A Prediction of mRNA Polyadenylation Sites in Human Genes

Jorng-Tzong Horng[1,3*], Li-Ching Wu[2], Shun-Kai Liu[1], Cheng-Wei Chang[4], Tsung-Ming Chao[5], Rong-Hwei Yeh[6], and Kuang-Fu Cheng[7]

***Abstract*—mRNA polyadenylation is an essential mechanism in human genes and is direct linked to the termination of transcription. Alternative polyadenylation changes the length of the mature mRNA's 3'UTR. Since 3'UTRs have been shown to contain regulatory elements that control mRNA functioning, alternative polyadenylation plays an important role in controlling the expression of human genes. Prediction of polyadenylation sites can help with the identification of genes and aid our understanding of the mechanisms of alternative polyadenylation. In this study, we constructed a system for mRNA polyadenylation site prediction in human genes using SVM and based on an analysis of the sequence alignment between pair-end diTags (PET) and genome sequences. The PET sequences were mapped to the reference genome more accurate compared to earlier methods. We also analyzed single-site type and multiple-site type sequences PET sequence datasets and found that the frequencies of each nucleotide were different when the single-site type and multiple-site type PET sequences were compared.**

I. Introduction

Polyadenylation of mammalian pre-mRNAs is a cellular process that occurs after transcription termination [1]. The process of polyadenylation consists of two stages. The first stage is recognition of the polyadenylation site (poly(A) site) on pre-mRNA and cleavage at a poly(A) site. The second stage is the addition of a poly(A) tail to the newly formed 3' end. The poly(A) tail consists of up to 250 adenosines.

The cleavage reaction requires a large set of proteins complexes, including the cleavage and polyadenylation specificity factor (CPSF), the cleavage stimulation factor (CstF), cleavage factors Ⅰ and Ⅱ (CF Ⅰ and CF Ⅱ), and the poly(A) polymerase (PAP) [2]. There are two major binding elements for these proteins and these are shown in Figure 1. CPSF binds to the highly conserved AAUAAA hexamer sequence or a close variant that is located between ten and forty nucleotides upstream of the poly(A) site. CstF binds to GU-rich or U-rich sequences that are located between ten and forty nucleotides downstream of poly(A) site [3]. The AAUAAA (or a close variant) hexamer sequence and the downstream GU/U-rich element have been previously recognized as the basic features needed for polyadenylation.

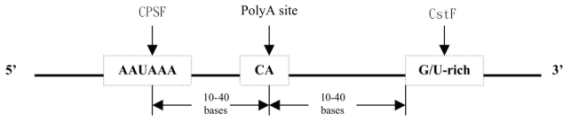


Figure 1. Schematic representation of the features present in human mRNA 3'end poly(A) sites [18]

A. Alternative polyadenylation

Many protein-coding genes have more than one polyadenylation site [4], which can leads to alternative polyadenylation. Polyadenylation at different positions along pre-mRNAs changes the length of the mature mRNA's 3'UTR. Since 3'UTRs have been shown to contain regulatory elements that control mRNA stability [5, 6], translational efficiency [7], and intracellular localization [8], alternative polyadenylation has been recognized as an important mechanism that affects the expression of human genes.

For a gene with a unique poly(A) site, the poly(A) site would always be found in the exon at the extreme 3' end of the pre-mRNA according to existing knowledge. When a gene has multiple poly(A) sites, the poly(A) sites may be located in different places within the gene, including various exons at the extreme 3' end of the pre-mRNA or other regions including internal exons or introns [4]. Alternative polyadenylation results in a range of different mRNAs that have a variable length 3'UTRs and distinct protein products.

