In Silico Biology 9, 0012 (2009); ©2009, Bioinformation Systems e.V.

Prediction of polyadenylation signals in human DNA sequences using nucleotide frequencies

Firoz Ahmed, Manish Kumar and Gajendra P. S. Raghava*

Bioinformatics Centre, Institute of Microbial Technology, Sector 39-A, Chandigarh, India
Email: firoz@imtech.res.in; manish@imtech.res.in

* Corresponding author

Email: raghava@imtech.res.in
Phone: +91-172-2690557; Fax: +91-172-2690632

Edited by E. Wingender; received July 27, 2008; revised November 24, 2008, and April 13, 2009; accepted April 13, 2009; published May 23, 2009

Abstract

The polyadenylation signal plays a key role in determining the site for addition of a polyadenylated tail to nascent mRNA and its mutation(s) are reported in many diseases. Thus, identifying poly(A) sites is important for understanding the regulation and stability of mRNA. In this study, Support Vector Machine (SVM) models have been developed for predicting poly(A) signals in a DNA sequence using 100 nucleotides, each upstream and downstream of this signal. Here, we introduced a novel split nucleotide frequency technique, and the models thus developed achieved maximum Matthews correlation coefficients (MCC) of 0.58, 0.69, 0.70 and 0.69 using mononucleotide, dinucleotide, trinucleotide, and tetranucleotide frequencies, respectively. Finally, a hybrid model developed using a combination of dinucleotide, 2nd order dinucleotide and tetranucleotide frequencies, achieved a maximum MCC of 0.72. Moreover, for independent datasets this model achieved a precision ranging from 75.8 - 95.7% with a sensitivity of 57%, which is better than any other known methods.

Keywords: polyadenylation signals, mRNA, Support Vector Machine (SVM), Matthews correlation coefficient (MCC), ROC plot, nucleotide frequency

Introduction

Mechanistically, polyadenylation is a tightly coupled two-step process. In the first step, cleavage and polyadenylation specificity factor (CPSF) forms an unstable complex with polyadenylation signal (PAS), whereafter cleavage stimulation factor (CstF) interacts with a GU/U-rich element situated downstream of the poly(A) signal through its 64 kDa subunit to form a loop in the pre-mRNA. Binding of cleavage factor I (CFI) and II (CFII) helps in stabilizing the mRNA-protein complex and, ultimately, promotes cleavage of mRNA with the help of an endonuclease, at a position ~35 nucleotides downstream of the poly(A) signal [1, 2]. The second step involves a rapid polyadenylation of the freshly generated 3'-end by poly(A) polymerase (PAP) which has already bound to the complex even before the endonuclease activity. The resulting poly(A) tail has been implicated in several aspects of RNA metabolism, such as efficiency of mRNA export from the nucleus, its stability and localization, and initiation and efficiency of translation [3, 4]. Previous studies have shown that mutation in the poly(A) signal, a hexamer, not only disrupts the normal site of transcriptional termination but also blocks mRNA polyadenylation [5]. In case of mild thalassemia, due to a mutation in the canonical poly(A) signal (AAUAAA) of the HBB gene renders it non-functional, so that a second AAUAAA pattern (present

downstream of mutated one) is perceived as the poly(A) signal. Thus, a larger mRNA transcript is produced, which is unstable [6]. These studies demonstrate the importance of poly(A) signals for proper termination of transcription followed by proper polyadenylation of the transcript [2, 5, 6]. Identification of poly(A) sites is important to determine the gene boundary like the last exon and 3'-UTR, which plays critical role in mRNA stability and localization [2, 7]. Alternative polyadenylation mechanism results in mRNA diversity with different 3'-UTR, in tissue or developmental stage specific manner, and can greatly affect the pattern of gene expression. Prediction of poly(A) signals and, thus, 3'-cleavage sites can provide useful information for developing methods for annotating genomes [8-11].

Poly(A) signals differ widely among different organisms. The highly conserved canonical poly(A) signal AAUAAA is found in animals whereas it is not conserved in plant and yeast [12-15]. In human, ~93% of all mRNAs have AAUAAA or single nucleotide variants of this motif as poly(A) signals. The most abundant among this group are AAUAAA and AUUAAA, which constitute ~53% and ~17%, respectively [14, 16]. In human, poly(A) signals are located 15-30 nucleotides (nt) upstream and GU/U-rich elements are located 20-40 nt downstream of the cleavage sites (Fig. S1). The cleavage site is characterized by the presence of different *cis*-regulatory elements in its proximity.

