African Journal of Biotechnology Vol. 11(14), pp. 3215-3234, 16 February, 2012
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB10.2539
ISSN 1684–5315 © 2012 Academic Journals

Full Length Research Paper

Comprehensive recognition of messenger RNA polyadenylation patterns in plants

Xiaohui Wu[1], Guoli Ji[1]*, Qingshun Quinn Li[2] and Sun Zhou[1]

[1]Department of Automation, Xiamen University, Xiamen, Fujian, China.
[2]Department of Botany, Miami University, Oxford, Ohio 45056, USA.

Accepted 27 October, 2011

The polyadenylation of messenger RNA (mRNA) in eukaryotes is an essential step in gene expression. Currently, with the in-depth sequencing, a considerable amount of alternative poly(A) sites have been found in the coding sequences and introns, while there was little study on these unconventional poly(A) sites and their signals. To study the signals of mRNA polyadenylation, an effective poly(A) signal pattern recognition model was established to select and analyze the nucleotide patterns in the poly(A) site-related regions from large scale sequences generated from Sanger and next generation sequencing technologies. Our model, integrating a pattern and an assembly analysis pipelines and several visualization methods could be applied to various species. Through recognition of poly(A) patterns in three species including rice, *Arabidopsis* and *Chlamydomonas reinhardtii*, the experimental results showed that this model was able to select effective poly(A) signal patterns for poly(A) sites and alternative poly(A) sites to compare the poly(A) signals in different species and different regions, and to enhance the accuracy of poly(A) sites recognition to a larger extent.

Key words: Polyadenylation signal, pattern recognition, alternative polyadenylation.

INTRODUCTION

Maturation of eukaryotic mRNA involves three major steps of post-transcriptional processing, including 5' capping, splicing of introns and 3' end formation. The 3' end formation of mRNA includes two steps: The cleavage of pre-mRNA in a specific location [that is., poly(A) site] of 3'-UTR (3' untranslated region) and the addition of a poly(A) tail to the site (also known as polyadenylation). Polyadenylation is guided by *cis*-acting elements surrounding the poly(A) site (Hu et al., 2005), collectively known as the poly(A) signals. The 3'-UTRs containing *cis*-acting elements that may interact with RNA binding proteins and small non-coding RNAs, thereby affecting the function of RNA, such as mRNA stability, exportation, localization and translatability (Bartel, 2009; Buratowski, 2005; Hammell, 2002; Holec et al, 2006; Moor et al., 2005; Wickens et al., 2002) . Poly(A) tail marks the end of a gene, thus identification of poly(A) sites can help improve the gene structure prediction (Kan et al., 2001). Since a poly(A) signal is possible in the vicinity of a poly(A) site (Beaudoing et al., 2000), the recognition of poly(A) signal could be an alternative solution to the problem of poly(A) site predic-tion. Many eukaryotic genes possess multiple poly(A) sites (Tian et al., 2005; Wu et al., 2011), and thus undergo alternative polyadenylation (APA). APA can alter the nature of the 3'-UTR harboring many potential poly(A) signals for gene expression regulation. Recent large-scale studies have suggested that APA is widespread in many species (Jan et al., 2011; Mangone et al., 2010; Shen et al., 2008a; Tian et al., 2005; Wu et al., 2011). It is shown that over 50% of genes in humans, ~30% of genes in mice, ~50% of rice genes and up to 70% of *Arabidopsis* genes contain APA sites (Shen et al., 2008a; Tian et al., 2005; Wu et al., 2011). The APA sites located in coding sequences (CDS) and introns can significantly alter transcript sequences and their encoding proteins. Recent study showed that in *Arabidopsis* numerous novel poly(A) sites were located in CDS (11%) and intron(5.6%) (Wu et al., 2011). These APA sites provide a unique way to examine potential poly (A)

*Corresponding author. E-mail. glji@xmu.edu.cn. Tel. 86+05922 181049.

signals systematically and comprehensively to further explore the complex mecha-nism of APA.

To study the signals for mRNA poly(A) tailing, it is necessary to analyze the nucleotide patterns in the poly(A) site-related signal regions and select useful features from a large number of nucleoside sequences. In mammals, AAUAAA and its 11 single nucleotide variants have been identified as important hexamer signals (Beaudoing et al., 2000; Hu et al., 2005), among them, AAUAAA (~50%, usage frequency) and AUUAAA (~15%) are the most dominant ones, and both their sequences and relative locations are highly conserved. Currently, there are many methods for poly(A) signal or poly(A) site recognition in human (Akhtar et al., 2010; Beaudoing et al., 2000; Hu et al., 2005; Legendre and Gautheret, 2003; Liu et al., 2003). Legendre and Gautheret (2003) developed a program called Erpin which used 2-gram position-specific nucleotide acid patterns to characterize the sequences around candidate poly(A) signals. Liu et al. (2003) selected k-grams by an entropy-based algorithm and utilized support vector machine SVM to classify poly(A) sites. Cheng et al. (2006) used position specific scoring matrix (PSSM) to characterize patterns and also used SVM predict poly(A) sites. Recently, Akhtar et al. (2010) classified poly(A) sites into three classes and developed POLYAR program for the prediction. By testing different datasets, these methods have reached a reasonable specificity of 66 to 93% and sensitivity of 56 to 84%.

When compared with mammals, the poly(A) signals in plants are much less conserved. The canonical hexamer AAUAAA only occurs in ~10% of transcripts in *Arabidopsis* (Loke et al., 2005) and ~7% in rice (Shen et al., 2008a) and none of the *cis*-elements are highly conserved at the nucleotide level (Loke et al., 2005; Shen et al., 2008a), leading to very limited knowledge of plant poly(A) signals at present. Till now, several computational methods have been developed to predict poly(A) sites in different species including grape (Cai et al., 2008), rice (Shen et al., 2008a), *Chlamydomonas reinhardtii* (Chlamy) (Ji et al., 2010c; Shen et al., 2008b) and *Arabidopsis* (Ji et al., 2010a, b, 2007a, b; Loke et al., 2005; Tzanis et al., 2011). Ji et al. (2010a) developed a program PASS based on generalized hidden Markov model (GHMM) to predict poly(A) sites in *Arabidopsis*, and Shen et al. (2008a) extended this model and developed PASS_Rice for the prediction of rice poly(A) sites. Later, another program PAC (Ji et al., 2000a) was developed based on a classification model, using several feature representation methods to describe the sequences around poly(A) sites. These methods reached a specificity of 0.96 at the sensitivity of 0.97. Lately, Tzanis et al. (2011) utilized a distance-based scoring method to characterize emerging patterns and adopted different classifiers to predict poly(A) sites in *Arabidopsis*. These methods have their own strengths for the target species; however, they were all species specific and could hardly be applied on other species. Moreover, since the main purpose of these poly(A) site prediction methods was recognizing poly(A) sites rather than poly(A) signals, they tended to rely on the nucleotide distribution of the sequences around poly(A) sites rather than effective signal patterns. Till now, there is no universal poly(A) signal recognition model specifically for plants. Fortunately, these poly(A) site prediction tools allow users setting their own model parameters to enhance the identification accuracy by assigning the weight of the signal patterns (Ji et al., 2010a) and constructing a first order heterogeneous matrix (Ji et al., 2007b). Therefore, the poly(A) patterns selected by our poly(A) pattern identification model can be used to optimize the parameters of such poly(A) site recognition models.

At present, many motif recognition methods are available for detection of highly representative patterns for DNA sequences (Bailey and Gribskov, 1998; Hertz and Stormo, 1999; Hertzberg et al., 2005; Nuel, 2008; Ribeca and Raineri, 2008; Robin et al., 2007; Zhang et al., 2007). These methods searched the overrepresented patterns using some statistical models such as position-weighted matrix (PWM) and finite Markov-chain imbedding (FMCI) based on the pattern frequency. Discovery motif in DNA sequences helps to clarify the evolutionary relationships between sequences and determine the functions of sequences, which is also functional but not entirely practical in poly(A) signal recognition. Most of these methods were the approximations to the statistical models or additional training data set was required (Ribeca and Raineri, 2008). They targeted DNA sequences were not suitable for poly(A) signal pattern recognition. Moreover, they were slow in computing speed (especially when the model was more complex), so they could only be applied to a small amount of sequences each time instead of a large quantity of data. With the development of the next-generation sequencing technologies, more and more poly(A) sites were discovered for further exploration. There were only 8,160 poly(A) sites (called 8k dataset) in *Arabidopsis* (Ji et al., 2007b; Loke et al., 2005) from ESTs, while recently more than 70,000 poly(A) sites were found from NGS data (Wu et al., 2011). Therefore, effective recognition method specific for poly(A) signals to find important signal patterns to characterize the poly(A) sites, especially the APA sites is an urgent need.

In this study, an effective poly(A) signal pattern recognition model was established. First, the patterns with low frequency of occurrences were removed and the existing effective motif searching tools were integrated to further filter patterns. In particular, the pattern recognition using control region made the selected patterns specific to the studied region. Then, the selected patterns were clustered into different assemblies (or clusters) based on their similarities. Finally, the position-specific scoring matrix (PSSM) was used to characterize each assembly, and the nucleotide composition of the assembly could

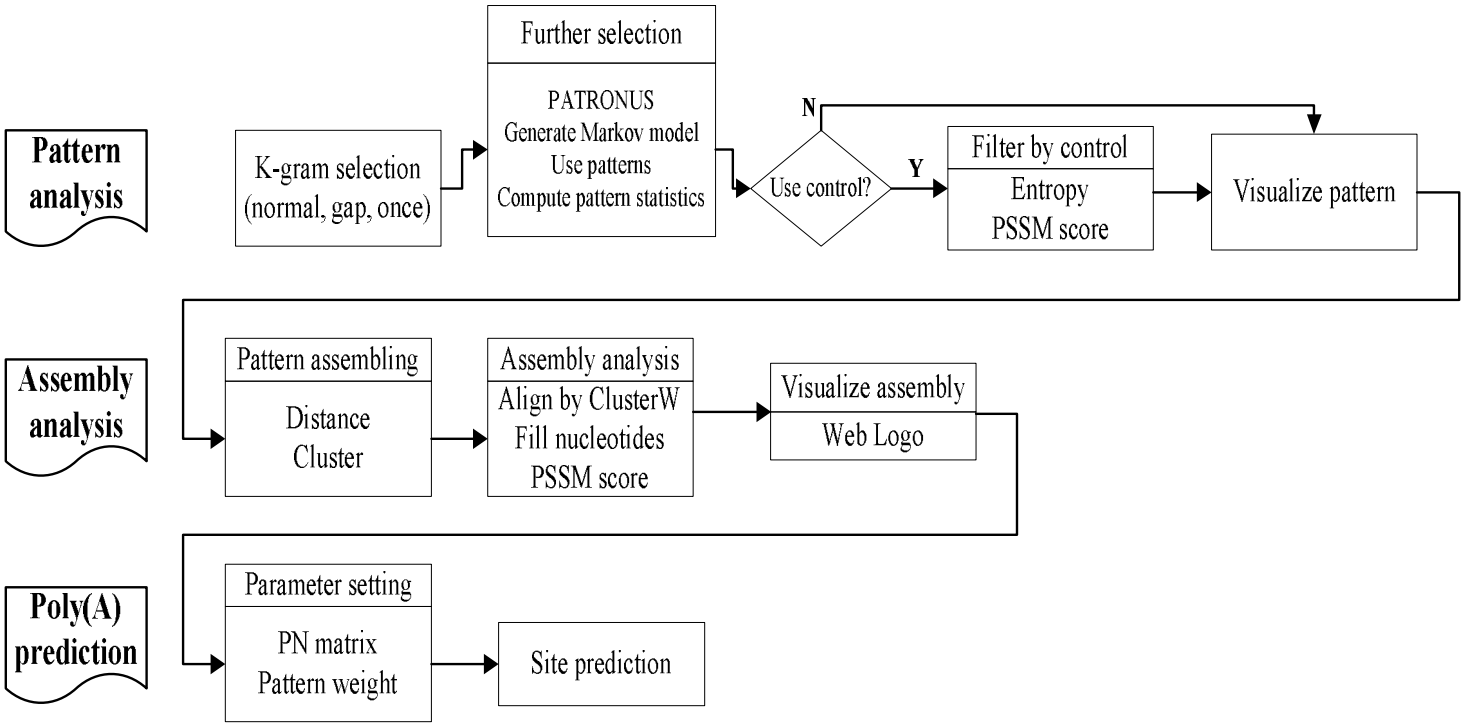


Figure 1. Process of the poly(A) pattern recognition model.

also be visualized by sequence LOGO.

