DNAFSMiner: A Web-Based Software Toolbox to Recognize Two Types of Functional Sites in DNA Sequences

Huiqing Liu Hao Han Jinyan Li Limsoon Wong

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{huiqing,hanhao,jinyan,limsoon}@i2r.a-star.edu.sg

INTRODUCTION

DNAFSMiner (DNA Functional Sites Miner) is a web-based software toolbox to recognize two types of functional sites in nucleic acid sequences. Functional sites such as translation initiation sites, splice sites, polyadenylation (cleavage) sites are associated with the primary structure of genes and play important roles in gene transcription and translation. Accurately identifying these biological functional sites is an important application of computational biology and bioinformatics. Currently DNAFSMiner has been implemented to predict translation initiation sites (TIS) in vertebrate DNA/mRNA/cDNA sequences by our TIS Miner, and to predict polyadenylation (poly(A)) signals in human DNA sequences by our Poly(A) Signal Miner. DNAFSMiner is available at http://sdmc.i2r.a-star.edu.sg/DNAFSMiner/.

TECHNOLOGY BACKGROUND

DNAFSMiner is built on statistical and data mining techniques. Our method for constructing the prediction model consists of three steps: (1) generating candidate features from the sequences; (2) selecting relevant features from the candidate features, and (3) integrating the selected features with a learning algorithm to build a classification and prediction system.

In the first step, candidate features are generated using *k*-gram nucleotide acid or amino acid patterns, which are patterns defined as *k* consecutive letters of nucleotide symbols or amino acid symbols. So, candidate features are these patterns. The *number of occurrences* of a pattern within certain bps upstream and downstream of a candidate functional site is used as the value of the feature. Then, in the framework of the new feature space, the original nucleotide sequences are transformed into data of the form of integer values.

In the second step, an entropy-based feature selection algorithm is applied to the training data to select important features that can discriminate between true functional sites and false ones sharply.

In the third step, a state-of-the-art learning algorithm *support vector machines* (SVM) is used to build our prediction model. An SVM can select a small number of critical boundary samples from each class of training data and then build a discriminant function that separates them as widely as possible. The decision function for a test sample *T* is usually defined as:

$$f(T) = \sum_i a_i y_i K(x_i, T) + b$$

where x_i are the training data points, y_i are the class labels (true functional site is mapped to 1 while non-functional site is mapped to -1) of these data points, b and a_i are parameters to be determined from training samples. $K(\bullet,\bullet)$ is the kernel function which defines an inner product. The kernel function is used by the SVM to map the training data into a higher dimensional space when the linear separation is impossible in the original one. Then, $f(T)>0$ if the sample T is more likely to be a functional site, and $f(T)<0$ if T is more likely to be a non-functional site. To normalize $f(T)$, we propose a transformation function $s(T)$ defined as:

$$s(T) = \frac{1}{1+e^{-f(T)}}$$

Thus, $f(T)$ is normalized by $s(T)$ into the range (0,1). For each candidate of the functional site, score $s(T)$ is used to give the prediction. Note that if $f(T)>0$ then $s(T)>0.5$, and if $f(T)<0$ then $s(T)<0.5$.

For more information about the technologies used in DNAFSMiner, please refer to our publications [4-6]

DATA

TIS Miner

The TIS Miner was trained on 3312 vertebrate mRNA sequences extracted from GenBank release 95). This set of data consists of sequences from Bos taurus (cow), Gallus gallus (chicken), Homo sapiens (man), Mus musculus (mouse), Oryctolagus cuniculus (rabbit), Rattus norvegicus (rat), Sus scrofa (pig), and Xenopus laevis (African clawed frog) [7]. The data was first analysed by Pedersen *et al* in [7] and it contains 3312 true TIS ATGs and 10063 non-TIS ATGs. The training accuracy of the classification model is 92.45% at 80.19% sensitivity and 96.48% specificity. The model has been tested on two sets of data.

The first validation set consists of 480 human cDNA sequences that were previously analyzed by Hatzigeorgiou in [1]. This set of data was collected from the protein database SWISS-PROT. All the human proteins whose N-terminal sites are sequenced at the amino acid level were selected and manually checked. Then the full-length mRNAs for these proteins, whose TIS had been indirectly experimentally verified, were retrieved. The testing accuracy on this data (after removing sequences that were similar to the training set) is 89.42% at 96.28% sensitivity and 89.15% specificity.