In the past, a number of methods have been developed for predicting poly(A) signals in a given nucleotide sequence. Salamov *et al.* developed a method POLYAH based on linear discriminant function of 8 variables from -100 to +100 nt region around poly(A) signals and achieved a MCC of 0.62 [17]. Tabaska and Zhang, 1999, developed a weight-matrix based method Polyadq for predicting poly(A) signals in a DNA sequence using two quadratic discriminant functions [18]. This program used only 100 nt sequence information downstream of a candidate poly(A) signal. In their test of finding poly(A) signals, they achieved MCC of 0.413 and 0.512 on full genes and on the 3'-terminal region of the test genes, respectively. In 2003, Legendre *et al*. used -300 nt to +300 nt genomic sequences around the poly(A) signals [19]. They developed a program called ERPIN that is based upon nucleotide profile and achieved a precision of 69% to 85% corresponding with a sensitivity of 56% on independent datasets [19]. A SVM based method has been developed for the first time using k-gram nucleotide composition of 200 nucleotides around poly(A) signals (100 upstream and 100 downstream); by this approach, a precision from 73% to 93% with a sensitivity of 56% was achieved on independent datasets [20]. Recently, Cheng *et al*. developed a method for the prediction of poly(A) sites in mRNA. First they extracted 15 important *cis*-regulatory elements around the poly(A) sites and then generated position-scoring matrices for these elements. They developed the program polya_svm using SVM and similarity score for 15-*cis* motifs which improved the sensitivity but specificity almost remained the same [21]. The precision was referred to as specificity in above studies [19-21]. However, prediction accuracies of these methods remain low. In this study, an attempt has been made to develop a new method for predicting poly(A) signals in human DNA by exploiting the discriminatory features of 200 nucleotides around poly(A) signals. In the first step, sequences around poly(A) signals were analyzed to understand the preference of particular nucleotide patterns. In the second step, models have been developed using different nucleotide frequencies as feature vectors by the popular machine learning technique SVM_light [22]. Finally, a hybrid model was developed by combination of more than one informative feature. Based on this study, a web-based server PolyApred has been developed which is available at http://www.imtech.res.in/raghava/polyapred/. This classifier can predict poly(A) signals of 13 different variants (AAUAAA, AUUAAA, UAUAAA, AGUAAA, AAGAAA, AAUAUA, AAUACA, CAUAAA, GAUAAA, AAUGAA, UUUAAA, ACUAAA and AAUAGA) that constitute ~93% of the whole poly(A) signals in human as well as putative new poly(A) signal in a sequence based on nucleotide frequencies [14].

Methods

The datasets for training

The training was performed on two different datasets: (a) A positive dataset containing 2327 sequences, each sequence 206 nt long having a poly(A) signal at the centre (101 to 106 nt). The positive dataset consists of 1632 "unique" and 695 "strong" poly(A) sites. Unique poly(A) sites: those sites from UTRs with a single EST-supported poly(A) site, strong poly(A) sites: those sites from UTR with at least 10 ESTs-supported poly(A) sites and more than 70% of ESTs of that gene are associated to this site [19]. (b) A negative dataset containing 2333 sequences, each sequence 206 nt long extracted from coding regions having AATAAA at the centre (101 to 106 nt). In this study the hexamer at the centre of a negative sequence is referred to as pseudo-PAS.

The datasets for independent testing

To evaluate the performance of the model developed on the training dataset, it was tested on the same independent datasets, which were earlier used by Legendre *et al*. and Liu *et al*. for evaluating their methods [19, 20]. These datasets contain (a) a positive dataset of 982 sequences containing annotated poly(A) signals from EMBL, and (b) four negative datasets of nearly equal size. These are (i) CDS sequences from coding regions, (ii) sequences from first introns, (iii) random simple shuffled sequences of 3'-UTRs and (vi) sequences generated with a 1^{st} order Markov model of 3'-UTRs. All these positive and negative testing set sequences are 206 nt long and contain a poly(A) signal in the positive set whereas a pseudo-PAS in the negative sets at the center. All these training and testing datasets were obtained from Legendre *et al*. and Huiqing Liu via personal communication [19, 20].

In order to determine the performance of our method on each variant of PAS signal we created another independent dataset. We randomly extracted 50 sequences for each variant of poly(A) signal from Human ATD release 2 [23]. Thus, this independent dataset consists of 650 sequences with a PAS hexamer in the middle of each sequence 206 nt long. We also randomly extracted a similar number of negative sequences from the EMBL-CDS database, which lack a poly(A)-like hexamer in the middle. Sequences highly similar to any sequence in the training dataset have been removed using BLAST at E-value 5e-6.

Five-fold cross validation

A five-fold cross validation technique was used to evaluate the performance of each module constructed in this study [24]. Here, the dataset was divided randomly into five sets. The classifier was trained on four sets and performance was assessed on the remaining fifth set. This process was repeated five times so that each set could be used once for testing. The predictive performance of classifiers was evaluated by threshold dependent parameters.

Models based on various features

We developed various models for predicting poly(A) signals using different type of nucleotide frequencies and binary pattern. The nucleotide sequence around poly(A) signals is used as input, i. e. 100 nucleotides upstream (UP100), and 100 nucleotides downstream (DW100). In case of split sequence, UP100 and DW100 are divided into parts (Fig. S2).

Binary pattern: In the case of binary pattern each nucleotide was represented by a vector of four dimensions such as A by [1,0,0,0], C by [0,1,0,0], G by [0,0,1,0] and T by [0,0,0,1]. Thus a sequence of 200 nucleotides was represented by a vector of 800 (4 × 200) dimensions, which means UP100 and DW100 were both represented by vectors of 400 (4 × 100) dimensions.

Simple nucleotide frequency: In this case we calculated nucleotide frequencies of 100 upstream (UP100) and 100 downstream (DW100) positions, relative to poly(A) signals, separately and further added them to one another so that the total dimension is double. For instance, the sequence of 100 upstream was represented by a vector of four dimensions using mononucleotide frequency (frequency of A, T, G and C). In the case of dinucleotide frequency (AA, AC, AG, CG, AT ..), the sequence was represented by a 16-dimensional vector. Similarly, the sequence was represented by a vector of 64 dimensions in case of trinucleotides and by a vector of 256 dimensions in the case of tetranucleotides. The same coding scheme was also used for 100 nucleotides downstream sequence, such that in case of mononucleotide frequency, the sequence was represented by a vector of 8 dimensions (4 for UP100 and 4 for DW100) (Tab. S1)

Split nucleotide frequency: Instead of taking whole UP100 or DW100 sequence we splitted it into different parts and nucleotide frequencies were calculated separately for each part. Initially, we splitted UP100 in two parts namely UP51_100 (from 51 to 100 upstream sequences from PAS) and UP1_50 (from 1 to 50 upstream sequences from PAS) and similarly DW100 into DW1_50 (from 1 to 50 downstream sequences from PAS) and DW51_100 (from 51 to 100 downstream sequences from PAS). Therefore, a vector of 16 dimensions (4 for each UP51_100, UP1_50, DW1_50, DW51_100) was used to represent a sequence in case of mononucleotide frequency. Similarly, we also divided UP100 and DW100 into three and four parts (Tab. S1).

