Prediction of mRNA Polyadenylation Sites by Support Vector Machine

Yiming Cheng[1,2] Robert M. Miura[1] and Bin Tian[2]

[1] Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102

[2] Department of Biochemistry and Molecular Biology, New Jersey Medical School, University of Medicine and Dentistry of New Jersey, Newark, NJ 07101

CAMS Report 0506-48, Spring 2006

Center for Applied Mathematics and Statistics

Prediction of mRNA Polyadenylation Sites by Support Vector Machine

Yiming Cheng[1,2], Robert M. Miura[1] and Bin Tian[2,*]

[1]Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102

[2]Department of Biochemistry and Molecular Biology, New Jersey Medical School, University of Medicine and Dentistry of New Jersey, Newark, NJ 07101

[*]Corresponding author
E-MAIL: btian@umdnj.edu
FAX: (973) 972-5594
Tel: (973) 972-3615

Running title: Poly(A) site prediction by SVM

ABSTRACT

Motivation: mRNA polyadenylation is responsible for the 3'end formation of most mRNAs in eukaryotic cells and is linked to termination of transcription. Prediction of mRNA polyadenylation sites [poly(A) sites] can help identify genes, define gene boundaries, and elucidate regulatory mechanisms. Current methods for poly(A) site prediction achieve moderate sensitivity and specificity.

Results: Here, we present a method using Support Vector Machine for poly(A) site prediction. Using 15 *cis*-regulatory elements that are over-represented in various regions surrounding poly(A) sites, this method achieves higher sensitivity and similar specificity when compared with polyadq, a common tool for poly(A) site prediction. In addition, we found that while the polyadenylation signal AAUAAA and U-rich elements are primary determinants for poly(A) site prediction, other elements contribute to both sensitivity and specificity of the prediction, indicating a combinatorial mechanism involving multiple elements when choosing poly(A) sites in human cells.

Availability: The method is implemented in the program polya_svm, which can be downloaded from http://exon.umdnj.edu/polya_svm.

Contact: btian@umdnj.edu

Key words: polyadenylation, *cis* elements, machine learning, Support Vector Machine

INTRODUCTION

mRNA polyadenylation is the cellular process that adds poly(A) tails to maturing mRNAs. The process of polyadenylation is composed of two tightly coupled steps (Colgan and Manley, 1997): an endonucleolytic cleavage at a polyadenylation site [poly(A) site] and subsequent polymerization of an adenosine tail at the 3' end of cleaved RNA. Polyadenylation is directly linked to the termination of transcription (Buratowski, 2005; Proudfoot, 2004). Malfunction of polyadenylation has been implicated in several human diseases (Bennett, et al., 2001; Gehring, et al., 2001).

Signals required for promoting polyadenylation reside near the poly(A) site. The genomic sequence surrounding a poly(A) site is referred to as a poly(A) region. The nucleotide composition of human poly(A) regions is generally U-rich (Legendre and Gautheret, 2003; Tian, et al., 2005). A hexamer AAUAAA or a close variant, usually referred to as the polyadenylation signal (PAS), is located 10-35 nucleotides (nt) upstream of most human poly(A) sites (Tian, et al., 2005). U/GU-rich sequences are located within ~40nt downstream of the poly(A) sites (Hu, et al., 2005; Zarudnaya, et al., 2003). In addition, a number of auxiliary upstream elements (USE) or downstream elements (DSE) have been identified in viral and cellular systems ((Hu, et al., 2005), and references therein). Yeast and plant genes utilize a distinct set of *cis*-regulatory elements, or *cis* elements, for polyadenylation (Graber, et al., 1999; Zhao, et al., 1999). While AAUAAA is a prominent hexamer located upstream of poly(A) sites in these species, it occurs to a much lesser extent than in mammalian systems. Other A-rich elements seem to be equivalent to AAUAAA. In addition, UAUA and UGUA elements are the efficiency elements (EEs) located 30-70 nt upstream of yeast poly(A) sites (Graber, 2003), which also have been found to be functional elements in human cells (Venkataraman, et al., 2005).

Prediction of poly(A) sites has been attempted by several groups during the last several years. An early approach by Salamov and Solovyev (Salamov and Solovyev, 1997) used linear discriminant function. A number of variables were used, including position

weight matrices for the upstream AAUAAA element and downstream U/GU-rich element, distance between AAUAAA and U/GU-rich elements, and hexamer and triplet compositions in both upstream and downstream regions. Tabaska and Zhang (Tabaska and Zhang, 1999) developed polyadq, which employed two quadratic discriminant functions for sequences containing AAUAAA and AUUAAA. The program also uses a position weight matrix for the downstream sequence, a weighted average of hit positions for downstream elements, and downstream dimer preferences. In addition, weight-matrix-only (Legendre and Gautheret, 2003) and Hidden Markov Model (HMM) approaches (Graber, et al., 2002; Hajarnavis, et al., 2004) also have been employed for poly(A) site prediction. Overall, current methods achieve moderate sensitivity and specificity.