The second validation set was constructed by extracting a number of annotated human genes of Chromosome X and Chromosome 21 from Human Genome Build30. Figure 1 shows the ROC curve of our model on the prediction of TISs in these genomic sequences.

Poly(A) Signal Miner

The Poly(A) Signal Miner was trained on 2327 terminal sequences including 1632 "unique" and 695 "strong" poly(A) sites. It was first used to train the system of *Erpin* [3]. Our training accuracy is 78.16% at 84.10% sensitivity and 71.54% specificity.

Then it was evaluated on a set of 982 *positive* sequences containing annotated poly(A) signals from EMBL and *four* sets of same-sized *negative* sequences: 982 CDS sequences, 982

intronic sequences of the first intron, 982 randomized UTR sequences of same 1st order Markov model as human 3' UTRs, and 982 randomized UTR sequences of same mono nucleotide composition as human 3' UTRs. These data sets were first analysed by Gautheret *et al* in [3] using Erpin. Figure 2 shows ROC curves of the Poly(A) Signal Miner on the validation sets.

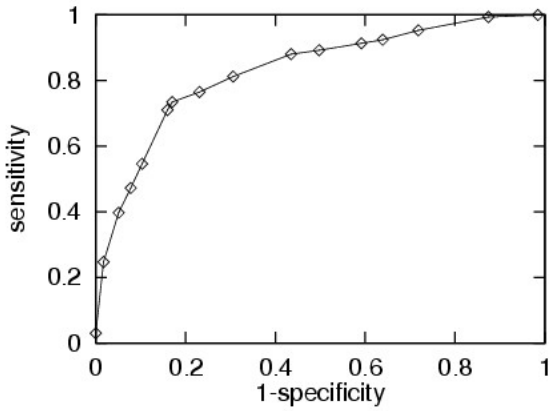


Figure 1: The ROC curve of TIS prediction on some genomic sequences given by the TIS Miner.

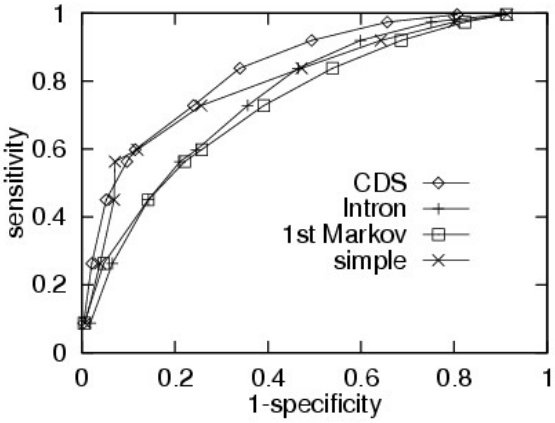


Figure 2: ROC curves of poly(A) signal prediction on some validation sequences given by the Poly(A) Signal Miner.

TOOLBOX DESCRIPTION

The main web page of DNAFSMiner is shown in Figure 3.

Input

"*TIS Miner*" and "*Poly(A) Signal Miner*" are invoked from the left pane of the main page. For prediction, a nucleic acid sequence is required which can be submitted either in raw or in FASTA format. A limit of maximum 50,000 bps per sequence per submission is set to avoid a long waiting time for users. The "*Number of predictions*" is the number of top scored candidates of the predicted functional site that is displayed in the result page (the default setting is 5). When

predicting poly(A) signals, users can also select the hexamer poly(A) signal consensus other than the default "AATAAA". The options include "ATTAAA" or any variant of "NNTANA"-type. Figure 4 shows the input pages of the TIS Miner and the Poly(A) Signal Miner, respectively.