2^{nd} order of dinucleotide frequency: In standard dinucleotide frequency, local order is considered where interaction between i^{th} and $(i + 1)^{th}$ nucleotide is taken into account. In this study, we also used second order

of dinucleotide frequency, where we considered interaction of the 1st nucleotide with the 3rd nucleotide in the sequence i. e. i + 2 was considered [25].

Performance measures

In order to assess the performance of models developed in this study, we calculated parameters like sensitivity, specificity, PPV (positive predictive value or precision or probability of correct prediction of poly(A) signal), accuracy (ACC), and Matthews correlation coefficient (MCC). These parameters were calculated using following equations [26].

$$Sensitivity = \frac{TP}{TP + FN} \times 100\%$$

$$Specificity = \frac{TN}{TN + FP} \times 100\%$$

$$PPV = \frac{TP}{TP + FP} \times 100\%$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\%$$

$$MCC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{[TP + FP][TP + FN][TN + FP][TN + FN]}}$$

where TP and FN refer to true positive and false negatives and TN and FP refer to true negatives and false positives, respectively.

Analysis of sequences

To identify the occurrence of region-specific bases associated with poly(A) signals sequence, t-test calculation was carried out using StatSoft version 8.0.

Support vector machine

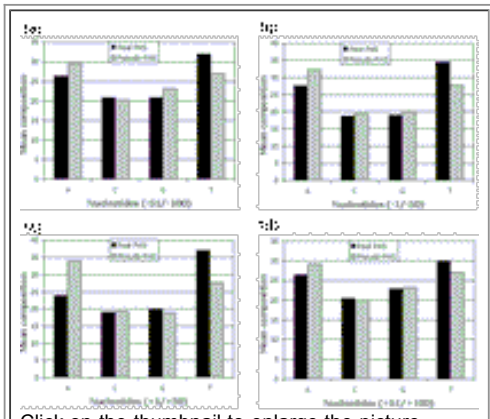
SVM is a kernel-based method used both for classification and regression tasks. SVM has been widely used for the prediction of important biological signals [27 - 30]. In this study, a freely downloadable package, SVM_light has been used [22]. The software enables the user to define a number of parameters as well as to select an in-build kernel function such as radial basis function (RBF) or polynomial kernel. In this study we also use the linear and polynomial kernel, but RBF kernel predict better than others to classify the data of PASes and pseudo-PASes (data not shown).

Results

Analysis of sequences

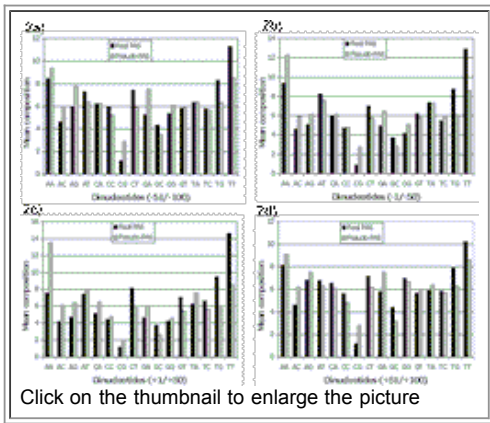
In order to understand the significance of nucleotides in sequence containing poly(A) signals, nucleotide (e. g. mono- and dinucleotide) compositions were computed in two different datasets, namely (a) a positive dataset (2327 sequences) comprising 206 nt long sequences having poly(A) signal at centre, which we refer to as real PAS containing sequences; and (b) a negative dataset (2333 sequences) comprising 206 nt long sequences

from coding region having AATAAA at centre, which we refer to as pseudo-PAS and the whole sequence as pseudo-PAS containing sequences (see "Methods"). In each sequence, 100 nucleotides upstream and downstream relative to signals were divided into two equal parts, for upstream region from -51 to -100 nt (UP51_100) and from -1 to -50 nt (UP1_50), and for downstream region from +1 to +50 nt (DW1_50) and from +51 to +100 nt (DW51_100). We calculated the mononucleotide and dinucleotide compositions of each of the four 50 nt long region excluding the middle hexamer. T-test were also carried to check the statistical significance. As shown in Fig. 1, composition of base T is significantly higher and A is significantly lower in real PAS sequences in comparison to pseudo-PAS sequences (Tab. S2). Nucleotide G is only higher in DW1_50 region and not in other regions. Fig. 2 shows dinucleotide composition of various regions; the real PAS sequences are found to have much higher composition of TT and TG as compared to pseudo-PAS sequences; this observation is more prominent in DW1_50 than in DW51_100. Lack of AA-rich region has been observed downstream of real PAS (DW1_50) as compared to pseudo-PAS. Statistical analysis showed that all the dinucleotides in the region DW1_50 have a high probability to contribute strongly to the identification of PAS because all of them are significantly different from their counterparts in pseudo-PAS (Tab. S3). Our analysis shows that region UP51_100 has very little sequence diversity; hence, it may be the least important among all regions.