Using a hexamer enrichment method called PROBE, we recently identified 15 *cis* elements in 4 regions surrounding human poly(A) sites (Hu, et al., 2005)(Figure 1). These *cis* elements were suggested to play enhancing roles in mRNA polyadenylation, as they 1) are over-represented in human poly(A) regions compared with random sequences, and 2) have higher frequency of occurrence for frequently used poly(A) sites than for less frequently used ones. Based on their locations, elements are named as follows: elements in the –100 to -41 nt region are called AUEs (auxiliary upstream elements, 4 in total); elements in the –40 to -1 region are called CUEs (core upstream elements, 2 in total); elements in the +1 to +40 nt region are called CDEs (core downstream elements, 4 in total); and elements in the +41 to +100 nt region are called ADEs (auxiliary upstream elements, 5 in total). Here, we present a method using Support Vector Machine (SVM) for poly(A) site prediction by these 15 *cis* elements. This method, which is implemented in a program called polya_svm, achieved higher sensitivity and similar specificity when compared with polyadq. In addition, we found that while PAS and U-rich elements are the most important determinants for prediction, other elements also contribute to the sensitivity and specificity of prediction, indicating a combinatorial mechanism involving multiple elements when utilizing poly(A) sites in human cells.

RESULTS AND DISCUSSION

Fifteen *cis* elements in human poly(A) regions

We have previously identified 15 candidate *cis* elements in 4 regions of human poly(A) sites using a hexamer enrichment approach (Hu, et al., 2005)(Figure 1). In the present study, we wished to address whether these elements could be used to predict poly(A) sites and how each one contributes to the prediction. On one hand, predication of poly(A) sites with high sensitivity and specificity can help identify genes in the genome, define gene boundaries, and elucidate regulatory mechanisms. On the other hand, successful prediction of poly(A) sites using these elements can validate the functions of these elements, and provide insights into how the polyadenylation machinery works in human cells. To this end, we first used position-specific scoring matrices (PSSMs) of 15 elements to search the –100 to +100 nt region of 29,283 human poly(A) sites listed in the polyA_DB database (Zhang, et al., 2005). These poly(A) sites were identified using poly(A/T)-tailed cDNA/EST sequences, are located in different regions of a gene, such as introns, internal exons, and 3'-most exons, and correspond to 13,942 genes (Tian, et al., 2005). As shown in Figure 2, different elements are present in poly(A) regions to various degrees. PAS element (CUE2), upstream and downstream U-rich elements (AUE2, CUE1, CDE2), UGUA element (AUE4), UGYCU element (CDE1), and G-rich elements (ADE3 and ADE5) tend to be present in poly(A) regions with higher frequency than others. Interestingly, the UGUG element (CDE3), which is the binding site for CstF-64 (Perez Canadillas and Varani, 2003), does not occur extensively.

To understand the relationship among *cis* elements, we applied a hierarchical clustering method to group the 15 *cis* elements based on their occurrence. Using Pearson Correlation as the metric and average linkage for tree building, the 15 elements can be largely divided into two groups (Figure 2). One group consisted of CUE1, CUE2, CDE2, AUE2, AUE3, and AUE4, which are all upstream elements except CDE2, and the other consisted of downstream elements except AUE1. This grouping is robust, as clustering using other parameters, such as Kendall's Tau Correlation and complete linkage also resulted in similar groupings (data not shown). Thus, upstream elements and

downstream elements in general have different profiles, indicating that they may compensate for each other during mRNA polyadenylation. In biochemical terms, this result suggests that weak upstream signals can be “helped” by strong downstream signals, and vice versa. However, further experiments are needed to confirm this model.

Prediction of poly(A) site by SVM

Naturally, 15 *cis* elements can be considered as 15 variables and used in machine-learning tools for poly(A) site prediction. In particular, methods that take into account interactions between the variables are most suitable for predicting poly(A) sites. This is because *cis* elements are recognized by RNA-binding proteins during mRNA polyadenylation, such as CPSF-160 binding to AAUAAA, CF I_m binding to the UGUA element, CstF-64 binding to the U-rich and UG-rich elements, hFip1 binding to the U-rich element, and hnRNP H family proteins binding to the G-rich element, and extensive protein-protein interactions have been reported for proteins in the polyadenylation machinery (Proudfoot, 2004). In light of this, we tested three discriminant analysis methods, namely linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and support vector machine (SVM). LDA finds a hyper-plane to separate two or more classes with linear combination of variables (Zhang, 2000). It assumes that the data distribution for each class is normal and that all classes have the same covariance. QDA uses a quadratic surface to separate classes (Zhang, 2000), and also makes the assumption of a normal distribution, but relaxes the requirement of covariance. SVM employs kernel functions to separate data by a hyperplane that is supported by vectors lying at the boundaries of classes (Cortes and Vapnik, 1995). Several formulations and kernels are available for SVM (Burges and Smola, 1998). All these methods have been used on biological sequences for identification of signals, such as splice site (Yeo, et al., 2004; Zhang, 2000; Zhang, et al., 2003).

To compare these methods, we randomly selected 2,000 poly(A) sites from the polyA_DB database, retrieved the –100/+100 nt genomic region surrounding each poly(A) site, and used them as a positive dataset. We then randomized the positive

sequences by a first-order Markov Chain (MC) model to obtain 2,000 negative sequences, each with 200 nt in length. Using LDA, QDA, and SVM functions in program R, we compared their performance for prediction of poly(A) sites with respect to sensitivity (SN), specificity (SP), and correlation coefficient (CC, see Materials and Methods for calculation of these values). As summarized in Table 1, of these three methods, QDA achieved the best sensitivity and SVM achieved the best specificity. Overall, SVM has the best performance judged by CC. Thus, we have selected SVM as the prediction method for further studies.

