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

Abstract

Summary: DNAFSMiner (DNA Functional Sites Miner) is a web-based software toolbox to recognize two types of functional sites in nucleic acid sequences. Currently it has been implemented to predict translation initiation sites (TIS) in vertebrate DNA/mRNA/cDNA sequences by TIS Miner, and to predict polyadenylation (poly(A)) signals in human DNA sequences by Poly(A) Signal Miner. DNAFSMiner is available free of charge for academic and non-profit organizations.
Availability: http://sdmc.i2r.a-star.edu.sg/DNAFSMiner/.
Contact: huiqing@i2r.a-star.edu.sg.

Introduction

DNA sequences are an important type of biomedical data that contains many biologically meaningful functional sites such as transcription start site, coding region, translation initiation site (TIS), splice site, polyadenylation signal (PAS) and so on. These functional 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.

There are several software programs that have been developed to detect TISs or PASes from DNA sequences. For example, ATGpr [6] is a web application to predict TIS in cDNA sequences using a linear discriminant function that combines some statistical features derived from the sequence. It can be accessed via interface `http://www.hri.co.jp/atgpr/`. *Polyadq* [7] and *Erpin* [2] are two programs to detect PASes in human DNA and mRNA sequences by analysing the characteristics of upstream and downstream sequence elements around PASes. *Polyadq* finds PASes using a pair of quadratic discriminant functions and is available at `http://argon.cshl.org/tabaska/polyadq_form.html`. *Erpin* was built on bioinformatics analysis of EST and genomic sequences to characterize biases in the regions encompassing 600 nucleotides around the cleavage site. The program can be found at `http://tagc.univ-mrs.fr/pub/erpin/`.

Our DNAFSMiner is a web-based toolbox to recognize TIS in vertebrate DNA/mRNA/cDNA sequences (via *TIS Miner*) as well as PASes in human DNA sequences (via *Poly(A) Signal Miner*). The software is built on statistical and data mining techniques. Our method for constructing the prediction models consists of three steps [3, 4]: (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. The prediction models are trained and validated on several different data sets, including public ones and our own extracted ones.

Toolbox Overview

Technologies. When constructing the prediction models of TIS Miner and Poly(A) Signal Miner, in the first step, we generated candidate features 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 base pairs 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 support vector machines (SVM) is used to build 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:

$$f(T) = \sum_i \alpha_i^0 y_i K(x_i, T) + b$$

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 α_i^0 are parameters to be determined. $K(\cdot)$ 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,

*To whom correspondence should be addressed.

$f(T) > 0$ 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 background technologies of our 3-step method and the data sets used for training and validating the models, please refer to our publications [5, 3, 4].

Input. For prediction, both TIS Miner and Poly(A) Signal Miner require a nucleic acid sequence which can be submitted either in raw or in FASTA format. A limit of maximum 50,000 base pairs 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 (default setting is 5). When predicting PAS, 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. Please refer to Figure 1 (a) for the input page of the Poly(A) Signal Miner.

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 1 (b) shows the output page of the TIS Miner.

(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 ith 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.

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

(4) *Identity to Kozak consensus* [**AG**]XXATG**C**. According to Kozak's weight matrix [1] 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 100 base pairs upstream?* This column indicates that whether an ATG exists within the 100 base pairs 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 base pairs of the downstream of the candidate.

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 [4].

Threshold	Accuracy	Sensitivity	Specificity	Precision
0.1	73.3%	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%

References

[1] Kozak, M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research*, **15**, 8125-8148.

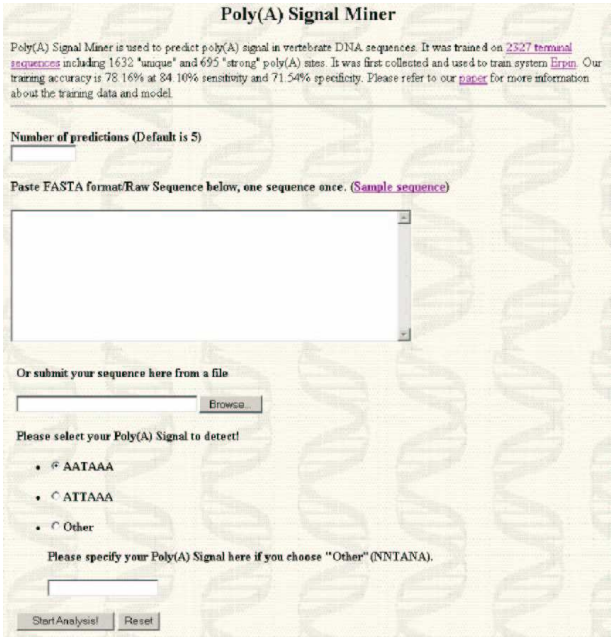
[2] Legendre, M. and Gautheret, D. (2003) Sequence determinants in human polyadenylation site selection, *BMC Genomics*, 4(1):7.

[3] 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.

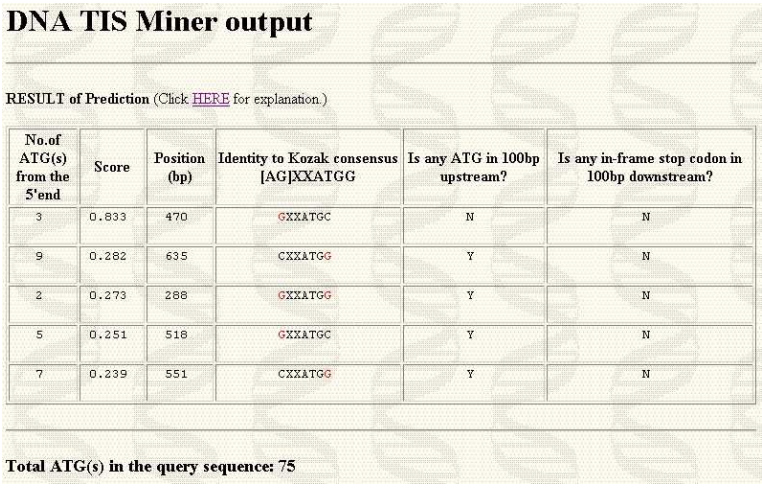
[4] Liu, H., Han, H., Li, J., and Wong, L. (2004) Using amino acid patterns to accurately predict translation initiation sites. *In-Silico Biology* 4, 0022. Published online at `http://www.bioinfo.de/isb/2004/04/0022/`.

[5] Liu, H., and Wong, L. (2003) Data Mining Tools for Biological Sequences. *Journal of Bioinformatics and Computational Biology*, **1(1)**, 139-168, Imperial College Press.

[6] Salamov, A.A., Nishikawa, T., and Swindells, M.A. (1998) Assessing protein coding region integrity in



(a) Input page of the Poly(A) Signal Miner



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

(b) Output page of the TIS Miner

Figure 1: Input and output of DNAFSMiner: (a) input page of the Poly(A) Signal Miner, (b) output page of the TIS Miner.

cDNA sequencing projects. *Bioinformatics*, **14**, 384-390.

[7] Tabaska, J.E. and Zhang, M.Q. (1999) Detection of polyadenylation signals in human DNA sequences, *Gene*, 231:77–86.