METHODS

General flow

The process of our poly(A) pattern recognition model is shown in Figure 1, which includes the pattern pipeline and assembly analysis pipeline and can be applied to various species. In the pattern analysis pipeline, the weak patterns with low number of occurrences were filtered out by first accelerating the subsequent processes. Then, PATRONUS (Ribeca and Raineri, 2008), an efficient motif identification tool, was integrated to further select patterns. Particularly, a control region was allowed to make the selected signal patterns specific to a given region, which is specially adapted for signal pattern identification for poly(A) site with multiple signal regions. In plants, there are three typical signal regions: FUE (far upstream element), NUE (near upstream element), CS (cleavage site) and CE (cleavage element, including CE-L and CE-R). To identify the NUE-specific patterns, the FUE and CE can be considered as control regions. Finishing the pattern analysis pipeline, the assembly pipeline was adopted to cluster similar patterns for further analysis. First, similar patterns were assembled based on their edit distance (or Levenshtein distance) (Navarro, 2001) into different assemblies. Then, the PSSM was used to represent the expression level of each assembly in the sequences and each assembly was visualized by sequence LOGO to display its nucleoside composition. Finally, the parameters of existing poly(A) site recognition models like PASS (Ji et al., 2007b) or PAC (Ji et al., 2010a) can be optimized to improve the recognition effect, using the selected assemblies or patterns to construct the heterogeneous formation of the first order Markov matrix (Ji et al., 2007b, 2010a) or to weigh special patterns (Ji et al., 2010a).

Pattern searching

Here, a Perl script was written to implement the pattern analysis pipeline, which integrates different filtering rules and the existing motif recognition tools. Pattern usually refers to the conserved DNA sequence fragment, appearing significantly more or less than mere chance (Ribeca and Raineri, 2008). Since many of the current motif recognition methods have limitations on the number of the input sequences and the computing speed, here, a preliminary filtering was applied to filter out useless k-grams (sub-sequence with k consecutive nucleotides without mismatch) to enhance the efficiency of further processing. In pattern recognition, the number of occurrences of a pattern is one of the most important indicators, which is directly or indirectly used in almost all the motif recognition models to determine the representative patterns. However, with the increase of the k-gram length, the number of the k-grams increases exponentially. For example, there are as many as 4096 hexamers, given k = 6, which will undoubtedly increase the calculation time greatly if all these k-grams are considered. Since the k-grams with low number of occurrences or 0 occurrence merely emerge from the background sequences, it is worth removing these background noises to accelerate subsequent analysis.

Here, three k-gram scanning modes were used to filter out useless k-grams (Figure 2). Normally, given a short region like the NUE where most k-grams only appear once, we can use the simplest overlapping-mode to scan the studied region and obtain the number of occurrences of all k-grams. Whereas, if a k-gram appears more than once in a given region of one sequence, the once-mode can be adopted to decide whether or not this k-gram is present, where each k-gram is counted once and the last location it appears in the studied region is recorded. In addition, we also provided another mode gap-mode for calculating the number of occurrences of k-grams in long sequences (example, 400 nt) to avoid overlapping matches for some periodic k-grams like ATATAT. Each occurrence of such periodic k-gram strongly favors additional

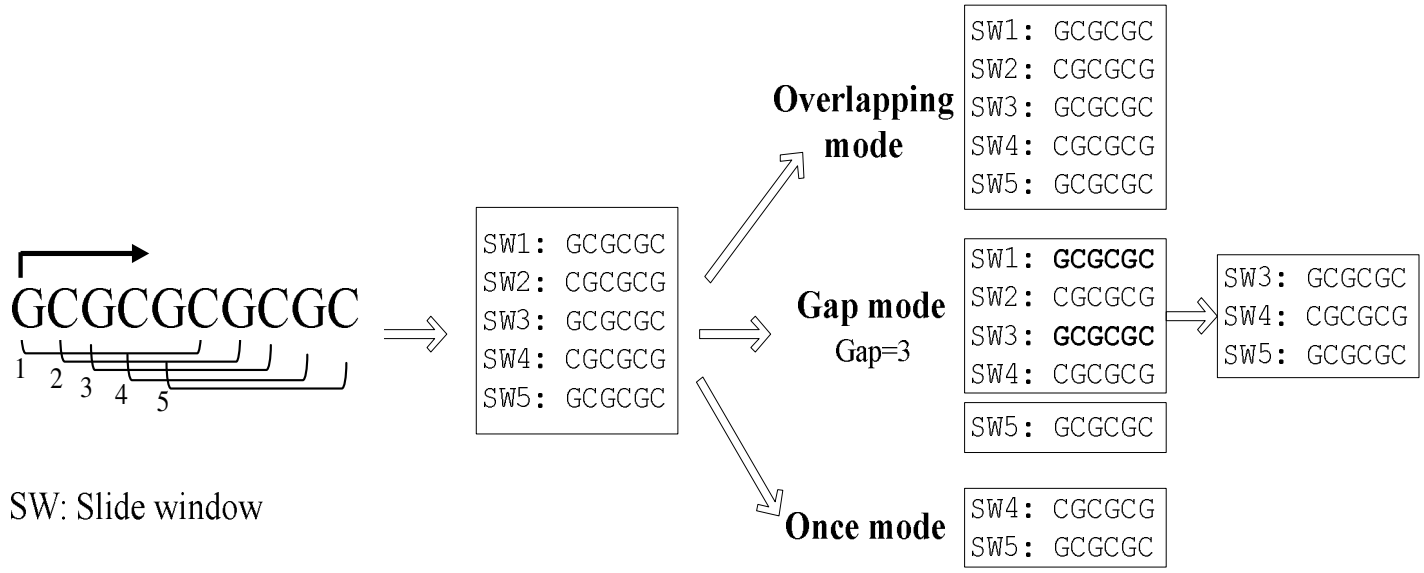


Figure 2. Three k-gram scanning modes.

occurrences in its immediate vicinity, which introduces a bias to most statistics (binomial, log-likelihood). The gap-mode can be adopted to correct this bias by preventing counting twice the mutually overlapping occurrences. For example, TATATATATATA would represent two occurrences of TATATA when self-overlap is prevented, but five occurrences of TATATA when self-overlap is allowed. Generally, for short sequences, these three modes return similar results when scanning k-grams with a certain length like pentamer or hexamer. While for long sequences, the once-mode or gap-mode may be more appropriate in that they can prevent counting too many times for the periodic k-grams, especially for a stretch of the same nucleotide like AAAAAAn. To examine the difference may be introduced by these three modes; we tested the NUE of *Arabidopsis* 8K dataset and selected the top 50 hexamers by each mode. As can be seen from Table S1, the number of occurrences of the same hexamer using overlapping-mode is much higher than using once-mode or gap-mode, especially for the A-stretch AAAAAA and T-stretch TTTTTT. While the selected hexamers and their frequencies from gap-mode or once-mode showed less difference. The once-mode was used for the following analysis since it was relatively simple and could get similar result as the gap-mode.

The top 300 patterns ranked by the frequency of occurrences calculated by once-mode were used to form the initial pattern space. The empirical number 300 was chosen according to previous studies (Loke et al., 2005; Shen et al., 2008a) where only top the 50 patterns were presented, which is sufficient to include all potential patterns. The next step is to filter statistically significant patterns using sophisticated motif recognition tool. To this end, another Perl script integrating a motif searching tool PATRONUS (Ribeca and Raineri, 2008) was implemented. At present, many algorithms or tools for motif searching are available (Bailey and Gribskov, 1998; Hertz and Stormo, 1999; Hertzberg et al., 2005; Nuel, 2008; Ribeca and Raineri, 2008; Robin et al., 2007; Zhang et al., 2007). PATRONUS was used here because it is far better and much faster than many of such tools (Ribeca and Raineri, 2008). Given a list of patterns with their numbers of occurrences from above k-gram searching flow, PATRONUS attempts to compute from its numerical estimate of the probability function one or both of the following indicators: the p-value and the z-value, which evaluate in different ways how improbable is the number of times the given k-gram is found in the sequence. Here, the final patterns with p-value<0.05 and z-value>0 were selected.

Pattern searching using control region

Pattern searching for DNA sequence is usually to find over-representative patterns by comparing the expected frequency of the background sequence with the frequencies of the candidate patterns. In this case, to find the patterns in a given signal region out of several signal regions, the difference between the background region and other signal regions will be ignored if only the given region is considered. For instance, given the NUE as the studied region, it is expected that the selected patterns are dominant in this region rather than other signal regions like FUE or CS. Therefore, to select patterns unique to a specific signal region, it is necessary to consider the background region as well as other signal regions as the control regions. Given a region, the aforementioned pattern searching process was adopted to get candidate patterns in this region first. Then, the background regions and other signal regions were set as control regions, and the pattern filtering process using the control region was applied on the candidate patterns to get patterns unique to the studied region. This pattern filtering process integrated the feature selection methods based on PSSM and entropy. First, the sequences of the given region and a control region were considered as two sample sets. Then, patterns were selected by the entropy-based method, and the PSSM scores of the candidate patterns in the given region were compared with those in the control region. If the PSSM score of a pattern in the studied region is higher than that in the control region, this pattern is recognized as the one specific to the studied region.

The followings are the steps of the entropy-based feature selection method. Given a sequence *S* of length *L* and a k-gram *G* of length *k*, the frequency of *G* in *S* is:

$$F(k) = \frac{O(k)}{W(k)}$$

Table S1. Top 50 hexamers in the NUE of *Arabidopsis* 8K dataset using three scanning modes.

Overlapping-mode		Gap-mode (gap=6)		Once-mode	
hexamer	**occu.**	**hexamer**	**occu.**	**hexamer**	**occu.**
AAUAAA	902	AAUAAA	880	AAUAAA	844
AUAAAA	608	AUAAAA	604	AUAAAA	589
AUAUAU	608	UAAUAA	574	UAAUAA	547
UAUAUA	603	AAAUAA	539	AAAUAA	519
UAAUAA	600	AUAAAU	509	AUAAAU	491
AAAUAA	551	AUAUAU	489	UAUAUA	438
AUAAAU	521	UAUAUA	479	AUAUAU	438
AAAAAA	521	AAUAAU	445	AAUAAU	423
UUUUUU	520	AUAAUA	442	AUAAUA	415
AAUAAU	479	UUUUAU	419	UUUUAU	407
AUAAUA	471	UAUAAA	409	AUAUAA	403
UUUUAU	420	AUAUAA	407	UAUAAA	400
UAUAAA	409	UUAUAU	393	UUAUAU	386
AUAUAA	408	UAAAAA	376	UUUUUA	372
UUAUAU	393	UUUAAU	374	UAAAAA	371
UAAAAA	377	UUUUUA	374	UUUAAU	369
UUUAAU	375	AAUAUA	367	AAUAUA	363
UUUUUA	374	AAAAAU	363	AAAAAU	354
AAUAUA	367	AAAUUU	360	AAAUUU	354
UUAUUU	363	UUAUUU	356	UAAUAU	347
AAAAAU	363	UAAUAU	352	UUAAUA	347
AAAUUU	360	UUAAUA	352	UAUAAU	345
UAAAUA	355	AAAAUA	349	UUAUUU	345
UAAUAU	354	UUUUAA	349	UUUUAA	344
UUUAUU	352	UAUAAU	347	AAAAUA	342
AAAAUA	352	UAAAUA	346	UUUAUU	339
UUAAUA	352	UUUAUU	344	UAAAUA	333
UUUUAA	349	UAAAAU	339	UAAAAU	331
UAUAAU	348	UAUUUU	337	AUUUAU	328
UAAAAU	340	AUUUAU	334	UUUAUA	327
AUUUAU	340	AAGAAA	331	UAUUUU	322
AAGAAA	338	UUUAUA	330	AUUUUU	320
UAUUUU	337	AAUGAA	327	AAGAAA	318
AAUGAA	334	AUUUUU	323	UUAUAA	318
UUUAUA	331	UUUGUU	321	AAUGAA	318
UUUGUU	326	AUGAAA	320	AAAUAU	314
AUUUUU	324	AAAAAA	318	UUUGUU	312
AUGAAA	322	UUAUAA	318	AUGAAA	312
UUAUAA	319	AAAUAU	316	AUUAAU	304
AAAUAU	316	AUUAAU	308	AAUUUU	302
AUUAAU	308	UUUUUU	305	AUUUUA	299
AUUUUA	303	AUUUUA	303	AAAAUU	299
UUUUGU	303	AAUUUU	302	UUUUGU	295
AAUUUU	302	AAAAUU	302	UAAAUU	293
AAAAUU	302	UUUUGU	300	UAUAUU	291
UAUAUU	295	UAAAUU	295	AUAUUU	290
UAAAUU	295	UAUAUU	294	AAAAAA	288
AUAUUU	293	AUAUUU	293	UAAUUU	287
UAAUUU	290	UAAUUU	290	UGUUUU	286
UGUUUU	290	UGUUUU	289	UUUUUU	285

Column 'Occu.' is the number of occurrences of the hexamer.