To examine how each element contributes to the prediction and whether or not we can reduce the number of variables, we conducted a leave-one-out experiment, where we left out one element at a time and calculated its effect on SN, SP, and CC in poly(A) site prediction. We reasoned that omission of important elements would significantly lower the performance of prediction, whereas omission of non-essential ones would not make much difference. As shown in Table 2, we found that CUE2 is the most important one, as omission of CUE2 led to substantial drop of both sensitivity and specificity, which is consistent with the notion that the AAUAAA element is critically important for mRNA polyadenylation. Omissions of CUE1 or CDE2 had similar effects, albeit to a lesser extent, indicating that U-rich elements surrounding the poly(A) sites are important determinants. In addition, omissions of AUE3, ADE3, or ADE5 caused drops in sensitivity, indicating their important roles in poly(A) site selection. For the rest of the 9 elements, leaving out any single element caused some decrease in prediction performance, while none of them appeared to be significant based on a t-test (p-value > 0.05, Table 2). However, leaving out all 9 elements made both sensitivity and specificity drop significantly. Thus, these 9 elements may contribute to poly(A) site recognition coordinately and some elements may be important for only a small subset of poly(A) sites. Taken together, these data indicate that the 15 elements are necessary for poly(A) site prediction, validating the functional importance of these elements for polyadenylation. In addition, the fact that multiple variables are required for poly(A) site prediction suggests a combinatorial mechanism for poly(A) site recognition in human cells.

Polya_svm

Encouraged by our initial results using SVM, we developed a stand-alone program named polya_svm for poly(A) prediction using the 15 *cis* elements and SVM. The program searches an input sequence and uses LIBSVM (www.csie.ntu.edu.tw/~cjlin/libsvm) to make predictions using support vectors derived from our experiments described above. Since the prediction has to be carried out at each position of a given sequence, a multiple testing issue arises when predicting the likelihood of a sequence containing a poly(A) site. Since polyadenylation cleavage is usually heterogeneous and occurs in a window rather than at a defined position (Tian, et al., 2005), we designed a window-based scoring scheme to address the multiple testing issue: We required a region M of *m* nt to have probability > 0.5 at each position in the region and another region N of *n* nt within M to have high probability values. M is called a positive region, and N is called a high probability region (HPR, Figure 1). Using different combinations of *m* and *n*, we found that *m*=30 and *n*=10 achieved the best performance when the product of the 10 probabilities for HPR was set to be greater than 0.5. Thus, on average, the probability of being positive for each position is greater than 0.933 in HPR. As shown in Figure 3, using this method, polya_svm can effectively eliminate false positive sites and accurately locate real poly(A) sites.

We tested polya_svm using all human poly(A) regions in the polyA_DB (29,283 in total) and compared its performance with polyadq, a commonly used tool for poly(A) site prediction (Tabaska and Zhang, 1999). For polya_svm, if the predicted location (middle of HPR) is within 24 nt from a real poly(A) site, the prediction was considered true positive (TP), and otherwise false negative (FN). For polyadq, since it uses the PAS location for poly(A) site prediction, we considered a prediction to be TP if a PAS is within 48 nt upstream of a real poly(A) site. As shown in Table 3, polya_svm is 33.8% more sensitive than polyadq (52.8% SN vs. 39.5% SN). We next divided poly(A) sites into different groups based on two criteria: their usage and location. For poly(A) site usage, we used the number of supporting cDNA/ESTs for a poly(A) site to determine its

frequency of usage. Poly(A) sites in genes with only one poly(A) site are called constitutive sites. Poly(A) sites in genes with multiple sites were grouped into strong, weak, and medium. A strong site is used more than 75% of the time based on supporting cDNA/ESTs. If a gene has a strong site, other sites are called weak sites. If a gene does not have a strong site, all sites are called medium sites. For poly(A) site location, we first separated poly(A) sites located in introns and internal exons (called upstream sites) from those in 3'-most exons, and then divided poly(A) sites in the 3'-most exons into three groups depending upon their location (Table 3). The 5'-most site is called the first site, the 3'-most site is called the last site, and sites in between the 5'-most and 3'-most sites are called middle sites. In addition, if a 3'-most exon contains only one poly(A) site, the site is called a single site. As shown in Table 3, polya_svm was over 50% more sensitive than polyadq for detecting medium and weak poly(A) sites, and about 19.5% and 7.2% more sensitive than polyadq for strong and constitutive poly(A) sites, respectively. As for poly(A) sites located in different regions of a gene, polya_svm also made more sensitive predictions than polyadq in all categories, particularly for poly(A) sites located upstream of the 3'-most poly(A) sites (62.2% and 86.1% more sensitive for the first and middle poly(A) sites in 3'-most exons). A more detailed analysis revealed that the high sensitivity of polya_svm is mainly ascribed to its capability to predict poly(A) sits without AAUAAA or AUUAAA, and sites with weak downstream signals (data not shown). Taken together, these data demonstrate that polya_svm is highly sensitive for poly(A) site prediction.