B. Paired-end diTags (PET) Sequencing

In this study, we defined the features of polyadenylation sites based on an analysis of sequences created by aligning Paired-end diTags and the human genome sequence. Paired-end diTags (PET) are a sequencing strategy by which short and paired tags are extracted from the ends of long DNA fragments for ultra-high-throughput sequencing [9]. Figure 2 shows the process for the creation of Paired-end diTag. The PET sequences can be accurately mapped to the reference genome, thus demarcating the genomic boundaries of

[1]Department of Computer Science and Information Engineering, National Central University, Taiwan; [2]Graduate Institute of System Biology and Bioinformatics, National Central University, Taiwan; [3]Department of Bioinformatics, Asia University, Taiwan; [4]Department of Information Management, Hsing Wu College, Taiwan; [5]Department of International Business, Ching Yun University, Taiwan; [6]Department of Photonics and Communication Engineering, Asia University, Taiwan; [7]Biostatistics Center and Department of Public Health, and Graduate Institute of Statistics, China Medical University

*Corresponding Author

978-0-7695-4391-8/11 $26.00 © 2011 IEEE
DOI 10.1109/BIBE.2011.64

PET-represented DNA fragments and revealing the identities of the target DNA elements. PET protocols have been developed for the analysis of transcriptomes, transcription factor binding sites, epigenetic sites such as histone modification sites, and genome structures. The specific advantage of the PET technology is its ability to uncover linkages between the two ends of DNA fragments. Using this unique feature, unconventional fusion transcripts, genome structural variations, and even molecular interactions between distant genomic elements can be unraveled by PET analysis. In this study, we used human genomic sequences mapping from PET data as the positive dataset.

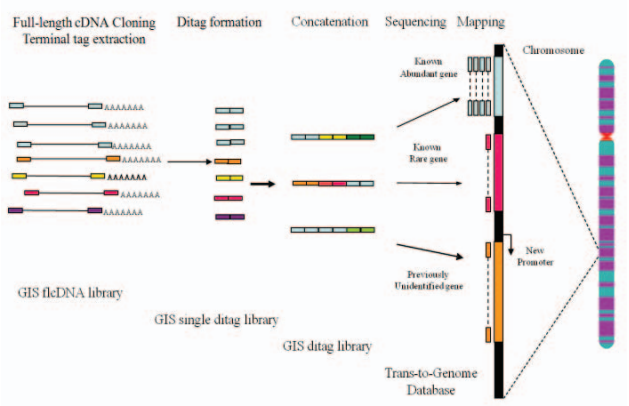


Figure 2. Process of Paired-end diTag sequencing and mapping [9]. The 3'end tag is located exactly before the poly(A) tail. Accurate mapping of the tag can help to identify poly(A) sites.

Alternative polyadenylation may lead to the development of human disease or produce proteins with new functionalities [10-14]; it thus plays an important role in human genes. However, current knowledge of alternative polyadenylation mechanisms is not really abundant, and current prediction tools have achieved only moderate sensitivity and specificity. An accurate identification of potential poly(A) sites would contribute significantly to studies of alternative polyadenylation. In this study, we wanted to develop a machine-learning methodology that would help pinpoint real polyadenylation signals and allow accurate poly(A) site prediction. Rather than using cDNA/ESTs alignment as has been done previously, we used next generation DNA sequencing information in the form of paired-end ditags (PET) and aligned these with the human genome sequence. It was expected that this would improve the identification of human gene polyadenylation signals and poly(A) sites and also increase accuracy.

A computational model was developed that used support vector machines (SVMs) for human poly(A) site prediction. The development of our model began with an analysis of the nucleotide composition of known poly(A) site sequences. These sequences were then aligning with the human genome sequence and paired-end ditags (PET). This analysis allowed us to identify various regions with obvious characteristic that could be used as features. The results show that our predictive model is able to achieve a comparable performance to current prediction tools and also allows the discovery of new patterns that are possibly involved in polyadenylation.