Where, $O(k)$ is the number of occurrences of *G* in *S*; $W(k) = L - k + 1$ is the number of sliding windows with length *k* in *S*. Here, the attribute of each k-gram was represented by its frequency. Given a sequence set, a signal region and a control region, the sequences of the signal region and the control region were trimmed as dataset 1 and 2, respectively. Given a k-gram *G*, its entropy value was calculated as follows:

(1) Initial setting. The dataset 1 and 2 are denoted as $\{D_i\}\ i = 1, 2$. The number of sequences in $\{D_i\}$ is N_i. The total number of sequences is $N_0 = N_1 + N_2$. There are two classes $C = \{C_1, C_2\}$. And the frequency value, ranging from 0 to 1, is divided into 50 intervals, having $N_f = 50$,

(2) Frequency calculation. The number of occurrences of *G* in each sequence of $\{D_i\}$ was calculated and denoted as $O_i(G)$. Then, the number of sliding windows in $\{D_i\}$ is counted as $W_i(G)$, and the frequency of *G* in $\{D_i\}$ is $O_i(G)/W_i(G)$.

(3) Probability calculation. Given a frequency interval $r(j)$ $(r(j) \in [0,1],\ j = 1, 2, \ldots, N_f,\ i = 1, 2)$, the number of sequences in $\{D_i\}$ where *G* is located and the frequency of *G* is in this interval is counted and denoted as $N(j,i)$. Then, the total number of sequences is $N(j) = N(1,j) + N(2,j)$ and the probability of sequence $S \in r(j)$ belonging to class C_i is $p(j,i) = N(j,i)/N(j)$. Finally, the probability of a sequence $S \in r(j)$ is calculated as $p(j) = N(j)/N_0$.

(4) Entropy calculation. The entropy value of *G* is: $H(G) = -\sum_{j=1}^{N_f} P(j) \sum_{i=1}^{2} P(j,i) \log_2 p(j,i)$.

Finally all the k-grams were ranked by their entropy values and the ones with entropy value less than a threshold were chosen for further analysis. Here, the threshold was determined empirically to get a reasonable number of k-grams (approximately 100) for further selection.

When the candidate patterns were selected by the entropy method, we then compared their PSSM score (Cheng et al., 2006; Hu et al., 2005) in the studied and control regions to further select valid patterns. First, the number of the occurrences of each candidate pattern was counted and a corresponding PSSM was generated for the given region. The PSSM has four rows and k columns, corresponding to the four bases $\{\mathrm{A, T, C, G}\}$ and the length of k-gram, respectively. Each element in the matrix is calculated as:

$$f_{i,j} = \frac{n_{i,j} + b/4}{\sum_{i=1}^{4} n_{i,j} + b}$$

Where, $f_{i,j}$ is the corrected relative frequency of nucleotide *i* at position *j*; $n_{i,j}$ is the number of occurrences of nucleotide *i* at position *j*; *b* is the pseudo weight (arbitrary, 1 in this case) to avoid the problem of zero entries in the frequency matrix and negative infinity in the log odds scoring matrix.

For a given sub-sequence with the length equal to the column number of the PSSM, its score is the sum of individual scores at all nucleotide positions:

$$S = \sum_{i=1..4} \sum_{j=1}^{k} \log_2(f_{i,j})$$

Higher score indicates the higher likelihood of the presence of a pattern similar to the k-gram represented by the PSSM. Finally, for each k-gram, we calculated its score in each location of the signal region and the control region. The final score of the k-gram in a region was the maximum score in this region. If the score of a k-gram in the signal region is larger than that in the control region, then this k-gram is constant, otherwise it is discarded since it is highly represented in the control region than in the signal region.

Assembling patterns into pattern-assembly

Through the aforementioned processes, the representative patterns in a given signal region were selected. To provide a more refined description of these patterns, they were further clustered into different pattern-assemblies based on their similarities. An assembly is a cluster of mutually overlapping patterns sharing similar nucleotide composition. Finally, these assemblies were characterized by PSSM (Hu et al., 2005) and visualized by sequence LOGO (Crooks et al., 2004).

First, the patterns were clustered into several assemblies based on their Levenshtein distance (edit distance) (Navarro, 2001). In information theory and computer science, the Levenshtein distance can be used to measure the difference between two sequences. The Levenshtein distance between two strings is defined as the minimum number of edits required to transform one string into the other. The allowable edit operations are insertion, deletion and substitution of a single character. This distance metric is equivalent to the negative of the score of a pairwise sequence alignment, where a match is 0, a mismatch is -1, the penalty for opening a gap is 0, and the penalty for extending a gap is -1. The dynamic programming algorithm based on the Needleman-Wunsch and Smith-Waterman algorithms can be used for global and local pairwise sequence alignments, respectively. This algorithm consumes memory and computation time proportional to the product of the length of the two strings. Here, the distance calculation was implemented in R (www.r-project.org) using 'stringDist' method in 'Biostrings' library.

Then, a hierarchical clustering method called 'Agnes' in 'cluster' library in R was adopted to compute agglomerative hierarchical clustering of the dataset using the earlier mentioned Levenshtein distance matrix. The Agnes algorithm constructs a hierarchy of clusters. At first, each observation is a small cluster by itself. Clusters merged until only one large cluster with all the observations remained. At each stage, the two nearest clusters are combined to form one larger cluster. An empirical cutoff 2.6 was used to group the patterns into pattern-assembly.

After clustering, patterns in the same cluster were aligned by ClustalW (http://www.ebi.ac. uk/clustalw). The gaps at both ends of the patterns after alignment were filled by nucleotides randomly generated based on the background nucleotides distribution in the studied region. Then, the weight of each filled pattern was set as the frequency of occurrences of that pattern in the studied region. Since there were large amount of sequences analyzed, a pattern usually occurs thousands of times, the file storing all redundant patterns will be too large to upload to WebLogo. Here, we used the relative frequency to replace the real frequency of each pattern to reduce the size of the output file to be visualized by Web Logo

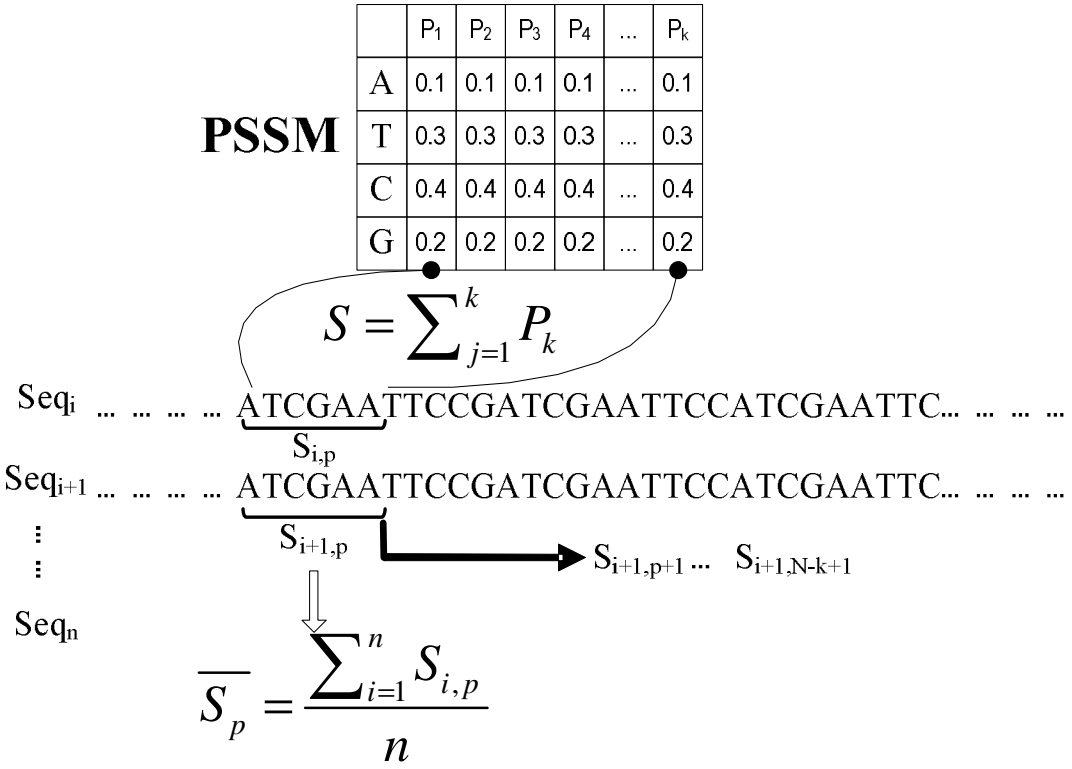


Figure 3. Calculation of PSSM for an assembly.

easier. To get the relative frequency, first, the minimum number of occurrences of the patterns in the assembly was divided by 100 to get a common divisor. Then, this divisor was applied on each pattern to get its relative number of occurrences. Each pattern in the assembly was written in a file for times of the relative number of occurrences. Finally, this file was uploaded and visualized by sequence logo using WebLogo (Crooks et al., 2004).

To detect whether an assembly was representative in the studied region for a given sequence set, the PSSM score (Hu et al., 2005) for this assembly was calculated. The flow to calculate the PSSM score of an assembly is shown in Figure 3. First, the given region of a sequence was scanned for the presence of the assembly and the score of the assembly was calculated, using the PSSM generated from that assembly. Then, the score of each position of the given region is the average of all positive scores in all sequences in the given dataset. Finally, the scores were smoothed by a sliding window with length 3. To calculate PSSM score, each aligned pattern-assembly was used to generate a PSSM with dimension 4**L*, where *L* is the length of the aligned pattern. For a given sub-sequence with the length equal to the column number of the PSSM, its score was the sum of individual scores at all nucleotide positions. Higher score indicates the higher likelihood of the presence of an assembly.

RESULTS

Datasets

We used sequences containing authenticated poly(A) sites from *Arabidopsis*, Chlamy and rice to test our model. Chlamy is a green algal species widely used to study photosynthesis and cellular movements' mechanisms (Mayfield, 2007; Wilson et al., 2008), and may be related to renewal energy production (Rupprecht, 2009). Rice is a dominant staple food crop and *Arabidopsis* is a widely studied model plant. The poly(A) sites of *Arabidopsis* includes 8160 sites from ESTs (called 8K) (Ji et al., 2007b). The Chlamy dataset contains 16,952 poly(A) sites (called 17K) (Shen et al., 2008b) and the dataset of rice contains 57,996 sites (called 55K) (Shen *e*t al., 2008a). All these poly(A) sites were mostly in 3'-UTRs. To analyze alternative poly(A) signals, the unconventional poly(A) sites from the next generation sequencing, including 3223 poly(A) sites in CDS and 4860 poly(A) sites in intron of *Arabidopsis* were used (Wu et al., 2011). For pattern recognition, the 180 nt sequences containing upstream 150 nt and downstream 30 nt (poly(A) sites included) around the poly(A) sites were trimmed. This range was chosen because it covers all plant poly(A) signal regions (Loke et al., 2005).

