variant PASes by use

TABLE 6. Results of prediction test for AAUAAA, comparing PolyFud and PolyFd to Erin and Polyadq

Program	Positive dataset			Negative dataset (#2 CDS)				Negative dataset (#4 CDS)				Negative dataset (#6 CDS)				Over all (#2,4,6 CDS)				CC Diff (%) for	
	TP	FN	SN (%)	TN	FP	SP(%)	CC	TN	FP	SP (%)	CC	TN	FP	SP (%)	CC	TN	FP	SP (%)	CC	Erpin	Polyadq
Erpin	842	418	66.8	626	98	83.2	0.543	391	65	82.4	0.535	327	51	83.2	0.544	1344	214	83.0	0.541	–	–
Polyadq	829	431	65.8	608	116	80.4	0.506	374	82	78.5	0.485	303	75	76.8	0.464	1285	273	79.0	0.490	–	–
PolyFd	926	334	73.5	528	196	73.1	0.464	350	106	76.0	0.503	295	83	77.0	0.516	1173	385	74.8	0.488	−9.9	−0.34
PolyFud	896	364	71.1	593	131	79.7	0.533	372	84	79.4	0.530	317	61	81.5	0.554	1282	276	80.1	0.537	−0.72	4.8

#2, #4, #6 CDS, coding sequence from human chromosome 2, 4, 6.

TABLE 7. Results of prediction test for AUUAAA, comparing PolyFud and PolyFd to Erin and Polyadq

Program	Positive dataset			Negative dataset (#2 CDS)				Negative dataset (#4 CDS)				Negative dataset (#6 CDS)				Over all (#2,4,6 CDS)				CC Diff (%) for	
	TP	FN	SN (%)	TN	FP	SP [%]	CC	TN	FP	SP (%)	CC	TN	FP	SP (%)	CC	TN	FP	SP (%)	CC	Erpin	Polyadq
Erpin	240	150	61.5	511	63	84.9	0.526	287	48	81.1	0.486	303	53	80.5	0.480	1101	164	82.6	0.502	–	–
Polyadq	216	174	55.4	531	43	88.1	0.516	311	24	88.5	0.520	331	25	88.7	0.522	1173	92	88.4	0.519	–	–
PolyFd	288	102	73.8	435	139	75.3	0.496	244	91	73.1	0.467	275	81	76.4	0.511	954	311	75.0	0.493	−1.9	−5.0
PolyFud	304	86	77.9	453	121	78.7	0.569	261	74	77.9	0.559	297	59	82.5	0.615	1011	254	79.5	0.579	15.2	11.6

#2, #4, #6 CDS, coding sequence from human chromosome 2, 4, 6.

of Fuzzy membership functions in Tables 2–4, which are generated by a limited number of datasets. The variant PASes were predicted with the accuracies at the same level as A(A/U)UAAA (Table 8). The number of variants to be examined is not a constraint in fuzzy logic system because the membership function is independently determined for every variant.

The prediction accuracies obtained by PolyFd for the 10 single-nucleotide variants were from 0.408 to 0.628 (the average CC was 0.468), which were approximately equivalent (−2.7% difference in average CC) to those for A(A/U)UAAA (the average CC was 0.481). For some of the variant PASes, the prediction accuracies were considerably higher than those for A(A/U)UAAA (Table 8). A low accuracy for AAUACA prediction was foreseen because the DSE distribution for AAUACA was relatively scattered compared to A(A/U)UAAA (see Kamasawa and Horiuchi (8), Fig. 6). Contrary to our projection, the prediction accuracy obtained by PolyFd for AAUACA was 0.628, which was 35.3% higher than AAUAAA, and 26.4% higher than AUUAAA respectively. Interestingly, PolyFud did not significantly improve the prediction accuracy for AAUACA (7.2%) compared to the gains for other PASes (the average gain in CC for the rest 11 PASes was 18.7%). This result suggested that AAUACA might be related more strongly with the DSE, but more weakly with the USE, than other PASes. On the other hand, PolyFud improved the prediction accuracies more than 20% for UAUAAA (25.5%), AGUAAA (20.8%), CAUAAA (28.3%), and ACUAAA (35.7%), suggesting that those PASes are related more strongly with the USE than other PASes. Overall, the average prediction accuracy by PolyFud was improved from 0.468 to 0.550 (17.5% gain in accuracy).