II. Related Works

A. *Predict tools for human poly(A) sites*

Traditional bioinformatic techniques aimed at predicting poly(A) sites have previously collected a large number of cDNA sequences and Expressed Sequenced Tags (ESTs) and then aligned the cDNA/ESTs with the genome sequence; this provided a systematic approach to identify poly(A) sites within genomes. A substantial amount of the data generated computationally via cDNA/ESTs alignment is considered valid and, consequently, it has served as an excellent resource for studies related to the polyadenylation machinery. The prediction of poly(A) sites took advantage of the availability of cDNA/ESTs datasets and hence became practical. In early studies, the problem of poly(A) site prediction was transformed into one of the identification of the putative polyadenylation signal, which was thought to primarily define the location of poly(A) sites [15]. Since PASs are highly conserved elements in the upstream region of poly(A) sites, a correctly identified PAS indicates a real poly(A) site is not far away. In view of this, recognition of the PAS is considered an alternative solution to the problem of poly(A) site prediction.

B. *POLYAH*

POLYAH is a web-base program developed to identify 3'-processing sites of human mRNA precursors [16]. The algorithm is based on a linear discriminant function (LDF) trained to discriminate real poly(A) signal regions from the other regions of human genes possessing the AATAAA sequence that are most likely to be non-functional. The parameters used by LDF consisted of various significant contextual characteristics of the sequences surrounding the AATAAA signals. The accuracy of method was estimated using a set of 131 poly(A) regions and 1466 regions of human genes having an AATAAA sequence. When the threshold was set to predict 86% of the poly(A) regions correctly, a specificity of 51% and a correlation coefficient of 0.62 were achieved.

C. *Polyadq*

Polyadq is a web-based program developed to detect polyadenylation signals in human DNA sequences [17]. The program finds poly(A) signals using two discriminant functions: one specific for AATAAA type poly(A) sites and the other for ATTAAA type poly(A) sites. In this study, the downstream sequence of the PAS was characterized by three features, namely the position weight matrix, the hit position frequency and di-nucleotide frequency. For sequences containing AATAAA or ATTAAA, two quadratic discriminant functions (QDFs) were correspondingly employed to score the candidate PASs. Polyadq decides whether a given AATAAA or ATTAAA hexamer is a true PAS by comparing the hexamer's QDF score to a cutoff value. Hexamers scoring above the cutoff are reported as true signals.

D. PASPrediction

In 2003 Liu et al. proposed a machine learning method to predict polyadenylation signals in human RNA sequences by analyzing the features around them [18]. The method consists of three steps: (1) Generating candidate features from the original sequence data using k-gram nucleotide patterns or amino acid patterns, (2) selecting relevant features using an entropy-based algorithm and (3) integrating the selected features by SVMs to build a system that recognizes poly(A) sites.

E. Polya_svm

Polya_svm is a program for human poly(A) site prediction [19]. The program is based on the identification of the cis-regulatory elements involved in human mRNA polyadenylation. In this study, the region flanking poly(A) sites was divided into four sub-regions. Using a hexamer enrichment method, numerous ver-represented hexamers were aligned and clustered into cis elements. As a result, fifteen cis elements were identified in the four sub-regions. Position-specific scoring matrices (PSSMs) of fifteen cis elements were then used to score each corresponding region of the input sequence. These scores were employed by the support vector machine (SVM) as features during the training phase and testing phases. Polya_svm achieved a higher sensitivity and a similar specificity when it was compared with Polyadq.

III. MATERIALS

A. Positive sequences

We first identified poly(A) sites based on PET data mapped to the human genome sequence. The PET data were obtained from UCSC ENCODE Pilot Project. We retrieved 95,792 human genomic sequences surrounding poly(A) sites (-200 to +200 nt) that correspond to 26,946 genes. A sequence is defined as single-site type if its associated poly(A) site is unique to the gene, otherwise it is defined as a multiple-site type [4]. We classified PET sequences into these two types to help with the further analysis of alternative polyadenylation sites. In total, 3,995 sequences were defined as single-site and 91,797 sequences were defined multiple-site; the results are presented in Table 1. We used these sequences as the positive dataset in order to train and test our prediction model.