We next examined polya_svm's performance for sequences without poly(A) sites, or negative sequences. A true negative sequence is difficult to obtain, as there is no extensive experimental evidence for negative sequences. Thus, we tested several types of sequences that were presumed to have very few poly(A) sites, including randomized poly(A) regions (-300/+300 nt), randomized genome sequences, mRNA coding sequences (CDS), and 5'untranslated regions (UTRs). As shown in Table 4, comparable false positives (FP) were predicted by polya_svm and polyadq for randomized sequences, but polya_svm predicts significantly more sites than polyadq in CDS and 5'UTR sequences (more than 2 fold). Interestingly, the difference was not

significant when randomized CDS and 5'UTR sequences were used (Table 4), suggesting that some of the false positives in CDS and 5'UTRs predicted by polya_svm may actually be true positives. Accordingly, it is tempting to speculate that there may, in fact, exist a large number of poly(A) sites in CDS and 5'UTRs. This would be consistent with previous findings of poly(A) sites in internal exons (Tian, et al., 2005; Yan and Marr, 2005). However, this hypothesis has yet to be tested by wet lab experiments. On the other hand, a highly sensitive method would aid in the identification of these sites, which are difficult to detect by cDNA/EST-based approaches, as many of these poly(A) sites would result in aberrant transcripts that may be rapidly degraded by cellular surveillance mechanisms, such as those without an in-frame stop codon (Frischmeyer, et al., 2002).

In summary, highly sensitive prediction of poly(A) sites was achieved by SVM using 15 *cis* elements surrounding the poly(A) sites. On the other hand, about 47% of positive poly(A) sequences in the polyA_DB database were still predicted to be negative. No obvious difference can be discerned between false negative and true positive sequences with respect to the usage of 15 *cis* elements (Figure 2), indicating other unidentified features may account for polyadenylation activity for those false negative sequences, such as RNA structure and genome location. These are to be explored in the future to further improve predictions.

MATERIALS AND METHODS

Datasets. We used 29, 283 human genomic sequences surrounding the poly(A) sites (-300 to +300 nt) in the polyA_DB database (Zhang, et al., 2005), which correspond to 13,942 genes. Poly(A) sites in the polyA_DB database were identified by aligning cDNA/ESTs with genome sequences using a method described in (Tian, et al., 2005). The training set for LDA, QDA and SVM consisted of 4,000 sequences, with 2,000 positive sequences randomly selected from polyA_DB and 2,000 negative sequences generated by the first-order Markov Chain model derived from positive sequences. To test if there is bias in choosing the training data, we randomly generated training data

10 times and did not observe any difference (data not shown). The chromosome 1 sequence of the human genome (hg17 version) was downloaded from the UCSC genome bioinformatics site (http://genome.ucsc.edu). RefSeq sequences (August 2005 version) were obtained from NCBI (Pruitt and Maglott, 2001). Poly(A) site groupings based on usage and location were carried out as in (Zhang, et al., 2005).

Scoring 15 *cis* elements. Position-specific scoring matrices (PSSMs) of 15 *cis* elements (Hu, et al., 2005) were used to search training or testing sequences, and the score for each matching sub-sequence of *n* nt was calculated by $S = \sum_{j=1}^{n} m_{i,j}$ where $m_{i,j}$ is the score of nucleotide *i* at position *j* in the PSSM. If there existed an infinite value, i.e. no possibility of occurrence, *S* was set to –15, as it was the lowest value we observed in our training data. For each element, we used the maximum score in its corresponding region for prediction. For example, for CUE2, we used its maximum score in the –40 to -1 nt region. Scores were scaled by $(S-\bar{S})/\sigma_S$ where $\bar{S}$ is the mean of *S* for all positive and negative sequences in the training data, and σ_S is the standard deviation of *S* for all positive and negative sequences in the training data.

Machine learning. Testing of LDA,QDA, and SVM was carried out in program R (http://www.r-project.org) with default settings. For prediction, we used the following equations for sensitivity (SN), specificity (SP), and Correlation Coefficient (CC):

Sensitivity: SN = $\frac{TP}{TP+FN}$, Specificity: SP = $\frac{TP}{TP+FP}$, and Correlation Coefficient: CC = $\frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$. TP is true positive; TN is true negative; FN is false negative; FP is false positive.

Polya_svm. For SVM predictions, we used the SVM library LIBSVM (www.csie.ntu.edu.tw/~cjlin/libsvm), and applied the *C*-support vector classification (*C*-SVC) method and the radial basis kernel function (RBF), with default settings, i.e. *C*=1 and *gamma*=1/15. For calculating probabilities, we applied a window-based adjustment

method using probability values generated by LIBSVM: The probability of having a poly(A) site at position *i* is called its E-value and is calculated by $E_i = 1-\prod_j \Pr(j)$ where *j* is a position relative to *i*, Pr(*j*) is the probability value obtained from LIBSVM for position *j*, and $j \in \{ -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$. Thus, the E-value is based on a 10 nt region centered at *i*, and the higher the probability, the lower the E-value. In addition, we also required that the region (-15/+15) adjacent to a positive site to have Pr(i)>0.5 at every position of its sequence.

REFERENCES

Bennett, C.L., Brunkow, M.E., Ramsdell, F., O'Briant, K.C., Zhu, Q., Fuleihan, R.L., Shigeoka, A.O., Ochs, H.D. and Chance, P.F. (2001) A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA-->AAUGAA) leads to the IPEX syndrome, *Immunogenetics*, **53**, 435-439.