TABLE 8. Results of prediction test for A(A/U)UAAA and non-canonical variant hexamers, comparing PolyFud to PolyFd

PAS motif	Algorithm	Positive dataset		Negative data (#2 CDS)					
		TP	FN	TN	FP	SN (%)	SP (%)	Accuracy (CC)	Gain[a] (%)
AAUAAA	PolyFd	926	334	528	196	73.5	73.1	0.464	
	PolyFud	896	364	593	131	71.1	79.8	0.534	15.1
AUUAAA	PolyFd	288	102	435	139	73.8	75.4	0.497	
	PolyFud	304	86	453	121	77.9	78.8	0.569	14.5
UAUAAA	PolyFd	144	51	322	120	73.8	73.5	0.470	
	PolyFud	160	35	340	102	82.1	78.0	0.591	25.5
AGUAAA	PolyFd	66	39	394	97	62.9	76.7	0.443	
	PolyFud	74	31	404	87	70.5	80.4	0.536	20.8
AAGAAA	PolyFd	93	52	1831	438	64.1	77.5	0.461	
	PolyFud	106	39	1793	476	73.1	77.9	0.524	13.7
AAUAUA	PolyFd	94	46	319	70	67.1	79.0	0.497	
	PolyFud	101	39	318	71	72.1	80.2	0.544	9.49
AAUACA	PolyFd	60	20	536	82	75.0	85.7	0.628	
	PolyFud	66	14	521	97	82.5	84.6	0.673	7.21
CAUAAA	PolyFd	40	20	265	89	66.7	72.7	0.414	
	PolyFud	44	16	280	74	73.3	78.6	0.531	28.3
GAUAAA	PolyFd	37	23	430	114	61.7	75.5	0.420	
	PolyFud	42	18	429	115	70.0	77.8	0.499	18.8
AAUGAA	PolyFd	71	34	839	265	67.6	74.0	0.437	
	PolyFud	76	29	862	242	72.4	76.8	0.503	15.1
ACUAAA	PolyFd	35	25	289	69	58.3	76.1	0.408	
	PolyFud	42	18	301	57	70.0	82.4	0.553	35.7
AAUAGA	PolyFd	34	16	279	63	68.0	79.1	0.501	
	PolyFud	34	16	308	52	68.0	82.9	0.545	8.90

[a] Gain is calculated as (CCpolyFud − CCpolyFd) / CCpolyFd.

Thus, PolyFud achieved successful prediction of non-A(A/U)UAAA variant PASes. As described in the previous section, our approach using fuzzy logic enables us to generate membership functions of variables by a limited number of datasets using a graphical representation. The number of datasets of the non-A(A/U)UAAA variant PASes used for program tuning in this study were between 18 (for AAUAGA) and 91 (for UAUAAA), which were extremely small compared to training sets generally required by machine learning methods (e.g., Polyadq used 280 mRNA sequences plus136 DNA sequences, Erpin used 2327 terminal sequences). Nevertheless, the prediction accuracy was satisfactory on average, indicating that a limited number of datasets was adequate for successful tuning of membership function used in PolyF. No correlation was observed between the amount of tuning data and the prediction accuracy achieved by PolyF. Moreover, the large amount of tuning data used for AAUAAA (1242 data) apparently did not enhance the prediction accuracies compared to those attained by small datasets for the variant PASes. This is due to a remarkable feature of PolyF in which human knowledge can be incorporated into its reasoning process to compensate for insufficient or limited data available for program tuning.

By comparing prediction accuracies by PolyF and PolyFud, the USE was shown to contribute significantly to improving the prediction accuracies. Legendre and Gautheret (5) claimed that the USE was not beneficial to prediction accuracy in their program, Erpin, whereas they observed significant sequence bias of the uracil content in the upstream regions of A(A/U)UAAA. Nevertheless, our result suggested that the USE is an essential element whose association with the authentic PASes is equal to that of the DSE.

Thus, by introducing Fuzzy logic, PolyF has conferred significant advances on variant PAS prediction based on a limited number of datasets and achieved high accuracy in the canonical A(A/U)UAAA prediction, but further study will be necessary to predict PASes with near perfect accuracy. The sequence features such as GC contents, dimmer preference, and Quadruplex G-rich Sequences (59) may be incorporated into PolyF algorithm to improve PAS prediction accuracy. However, the accuracy of PAS prediction, when it relies solely on the primary structure of terminal sequences, seems to be approaching the limit. To break through this limit, it is necessary to characterize and utilize the secondary and higher-order structure of the sequences around PASes for prediction, as Loke et al. and Zarudnaya et al. suggested (60, 61).