Table 1. Dataset

Dataset	#Sequence	#Genes
Positive		
Single-type	4,512	4,512
Multiple-type	101,362	15,467
Total	105,874	19,979
Negative		
coding sequences	37,460	-
5'-UTRs sequence	19,140	-
Total	56,600	

#Sequences, the number of sequences. #Genes, the number of genes that poly(A) sites are related to.

B. Negative sequences

A true negative sequence at a 3'UTR is difficult to obtain, as there is no extensive experimental evidence defining a negative sequence. We used two types of sequences that are presumed to have very few poly(A) sites, namely human mRNA coding sequences (CDS) and human 5'-untranslated region (5'-UTR) sequences. These were used as the negative dataset to train and test our prediction model and these are presented in Table 2. It can be presumed that these types of sequence have no or very few poly(A) sites. The human mRNA coding sequences (CDS) were obtained from NCBI Build 36. The human 5'-untranslated regions (5'-UTRs) sequences were obtained from UTRdb release 22.

Table 2. Feature definition

Feature		position
Upstream		
1	A-rich	-22~-11
2	A-rich	-5~-2
3	U-rich	-9~0
4	Non-G-rich	-40~-0
Downstream		
5	Minimum value of A	0
6	A-rich	1~7
7	U-rich	10~99
8	G-rich	2~36
9	Peak value of C	0
10	Minimum value of C	1

Features 1,2,3,4 are defined by upstream of poly(A) site; Feature 5,6,7,8,9,10 are defined by downstream of poly(A) site. Position, position relative to poly(A) site.

IV. METHODS

Our system is divided into four phases, the pre-processing phase, the feature selection phase, the training phase and the testing phase. The pre-processing phase includes data source collection and sequence dataset generation. The feature selection phase includes nucleotide composition analysis and feature definition. The training phase includes the choice of training set, feature extraction from the training set, and SVM model construction. The testing phase includes the choice of testing set, feature extraction from the testing set, classification by the SVM, and performance evaluation. Figure 3 shows the overview of our system flow.

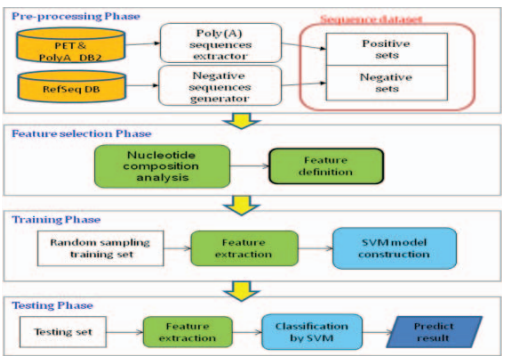
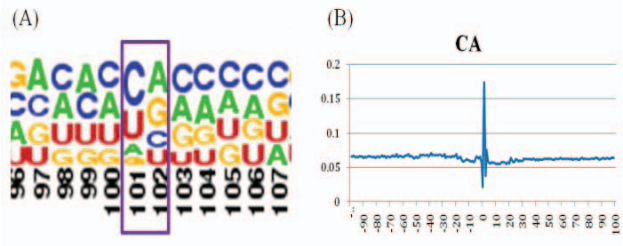


Figure 3. Overview of the system flow

A. *Feature definition*

As the first step, we built the sequence dataset by extracting the positive and negative sequences from corresponding databases. We assumed the middle region (-25 to 25 nt) of a positive sequence was the potentially poly(A) site region. Next we analyzed the nucleotide composition of sequences in datasets. We discovered that some of these regions correspond to features previously identified as surrounding the poly(A) site by previous studies. These include the following: the poly(A) site has high occurrence ratio for 5'-CA-3' [20] (Figure 4) and the poly(A) site is followed 10-40 nt later by a U-rich or GU-rich region (Figure 5). The latter GU-rich region is the binding site for cleavage stimulation factor (CstF). It is assumed that sequences around these two regions can be also using as features for poly(A) site prediction. These regions were then explored as potential machine learning features, as shown in Table 2 and Figure 6.