Patterns in the FUE and NUE

To show the effectiveness of our pattern recognition model, first we analyzed the patterns in the FUE and NUE. The NUE is the most conserved signal region in plant, where AAUAAA is the dominant pattern in

Arabidopsis as well as in rice (Loke et al., 2005; Shen et al., 2008a) and UGUAA is dominant in Chlamy (Shen et al., 2008b). Though the FUE is much less conserved than the NUE, it is still a valid signal region detected by conventional genetic mutagenesis experiments (Li and Hunt, 1997; Rothnie, 1996; Rothnie et al., 2001). Thus, here we chose the FUE and NUE to analyze the patterns and compare the difference of signals in the FUE with those in the NUE for each species, and also to compare the patterns in the same signal region among the three species.

Here, the patterns were selected by the pattern recognition flow with a control region. For each species, the FUE was considered as control region when the NUE was targeted, and vice versa. Based on previous studies (Loke et al., 2005; Shen et al., 2008a, b), the hexamer was analyzed for *Arabidopsis* and rice, while the pentamer was used for Chlamy. We listed top three assemblies and their patterns for the FUE and NUE for each species (Tables 1, S2 and S3).

In order to examine the nucleoside composition of each assembly, we also generated sequence LOGO. It is noteworthy that some dominant patterns, such as AAUAAA, may not be clearly shown, because there were other patterns in the same assembly and the LOGO only displays the nucleotide composition in each position of the assembly. It can be seen from Table 1, in *Arabidopsis*, the FUE is UC-rich, while the NUE is UA-rich. In rice, the FUE is UG-rich and UA-rich, while the NUE is UG-rich. In Chlamy, the FUE is UG-rich and UGUAA clearly appears in the LOGO of the NUE (46%, Table S2), suggesting that UGUAA is highly conserved. The UGUAA also exists in *Arabidopsis* (17%) and rice (16%) (Table S3), while it is mainly in the FUE rather than NUE, suggesting a shift in function. These results are consistent with the published results (Loke et al., 2005; Shen et al., 2008a, b), which also demonstrates the effectiveness of our method. Using the selected patterns and the assemblies, we also compared difference of the signals among these three species, where the signals were similar in *Arabidopsis* and rice, but the signals in Chlamy are significantly different from the other two species. Figure S1 shows the poly(A) signals in the FUE and NUE for a typical poly(A) site in these three species, where the most dominant assembly from Table 1 was used. Figure S1 also clearly shows the similarity of the poly(A) signals between *Arabidopsis* and rice, as well as the shift of the UGUAA assembly from their FUE to the NUE of Chlamy.

Since some patterns are of positional propensity, for example, the NUE signals are usually located upstream -10 to -30 nt of a poly(A) site, we also provided another pattern selection flow according to the maximum number or the total number of the occurrences of the patterns in the studied region. After the poly(A) signal patterns were obtained, a Perl script was used to count the number of occurrences of each individual pattern at every location of the given region by once-mode. Normally, for a given region like the NUE, most of the patterns only appeared once in one sequence, whereas if a pattern appears more than once, then it is counted for only one time and its position is the last location. As shown in Figure 4, AAUAAA is the most dominant in the NUE of *Arabidopsis* and rice, while other patterns are not so apparent, which also shows the low conservation of plant poly(A) signals. In contrast, UGUAA in Chlamy is significantly higher than other patterns. In the FUE of rice, the number of occurrences of three patterns (UGUAAU, UUGUAA and UGUAAA) are dramatically increased in the vicinity of the NUE, which demonstrates that this kind of analysis is conducive for searching the dominant position of the pattern.

To reflect whether the selected assemblies could well characterize the studied region, we calculated the PSSM scores along the 180 nt sequence. As shown in Figure 5, for the assemblies specific to the FUE, their scores in the NUE were significantly reduced while the scores were distributed uniformly in the FUE, indicating that there was no obvious positional propensity for patterns in the FUE. Similarly, the scores of the assemblies in the NUE were significantly higher in the NUE than in the FUE. In particular, in Chlamy, the scores of the FUE assemblies were reduced dramatically near the NUE. Moreover, the scores of the NUE assemblies in the NUE were the most distinct from the rest regions among the three species, especially for the assembly containing UGUAA. This is consistent with the fact that the UGUAA is accounted for 50% in Chlamy and also demonstrates that the score curve can reflect the likelihood of the presence of an assembly. In addition, there is an apparent peak in the NUE of the curve while the curve is rather smooth outside the NUE, indicating that the signals in the NUE are the most conserved. In contrast, though the length of FUE region is relatively long, and there is no significant fluctuation along this region, the scores of the FUE assemblies in the FUE were still significantly higher than in other regions.

Patterns for alternative poly(A) sites

With the development of next generation sequencing technologies, quite a few novel poly(A) sites were found to be in CDS and introns. Here, we also identified the poly(A) signals of these unconventional poly(A) sites to see whether a different group of signal patterns were used in these APA sites from the 3'-UTR poly(A) sites. To this end, we compared the patterns in the NUE since the NUE is the most conserved region.

As shown in Figure 6A and B, there is no significant assembly of CDS APA sites which is highly conserved in the NUE, while there are some weak peaks in the NUE in the score curves of some assemblies of intron APA sites. This result indicates that the poly(A) signals of poly(A) sites in CDS or introns may be less conserved than in 3'-

Table 1. Patterns and assemblies in the FUE and NUE of Chlamy, *Arabidopsis* and rice. Column 'Pattern' is the top three patterns of the aligned patterns of that assembly.

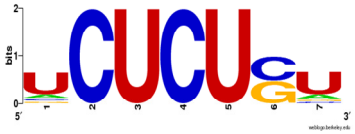
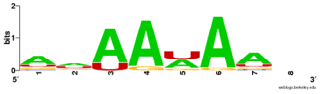
Species	S/N	FUE			NUE		
		Pattern	**Occu.**	**Logo**	**Pattern**	**Occu.**	**Logo**
Arab	1	--UUGUAA----GUAAA-UAUGUA--	1423 1016 1047		--AAUAAA----UAAUAA--AUAAAA-	844 547 589	
	2	UUCUUC---UCUUCU---CUUCUU	1069 1151 1013		-UCAAUAAUCAAU--UUAAUC	243 247 182	
	3	UCUCUG--CUCUCUUCUCUC-	656 742 678		UUUAGUUUUGGU	130 164	
Rice	1	---UGUAAU---UUGUAA-----GUAAA-	9237 8066 7271		AAUAAA---AAAUAA-AUAAAA--	3714 2432 2291	
	2	-UAUAUGGUAUAU--UAUAUU	4779 4188 4327		UAUAUA--AUAUAUAAUAUA-	2128 2669 2043	
	3	AUGCAU--UGCAUG-UGCUUG	4137 4533 4698		-GAUGAUUGAUGA--GAUGAA	1015 1303 1134	

Table 1. Continue

Chlamy	1	GGAUG-GAUG--GAUGG	3619 3218 3402		UGUAA-UGCAA--GCAAG	8596 1415 783	
	2	CGCUG--GCUGGUGCUG-	3344 4578 4710		CGUGUGCGUG-ACGUG-	1475 915 736	
	3	CAUGGCAUGC	3596 3630		UUGUAUUGAAUUGCA	2114 452 801	

Column 'Occu.' is the number of occurrences of each pattern in that assembly.

Table S2. The number of occurrences of each pattern of each assembly in the NUE.

Arabidopsis		Rice			Chlamy						
No.	Pat.	Pat#	Pat (%)	No.	Pat.	Pat#	Pat (%)	No.	Pat.	Pat#	Pat (%)
1	**AAUAAA**	844	10	1	**UGUAA**	8596	105	1	**AAUAAA**	3714	46
1	AUAAAA	589	7	1	UGCAA	1415	17	1	AAAUAA	2432	30
1	UAAUAA	547	7	1	GCAAG	783	10	1	AUAAAA	2291	28
1	AAGAAA	318	4	1	UGAAA	570	7	1	UAAUAA	1923	24
1	CAAUAA	277	3	2	CGUGU	1475	18	1	AUAAUA	1870	23
1	GAAUAA	269	3	2	GCGUG	915	11	1	GAAUAA	1757	22
1	UAAUGA	253	3	2	ACGUG	736	9	2	AUAUAU	2669	33
1	UGUAAU	250	3	2	CCGUG	658	8	2	UAUAUA	2128	26
1	AUAAAG	242	3	2	UGGUG	624	8	2	AAUAUA	2043	25
1	GUAAUA	219	3	2	UCGUG	549	7	3	UGAUGA	1303	16
1	UAAAAC	174	2	3	UUGUA	2114	26	3	GAUGAA	1134	14
2	AUCAAU	247	3	3	UUGCA	801	10	3	GAUGAU	1015	12
2	UCAAUA	243	3	3	UUGAA	452	6	3	AGAUGA	752	9
2	UUAAUC	182	2								
2	UCAAUG	145	2								
2	UGAAUC	141	2								
3	UUUGGU	164	2								
3	UUUAGU	130	2								

Column 'No.' is the index of the assembly. 'Pat' is pattern.

Table S3. The number of occurrences of each pattern of each assembly in the FUE.

Arabidopsis	Rice	Chlamy	Pat (%)	No.	Pat.	Pat#	Pat (%)	No.	Pat.	Pat#	Pat (%)
No.	Pat.	Pat#									
1	U**UGUAA**	1423	17	1	**UGUAA**U	9237	16	1	GGAUG	3619	21
1	UAUGUA	1047	13	1	UUGUAA	8066	14	1	GAUGG	3402	20
1	UGUAAA	1016	12	1	UGUAAA	7271	13	1	UGAUG	3218	19
1	AUGUAA	1007	12	1	AUGUAA	7171	12	1	GAUGC	3209	19
1	UAAUGU	822	10	1	UAUGUA	6698	12	2	UGCUG	4710	28
1	AAUGUA	784	10	1	UGUACU	6353	11	2	GCUGG	4578	27
1	UGUAAC	581	7	1	GUUGUA	6007	10	2	CGCUG	3344	20
1	GUGUAA	569	7	1	AUAUGU	5869	10	2	GCUGA	3139	19
2	UCUUCU	1151	14	1	UGUACA	5652	10	3	CAUGC	3630	21
2	UUCUUC	1069	13	2	UAUAUG	4779	8	3	CAUGG	3596	21
2	CUUCUU	1013	12	2	UAUAUU	4327	7				
2	CCUUUU	737	9	2	GUAUAU	4188	7				
2	UCCUUU	667	8	2	GUAAAU	4176	7				
2	UUCCUU	630	8	3	UGCUUG	4698	8				
2	UUUCCU	604	7	3	UGCAUG	4533	8				
3	CUCUCU	742	9	3	AUGCAU	4137	7				
3	UCUCUC	678	8	3	GUACAU	3459	6				
3	UCUCUG	656	8								
3	CUCUGU	641	8								

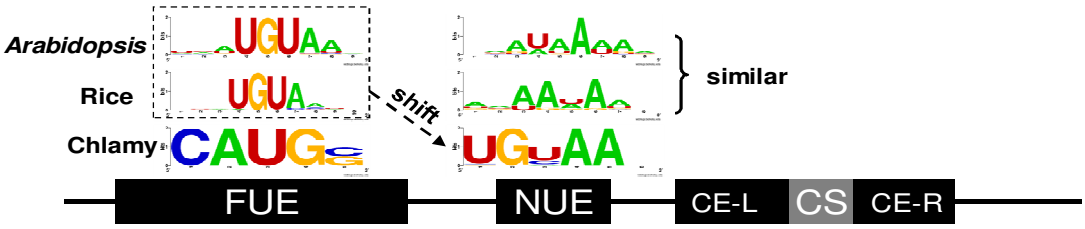


Figure S1. Poly(A) signals in the FUE and NUE for a typical poly (A) sitw in 3'-UTR.