Classification of poly(A) sites and their characteristics Classification of human poly(A) sites has been proposed in previous studies, where poly(A) sites were divided into several categories based on their frequency of usage, as measured by relative EST counts (5, 24). For genes with multiple poly(A) sites, poly(A) sites could be generally grouped into "strong" and "weak". Cheng and colleagues reported that strong sites can be predicted more sensitively compared with weak sites using their algorithms, PolyA_SVM, which is one of the most recently developed algorithms (24). PolyA_SVM is unique among other PAS finder programs because it is designed not to predict PASes but to predict poly(A) sites with 15-*cis* regulatory elements using SVM. We examined whether prediction performance of PolyFud is subjective to those two poly(A) classes, or is consistent in comparison with PolyA_SVM even though it is difficult to precisely compare performance of PolyFud with PolyA_SVM because PolyFud predicts PASes appearing upstream of poly(A) sites, whereas PolyA_SVM predicts poly(A) sites downstream of PASes. The poly(A) sites, which were identified using EST-clustering (see Materials and methods), were divided into strong and weak sites based on the same criteria used by Cheng et al. (24). The −150/+150 nt regions surrounding poly(A) sites were stored as positive sequences. From those sequences, each 200 sequences with strong and with weak sites were randomly selected respectively as test sequences for PolyFud and PolyA_SVM. For PolyFud, a sequence was considered to be a true positive if a PAS was predicted positive within 32 nt upstream of the poly(A) site. For PolyA_SVM, a sequence was considered a true positive if a predicted poly(A) site (middle of HPR, default setting of 32 nt was used) was within 16 nt from a real poly(A) site. Interestingly, PolyFud predicted the strong sites with 54.5% SN whereas the weak sites were predicted with 61.5% SN, which is 12.8% more sensitive than for the strong sites (Table 9). Using the same sequences, PolyA_SVM predicted the strong sites as 67.1% more sensitive than for the weak sites (71.0% vs. 42.5% in SN), which is consistent with the results in their original report. It is apparently legitimate that strong sites could be detected more sensitively than weak sites because strong sites were reported to have DSEs with higher uracil content than weak site (5). We examined frequencies of PASes that appear upstream of poly(A) sites, and found that the most canonical AAUAAA dominated PASes upstream of strong sites (Table 10). It is surprising that 80% of the strong sites were associated with AAUAAA within 32 nt upstream regions, whereas only 42.5% of the weak sites harbor AAUAAA. Moreover, 65.1% of total PASes detected upstream of the weak sites were single-nucleotide variants of AAUAAA, while only 34% were occupied by those variants for the strong sites. To verify whether PolyA_SVM detects AAUAAA more sensitively than AUUAAA, each 200 sequences that contain AAUAAA and AUUAAA upstream of the poly(A) sites were randomly selected respectively from the positive sequences listed in Table 1. As a result, PolyA_SVM detected AAUAAA (78.0% SN) with much higher sensitivity than for AUUAAA (23.0% SN), while PolyFud successfully detected AUUAAA with rather higher sensitivity than AAUAAA (78.5% vs. 72.0% in SN) (Table 9). We concluded that the higher sensitivity in detection for strong sites than for weak sites by PolyA_SVM was attributed to the following: high frequency of AAUAAA, which dominated PASes for strong sites; and low sensitivity in detection for variant PASes represented by AUUAAA. On the other hand, PolyFud was found to consistently detect poly(A) sites regardless of their class, suggesting that sequence features (e.g., uracil content of DSE) around poly(A) sites are not substantially different between strong and weak sites.