Click on the thumbnail to enlarge the picture

Figure 1: Mononucleotide composition of sequence having PAS and pseudo-PAS; for regions (a) upstream from 51 to 100 (UP51_100), (b) upstream from 1 to 50 (UP1_50), (c) downstream from 1 to 50 (DW1_50) and (c) downstream from 51 to 100 (DW51_100).



Click on the thumbnail to enlarge the picture

Figure 2: Dinucleotide composition of sequence having PAS and pseudo-PAS; for regions (a) upstream from 51 to 100 (UP51_100), (b) upstream from 1 to 50 (UP1_50), (c) downstream from 1 to 50 (DW1_50) and (d) downstream from 51 to 100 (DW51_100).

Prediction of poly(A) signals

In order to discriminate between the sequences having poly(A) signals and pseudo-PAS signal, SVM models have been developed using different features. These models were trained and tested on a dataset of 2327 positive and 2333 negative examples, and were evaluated using five-fold cross validation technique. We only considered 200 nt around poly(A) signals excluding the middle hexamer as an input features.

Binary pattern: We first developed a model based on binary pattern feature, which reveals the occurrence of position specific nucleotides. This model contains vector of dimension 800 (one nucleotide is represented by a vector of dimension four) and achieved a maximum MCC of 0.54.

Simple frequency: Further SVM models have been developed using 100 nucleotides upstream (UP100) and 100 nucleotides downstream (DW100) of poly(A) signals (excluding PAS hexamer) on main dataset. In this

case, frequency of downstream hundred nucleotides is added to that of the upstream one, which doubles the total dimension, e. g.: for mononucleotide, 4 (upstream) + 4 (downstream) = 8. As shown in Tab. 1, performance of these models improved from an MCC of 0.51 to 0.68 when content of information increases from mononucleotide to tetranucleotide (Tab. S4 - S7). It is interesting to note that simple SVM model based on dinucleotide frequency performed better than model based on binary pattern. This indicates that position specific nucleotide is not acting as an important feature as the dinucleotide.

Table 1: The performance of frequency based SVM modules using various features of sequence around PASes.

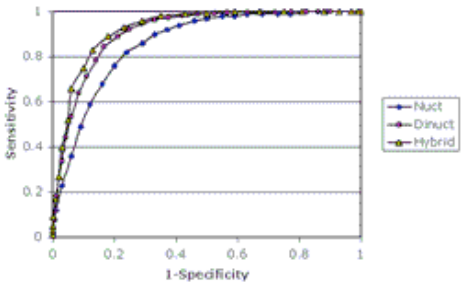
Type	100 nt around PAS				(50+50) nt around PAS				(33+33+34) nt around PAS				(25nt×4) nt around PAS			
	SN	*SP*	*ACC*	*MCC*	*SN*	*SP*	*ACC*	*MCC*	*SN*	*SP*	*ACC*	*MCC*	*SN*	*SP*	*ACC*	*MCC*
Nuct	79.93	70.90	75.41	0.51	81.82	75.65	78.73	0.58	79.63	76.94	78.28	0.57	79.50	78.74	79.12	0.58
Dinuct	84.01	78.18	81.09	0.62	87.02	81.57	84.29	0.69	85.43	81.35	83.39	0.67	84.40	83.24	83.82	0.68
Trinuct	86.42	80.50	83.45	0.67	87.88	81.53	84.70	0.70	85.73	81.78	83.76	0.68	87.80	81.40	84.59	0.69
Tetnuct	85.30	82.30	83.80	0.68	87.32	82.08	84.70	0.69	86.46	81.18	83.82	0.68	85.60	82.17	83.88	0.68

Nuct: Monoucleotide; Dinuct: Dinucleotide; Trinuct: Trinucleotide; Tetnuct: Tetranucleotide; SN: Sensitivity; SP: Specificity; ACC: Accuracy; MCC: Matthews correlation coefficient.

Split nucleotide frequency: As Figs. 1 and 2 reveal that difference in nucleotide frequencies is more significant in regions that are near PAS (UP1_51, DW1_50) than in those far from it (UP51_100, DW51_100). Thus, we developed SVM models using individual frequencies of the two 50 nt long regions (UP100/DW100 divided into two equal parts). In this case, the frequency of each region is added to the frequency of the other region so that the total dimension is quadrupled, e. g.: for mononucleotide, 8 (upstream) + 8 (downstream) = 16. Interestingly, the performance improves significantly from a MCC of 0.51 to 0.58 and 0.62 to 0.69 for models based on mononucleotide and dinucleotide frequency, respectively. However, the performance of trinucleotide and tetranucleotide is nearly similar (Tab. 1). This clearly shows that information in parts improves the performance; this is probably due to the presence of region specific nucleotide pattern [15, 31]. In order to see further effects, we divided UP100 and DW100 into three parts and calculated the frequencies of each part separately. However, as shown in Tab. 1, performance does not improve any further. Similarly, we developed models using frequencies of each part after dividing UP100 and DW100 in four equal parts and achieved a MCC of 0.58, 0.68, 0.69 and 0.68 for mononucleotide, dinucleotide trinucleotide and tetranucleotide respectively (Tab. 1; for details see the Tabs. S8 - S19). These results suggest that dividing the upstream and downstream into 2 parts gave better information about the region specific nucleotide features whereas further divisions in the part dilute the discriminatory features. We also developed different SVM modules using features of base composition rather than frequency and found that prediction of poly(A) signals based on frequency were marginally better than those based on composition (Tab. S20).