UTRs . As shown in Figure 6D, for intron poly(A) sites, AAUAAA is still the most dominant pattern in the NUE, while it is much less significant than that of 3'-UTR poly(A) sites. For CDS poly(A) sites, AGAAGA is the most apparent, but more conserved in the CS than in the NUE (Figure 6C). In Figure 6E and F, the NUE of CDS poly(A) sites is rich in U or C, while the NUE of intron poly(A) sites is U/A-rich. The corresponding patterns of these assemblies are listed in Table S4.

Improvement of poly(A) site prediction using selected patterns

There are several poly(A) site prediction tools such as PASS_Rice (Shen et al., 2008a), PASS (Ji et al., 2007b) and PAC (Ji et al., 2010a), which allow users to set their own model parameters for specific site recognition for a given species. Normally, we can set parameters like the weights of the signal patterns (Ji et al., 2010a) and a first order heterogeneous matrix (Ji et al., 2007b). Here, the identified poly(A) patterns were used to optimize the model parameters of PASS_Rice (Shen et al., 2008a) for rice to improve the accuracy of poly(A) site prediction. To improve the parameters of poly(A) site prediction model, the aforementioned selected patterns in the NUE and FUE of rice were used to set the weights of patterns and to construct the first-order Markov matrix (Ji et al., 2007b). To set the weights of patterns, the frequencies of the selected patterns were used as the weights. To construct the Markov matrix, first, the selected patterns were used to form a list of vectors $V_1,...,V_k$, where *k* is the length of the pattern. Here the $V_1,...,V_k$ is the same as the first-order Markov matrix. V_1 is a vector storing the probabilities

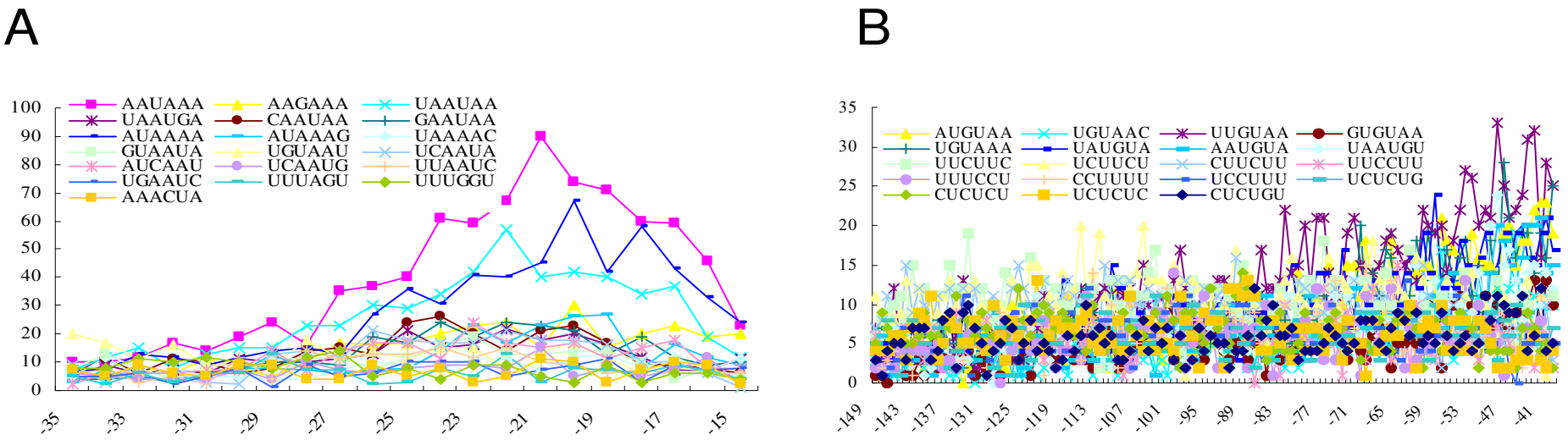


Figure 4. Occurrences of individual pattern at every location. X-axis is the location, Y-axis is the number of occurrences of each pattern in the given region. (A) NUE of *Arabidopsis*; (B) FUE of *Arabidopsis*; (C, D) same as (A) and (B) except for rice; (E, F) same as (A) and (B) except for Chlamy. The poly(A) site is at position -1. The upstream sequence of the poly(A) site is with '-' designation, and the downstream sequence is in '+' designation.

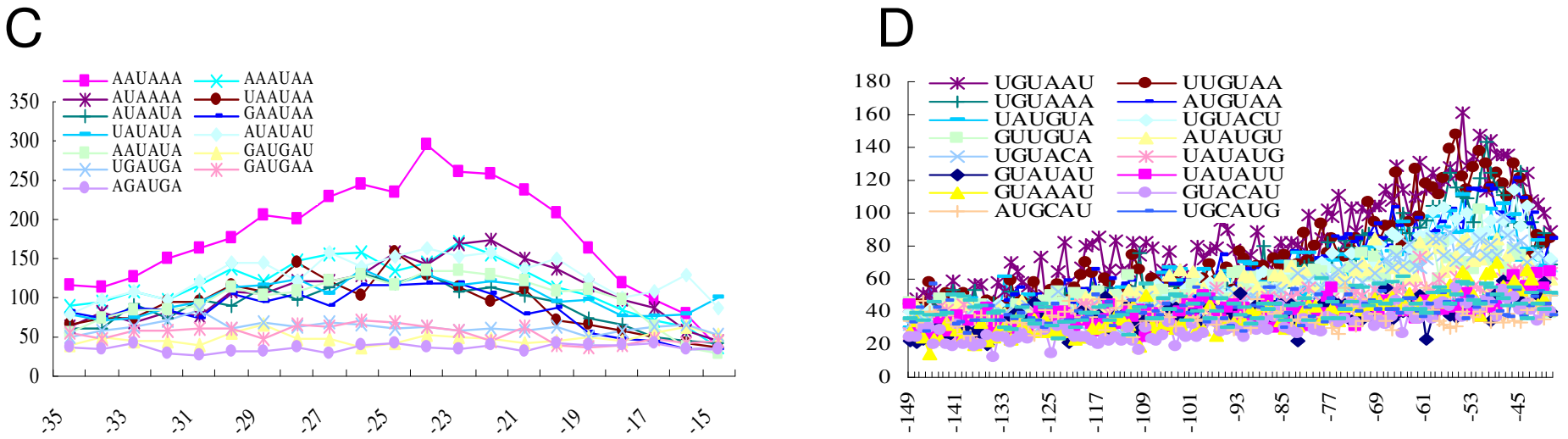


Figure 4. Contd.

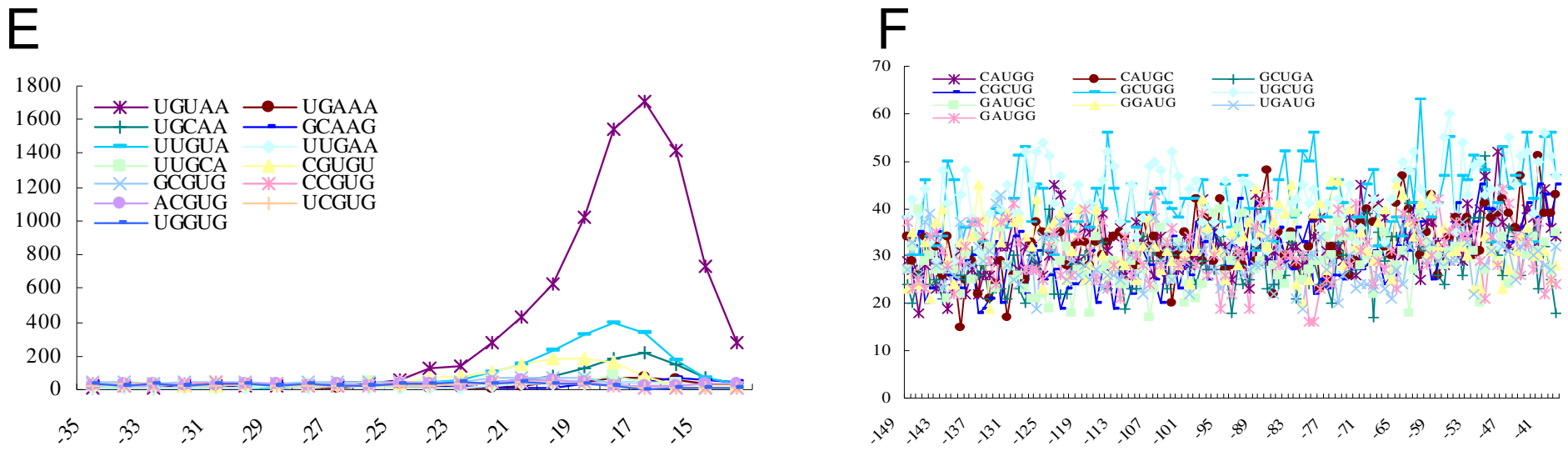


Figure 4. Contd.

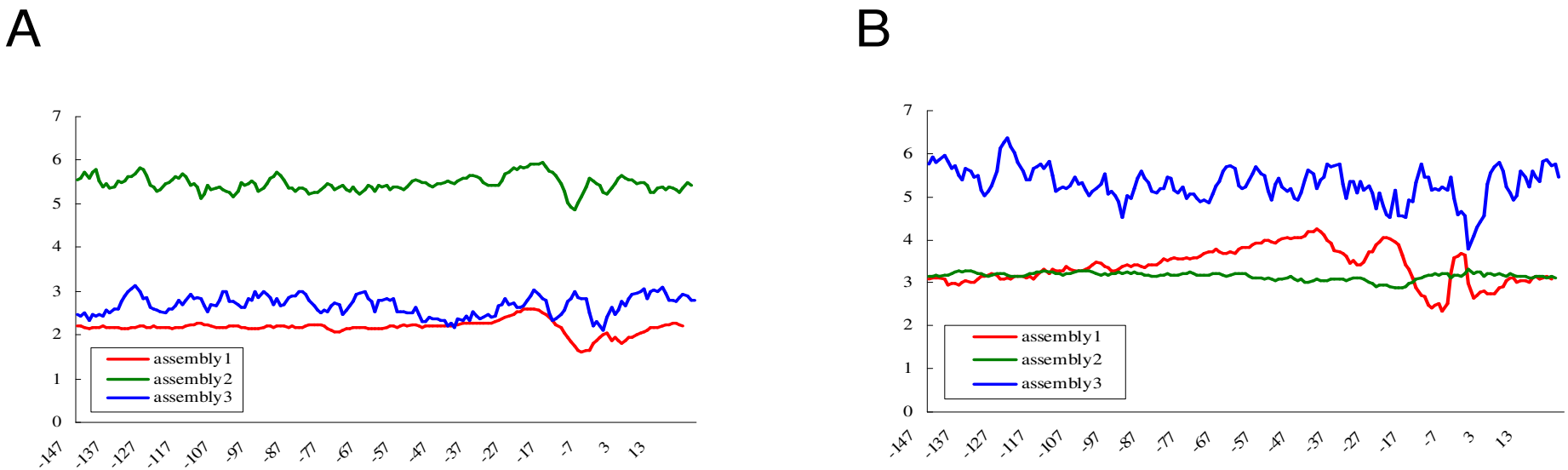
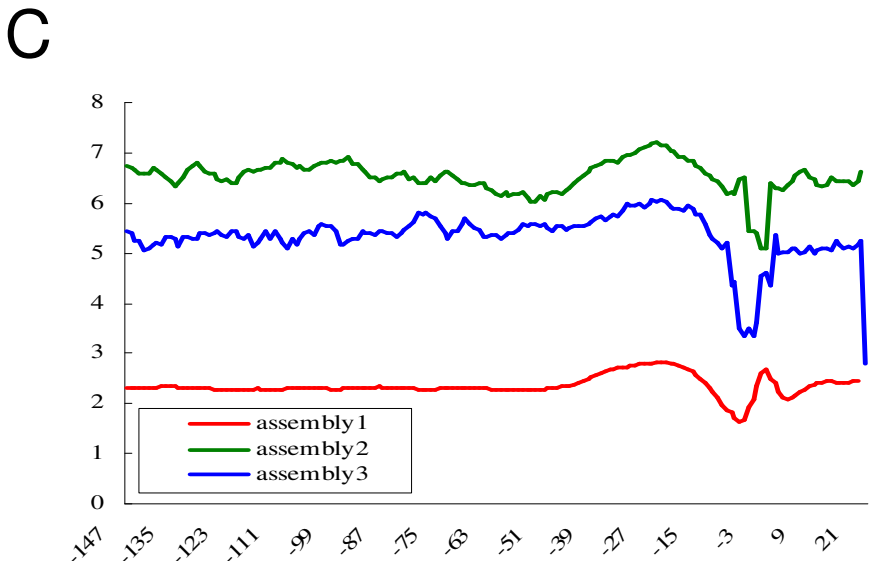


Figure 5. PSSM scores of assemblies in the FUE and NUE. X-axis is the location, Y-axis is the PSSM score of each of the top three assembly in the given region. (A) NUE of *Arabidopsis*; (B) FUE of *Arabidopsis*; (C, D) same as (A) and (B) except for rice; (E, F) same as (A) and (B) except for Chlamy.