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

Burges, C.J.C. and Smola, A.J. (1998) *Advances in Kernel Methods-Support Vector Learning*. MIT Press, Cambridge, MA.

Colgan, D.F. and Manley, J.L. (1997) Mechanism and regulation of mRNA polyadenylation, *Genes Dev*, **11**, 2755-2766.

Cortes, C. and Vapnik, V. (1995) Support-vector networks, *Machine Learning*, **20**, 273-297.

Frischmeyer, P.A., van Hoof, A., O'Donnell, K., Guerrerio, A.L., Parker, R. and Dietz, H.C. (2002) An mRNA surveillance mechanism that eliminates transcripts lacking termination codons, *Science*, **295**, 2258-2261.

Gehring, N.H., Frede, U., Neu-Yilik, G., Hundsdoerfer, P., Vetter, B., Hentze, M.W. and Kulozik, A.E. (2001) Increased efficiency of mRNA 3' end formation: a new genetic mechanism contributing to hereditary thrombophilia, *Nat Genet*, **28**, 389-392.

Graber, J.H. (2003) Variations in yeast 3'-processing cis-elements correlate with transcript stability, *Trends Genet*, **19**, 473-476.

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 U S A*, **96**, 14055-14060.

Graber, J.H., McAllister, G.D. and Smith, T.F. (2002) Probabilistic prediction of Saccharomyces cerevisiae mRNA 3'-processing sites, *Nucleic Acids Res*, **30**, 1851-1858.

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

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.
Legendre, M. and Gautheret, D. (2003) Sequence determinants in human polyadenylation site selection, *BMC Genomics*, **4**, 7.
Perez Canadillas, J.M. and Varani, G. (2003) Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein, *EMBO J*, **22**, 2821-2830.
Proudfoot, N. (2004) New perspectives on connecting messenger RNA 3' end formation to transcription, *Curr Opin Cell Biol*, **16**, 272-278.
Pruitt, K.D. and Maglott, D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources, *Nucleic Acids Res*, **29**, 137-140.
Salamov, A.A. and Solovyev, V.V. (1997) Recognition of 3'-processing sites of human mRNA precursors, *Comput Appl Biosci*, **13**, 23-28.
Tabaska, J.E. and Zhang, M.Q. (1999) Detection of polyadenylation signals in human DNA sequences, *Gene*, **231**, 77-86.
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.
Venkataraman, K., Brown, K.M. and Gilmartin, G.M. (2005) Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition, *Genes Dev*, **19**, 1315-1327.
Yan, J. and Marr, T.G. (2005) Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat, *Genome Res*, **15**, 369-375.
Yeo, G., Holste, D., Kreiman, G. and Burge, C.B. (2004) Variation in alternative splicing across human tissues, *Genome Biol*, **5**, R74.
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.
Zhang, H., Hu, J., Recce, M. and Tian, B. (2005) PolyA_DB: a database for mammalian mRNA polyadenylation, *Nucleic Acids Res*, **33 Database Issue**, D116-120.
Zhang, H., Lee, J.Y. and Tian, B. (2005) Biased alternative polyadenylation in human tissues, *Genome Biol*, **6:R100**.
Zhang, M.Q. (2000) Discriminant analysis and its application in DNA sequence motif recognition, *Brief Bioinform*, **1**, 331-342.
Zhang, X.H., Heller, K.A., Hefter, I., Leslie, C.S. and Chasin, L.A. (2003) Sequence information for the splicing of human pre-mRNA identified by support vector machine classification, *Genome Res*, **13**, 2637-2650.
Zhao, J., Hyman, L. and Moore, C. (1999) Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis, *Microbiol Mol Biol Rev*, **63**, 405-445.

ACKNOWLEDGEMENTS

We thank Michael Zhang for inspirational discussions of the results, and Carol Lutz, Melissa Rogers, and members of the B.T. lab for critical reading of the manuscript. Y.C. was supported by an NJIT Presidential Strategic Initiative Scholar Award. B.T. was supported by The Foundation of the University of Medicine and Dentistry of New Jersey.

TABLES

Table 1. Comparison of LDA, QDA and SVM.

Method	SN		SP		CC	
	Mean	P-value	Mean	P-value	Mean	P-value
LDA	83.0%	3.0E-12	78.5%	1.2E-99	0.603	1.6E-87
QDA	86.3%	1.3E-24	78.4%	1.1E-102	0.628	3.8E-66
SVM	84.3%	-	84.8%	-	0.693	-

LDA, Linear Discriminant Analysis; QDA, Quadratic Discriminant Analysis; SVM, Support Vector Machine. The mean values are based on 100 random tests, with 1,000 positive sequences and 1,000 negative sequences in each test. The –100 to +100 nt poly(A) region was used for predictions. Negative sequences were generated by the first-order Markov Chain model of the positive sequences. Default settings were used for each program in R. P-values are based on two-tailed t-tests comparing 100 values from LDA or QDA with those of SVM. SN, sensitivity; SP, specificity; CC, correlation coefficient.