Hybrid modules: It has been shown, regarding classification of proteins, that hybrid models, which combine two or more features at a time, perform better than models based on single feature [25, 26, 28]. In this study, hybrid SVM models have been developed using combination of various types of nucleotide frequencies and binary pattern. Initially a SVM based hybrid model has been developed using all features of simple frequency, mononucleotide, dinucleotide, trinucleotide, tetranucleotide composition and binary pattern (size of input vector 8 + 32 + 128 + 512 + 800 = 1480), and achieved a maximum MCC of 0.68. Following this a number of hybrid models have been developed using combinations of different features of split nucleotide frequency (data not shown). Finally, we developed a hybrid model that achieved a maximum MCC of 0.72 with accuracy of 85.75% at 89.34% specificity and 82.17% sensitivity by using combination of features, like dinucleotide, 2nd order dinucleotide and tetranucleotide frequencies (Tab. S21). In this hybrid model UP100 and DW100 were divided into two equal parts and frequency of each part was calculated separately and then combined. Thus the dimension of input vector for this model is: 16 (for dinucleotide) + 16 (for 2nd order dinucleotide) + 256 (for tetranucleotide) * 4 (four regions) = 1152. The SVM parameters which gave maximum MCC are g: 0.001, c: 2, j: 1. The performance of this hybrid model along with other models, generated by similar four equal parts (50 nt long) split frequency, nucleotide and dinucleotide are shown in Fig. 3 as receiver operating characteristic (ROC) plot [32]. In a ROC curve the true positive rate (sensitivity) is plotted as a function of the false positive rate (1−specificity) for different thresholds. At a false positive rate 0.2 the true positive rate of nucleotide, dinucleotide and hybrid is 0.76, 0.87 and 0.91 respectively. As evident from Fig. 3, area under the curve in

ROC plot is more in case of hybrid model, which indicate the high accuracy of this model. Now we have taken this hybrid model for our further study.



Click on the thumbnail to enlarge the picture

Figure 3: ROC plot depict the performance of different models: hybrid (PolyApred), Nuct: mononucleotide and Dinuct: dinucleotide.

Performance on independent data

The performances of our models, designed so far, were assessed using five-fold cross-validation techniques on the main dataset. In order to make evaluation more realistic, we evaluated performance of our hybrid model on independent datasets, which were not used in training or testing of our models. The independent datasets have 982 positive sequences and four nearly equal datasets of negative sequences. The performance of our best hybrid model (used to develop the PolyApred method) combination of dinucleotide, 2nd order dinucleotide and tetranucleotide frequencies on independent datasets are shown in Tab. 2. As this table shows our method achieved PPV (positive predictive value or precision) of 75.8 to 95.7% with a sensitivity of 57%. We have also compared our method with the previous algorithms on the same datasets and observed that over all out method outperforms the other available methods (Tab. 2).

Table 2: Assessing the performance of PolyApred with other methods on independent datasets.

Program	Positive	CDS		Introns		Simple shuffling		Markov 1st order	
	SN	*PPV*	*MCC*	*PPV*	*MCC*	*PPV*	*MCC*	*PPV*	*MCC*
ERPIN	55.9	84.3	0.483	69.5	0.320	85.4	0.494	72.3	0.354
Polyadq	55.7	82.0	0.459	67.5	0.293	77.8	0.415	68.7	0.309
Liu *et al.* [20]	56.3	85.4	0.497	72.8	0.363	93.3	0.570	71.9	0.351
Polya_svm	55.2	62.2	0.216	54.5	0.066	57.7	0.142	60.7	0.192
PolyApred	57.0	89.7	0.542	78.8	0.424	95.7	0.594	75.8	0.399

In case of ERPIN [18], Polyadq [19] and *Liu et al*. [20], performance was obtained from literature, whereas performance of Polya_svm [21] and PolyApred were calculated in terms of SN (Sensitivity), PPV (Positive predictive value), and MCC (Matthews correlation coefficient).

Among the previous algorithms, the performance of Liu *et al*. [20] was better on coding sequence (CDS), introns and simple shuffling sequences whereas the performance of ERPIN was better on 1st order Markov sequences. As Tab. 2 clearly showed that the performance of PolyApred was far better than a recently developed SVM based method polya_svm. The output of the performance of polya_svm (1.0) on these datasets using default parameters is available at http://www.imtech.res.in/raghava/polyapred/data.html.

Performance on 13 variants of PASes and CDS sequences lacking PAS like hexamer

In this study, we also evaluated the performance of our method for 13 different poly(A) signals. We extracted 50 sequences for each variant of poly(A) signal from Human ATD release 2 [23]. At default threshold our method correctly predicted 45 AATAAA, 44 ATTAAA, 44 TATAAA, 45 AGTAAA, 38 AAGAAA, 44 AATATA, 42 AATACA, 40 CATAAA, 41 GATAAA, 32 AATGAA, 34 TTTAAA, 33 ACTAAA and 38 AATAGA. This result shows that highly occurring poly(A) signals were predicted with high sensitivity up to 90%. However over all we achieved a sensitivity of 79.38% for predicting 13 variants of poly(A) signals (Tab. 3).

Table 3: The performance of polyApred on independent dataset of 650 positive sequences and 650 negative sequences.