(C)

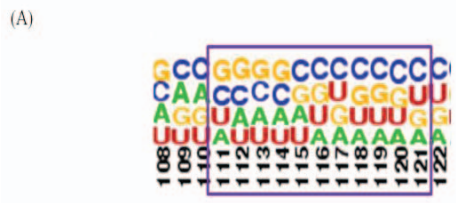
C at position 0

	Position 0	Other position
Positive set	43.76%	26.77%
Negative set	28.14%	27.27%

A at position 1

	Position 1	Other position
Positive set	38.3%	23.87%
Negative set	22.48%	23.9%

Figure 4. 5'-CA-3' in the middle of poly(A) sequence. (A) SequenceLogo of the middle region of the poly(A) sequences created by RNALogo [24]. (B) CA dinucleotide composition (C) Frequency of nucleotide C at position 0 and A at position 1. We found that 5'-CA-3' is the most conservative pattern around the poly(A) site.



(B)

GU at Downstream 10-40nt

	Downstream 10-40nt	Other position
Positive set	5.00%	4.34%
Negative set	4.46%	4.44%

U at Downstream 10-40nt

	Downstream 10-40nt	Other position
Positive set	24.99%	22.43%
Negative set	21.08%	20.93%

Figure 5. Downstream GU/U-rich region. (A) SequenceLogo of the middle region of poly(A) sequences. (B) Frequency of downstream GU and U. We can found that GU and U are more conservative than other patterns.

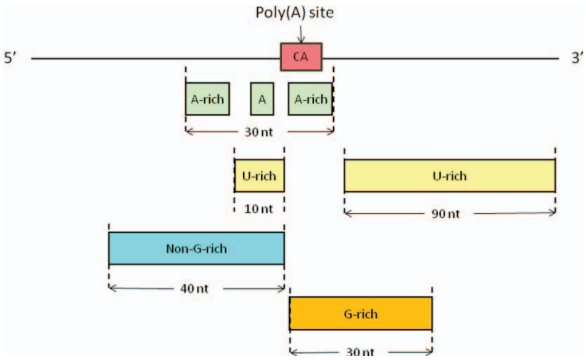


Figure 6. Regional features defined by the analysis.

B. *Support Vector Machine*

In this study, we used the support vector machine (SVM) to predict the presence or not of a poly(A) site in a genomic sequence. Support vector machine (SVM) is a supervised learning method that has been used widely to solve classification problems. SVM has shown excellent empirical performance in prediction tasks. The classification ideas of SVM involve mapping the input vectors into feature space, and then construct a hyperplane that separates different groups of input vectors with a maximum margin. We used LIBSVM for classification and then applied the C-support vector classification (C-SVC) method and the radial basis kernel function (RBF) with default setting.

C. *Training Phase*

We used the features described above to build a SVM training model. Since the feature regions were concentrated from -100 to +100 nucleotides relative to the poly(A) sites, we extracted the sequences around poly(A) sites that were 200 nucleotides in length. The training set consisted of 4,000 sequences, with 2,000 positive sequences randomly selected from the positive dataset and 2,000 negative sequences randomly selected from the negative dataset. To avoid bias due to a selected training set, we randomly generated thirty different training sets. The prediction performance result is the mean value obtained for all 30 training models. After the generation of the training set, we extracted the region features and PAS hexamers described above and construct the training model by LIBSVM.

D. *Testing Phase and Performance evaluation*

In this study, we tested our prediction model by using it with all positive and negative sequences. When testing the positive sequences, if the reported site is in the potentially poly(A) site region (±25 nt from real poly(A) site), the system is considered to have identified a true positive, otherwise it has identified a false positive.

We then compared our result with Polya_svm after calculating the predictive accuracy as follows:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad \text{Specificity} = \frac{TN}{TN+FP}$$

where TP is true positive, a poly(A) site correctly identified as poly(A) site; TN is true negative, a none poly(A) site correctly identified as a none poly(A) site; FN is false negative, a poly(A) site incorrectly identified as a none poly(A) site and FP is a false positive, a none poly(A) site incorrectly identified as a poly(A) site.