Possible role of variant PASes in post-transcriptional gene regulation Our prediction test, in which the 10 non-canonical variants could be successfully identified by PolyF, suggested that those variants are also tightly associated with the DSE and the USE, as well as A(A/U)UAAA and the alternative polyadenylation processed by those non-canonical variant PASes could be a more widespread phenomenon than ever reported. To date, there are a few reports on variant PASes that have been verified by biochemical experiments. However, most of the single-nucleotide variants were experimentally verified to be functional in vertebrate cells (summarized in (3)). The importance of variant PASes in polyadenylation also is supported by recent large-scale analyses of variant PAS usage in human genomic sequences (3, 7, 8, 62). The variant PASes occupy a considerable fraction of total PASes and could produce isoforms of an mRNA with various lengths of 3′-UTR from a single gene as a result of alternative polyadenylation. Through the heterogeneity of 3′-UTR, each gene could be regulated differently by various mechanisms of post-transcriptional gene regulation. Among those mechanisms, micro RNAs (miRNAs) have emerged as key post-transcriptional regulatory elements in vertebrate cells (reviewed by Bartel, (63)). MiRNAs are small non-coding, endogenous RNAs (typically 21–22 nt long) that bind to target sites in 3′-UTRs of mRNAs for translational repression. MiRNAs are now recognized as a widespread and essential mechanism that plays crucial roles in gene regulation of versatile tissues or organs, including in the central nervous system (64, 65, 66). Legendre et al. (67) suggested that among isoforms of an mRNA with 3′-UTRs of different lengths attributed to alternative polyadenylation, only target-containing isoforms are sensitive to control by a cognate miRNA. Hence, the heterogeneity of the 3′-UTRs could directly affect gene regulation by miRNAs, and variant PASes possibly contribute to this newly emerged, extensive mechanism of gene regulation, which induces tissue-, cell-type-, disease- and developmental stage-specific patterns of expressions. To identify target sites of miRNAs, computational approaches have been made (68, 69, 70, 71, 72) in which the length of 3′-UTR of a gene of interest is crucial for identifying the target sites of miRNAs. However, the length of 3′-UTR is often unknown because the annotation of poly(A) sites are inadequate or completely unavailable in most of the genomic databases. For example, John et al. (72) arbitrarily took a 4000 nt sequence flanking 3′ end of the last exon for a transcript with no 3′-UTR annotation to predict the target sites. Lewis et al. (68) simply extended each annotated 3′-UTR with a 3′ flanking sequence of 2000 nt for their prediction experiment. PAS prediction algorithm may contribute to enhancing the accuracy of computational identification of miRNA target sites by providing precise information about lengths of the 3′-UTRs and their variability.

In conclusion, PolyF, a newly developed algorithm based on fuzzy logic, is the first PAS finder program that is able to predict the non-canonical variant PASes consistently. Moreover, PolyF outperformed Erpin and Polyadq for A(A/U)UAAA prediction when compared on the same datasets. By use of fuzzy logic, the more precise prediction of PASes has been achieved with our simple computational algorithm. We emphasize that PolyF appears to have revealed previously unknown features of the polyadenylation mechanism, which, when precisely identified, may be used to further improve prediction accuracy, as well as to design new investigational approaches pertinent to gene regulation. Thus, PolyF has remarkable features as a PAS prediction algorithm, which other machine learning methodologies have not exhibited. In addition, we demonstrated that strong poly(A) sites

TABLE 9. Comparison of PolyFud and PolyA_SVM for detection of different classes of poly(A) sites

	PolyFud			PolyA_SVM		
	TP	FN	SN (%)	TP	FN	SN (%)
Strong sites	109	91	54.5	142	58	71.0
Weak sites	123	77	61.5	85	115	42.5
Sites with AAUAAA[a]	144	56	72.0	156	44	78.0
Sites with AUUAAA[b]	157	43	78.5	46	154	23.0

[a,b] Randomly selected 200 sequences from the positive dataset for testing listed in Table 1.

TABLE 10. Frequency of the most canonical AAUAAA associated with strong and weak poly(A) site

poly(A) site classification	Strong site	Weak site
Number of poly(A) sites	200 (100%)	200 (100%)
Poly (A) sites with AAUAAA[a]	160 (80.0%)	85 (42.5%)
Poly(A) sites without AAUAAA	40 (20.0%)	115 (57.5%)
PASes within 32 nt from poly(A) site[b]	250 (100%)	258 (100%)
Fraction of AAUAAA	165 (66.0%)	90 (34.8%)
Fraction of 1 nt variant PASes	85 (34.0%)	168 (65.1%)

[a] Number of AAUAAA that appeared within 32 nt upstream of poly(A) sites.