Table 2. Effect of leaving out some *cis* elements

Element(s) left out	SN		SP		CC	
	Mean	P-value	Mean	P-value	Mean	P-value
None	84.3%	-	84.8%	-	0.693	-
AUE1*	84.2%	0.33	84.8%	0.21	0.691	0.22
AUE2*	84.2%	0.35	84.7%	0.11	0.692	0.29
AUE3	84.0%	0.03	84.7%	0.16	0.689	0.04
AUE4*	84.0%	0.07	84.9%	0.09	0.693	0.44
CUE1	83.9%	0.01	84.6%	0.03	0.687	8.5E-3
CUE2	65.3%	9.9E-161	71.5%	2.7E-153	0.395	2.8E-175
CDE1*	84.1%	0.14	84.8%	0.25	0.691	0.16
CDE2	82.6%	9.1E-17	83.8%	3.2E-10	0.668	1.2E-19
CDE3*	84.1%	0.08	85.0%	0.07	0.693	0.44
CDE4*	84.2%	0.31	84.8%	0.20	0.692	0.35
ADE1*	84.2%	0.22	84.7%	0.18	0.692	0.26
ADE2*	84.0%	0.07	84.8%	0.24	0.690	0.10
ADE3	83.7%	4.5E-4	84.8%	0.22	0.688	0.02
ADE4*	84.3%	0.45	84.7%	0.17	0.692	0.28
ADE5	83.9%	6.6E-3	84.6%	0.06	0.687	0.01
9 elements	82.1%	1.1E-24	84.5%	0.02	0.671	6.3E-17

Element(s) were left out at both the training and testing stages. None, no elements deleted; 9 elements, leaving out 9 elements (those marked with asterisk in the table) whose individual omission did not lead to significant change (p-value > 0.05) of SN, SP, or CC. Mean and p-value calculations are as in Table 1. SN, sensitivity; SP, specificity; CC, correlation coefficient.

Table 3. Comparison of polya_svm with polyadq for different types of poly(A) sites.

Poly(A) site type		polya_svm			polyadq			SN Diff
		TP	FN	SN	TP	FN	SN	
Total		15,469	13,814	52.8%	11,563	17,720	39.5%	+33.8%
poly(A) site usage	Strong	1,602	655	71.0%	1,341	916	59.4%	+19.5%
	Medium	8,301	8,221	50.2%	5,488	11,034	33.2%	+51.3%
	Weak	1,521	2,565	37.2%	961	3,125	23.5%	+58.3%
	Constitutive	4,045	2,373	63.0%	3,773	2,645	58.8%	+7.2%
poly(A) site location	Single in 3'-most exons	4,941	2,853	63.4%	4,516	3,278	57.9%	+9.4%
	First in 3'-most exons	2,469	3,583	40.8%	1,522	4,530	25.2%	+62.2%
	Middle in 3'-most exons	2,256	2,479	46.7%	1,212	3,523	25.6%	+86.1%
	Last in 3'-most exons	3,763	2,289	62.2%	2,839	3,213	46.9%	+32.6%
	Intron and internal exons	2,040	2,610	43.9%	1,474	3,176	31.7%	+38.4%

For polya_svm, a sequence is predicted to be true positive if a predicted poly(A) site (the middle position of HPR) is within 24 nt from a real poly(A) site. For polyadq, a sequence is considered true positive, if the sequence is predicted to be positive and the real poly(A) site is within 48 nt downstream of a poly(A) signal AAUAAA or AUUAAA. TP, true positive; FN, false negative; SN, sensitivity; SN Diff is calculated as $(SN_{polya\_svm}–SN_{polyadq})/SN_{polya\_svm}$. See text for description of different types of poly(A) site.

Table 4. Comparison of polya_svm with polyadq for different negative sequences

Negative Set	polya_svm				Polyadq				SP Diff	CC Diff
	TN	FP	SP	CC	TN	FP	SP	CC		
Poly(A) region first-order MC	420	80	76.7%	0.387	426	74	72.8%	0.279	+5.1%	+27.9%
Genome first-order MC	446	54	83.0%	0.451	447	53	78.9%	0.334	+4.9%	+25.9%
CDS	410	90	74.6%	0.364	458	42	82.5%	0.365	-10.6%	-0.3%
CDS first-order MC	487	13	95.3%	0.561	485	15	93.0%	0.447	+2.4%	+20.3%
5' UTR	445	55	82.8%	0.448	482	18	91.7%	0.437	-10.7%	+2.5%
5' UTR first-order MC	491	9	96.7%	0.572	492	8	96.1%	0.470	+0.6%	+17.8%

MC, Markov Chain. Poly(A) region first-order MC, randomized –300 to +300 nt sequences surrounding poly(A) sites; genome first-order MC, randomized human chromosome 1 sequences; CDS, coding region sequences of human RefSeq sequence; CDS first-order MC, randomized CDS; 5'UTR, 5'UTR sequences of human RefSeq sequences; 5'UTR first-order MC, randomized 5'UTRs. For each negative set, 500 sequences were generated and predicted by polya_svm and polyadq. The process was repeated 10 times, and mean values are presented in the table. TP and FN of Table 3 for all poly(A) sites in the polyA_DB database were scaled and used to calculate CC. Thus, for polya_svm, TP = 264 and FN = 236, and for polyadq, TP = 198 and FN = 302. SP Diff and CC Diff were calculated a similar way as SN Diff in Table 3.

FIGURE LEGENDS

Figure 1. Schematic of 15 *cis* elements in the poly(A) region and the search algorithm of polya_svm. A poly(A) site is indicated by an arrow. Fifteen *cis* elements were previously identified in four regions surrounding the poly(A) site, namely –100 to -41 nt, -40 to -1 nt, +1 to +40 nt, and +41 to +100 nt. A positive region (30 nt) is a sequence in which every position has the probability of having a poly(A) site greater than 0.5. A high probability region (10 nt) is a sequence in which the product of all predicted probabilities is greater than 0.5.