V. Results

A. Predicted performance compared with previous approaches

We constructed a system of mRNA polyadenylation sites prediction for human genes using SVM based on the system described earlier. To examine the performance of our new approach, we tested our system using all positive sequences (105,874 in total) and all negative sequences (56,600 in total) within our dataset and compared the results with those obtained by polya_svm version 2.2 [19] at the default setting. For our system, if there is a poly(A) site predicted within 25 nt from a real poly(A) site, the prediction was considered a true positive (TP), otherwise it is a false negative (FN). As shown in Table 3, our predictive system is more sensitive than Polya_svm, and has a similar specificity.

Table 3. Comparison of our prediction system with Polya_svm

	Our model			Polya_svm		
Positive set	TP	FN	SN(%)	TP	FN	SN(%)
PET	76,356	29,518	72.12	54,345	51,529	51.33
Negative set	TN	FP	SP(%)	TN	FP	SP(%)
CDS	27,605	9,855	73.95	27,151	10,309	72.48
5'UTR	13,367	5,773	69.84	14,242	4,898	74.41

PET, Paired-End diTags base sequences; CDS, coding region sequences; 5'UTR, 5'UTR sequences. TP, true positives; FN, false positives; TN, true negatives; FP, false positives. SN, sensitivity; SP, specificity.

.

VI. Discussion

In this study, we have described a machine learning methodology for recognition of human polyadenylation sites. We found various potential patterns for the polyadenylation sites based on the analysis of PET sequences rather than cDNA/ESTs sequences, which have been used in the traditional poly(A) site prediction methodology. Some of these regions match the current knowledge base of polyadenylation signals, while others do not have clearly evidence to prove that they contribute to polyadenylation. We used these features to build a predictive system, and this gave improved prediction results when compared to other studies.

We also analyze sequences obtained from PolyA_DB 2, which uses alignments between cDNA/ESTs and genome sequences. It was found that the PET sequences and PolyA_DB 2 sequences gave a similar pattern in terms of nucleotide composition, but the characteristics of the PET sequences are weaker than those of the PolyA_DB 2 sequences. This means it is likely that when we build a predictive model using the PET sequences, we should obtain a higher sensitivity, but a lower specificity. In actuality, our system gave a better performance in terms of sensitivity, and the same level of specificity compare previous studies. This result may be attributable to connections between features, which is supported by our results based on deleting features. Only five features when deleted gave a lower specificity. However, even though deleting these features individually had little effect, when they were removed as a group then the specificity decreased. This is similar to the findings of 2008 Tzanis et al. [23] where they developed a tool that for mining emerging patterns of interest from mRNA sequences in order to predict polyadenylation sites. They focused on finding the association rules between patterns, and suggested that associations among patterns may play an important role in poly(A) site prediction. In the future, we would like to obtain more evidence to support the importance of these patterns to polyadenylation, because if we can pinpoint the biological implications of these patterns, it will help us to understand the mechanism of alternative polyadenylation better.

Acknowledgement

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. 99-2221-E-008-083.

References

[1] Maniatis, T. and R. Reed, An extensive network of coupling among gene expression machines. Nature, 2002. 416(6880): p. 499-506.

[2] Keller, W. and L. Minvielle-Sebastia, A comparison of mammalian and yeast pre-mRNA 3'-end processing. Curr Opin Cell Biol, 1997. 9(3): p. 329-36.

[3] Zarudnaya, M.I., et al., Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. Nucleic Acids Res, 2003. 31(5): p. 1375-86.

[4] Tian, B., et al., A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res, 2005. 33(1): p. 201-12.

[5] Dreyfus, M. and P. Regnier, The poly(A) tail of mRNAs: bodyguard in eukaryotes, scavenger in bacteria. Cell, 2002. 111(5): p. 611-3.