[b] Number of PASes (AAUAAA and 11 of 1nt variants) that appeared within 32 nt upstream of poly(A) sites.

are dominated by the most canonical AAUAAA for polyadenylation while the single nucleotide variants including AUUAAA dominates PASes for the weak sites. Those weak sites possibly play a large part in alternative polyadenylation among other classes of poly(A) sites because weak sites are likely utilized with a tissue-type, cell-type, disease-type, or developmental-stage specific manner. Compared to the strong sites, weak sites are supposed to be difficult to be detected due to their weak expressions, which result in few EST counts. Therefore it is crucial for PAS finder programs to detect variant PASes that could process those weak sites.

ACKNOWLEDGMENT

We thank Prof. John E. Rash of Colorado State University and James L. Rash, mathematician, for useful discussions regarding this manuscript.

References

1. **Colgan, D. F. and Manley, J. L.:** Mechanism and regulation of mRNA polyadenylation, Genes Dev., **11**, 2755–2766 (1997).
2. **Mignone, F., Gissi, C., Liuni, S., and Pesole, G.:** Untranslated regions of mRNAs, Genome Biol., **3**, REVIEWS0004.
3. **Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M., and Gautheret, D.:** Patterns of variant polyadenylation signal usage in human genes, Genome Res., **10**, 1001–1010 (2000).
4. **Tian, B., Hu, J., Zhang, H., and Lutz, C. S.:** A large-scale analysis of mRNA polyadenylation of human and mouse genes, Nucleic Acids Res., **33**, 201–212 (2005).
5. **Legendre, M. and Gautheret, D.:** Sequence determinants in human polyadenylation site selection, BMC Genomics, **4**, 7 (2003).
6. **Beaudoing, E. and Gautheret, D.:** Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data, Genome Res., **11**, 1520–1526 (2001).
7. **Venkataraman, K., Brown, K. M., and Gilmartin, G. M.:** Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition, Genes Dev., **19**, 1315–1327 (2005).
8. **Kamasawa, M. and Horiuchi, J.:** Identification and characterization of polyadenylation signal (PAS) variants in human genomic sequences based on modified EST clustering, In Silico Biol., **8**, 0028.
9. **Edwalds-Gilbert, G., Veraldi, K. L., and Milcarek, C.:** Alternative poly(A) site selection in complex transcription units: means to an end? Nucleic Acids Res., **25**, 2547–2561 (1997).
10. **Beyer, K., Dandekar, T., and Keller, W.:** RNA ligands selected by cleavage stimulation factor contain distinct sequence motifs that function as downstream elements in 3′-end processing of pre-mRNA, J. Biol. Chem., **272**, 26769–26779 (1997).
11. **Chen, F., MacDonald, C. C., and Wilusz, J.:** Cleavage site determinants in the mammalian polyadenylation signal, Nucleic Acids Res., **23**, 2614–2620 (1995).
12. **Hu, J., Lutz, C. S., Wilusz, J., and Tian, B.:** Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation, RNA., **11**, 1485–1493 (2005).
13. **Wahle, E. and Keller, W.:** The biochemistry of polyadenylation, Trends Biochem. Sci., **21**, 247–250 (1996).
14. **Graber, J. H., Cantor, C. R., Mohr, S. C., and Smith, T. F.:** In silico detection of control signals: mRNA 3′-end-processing sequences in diverse species, Proc. Natl. Acad. Sci. U. S. A., **96**, 14055–14060 (1999).