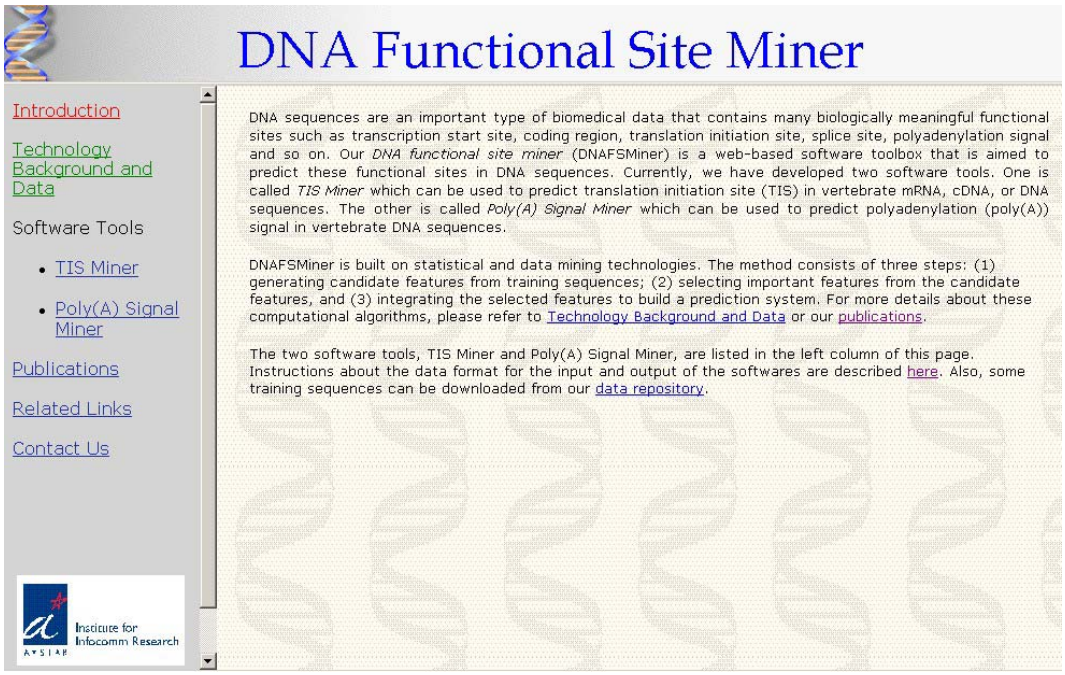
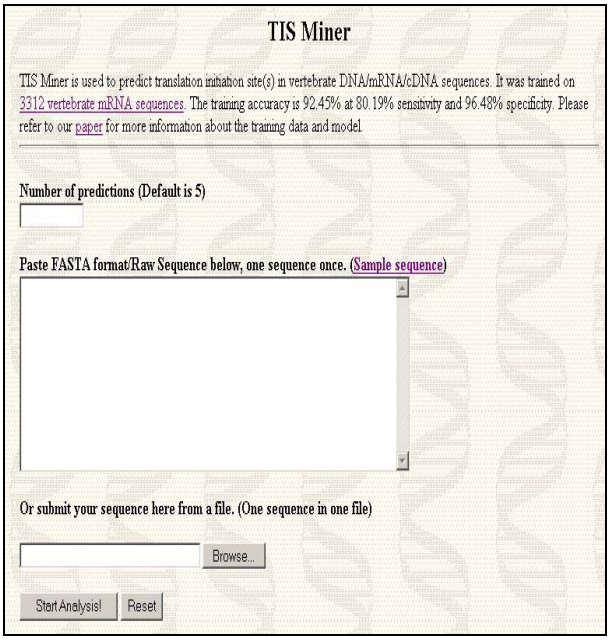
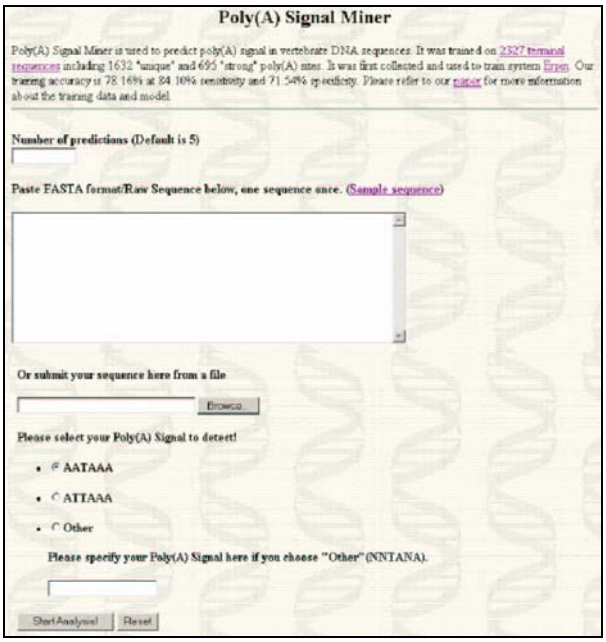


Figure 3: The main web page of DNAFSMiner



(a)



(b)

Figure 4: Input pages of DNAFSMiner: (a) TIS Miner, (b) Poly(A) Signal Miner

Range of computations

Currently, DNAFSMiner can be used for predicting: (1) translation initiation sites, which is in the form of ATG in most cases, in vertebrate DNA/mRNA/cDNA sequences; and (2) poly(A) signals in human DNA sequences.