[6] Touriol, C., et al., Expression of human fibroblast growth factor 2 mRNA is post-transcriptionally controlled by a

unique destabilizing element present in the 3'-untranslated region between alternative polyadenylation sites. J Biol Chem, 1999. 274(30): p. 21402-8.
[7] Knirsch, L. and L.B. Clerch, A region in the 3' UTR of MnSOD RNA enhances translation of a heterologous RNA. Biochem Biophys Res Commun, 2000. 272(1): p. 164-8.
[8] Kislauskis, E.H., X.C. Zhu, and R.H. Singer, Sequences Responsible for Intracellular-Localization of Beta-Actin Messenger-Rna Also Affect Cell Phenotype. Journal of Cell Biology, 1994. 127(2): p. 441-451.
[9] Chiu, K.P., et al., PET-Tool: a software suite for comprehensive processing and managing of Paired-End diTag (PET) sequence data. BMC Bioinformatics, 2006. 7: p. 390.
[10] Bennett, C.L., et al., A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA -> AAUGAA) leads to the IPEX syndrome. Immunogenetics, 2001. 53(6): p. 435-439.
[11] Brown, P.H., L.S. Tiley, and B.R. Cullen, Efficient Polyadenylation within the Human-Immunodeficiency-Virus Type-1 Long Terminal Repeat Requires Flanking U3-Specific Sequences. Journal of Virology, 1991. 65(6): p. 3340-3343.
[12] Carswell, S. and J.C. Alwine, Efficiency of utilization of the simian virus 40 late polyadenylation site: effects of upstream sequences. Mol Cell Biol, 1989. 9(10): p. 4248-58.
[13] Hall-Pogar, T., et al., Alternative polyadenylation of cyclooxygenase-2. Nucleic Acids Res, 2005. 33(8): p. 2565-79.
[14] Valsamakis, A., et al., The Human- Immunodeficiency-Virus Type-1 Polyadenylylation Signal - a 3' Long Terminal Repeat Element Upstream of the Aauaaa Necessary for Efficient Polyadenylylation. Proceedings of the National Academy of Sciences of the United States of America, 1991. 88(6): p. 2108-2112.
[15] Legendre, M. and D. Gautheret, Sequence determinants in human polyadenylation site selection. BMC Genomics, 2003. 4(1): p. 7.
[16] Salamov, A.A. and V.V. Solovyev, Recognition of 3'-processing sites of human mRNA precursors. Comput Appl Biosci, 1997. 13(1): p. 23-8.
[17] Tabaska, J.E. and M.Q. Zhang, Detection of polyadenylation signals in human DNA sequences. Gene, 1999. 231(1-2): p. 77-86.
[18] Liu, H., et al., An in-silico method for prediction of polyadenylation signals in human sequences. Genome Inform, 2003. 14: p. 84-93.
[19] Cheng, Y., R.M. Miura, and B. Tian, Prediction of mRNA polyadenylation sites by support vector machine. Bioinformatics, 2006. 22(19): p. 2320-5.
[20] Chen, F., C.C. Macdonald, and J. Wilusz, Cleavage Site Determinants in the Mammalian Polyadenylation Signal. Nucleic Acids Research, 1995. 23(14): p. 2614-2620.
[21] Beaudoing, E., et al., Patterns of variant polyadenylation signal usage in human genes. Genome Res, 2000. 10(7): p. 1001-10.
[22] Lee, J.Y., et al., PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. Nucleic Acids Res, 2007. 35(Database issue): p. D165-8.
[23] Tzanis, G., I. Kavakiotis, and I. Vlahavas, Polyadenylation Site Prediction Using Interesting Emerging Patterns. IEEE International Conference on BioInformatics and BioEngineering, 2008. 8.
[24] Chang, T.H., J.T. Horng, and H.D. Huang, RNALogo: a new approach to display structural RNA alignment. Nucleic Acids Res, 2008. 36(Web Server issue): p. W91-6.