Threshold	Sensitivity	Specificity	Accuracy	*MCC*
−1	94.77	18.31	56.54	0.20
−0.9	93.69	20.77	57.23	0.21
−0.8	93.23	25.23	59.23	0.25
−0.7	92.46	28.46	60.46	0.27
−0.6	90.77	32.00	61.38	0.28
−0.5	88.62	35.69	62.15	0.29
−0.4	87.85	40.31	64.08	0.32
−0.3	85.85	46.00	65.92	0.35
−0.2	83.85	52.00	67.92	0.38
−0.1	82.15	58.15	70.15	0.42
0	79.38	61.85	70.62	0.42
0.1	75.08	67.08	71.08	0.42
0.2	70.92	69.85	70.38	0.41
0.3	65.23	73.54	69.38	0.39
0.4	61.38	78.31	69.85	0.40
0.5	57.23	81.23	69.23	0.40
0.6	53.23	83.69	68.46	0.39
0.7	48.46	87.38	67.92	0.39
0.8	41.54	90.15	65.85	0.36
0.9	34.15	93.38	63.77	0.34
1	29.69	95.54	62.62	0.34

We also evaluated the performance of our method on 650 CDS as negative sequences that have no variants of PAS or pseudo PAS signal, taken from the EMBL-CDS database. On this dataset we achieved a specificity of 61.84% at default threshold (Tab. 3). The reason of poor performance was due to the fact that the negative datasets used here for training have AATAAA in the centre, whereas independent datasets have no AATAAA in the centre. Thus we trained and evaluated our models on new datasets using 5-fold cross validation and achieved highest accuracy of 89.69% (Tab. S22). Further we have taken the same input features for training and testing on different datasets to confirm the best input feature as split/hybrid (Tabs. S23-26).

Description of the server

Based on this study, the hybrid model was used to develop the web server PolyApred for predicting poly(A) signals in a DNA sequence. This is a user-friendly server developed on SUN server under Solaris environment using HTML, PERL and CGI-PERL. Users can paste or upload their nucleotide sequence in standard FASTA format. The server also allows the user to specify the threshold. This algorithm extracts 206 nt long sliding sequence from input sequence and frequency was calculated excluding the central hexamer. If the score of the sequence exceeds the threshold the middle hexamer is predicted as a poly(A) signal.

In order to provide high weightage to poly(A) signals with known pattern, our web server generates three tables. The first table contains predicted poly(A) signals matching the well known signature (A[A/U]UAAA), whose frequency in human is ~70%, the second table contains poly(A) signals having any of the 11 known poly(A) signal signatures, whose frequency in human is ~23%, and the third table contains predicted poly(A) signals having no known signatures; these may be new/novel signatures, in this case the user needs to validate the poly(A) signals. In each table overlapping PAS-like hexamers having low score were removed. One of the powerful features of this server is that it allows predicting poly(A) signals with one of the known human PAS signatures as well as poly(A) signals with a putative novel PAS signature. This feature will help researchers to discover novel PAS signatures.

Discussion

In the last decade, the genomes of a large number of organisms have either been completely sequenced or are in the final stage. Thus, the major challenge in the present era is to annotate the genomic data in order to extract biologically meaningful information. The first most important step in genome annotation is to predict gene-coding regions. Numerous methods have been developed to predict prokaryotic and eukaryotic genes including gene structure [8-10]. One of the important factors missing in most of the genome annotation tools is prediction of poly(A) signal. The poly(A) signal plays an important role in deciding the boundary of a gene. In this study, we have made a systematic attempt to understand the characteristics of sequences containing poly(A) signals. Sequences having PAS and pseudo-PAS signals were analyzed to understand the constitution of nucleotides around them. We compared nucleotide compositions of downstream and upstream regions around PAS and pseudo-PAS signals (Figs. 1 and 2). As we have observed, the regions closer to poly(A) signal (UP1_50 and DW1_50) are more significant than regions far from poly(A) signal (UP51_100 and DW51_100) and can be efficiently utilized for discerning from the sequence having PAS and pseudo-PAS signal. A previous study has shown that distance between poly(A) signal and the downstream GU/U rich elements are 25-50 nt [33]. Thus, GU/U rich sequence lie in our DW1_50 region. Our study supports the previous finding of GU/U rich region and lack of G repeats, which are characteristic features for the binding of 64kDa subunit of CstF [34]. Cleavage-polyadenylation apparatus is associated with RNA polymerase II elongation complex and responsible for processing of pre-mRNA. Some studies have shown that pausing of polymerase downstream of polyadenylation sites triggers termination [35]. Transcription of specific sequence of DNA may cause pausing. In this study, finding of CC, GC and GG rich motif in the downstream region of poly(A) signal (DW51_100) is important and may be acting as a road blocker for polymerase.

Furthermore, we developed SVM models for predicting poly(A) signals using different input features. Initially binary patterns were used as features and got an MCC of 0.54, while using mononucleotide frequency the MCC is decreased to 0.51. Interestingly, the accuracy was increasing by increasing the contents of information, e. g. di-,tri-, tetranucleotide frequencies (Tab. 1). This is because the order of nucleotides is not a very important feature in determining poly(A) signal by machine learning techniques, instead it is the density of certain types of motifs of di-,tri-, tetranucleotides on either side of poly(A) signal. It is interesting to note here that simple composition based (di-/tri-nucleotide) models outperform binary based models, whereas binary patterns provide more information including the order of nucleotides. This is a good example of "curse of dimensionality" where optimization/learning is effected by the dimension of the input vector. In the above example, the input vector is of dimension 800 (200 × 4) for binary pattern which is difficult for SVM to handle efficiently. Thus performance of the binary-based SVM model was lower than that of the simple di-nucleotide composition based method, where the dimension of the input vector is 32 (16+16). There is need to balance between information and dimension.