Output

The output of the TIS Miner is displayed in a table with 6 columns described below while the output of the Poly(A) Signal Miner is also a table but with only 3 columns, i.e. the column (1), (2) and (3) of the following description. Figure 5(a) and (b) show the out page of the TIS Miner and the Poly(A) Signal Miner, respectively

(1) *No. of ATG(s)/AATAAA(s) from the 5' end*. The number *i* in this column of the table indicates that the corresponding candidate is the *i*th candidate functional site from the 5' end. Generally, a sequence may contain multiple candidates of the functional site (e.g. *ATG* for TIS and *AATAAA* for poly(A) signal).

(2) *Score*. This column shows the score (ranging in [0,1]) of the prediction that "the corresponding candidate is a true functional site". It is given by the prediction model built by SVM on the training sequences. The higher the score is, the more likely the corresponding candidate is a true functional site. We also provide the information of accuracy, sensitivity, specificity and precision under different thresholds of the score based on our validation results, for both the TIS Miner and the Poly(A) Signal Miner. Table 1 is a summary of the information of the TIS Miner. For example, if the threshold is set as 0.6 (i.e. if the prediction score of a candidate is greater than 0.6, then it will be predicted as a true TIS; otherwise, it will be predicted as a non-TIS), the accuracy, sensitivity, specificity and precision are 72.2%, 54.6%, 89.7% and 84.1%, respectively.

Threshold	Accuracy	Sensitivity	Specificity	Precision
0.1	72.3%	88.1%	56.5%	66.9%
0.2	75.3%	81.2%	69.4%	72.6%
0.3	76.7%	76.5%	76.9%	76.8%
0.4	78.2%	73.4%	83.0%	81.2%
0.5	77.5%	71.0%	84.0%	81.6%
0.6	72.2%	54.6%	89.7%	84.1%
0.7	69.7%	47.3%	92.2%	85.9%
0.8	67.3%	39.7%	94.9%	88.6%
0.9	61.5%	24.7%	98.3%	93.4%

Table 1: TIS Miner --- overall accuracy, sensitivity, specificity and precision under different thresholds of the score based on the validation results on Human Chromosome data.

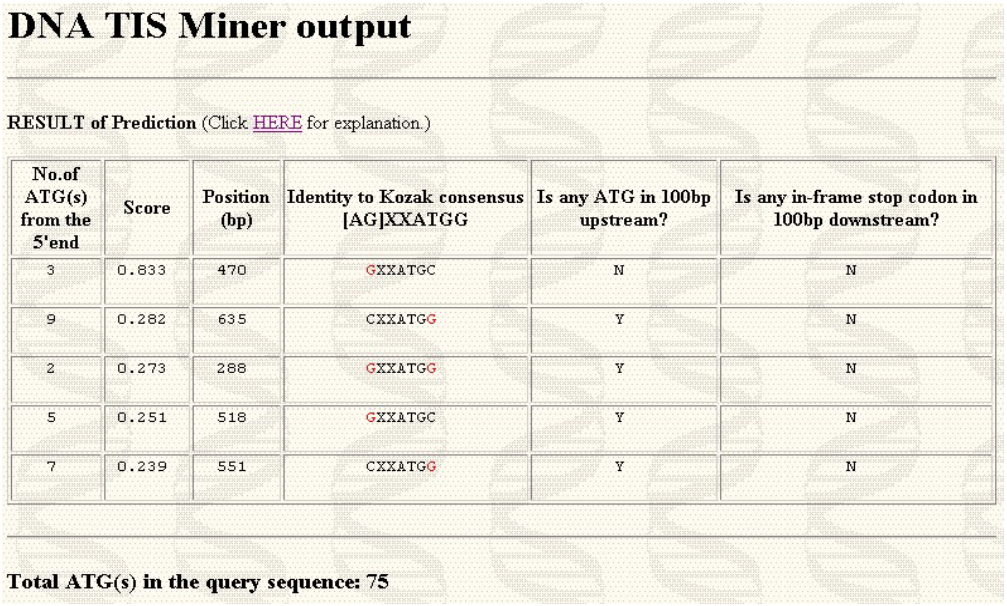
(3) *Position(bp)*. This column is the position of the corresponding candidate in the submitted nucleic acid sequence.

(4) *Identity to Kozak consensus* **[AG]XXATGC**. According to Kozak's weight matrix [2] developed for TIS prediction, a G residue tends to follow a true TIS while an A or G

residue tends to be found 3 nucleotides upstream of a true TIS. This column shows how the candidate ATG fits this consensus.