Figure 2. Clustering of 15 *cis* elements using poly(A) sites from polyA_DB. PSSMs of 15 *cis* elements were used to search human poly(A) sites from polyA_DB in their respective regions. The maximum score for each element is represented in a grayscale heapmap according to the scale shown at the bottom. Each row is a poly(A) site (29,283 in total), and each column is an element. Hierarchical clustering using Pearson Correlation was employed to cluster *cis* elements and the resulting tree is shown on top of the heatmap. Rows are ordered such that those predicted to be positive by polya_svm are listed at the top, and those negative at the bottom. Positive and negative predictions also are indicated by '+' and '-' next to the graph.

Figure 3. Prediction of positive and negative sequences. E-values of 1,000 positive and 1,000 negative sequences are shown in a heatmap according to the gray scale shown at the bottom. Each sequence is 600 nt in length. All positive sequences have a poly(A) site in the middle, and some may have multiple sites. Poly(A) site prediction was carried out for the 101 to 500 nt region. Thus, each row contains 400 E-values. An E-value is the product of probabilities in the surrounding 10 nt region. The x-axis is position in the sequence.

Figure 1.

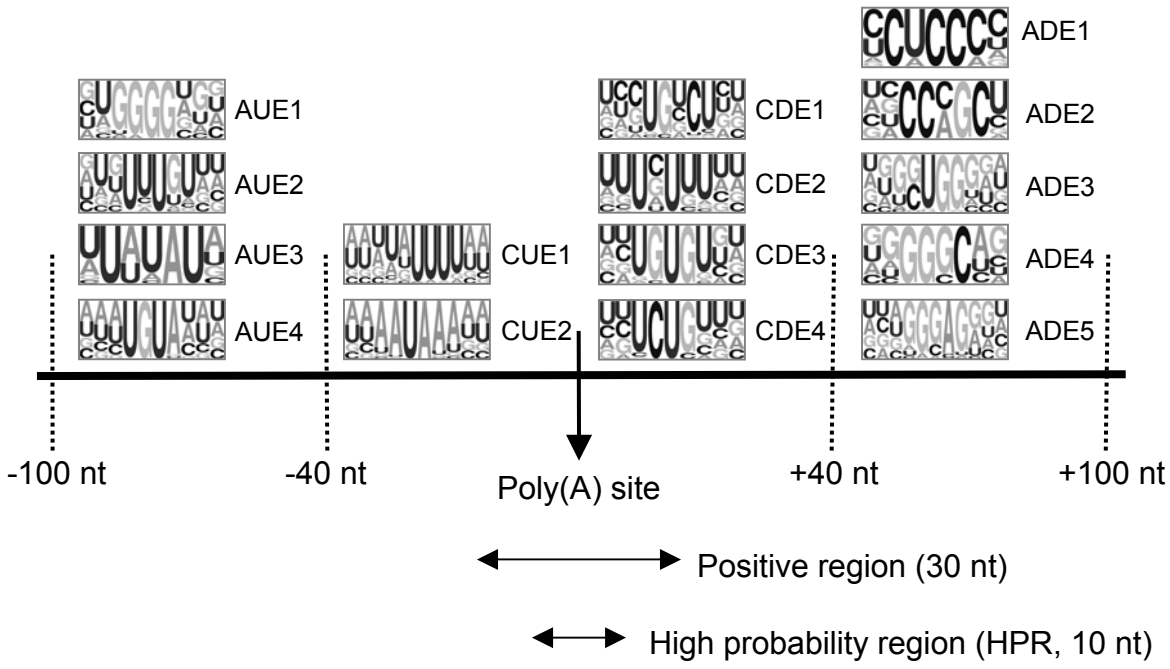

AUE1
AUE2
AUE3
AUE4
CUE1
CUE2
CDE1
CDE2
CDE3
CDE4
ADE1
ADE2
ADE3
ADE4
ADE5
-100 nt
-40 nt
Poly(A) site
+40 nt
+100 nt
Positive region (30 nt)
High probability region (HPR, 10 nt)

Figure 2.