Subsequently, we used split nucleotide methods for developing models, where upstream and downstream sequences were divided into 2 equal parts that further increased the accuracy up to 84.7% by using trinucleotide as well as tetranucleotide frequencies. However, further splitting of this up- and downstream sequence in 3 and 4 parts didn't improve the performance, instead, marginally decreased. These results demonstrate the presence of a specific key motif in a particular region as well as the significance of length of each region around the PAS. Splitting the sequences into 50 nt long regions gave high accuracy since it provides optimal information for discriminatory features whereas ~33 nt and 25 nt long regions deteriorate some vital information for SVM technique. This finding also supports the previous studies that show the presence of region specific different motifs around poly(A) sites [15, 31].

Finally, we used hybrid models to increase the accuracy by combining more than one feature. In the first case, we achieved a maximum MCC of 0.68 on simple frequency by combining mononucleotide, dinucleotide, trinucleotide, tetranucleotide composition and binary pattern. In spite of containing more features, the performance of this hybrid is nearly equal to that of dinucleotide frequency. This result also reveals that splitting sequence is effective to extract explicit information for better prediction with reduced number of features by SVM. Therefore, finally a hybrid model was developed on split nucleotide frequency and achieved a maximum accuracy of 85.8% and was used to develop the method PolyApred. This hybrid model is the combination of dinucleotide that gives the information about two consecutive nucleotides, 2^{nd} order dinucleotide that gives the relation of 1^{st} nucleotide with 3^{rd} nucleotide and tetranucleotides that give the information about four consecutive nucleotide pattern.

In this way, our method is simpler, considers the features of various sizes of motifs in region-specific manner and results in higher accuracy than other existing methods tested on the same independent dataset. Recently,

Cheng *et al*. developed a SVM based method, polya_svm, which searches 15 *cis*-regulatory signals and position-specific scoring matrices of these regulatory signals were used as features to predict poly(A) sites [21]. We also compared the performance of our method with polya_svm and found that our method gave better results in our benchmark test (Tab. 2). Since polya_svm predicts poly(A) sites, we considered a prediction would be true positive (TP) if a poly(A) site lies within 48 nt downstream of real poly(A) signal, otherwise we considered it FN. We tested the false positives (FP) in the negative datasets. Although the polya_svm used a novel approach for searching new poly(A) sites, our method showed better performance; this may be because of the following reasons: (1) to develop a SVM model the negative training dataset is equally important as the positive training dataset, in polya_svm negative sequences were generated from positive sequence by a first order Markov Chain model and thus it is not representing naturally occurring true negative sequences. However, our method used real negative sequences, which were taken from coding regions of genes. (2) Polya_svm used only 15 *cis*-elements and the length of each element is greater than 6 nt. As a result this algorithm works with a limited number of features and potential motifs of less than 6 nt may not be detected efficiently. Although our method used the frequency of dinucleotides, 2^{nd} order dinucleotides and tetranucleotides, which results in a vector of 1152, these features can cover important motifs of different lengths, which were not even considered by polya_svm. (3) Polya_svm considered region specific *cis*-elements and that additional feature was also considered in our method by split nucleotide techniques. Even the better performance of our method over Liu *et al*. is due to considering the region-specific nucleotides and more features like 2^{nd} order dinucleotide and tetranucleotide, which were not considered by Liu *et al.* [20]. After all, we have developed a user-friendly web server based on our investigation for scientists working in the field of genome annotation. This means that our method is able to exploit the potential features of poly(A) sites better than other SVM based methods such as Liu *et al*. [20], and polya_svm [21]. It is interesting to note from this study that simple composition/frequency-based methods perform better than position-specific nucleotide based methods (SVM module using binary pattern). This suggests that density of certain types of nucleotides, and not the order of nucleotides, is significant in determining poly(A) signals by using SVM. This is similar to subcellular localization of proteins where simple composition-based methods perform better than similarity based methods [25, 36, 37]. The newly developed SVM-based method, PolyApred, is freely accessible at http://www.imtech.res.in/raghava/polyapred/. It will be helpful to determine the 3'-end of a gene which is highly relevant to understanding the posttranscriptional gene regulation.

Acknowledgements

The authors are grateful to Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), government of India, for financial assistance. This manuscript has IMTECH communication number 053/2006.

References

1. Addepalli, B. and Hunt, A. G. (2007). A novel endonuclease activity associated with the *Arabidopsis* ortholog of the 30-kDa subunit of cleavage and polyadenylation specificity factor. Nucleic Acids Res. **35**, 4453-4463.
2. Danckwardt, S., Hentze, M. W. and Kulozik, A. E. (2008). 3' end mRNA processing: molecular mechanisms and implications for health and disease. EMBO J. **27**, 482-498.
3. Buratowski, S. (2005). Connections between mRNA 3' end processing and transcription termination. Curr. Opin. Cell Biol. **17**, 257-261.
4. Reed, R. and Hurt, E. (2002). A conserved mRNA export machinery coupled to pre-mRNA splicing. Cell **108**, 523-531.
5. Proudfoot, N. J., Furger, A. and Dye, M. J. (2002). Integrating mRNA processing with transcription. Cell **108**, 501-512.
6. Orkin, S. H., Cheng, T. C., Antonarakis, S. E. and Kazazian, H. H., Jr. (1985). Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene. EMBO J. **4**, 453-456.
7. Jansen, R. P. (2001). mRNA localization: message on the move. Nat. Rev. Mol. Cell Biol. **2**, 247-256.
8. Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol.