(5) *Is any ATG in 100bp* upstream? This column indicates that whether an ATG exists within the 100 bps of the upstream of the candidate.

(6) *Is any in-frame stop codon in 100bp downstream*? This column answers that whether an in-frame stop codon is found within the 100 bps of the downstream of the candidate.



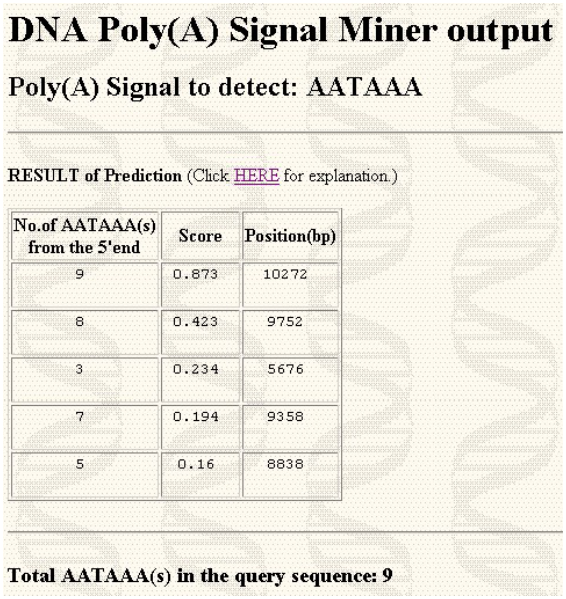
DNA TIS Miner output

RESULT of Prediction (Click HERE for explanation.)

No.of ATG(s) from the 5'end	Score	Position (bp)	Identity to Kozak consensus [AG]XXATGG	Is any ATG in 100bp upstream?	Is any in-frame stop codon in 100bp downstream?
3	0.833	470	GXXATGC	N	N
9	0.282	635	CXXATGG	Y	N
2	0.273	288	GXXATGG	Y	N
5	0.251	518	GXXATGC	Y	N
7	0.239	551	CXXATGG	Y	N

Total ATG(s) in the query sequence: 75

(a)



DNA Poly(A) Signal Miner output

Poly(A) Signal to detect: AATAAA

RESULT of Prediction (Click HERE for explanation.)

No.of AATAAA(s) from the 5'end	Score	Position(bp)
9	0.873	10272
8	0.423	9752
3	0.234	5676
7	0.194	9358
5	0.16	8838

Total AATAAA(s) in the query sequence: 9

(b)

Figure 5: Output pages of DNAFSMiner: (a) TIS Miner, (b) Poly(A) Signal Miner.

Others

In the main web page, we also provide some other useful information. For example:

- A publication list of our relevant papers about the method and technologies used to build this on-line application.
- Related links of other published programs for TIS predictions or poly(A) signal predictions, such as *ATGpr* [8], *Polyadq* [9], *Erpin* [3] and so on.
- A link to our data repository where the DNA sequences used to train our TIS Miner and Poly(A) Signal Miner can be found, as well as a large number of other types of biological data are stored.

REFERENCE

1. Hatzigeorgiou, A.G. (2002) Translation initiation start prediction in human cDNAs with high accuracy. *Bioinformatics*, **18**, 343-350.
2. Kozak, M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research*, **15**, 8125-8148.
3. Legendre, M. and Gautheret, D. (2003) Sequence determinants in human polyadenylation site selection. *BMC Genomics*, **4**(1):7.
4. Liu, H., Han, H., Li, J., and Wong, L. (2003) An In-silico method for prediction of polyadenylation signals in human sequences. *Proceedings of 14th International Conference on Genome Informatics* (GIW 2003), 84-93.
5. Liu, H., Han, H., Li, J., and Wong, L. (2004) Using amino acid patterns to accurately predict translation initiation sites. *In-Silico Biology* 4, 0022. Published online at http://www.bioinfo.de/isb/2004/04/0022/.
6. Liu, H., and Wong, L. (2003) Data mining tools for biological sequences. *Journal of Bioinformatics and Computational Biology*, **1**(1): 139-168.
7. Pedersen, A.G., and Nielsen, H. (1997) Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, 226-233.
8. Salamov. A.A., Nishikawa, T., and Swindells, M.A. (1998) Assessing protein coding region integrity in cDNA sequencing projects. *Bioinformatics,* **14**: 384-390.
9. Tabaska, J.E. and Zhang, M.Q. (1999) Detection of polyadenylation signals in human DNA sequences. *Gene*, **231**: 77-86.