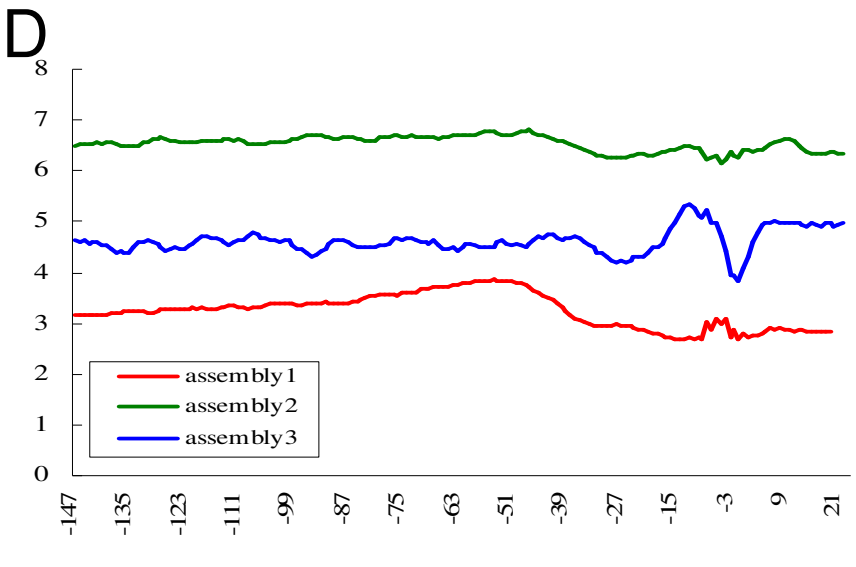
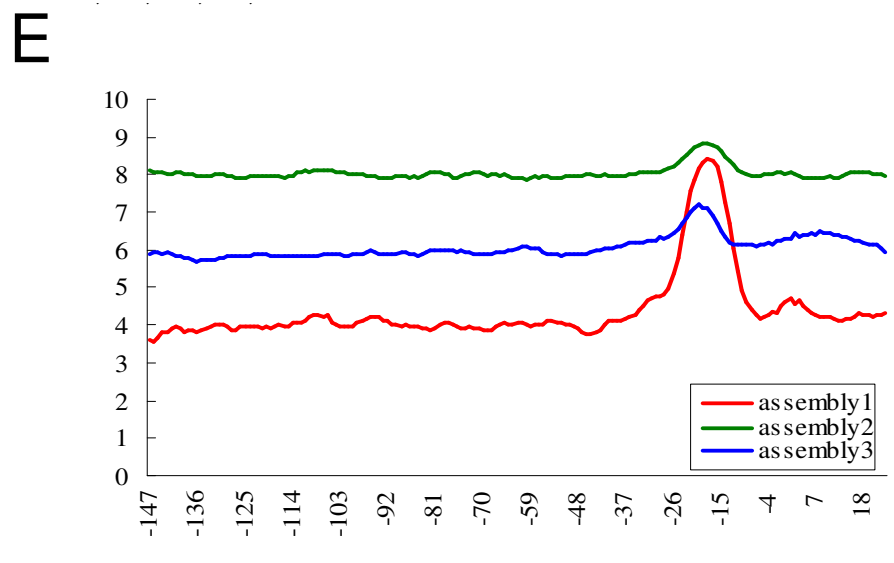


Figure 5. Contd.



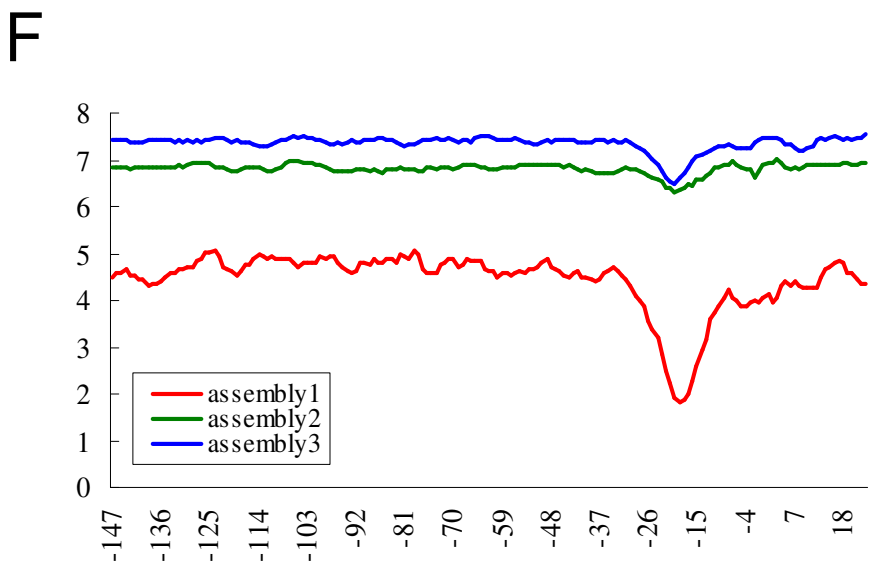


Figure 5. Contd

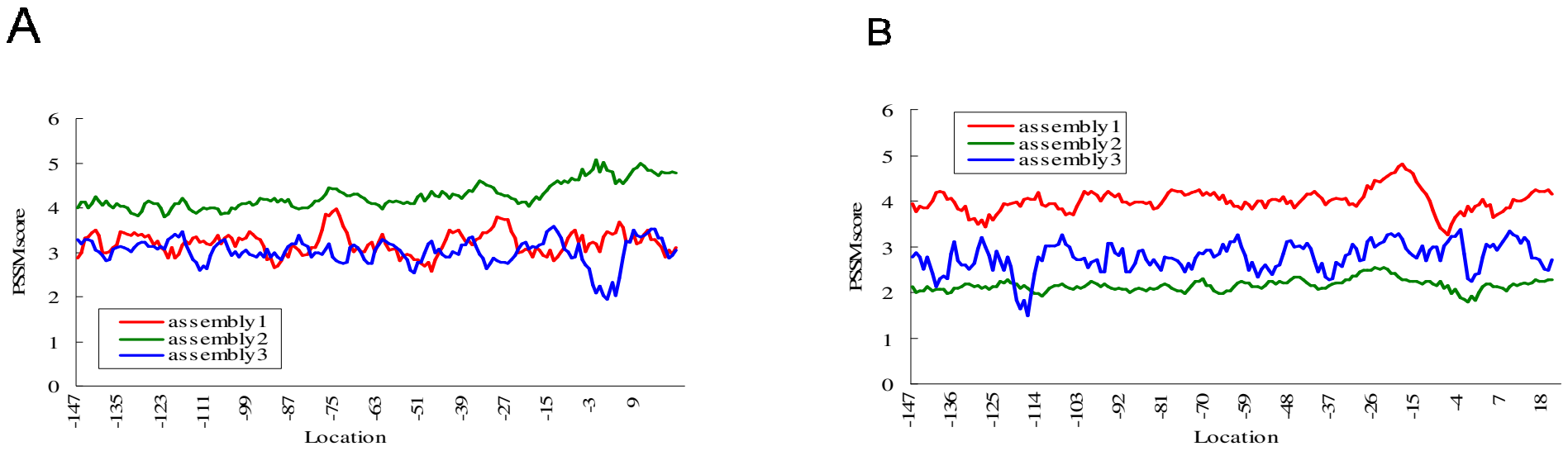


Figure 6. Poly(A) signal patterns of poly(A) sites in CDS and introns. (A) PSSM scores of the top three assemblies in the NUE of CDS poly(A) sites. (B) Same as (A) but for intron poly(A) sites; (C) Number of occurrences of individual pattern at every location in CDS poly(A) sites; (D) Same as (A) but for intron poly(A) sites; (E) Sequence LOGO for assemblies in CDS poly(A) sites; (F) same as (E) but for intron poly(A) sites.