15. **Aissouni, Y., Perez, C., Calmels, B., and Benech, P. D.:** The cleavage/polyadenylation activity triggered by a U-rich motif sequence is differently required depending on the poly(A) site location at either the first or last 3′-terminal exon of the 2′–5′ oligo (A) synthetase gene, J. Biol. Chem., **277**, 35808–35814 (2002).
16. **Brackenridge, S. and Proudfoot, N. J.:** Recruitment of a basal polyadenylation factor by the upstream sequence element of the human lamin B2 polyadenylation signal, Mol. Cell. Biol., **20**, 2660–2669 (2000).
17. **Moreira, A., Takagaki, Y., Brackenridge, S., Wollerton, M., Manley, J. L., and Proudfoot, N. J.:** The upstream sequence element of the C2 complement poly(A) signal activates mRNA 3′ end formation by two distinct mechanisms, Genes Dev., **12**, 2522–2534 (1998).
18. **Matis, S., Xu, Y., Shah, M., Guan, X., Einstein, J. R., Mural, R., and Uberbacher, E.:** Detection of RNA polymerase II promoters and polyadenylation sites in human DNA sequence, Comput. Chem., **20**, 135–140 (1996).
19. **Salamov, A. A. and Solovyev, V. V.:** Recognition of 3′-processing sites of human mRNA precursors, Comput. Appl. Biosci., **13**, 23–28 (1997).
20. **Tabaska, J. E. and Zhang, M. Q.:** Detection of polyadenylation signals in human DNA sequences, Gene, **231**, 77–86 (1999).
21. **Graber, J. H., McAllister, G. D., and Smith, T. F.:** Probabilistic prediction of *Saccharomyces cerevisiae* mRNA 3′-processing sites, Nucleic Acids Res., **30**, 1851–1858 (2002).
22. **Hajarnavis, A., Korf, I., and Durbin, R.:** A probabilistic model of 3′ end formation in *Caenorhabditis elegans*, Nucleic Acids Res., **32**, 3392–3399 (2004).
23. **Liu, H., Han, H., Li, J., and Wong, L.:** An in-silico method for prediction of polyadenylation signals in human sequences, Genome Informatics, **14**, 84–93 (2003).
24. **Cheng, Y., Miura, R. M., and Tian, B.:** Prediction of mRNA polyadenylation sites by support vector machine, Bioinformatics, **22**, 2320–2325 (2006).
25. **Yoshimura, S., Suemizu, H., Taniguchi, Y., Arimori, K., Kawabe, N., and Moriuchi, T.:** The human plasma glutathione peroxidase-encoding gene: organization, sequence and localization to chromosome 5q32, Gene, **145**, 293–297 (1994).
26. **Plant, M. H. and Laneuville, O.:** Characterization of a novel transcript of prostaglandin endoperoxide H synthase 1 with a tissue-specific profile of expression, Biochem. J., **344**, (Pt 3) 677–685 (1999).
27. **Martins, A. S., Greene, L. J., Yoho, L. L., and Milsted, A.:** The cDNA encoding canine dihydrolipoamide dehydrogenase contains multiple termination signals, Gene, **161**, 253–257 (1995).
28. **Silver Key, S. C. and Pagano, J. S.:** A noncanonical poly(A) signal, UAUAAA, and flanking elements in Epstein–Barr virus DNA polymerase mRNA function in cleavage and polyadenylation assays, Virology, **234**, 147–159 (1997).
29. **Sasaki, K., Kitagawa, Y., Shima, H., Irino, S., Sugimura, T., and Nagao, M.:** Production of shorter mRNA for protein phosphatase 2A beta by alternative poly(A) addition, Biochem. Biophys. Res. Commun., **170**, 169–175 (1990).
30. **Hu, Z. Z., Buczko, E., Zhuang, L., and Dufau, M. L.:** Sequence of the 3′-noncoding region of the luteinizing hormone receptor gene and identification of two polyadenylation domains that generate the major mRNA forms, Biochim. Biophys. Acta, **1220**, 333–337 (1994).