Biol. **268**, 78-94.

9. Issac, B., Singh, H., Kaur, H. and Raghava, G. P. (2002). Locating probable genes using Fourier transform approach. Bioinformatics **18**, 196-197.

10. Issac, B. and Raghava, G. P. (2004). EGPred: prediction of eukaryotic genes using ab initio methods after combining with sequence similarity approaches. Genome Res. **14**, 1756-1766.

11. Sharma, D., Issac, B., Raghava, G. P. and Ramaswamy, R. (2004). Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. Bioinformatics **20**, 1405-1412.

12. Graber, J. H., Cantor, C. R., Mohr, S. C. and Smith, T. F. (1999). *In silico* detection of control signals: mRNA 3'-end-processing sequences in diverse species. Proc. Natl. Acad. Sci. USA **96**, 14055-14060.

13. Hajarnavis, A., Korf, I. and Durbin, R. (2004). A probabilistic model of 3' end formation in *Caenorhabditis elegans*. Nucleic Acids Res. **32**, 3392-3399.

14. Tian, B., Hu, J., Zhang, H. and Lutz, C. S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. **33**, 201-212.

15. Hu, J., Lutz, C. S., Wilusz, J. and Tian, B. (2005). Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. RNA **11**, 1485-1493.

16. Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. Genome Res. **10**, 1001-1010.

17. Salamov, A. A. and Solovyev, V. V. (1997). Recognition of 3'-processing sites of human mRNA precursors. Comput. Appl. Biosci. **13**, 23-28.

18. Tabaska, J. E. and Zhang, M. Q. (1999). Detection of polyadenylation signals in human DNA sequences. Gene **231**, 77-86.

19. Legendre, M. and Gautheret, D. (2003). Sequence determinants in human polyadenylation site selection. BMC Genomics **4**, 7.

20. Liu, H., Han, H., Li, J. and Wong, L. (2003). An in-silico method for prediction of polyadenylation signals in human sequences. Genome Inform. **14**, 84-93.

21. Cheng, Y., Miura, R. M. and Tian, B. (2006). Prediction of mRNA polyadenylation sites by support vector machine. Bioinformatics **22**, 2320-2325.

22. Joachims, T. (1999). Making large-scale support vector machine learning practical. In: Advances in kernel methods: support vector learning. MIT Press, pp. 169-184.

23. Le Texier, V., Riethoven, J. J., Kumanduri, V., Gopalakrishnan, C., Lopez, F., Gautheret, D. and Thanaraj, T. A. (2006). AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. BMC Bioinformatics **7**, 169.

24. Kaur, H. and Raghava, G. P. (2003). A neural-network based method for prediction of gamma-turns in proteins from multiple sequence alignment. Protein Sci. **12**, 923-929.

25. Garg, A., Bhasin, M. and Raghava, G. P. (2005). Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J. Biol. Chem. **280**, 14427-14432.

26. Saha, S. and Raghava, G. P. (2006). AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. Nucleic Acids Res. **34**, W202-209.

27. Kumar, M., Bhasin, M., Natt, N. K. and Raghava, G. P. (2005). BhairPred: prediction of beta-hairpins in a protein from multiple alignment information using ANN and SVM techniques. Nucleic Acids Res. **33**, W154-159.

28. Kumar, M., Verma, R. and Raghava, G. P. (2006). Prediction of mitochondrial proteins using support vector machine and hidden Markov model. J. Biol. Chem. **281**, 5357-5363.

29. Kaundal, R., Kapoor, A. S. and Raghava, G. P. (2006). Machine learning techniques in disease forecasting: a case study on rice blast prediction. BMC Bioinformatics **7**, 485.

30. Saha, S. and Raghava, G. P. (2007). Prediction of neurotoxins based on their function and source. In Silico Biol. **7**, 369-387.

31. Shen, Y., Ji, G., Haas, B. J., Wu, X., Zheng, J., Reese, G. J. and Li, Q. Q. (2008). Genome level

analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. Nucleic Acids Res. **36**, 3150-3161.

32. Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. Science **240**, 1285-1293.

33. Zarudnaya, M. I., Kolomiets, I. M., Potyahaylo, A. L. and Hovorun, D. M. (2003). Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. Nucleic Acids Res. **31**, 1375-1386.

34. Takagaki, Y. and Manley, J. L. (1997). RNA recognition by the human polyadenylation factor CstF. Mol. Cell Biol. **17**, 3907-3914.

35. Yonaha, M. and Proudfoot, N. J. (1999). Specific transcriptional pausing activates polyadenylation in a coupled in vitro system. Mol. Cell. **3**, 593-600.

36. Bhasin, M. and Raghava, G. P. (2004). ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res. **32**, W414-419.

37. Bhasin, M., Garg, A. and Raghava, G. P. (2005). PSLpred: prediction of subcellular localization of bacterial proteins. Bioinformatics **21**, 2522-2524.

Supplementary file

Description: Supplementary file carries the detailed of nucleotide composition with p-value, performance of different SVM models with their parameters, dimension of vectors in each model, schematic diagram of 3 -end of pre-mRNA, and representation of different input patterns.

Supplementary file available at http://www.imtech.res.in/raghava/polyapred/Supplementary.pdf