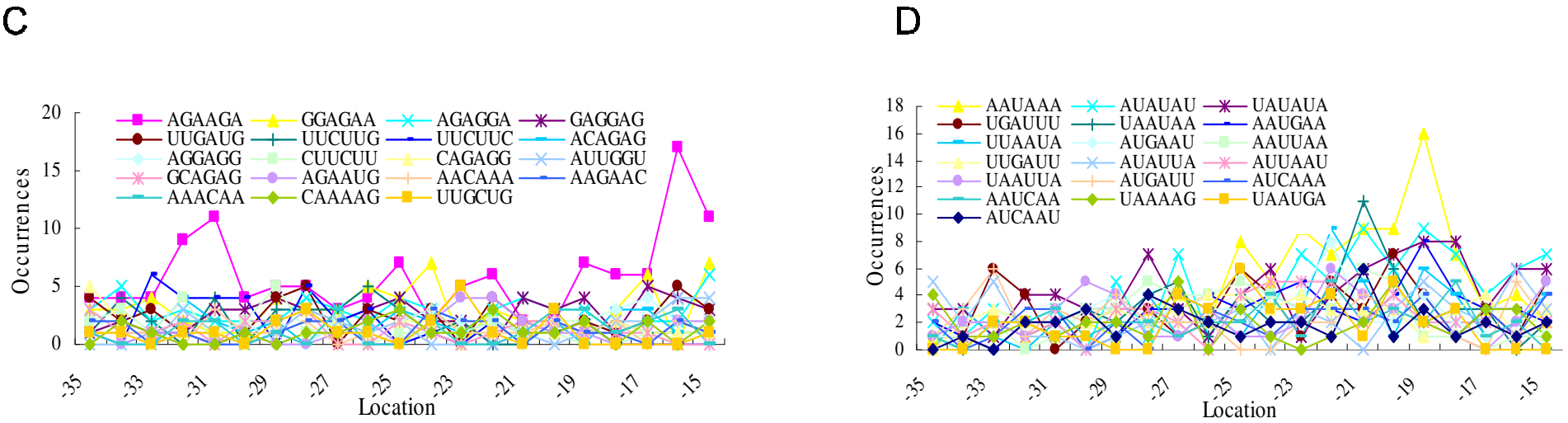


Figure 6. Contd

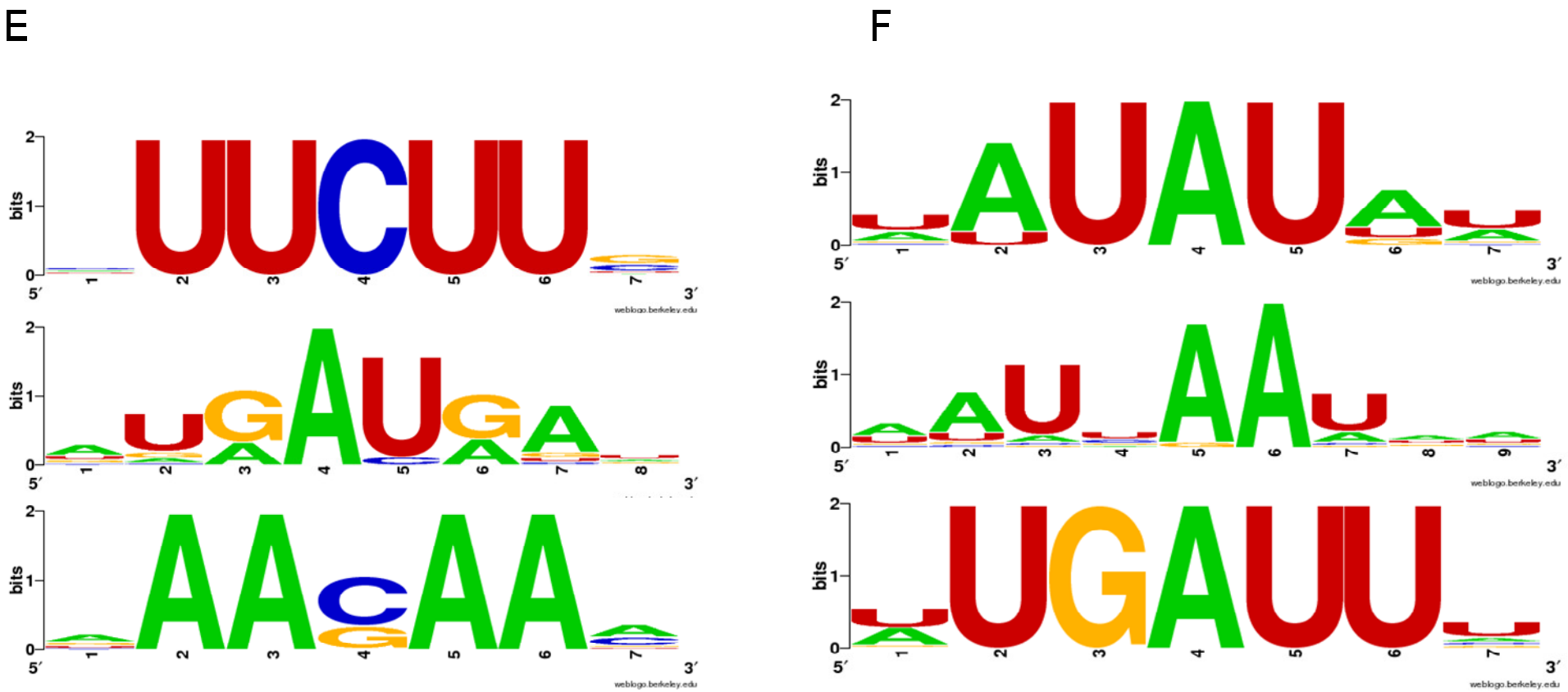


Figure 6. Contd

of the four bases A, T, C, and G.

$V_k\ (k>1)$ is a two-dimensional vector holding the transition probability of each base from one position to the next position. The frequency of each base at the first position of the aligned patterns was used to calculate V_1:

$$V_1[i] = N_i / \sum_i N_i \ \ (i = A,T,C,G)$$

$V_k(k>1)$ was calculated based on the frequency of the di-nucleotide at position *k*-1 and *k*:

$$V_k[i,j] = N_{i,j} / \sum_i N_{i,j} \ \ (i = A,T,C,G;\ j = A,T,C,G)$$

Given a sub-sequence with length *k* $S = s_1,...,s_k$ and the vectors $V_1,...,V_k$, the probability of *S* presented in the vectors is:

$$p = V_1[s_1] \cdot V_2[s_2,s_3] \cdot ... \cdot V_k[s_k,s_{k-1}]$$

Then, the poly(A) site recognition result after improving the parameters was compared with the result without using any pattern in the model parameters. Using PASS_Rice, each location of the input nucleotide sequence will be assigned a score representing its possibility being a poly(A) site. We then examined the distribution of the average scores with or without using the improved

Table S4. The number of occurrences of each pattern of each assembly in the NUE for APA sites in CDS and introns.

NUE of CDS poly(A) sites				NUE of intron poly(A) sites			
No.	Pat.	Pat#	Pat (%)	No.	Pat.	Pat#	Pat (%)
1	UUCUUG	39	2	1	AUAUAU	92	5
1	UUCUUC	36	2	1	UAUAUA	91	5
1	CUUCUU	30	2	1	AUAUUA	51	3
2	UGAUGA	69	4	1	AUUAUG	36	2
2	AUGAUG	51	3	2	AAUAAA	97	6
2	GAUGAU	47	2	2	UAAUAA	60	3
2	UAAUGA	30	2	2	AAUGAA	59	3
2	AUGACA	28	1	2	UUAAUA	58	3
2	UAAUAA	28	1	2	AUGAAU	57	3
2	AUAAUA	26	1	2	AAUUAA	54	3
2	GAUAAU	23	1	2	AUUAAU	51	3
2	AAUAAU	23	1	2	UAAUUA	50	3
2	GGAUGA	22	1	2	AUCAAA	45	3
3	AACAAA	25	1	2	AAUCAA	43	2
3	AAGAAC	24	1	2	UAAUGA	40	2
3	AAACAA	24	1	2	AUCAAU	40	2
				2	UUAAUG	39	2
				2	UAUCAA	38	2
				2	UGAUGA	37	2
				2	UCAAUA	36	2
				3	UGAUUU	63	4
				3	UUGAUU	54	3
				3	AUGAUU	47	3

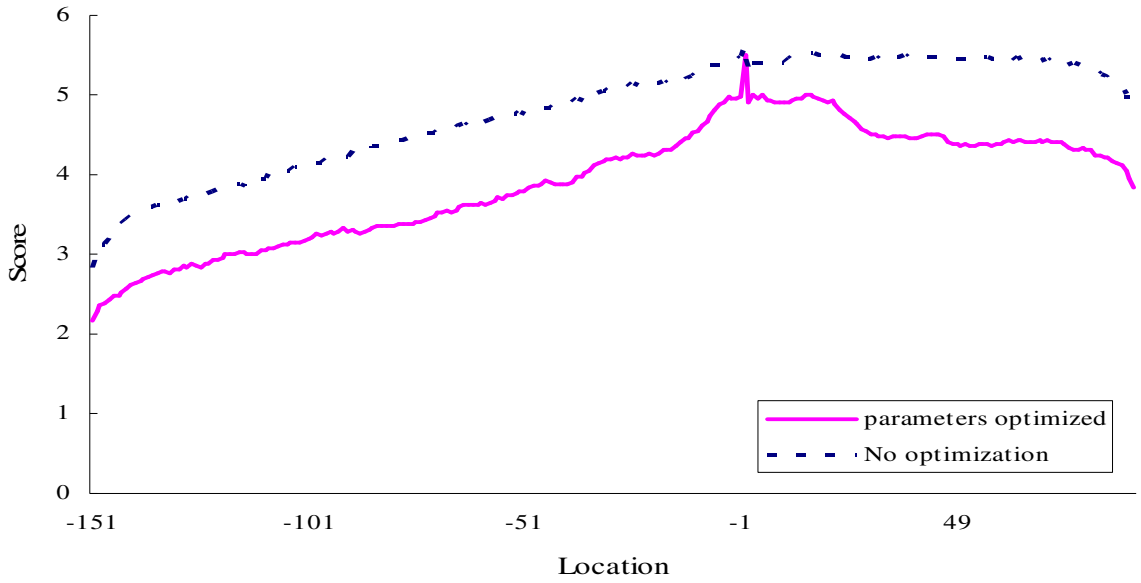


Figure 7. Average score using improved parameters or not.

parameters. As shown in Figure 7, the score curve not using selected patterns is more even and there is no particular prominent peak. In contrast, in the score curve with optimized parameters, the score of the poly(A) site (location -1) is significantly higher than the scores of other positions. This result demonstrated the efficacy of

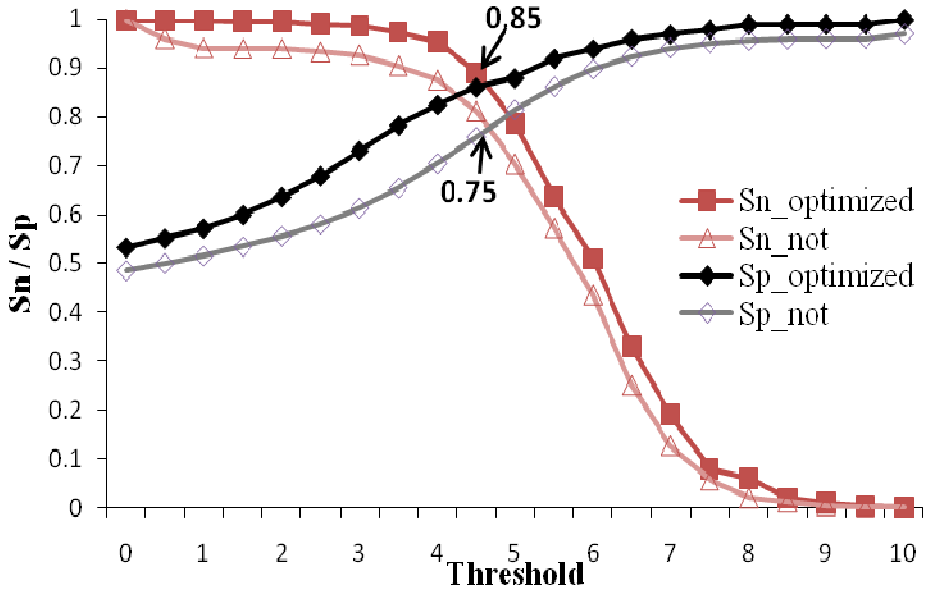


Figure S2. The Sn and Sp using improved parameters (Sn_optimized andSp_optimized) and Sn and Sp without improving parameters (Sn_not and Sp_not). The arrows mark the crossing value of Sn and Sp.

of our pattern identification model in that the selected patterns could make the poly(A) site more presentable.

Here, we also explored the sensitivity (Sn) and specificity (Sp) to evaluate the prediction result. The positive sequences with poly(A) sites were used to calculate Sn. Since recent study have shown that there were quite a number of novel poly(A) sites in CDS and introns (Wu et al., 2011), here the 5'-UTR sequences without any poly(A) site from previous study (Shen et al., 2008a) were used as negative dataset to calculate Sp. The PASS_Rice was adopted to test the positive and negative sequences, with model parameters improved by the selected patterns or not. As shown in Figure S2, the Sn and Sp after improving model parameters (Sn_optimized and Sp_optimized) were significantly higher than the Sn and Sp without improving parameters (Sn_not and Sp_not). The improvement is statistically significant at 90% confidence level (the p-value of Wilcoxon rank sum test is 2.9e-05 for Sn and 0.07 for Sp). The crossing value of Sn and Sp was considered as an overall merit of the prediction result (Ji et al., 2007b; Shen et al., 2008a). The crossing value of Sn_optimized and Sp_optimized (0.85) was -10% higher than Sn_not and Sp_not (0.75), demonstrating that the selected patterns could enhance the poly(A) site prediction result greatly.

It is widely known that AAUAAA is the most important pattern in the NUE in plants and has been verified by biological experiments. Although not all the patterns selected by our model were verified by biological experiments, these patterns also played an important role in the poly(A) site recognition from the perspective of biological computing. Figure 8 shows the output scores using or not using the improved parameters and the selected patterns for the NUE and FUE of a rice sequence. As shown in Figure 8B, there is no AAUAAA in the rice sequence, but another hexamer TAATAA (position is 274) exists in the NUE and TAATGT appears in the FUE (position is 185). The scores calculated using the selected patterns are apparently higher in the region around poly(A) site and dramatically decrease in other regions. While the score curve without using selected patterns is very flat and the potential region of the poly(A) site can be hardly determined.

DISCUSSION

Due to the absence of highly conserved signals around the poly(A) site, computational reorganization of plant poly(A) signals is still a challenging problem. The model established here is applicable to the poly(A) signal pattern recognition for a specific region for various species in plants. Studies (Loke et al., 2005; Shen et al., 2008a, b) have shown several signal regions around poly(A) site, here the background region and other signal regions can be considered as control regions to identify the patterns unique to a specific signal region.