31. **Sugimoto, Y., Kusakabe, T., Kai, T., Okamura, T., Koga, K., and Hori, K.:** Analysis of the in vitro translation product of a novel-type *Drosophila melanogaster* aldolase mRNA in which two carboxyl-terminal exons remain unspliced, Arch. Biochem. Biophys., **323**, 361–366 (1995).
32. **Mans, B. J., Louw, A. I., and Neitz, A. W.:** Amino acid sequence and structure modeling of savignin, a thrombin inhibitor from the tick, *Ornithodoros savignyi*, Insect. Biochem. Mol. Biol., **32**, 821–828 (2002).
33. **Bishop, D. F., Calhoun, D. H., Bernstein, H. S., Hantzopoulos, P., Quinn, M., and Desnick, R. J.:** Human alpha-galactosidase A: nucleotide sequence of a cDNA clone encoding the mature enzyme, Proc. Natl. Acad. Sci. U. S. A., **83**, 4859–4863 (1986).
34. **Kim, K. H., Lee, K., Moon, Y. S., and Sul, H. S.:** A cysteine-rich adipose tissue-specific secretory factor inhibits adipocyte differentiation, J. Biol. Chem., **276**, 11252–11256 (2001).
35. **Schutz, T., Kairat, A., and Schroder, C. H.:** DNA sequence requirements for the activation of a CATAAA polyadenylation signal within the hepatitis B virus X reading frame: rapid detection of truncated transcripts, Virology, **223**, 401–405 (1996).
36. **Rabbitts, K. G. and Morgan, G. T.:** Alternative 3′ processing of *Xenopus* alpha-tubulin mRNAs; efficient use of a CAUAAA polyadenylation signal, Nucleic Acids Res., **20**, 2947–2953 (1992).
37. **Lay, J. M., Jenkins, C., Friis-Hansen, L., and Samuelson, L. C.:** Structure and developmental expression of the mouse CCK-B receptor gene, Biochem. Biophys. Res. Commun., **272**, 837–842 (2000).
38. **Yan, Y., Smant, G., Stokkermans, J., Qin, L., Helder, J., Baum, T., Schots, A., and Davis, E.:** Genomic organization of four beta-1,4-endoglucanase genes in plant-parasitic cyst nematodes and its evolutionary implications, Gene, **220**, 61–70 (1998).
39. **Wu, L., Ueda, T., and Messing, J.:** 3′-end processing of the maize 27 kDa zein mRNA, Plant J., **4**, 535–544 (1993).
40. **Trowsdale, J. and Kelly, A.:** The human HLA class II alpha chain gene DZ alpha is distinct from genes in the DP, DQ and DR subregions, EMBO J., **4**, 2231–2237 (1985).
41. **Wahlberg, M. H. and Johnson, M. S.:** Isolation and characterization of five actin cDNAs from the cestode *Diphyllobothrium dendriticum*: a phylogenetic study of the multigene family, J. Mol. Evol., **44**, 159–168 (1997).
42. **Zadeh, L. H.:** Fuzzy sets, Inf. Control., **8**, 338–353 (1965).
43. **Zadeh, L. H.:** Fuzzy algorithms, Inf. Control., **12**, 94–102 (1968).
44. **Horiuchi, J. and Kishimoto, M.:** Application of fuzzy control to industrial bioprocess in Japan, Fuzzy Sets and Sys., **128**, 117–124 (2002).
45. **Horiuchi, J.:** Fuzzy modeling and control of biological processes, J. Biosci. Bioeng., **94**, 574–578 (2002).
46. **Kishimoto, M., Kitta, Y., Takeuchi, S., Nakajima, M., and Yoshida, T.:** Computer control of glutamic acid production based on fuzzy clusterization of culture phases, J. Ferm. and Bioeng., **72**, 110–114 (1991).
47. **Shimizu, H., Miura, K., Shioya, S., and Suga, K.:** On-line state recognition in a yeast fed-batch culture using error vectors, Biotechnol. Bioeng., **47**, 165–173 (1995).
48. **Noguchi, H., Hanai, T., Honda, H., Harrison, L. C., and Kobayashi, T.:** Fuzzy neural network-based prediction of the motif for MHC class II binding peptides, J. Biosci. Bioeng., **92**, 227–231 (2001).