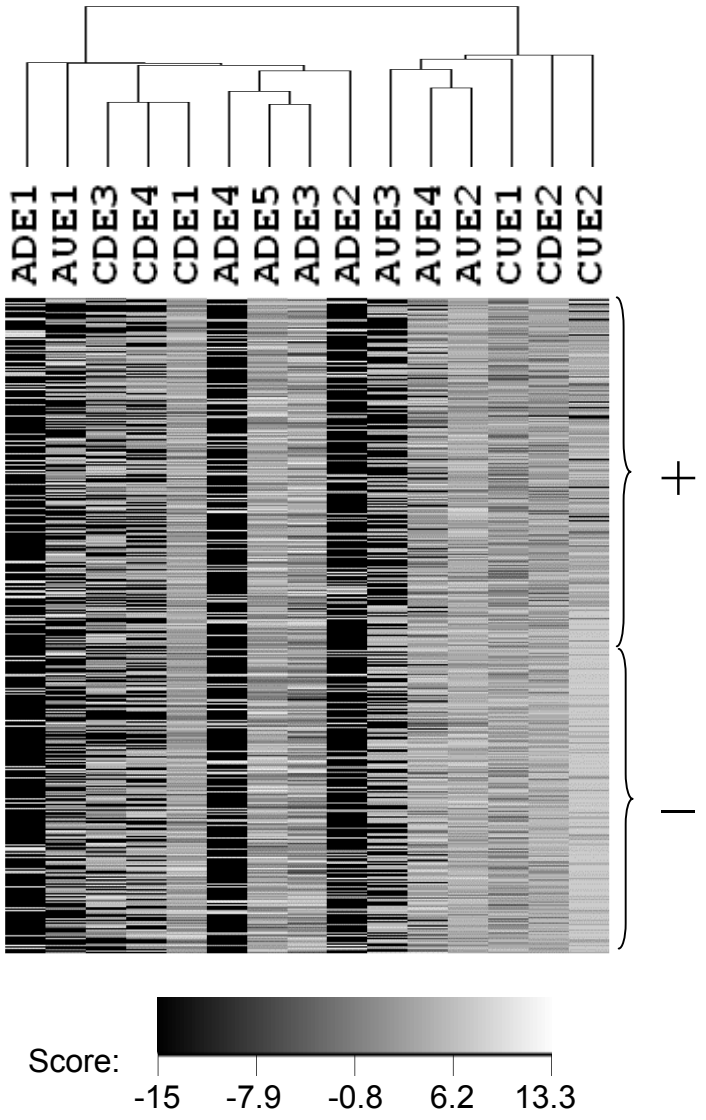
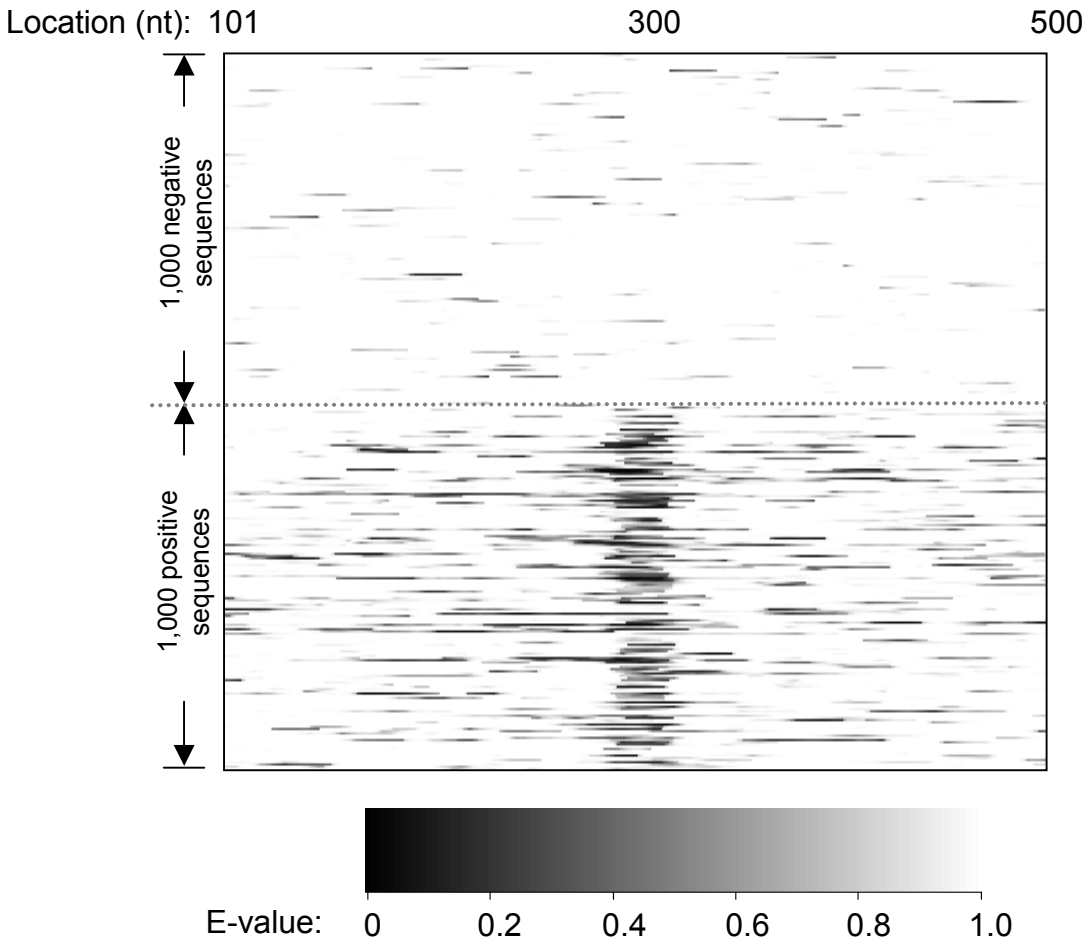

ADE1
AUE1
CDE3
CDE4
CDE1
ADE4
ADE5
ADE3
ADE2
AUE3
AUE4
AUE2
CUE1
CDE2
CUE2
+
−
Score:
-15
-7.9
-0.8
6.2
13.3

Figure 3.


Location (nt): 101
300
500
1,000 negative sequences
1,000 positive sequences
E-value:
0
0.2
0.4
0.6
0.8
1.0