One of the great challenge in bioinformatics is to visualize the poly(A) signals in a user-friendly manner to general biologists. Here, the poly(A) signal patterns could be displayed in a variety of ways including the PSSM scores, the location of the patterns and the sequence

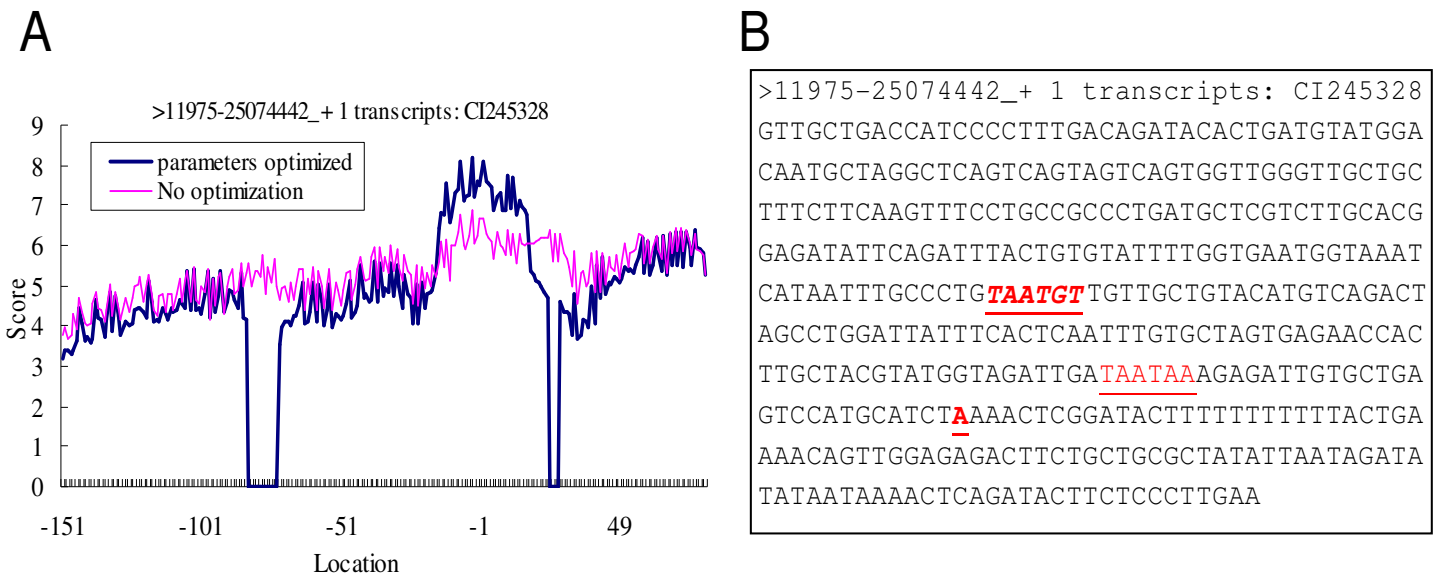


Figure 8. Poly(A) site recognition of a rice sequence. (A) Score curves; (B) the nucleotide sequence and the patterns in the FUE, NUE and CS. 'TAATGT' is the FUE pattern, 'TAATAA' is the NUE pattern, underlined 'A' is the poly(A) site.

LOGO, making the results easy to be understood. Through analyses of three different species including rice, *Arabidopsis* and Chlamy, useful patterns for poly(A) sites were selected and visualized and the poly(A) signals among different groups of poly(A) sites and species were compared. In particular, the poly(A) signals of the newly discovered APA sites in CDS and introns of *Arabidopsis* were explored, indicating a completely different set of poly(A) signals used in CDS poly(A) sites. The recently discovered phenomenon of antisense polyadenylation regulation of the sense gene transcript in plants (Liu et al., 2010; Wu et al., 2011) offers some clues on previous unknown gene regulation mechanisms. The accurate reorganization of the poly(A) signals of such antisense poly(A) sites will undoubtedly promote such research. We are working hard along these lines.

The emerging poly(A) site prediction is focused on discovering new patterns before predicting the poly(A) site (Akhtar et al., 2010; Tzanis et al., 2011). Thus, the approach proposed here may contribute to the problem of poly(A) site prediction. We used the selected patterns to optimize the parameters of existing poly(A) site prediction program PASS_Rice to predict poly(A) sites in rice, the effectiveness of our pattern recognition model was demonstrated by the 10% higher Sn and Sp. However, the model used in PASS_Rice was hardly altered, and only a part of the parameters of the model was modified, thus the potential performance might not be fully expounded. This study aims to find potential poly(A) signal patterns in plants, and attempts to apply these patterns on poly(A) site prediction. Efforts are also underway to develop or utilize some appropriate poly(A) site prediction model which can be seamlessly integrated with our pattern recognition model.

ACKNOWLEDGEMENTS

This project was funded by the National Natural Science Foundation of China (No. 61174161), the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20090121110022), the Fundamental Research Funds for the Central Universities of Xiamen University (Nos. 2011121047, 201112G018 and CXB2011035), the Key Research Project of Fujian Province of China (No. 2009H0044) and Xiamen University National 211 3rd Period Project of China (No. 0630-E72000).

REFERENCES

Akhtar MN, Bukhari SA, Fazal Z, Qamar R, Shahmuradov I (2010). POLYAR, a new computer program for prediction of poly (A) sites in human sequences. BMC Genomics, 11(1): 646.

Bailey TL, Gribskov M (1998). Combining evidence using p-values: application to sequence homology searches. Bioinformatics, 14(1): 48-54.

Bartel DP (2009). MicroRNAs: Target Recognition and Regulatory Functions. Cell, 136(2): 215-233.

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D (2000). Patterns of variant polyadenylation signal usage in human genes. Genome Res. 10(7): 1001-1010.

Buratowski S (2005). Connections between mRNA 3' end processing and transcription termination. Curr. Opin. Cell Biol. 17: 257-261.

Cai B, Peng RH, Xiong AS, Zhou J, Liu JG, Xu F, Zhang Z, Yao QH (2008). Identification of polyadenylation signals and alternative polyadenylation in Vitis vinifera based on ESTs data. Sci. Hortic-Amsterdam, 115(3): 292-300.

Cheng Y, Miura RM, Tian B (2006). Prediction of mRNA polyadenylation sites by support vector machine. Bioinformatics, 22: 2320-2325.

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004). WebLogo: A sequence logo generator. Genome Res. 14(6): 1188-1190.

Hammell CM GS, Zenklusen D, Heath CV, Stutz F, Moore C, Cole CN (2002). Coupling of termination, 3' processing, and mRNA export.

Mol. Cell. Biol. 22: 6441–6457.
Hertz GZ, Stormo GD (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics, 15(7-8): 563-577.
Hertzberg L, Zuk O, Getz G, Domany E (2005). Finding motifs in promoter regions. J. Comput. Biol. 12(3): 314-330.
Holec S LH, Kuhn K, Alioua M, Borner T, Gagliardi D (2006). Relaxed transcription in Arabidopsis mitochondria is counterbalanced by RNA stability control mediated by polyadenylation and polynucleotide phosphorylase. Mol. Cell. Biol. 26: 2869–2876.
Hu J, Lutz CS, Wilusz J, Tian B (2005). Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. RNA, 11: 1485-1493.
Jan CH, Friedman RC, Ruby JG, Bartel DP (2011). Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. Nature, 469(7328): 97-101.
Ji G, Wu X, Huang J, Li QQ (2010a). A Classification-Based Prediction Model of messenger RNA Polyadenylation Sites. J. Theor. Biol. 265(3): 287-296.
Ji G, Wu X, Huang J, Li QQ (2010b). Implementation of a Classification-Based Prediction Model for Plant mRNA Poly(A)Sites. J. Comput. Theor. Nanosci. 7(5): 927-932.
Ji G, Wu X, Li Q, Zheng J (2010c). Messenger RNA Polyadenylation Site Recognition in Green Alga Chlamydomonas Reinhardtii. Lecture Notes in Comput. Sci. 6063: 17-26.
Ji G, Wu X, Zheng J, Shen Y, Li QQ (2007a). Modeling Plant mRNA Poly(A) sites: Software Design and Implementation. J. Comput. Theor. Nanosci. 4: 1365-1368.
Ji G, Zheng J, Shen Y, Wu X, Jiang R, Lin Y, Loke JC, Davis KM, Reese GJ, Li QQ (2007b). Predictive modeling of plant messenger RNA polyadenylation sites. BMC Bioinform. 8: 43.
Kan ZY, Rouchka EC, Gish WR, States DJ (2001). Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res. 11(5): 889-900.
Legendre M, Gautheret D (2003). Sequence determinants in human polyadenylation site selection. BMC Genomics, 4(1): 7.
Li QQ, Hunt AG (1997). The Polyadenylation of RNA in Plants. Plant Physiol. 115: 321-325.
Liu FQ, Marquardt S, Lister C, Swiezewski S, Dean C (2010). Targeted 3 ' processing of antisense transcripts triggers Arabidopsis FLC chromatin silencing. Science, 327(5961): 94-97.
Liu H, Han H, Li J, Wong L (2003). An in-silico method for prediction of polyadenylation signals in human sequences. Genome Inform. 14: 84-93.
Loke JC, Stahlberg EA, Strenski DG, Haas BJ, Wood PC, QQ L (2005). Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures. Plant Physiol. (138): 1457-1468.
Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V, Attie O, Chen K, Salehi-Ashtiani K, Vidal M, Harkins TT, Bouffard P, Suzuki Y, Sugano S, Kohara Y, Rajewsky N, Piano F, Gunsalus KC, Kim JK (2010). The landscape of C. elegans 3'UTRs. Science, 329(5990): 432-435.
Mayfield SP MA, Chen S, Wu J, Tran M, Siefker D, Muto M, Marin-Navarro J. (2007). Chlamydomonas reinhardtii chloroplasts as protein factories. Curr. Opin. Biotechnol. 18(2): 126-133.
Moor CH, Meijer H, Lissenden S (2005). Mechanisms of translational control by the 3' UTR in development and differentiation. Semin. Cell Dev. Biol. 16(1): 49-58.
Navarro G (2001). A guided tour to approximate string matching. ACM Comput. Surv. 33(1): 31-88.
Nuel G (2008). Pattern Markov chains: Optimal Markov chain embedding through deterministic finite automata. J. Appl. Probab. 45(1): 226-243.
Ribeca P, Raineri E (2008). Faster exact Markovian probability functions for motif occurrences: a DFA-only approach. Bioinformatics, 24(24): 2839-2848.
Robin S, Schbath S, Vandewalle V (2007). Statistical tests to compare motif count exceptionalities. BMC Bioinform. 8(1): 84.
Rothnie HM (1996). Plant mRNA 3'-end formation. Plant Mol. Biol. 32: 43-61.
Rothnie HM, Chen G, Futterer J, Hohn T (2001). Polyadenylation in rice tungro bacilliform virus: cis-acting signals and regulation. J. Virol. 75: 4184-4194.
Rupprecht J (2009). From systems biology to fuel--Chlamydomonas reinhardtii as a model for a systems biology approach to improve biohydrogen production. J. Biotechnol. 142(1): 10-20.
Shen Y, Ji G, Haas BJ, Wu X, Zheng J, Reese GJ, Li QQ (2008a). Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. Nucleic Acids Res. 36(9): 3150-3161.
Shen Y, Liu Y, Liu L, Liang C, Li QQ (2008b). Unique features of nuclear mRNA poly(A) signals and alternative polyadenylation in Chlamydomonas reinhardtii. Genetics, 179: 167-176.
Tian B, Hu J, Zhang HB, Lutz CS (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. 33(1): 201-212.
Tzanis G, Kavakiotis I, Vlahavas I (2011). PolyA-iEP: A data mining method for the effective prediction of polyadenylation sites. Expert Syst. Appl. 38(10): 12398-12408.
Wickens M, Bernstein DS, Kimble J, Parker R (2002). A PUF family portrait: 3 ' UTR regulation as a way of life. Trends Genet. 18(3): 150-157.
Wilson NF IJ, Buchheim JA, Meek W (2008). Regulation of flagellar length in Chlamydomonas. Semin Cell Dev Biol. 19(6): 494-501.
Wu X, Liu M, Downie B, Liang C, Ji G, Li QQ, Hunt AG (2011). Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. Proc. Natl. Acad. Sci. USA. 108(30): 12533-12538.
Zhang J, Jiang B, Li M, Tromp J, Zhang XG, Zhang MQ (2007). Computing exact P-values for DNA motifs. Bioinformatics, 23(5): 531-537.