49. **Takahashi, H. and Honda, H.:** Prediction of peptide binding to major histocompatibility complex class II molecules through use of boosted fuzzy classifier with SWEEP operator method, J. Biosci. Bioeng., **101**, 137–141 (2006).
50. **Arima, C., Hakamada, K., Okamoto, M., and Hanai, T.:** Modified fuzzy gap statistic for estimating preferable number of clusters in fuzzy k-means clustering, J. Biosci. Bioeng., **105**, 273–281 (2008).
51. **Shimizu, K., Hayashi, S., Doukyu, N., Kobayashi, T., and Honda, H.:** Time-course data analysis of gene expression profiles reveals purR regulon concerns in organic solvent tolerance in *Escherichia coli*, J. Biosci. Bioeng., **99**, 72–74 (2005).
52. **Kim, S. Y., Lee, J. W., and Bae, J. S.:** Effect of data normalization on fuzzy clustering of DNA microarray data, BMC Bioinformatics, **7**, 134 (2006).
53. **Ando, T., Suguro, M., Kobayashi, T., Seto, M., and Honda, H.:** Selection of causal gene sets for lymphoma prognostication from expression profiling and construction of prognostic fuzzy neural network models, J. Biosci. Bioeng., **96**, 161–167 (2003).
54. **Takahashi, H., Aoyagi, K., Nakanishi, Y., Sasaki, H., Yoshida, T., and Honda, H.:** Classification of intramural metastases and lymph node metastases of esophageal cancer from gene expression based on boosting and projective adaptive resonance theory, J. Biosci. Bioeng., **102**, 46–52 (2006).
55. **Czernicki, T., Zegarska, J., Paczek, L., Cukrowska, B., Grajkowska, W., Zajaczkowska, A., Brudzewski, K., Ulaczyk, J., and Marchel, A.:** Gene expression profile as a prognostic factor in high-grade gliomas, Int. J. Oncol., **30**, 55–64 (2007).
56. **Takahashi, H., Tomida, S., Kobayashi, T., and Honda, H.:** Inference of common genetic network using fuzzy adaptive resonance theory associated matrix method, J. Biosci. Bioeng., **96**, 154–160 (2003).
57. **Sokhansanj, B. A., Fitch, J. P., Quong, J. N., and Quong, A. A.:** Linear fuzzy gene network models obtained from microarray data by exhaustive search, BMC Bioinformatics, **5**, 108 (2004).
58. **Chou, Z. F., Chen, F., and Wilusz, J.:** Sequence and position requirements for uridylate-rich downstream elements of polyadenylation signals, Nucleic Acids Res., **22**, 2525–2531 (1994).
59. **Kikin, O., Zappala, Z., D'Antonio, L., and Bagga, P. S.:** GRSDB2 and GRS_UTRdb: databases of quadruplex forming G-rich sequences in pre-mRNAs and mRNAs, Nucleic Acids Res., **36**, D141–148 (2008).
60. **Zarudnaya, M. I., Kolomiets, I. M., Potyahaylo, A. L., and Hovorun, D. M.:** Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures, Nucleic Acids Res., **31**, 1375–1386 (2003).
61. **Loke, J. C., Stahlberg, E. A., Strenski, D. G., Haas, B. J., Wood, P. C., and Li, Q. Q.:** Compilation of mRNA polyadenylation signals in *Arabidopsis* revealed a new signal element and potential secondary structures, Plant Physiol., **138**, 1457–1468 (2005).
62. **Gautheret, D., Poirot, O., Lopez, F., Audic, S., and Claverie, J. M.:** Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering, Genome Res., **8**, 524–530 (1998).
63. **Bartel, D. P.:** MicroRNAs: genomics, biogenesis, mechanism, and function, Cell., **116**, 281–297 (2004).
64. **Rash, J. E., Davidson, K. G., Kamasawa, N., Yasumura, T., Kamasawa, M., Zhang, C., Michaels, R., Restrepo, D., Ottersen, O. P., Olson, C. O., and Nagy, J. I.:** Ultrastructural localization of connexins (Cx36, Cx43, Cx45), glutamate receptors and aquaporin-4 in rodent olfactory mucosa, olfactory nerve and olfactory bulb, J. Neurocytol., **34**, 307–341 (2005).
65. **Krichevsky, A. M., King, K. S., Donahue, C. P., Khrapko, K., and Kosik, K. S.:** A microRNA array reveals extensive regulation of microRNAs during brain development, RNA, **9**, 1274–1281 (2003).
66. **Makeyev, E. V., Zhang, J., Carrasco, M. A., and Maniatis, T.:** The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing, Mol. Cell., **27**, 435–448 (2007).
67. **Legendre, M., Ritchie, W., Lopez, F., and Gautheret, D.:** Differential repression of alternative transcripts: a screen for miRNA targets, PLoS Comput Biol., **2**, e43 (2006).
68. **Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B.:** Prediction of mammalian microRNA targets, Cell, **115**, 787–798 (2003).
69. **John, B., Sander, C., and Marks, D. S.:** Prediction of human microRNA targets, Methods Mol. Biol., **342**, 101–113 (2006).
70. **Wang, X.:** miRDB: a microRNA target prediction and functional annotation database with a wiki interface, RNA, **14**, 1012–1017 (2008).
71. **Rajewsky, N. and Socci, N. D.:** Computational identification of microRNA targets, Dev. Biol., **267**, 529–535 (2004).
72. **John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S.:** Human MicroRNA targets, PLoS Biol., **2**, e363 (2004).