BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 1 2013, pages 117–118
doi:10.1093/bioinformatics/bts638

Genome analysis

Advance Access publication October 30, 2012

Dragon TIS Spotter: an Arabidopsis-derived predictor of translation initiation sites in plants

Arturo Magana-Mora[1,†], Haitham Ashoor[1,†], Boris R. Jankovic[1,†], Allan Kamau[1], Karim Awara[1], Rajesh Chowdhary[2], John A.C. Archer[1] and Vladimir B. Bajic[1,*]

[1]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center, Thuwal 23955-6900, Saudi Arabia and [2]Biomedical formatics Research Center, MCRF, Marshfield Clinic, Marshfield, WI 54449, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: In higher eukaryotes, the identification of translation initiation sites (TISs) has been focused on finding these signals in cDNA or mRNA sequences. Using *Arabidopsis thaliana* (*A.t.*) information, we developed a prediction tool for signals within genomic sequences of plants that correspond to TISs. Our tool requires only genome sequence, not expressed sequences. Its sensitivity/specificity is for *A.t.* (90.75%/92.2%), for *Vitis vinifera* (66.8%/94.4%) and for *Populus trichocarpa* (81.6%/94.4%), which suggests that our tool can be used in annotation of different plant genomes. We provide a list of features used in our model. Further study of these features may improve our understanding of mechanisms of the translation initiation.

Availability and implementation: Our tool is implemented as an artificial neural network. It is available as a web-based tool and, together with the source code, the list of features, and data used for model development, is accessible at http://cbrc.kaust.edu.sa/dts.

Contact: vladimir.bajic@kaust.edu.sa

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 22, 2012; revised on October 8, 2012; accepted on October 22, 2012

1 INTRODUCTION

One of the objectives of bioinformatics is to identify important biological signals in genomic sequences. The translation initiation site (TIS) is one such signal in mRNA that denotes the start codon at which translation initiates. Accurate recognition of TIS signals can help in discovery of protein-coding genes and improve annotation of gene loci (Do and Choi, 2006; Preiss and Hentze, 2003). In genomic DNA, signals that correspond to TISs consist of the ATG triplet of nucleotides, except for the rare cases of ACG or CTG triplets (Hann, 1994; Kozak, 1989). In this study, we focus solely on the recognition of ATG motifs within DNA that correspond to genuine TIS signals. We will refer to the ATG triplets as TIS motifs. Our study addresses prediction of TIS motifs in plant species. Recognizing DNA motifs in genomic sequences that correspond to genuine TIS signals is much more complex than recognizing them in mRNA or cDNA sequences, which was the main focus so far. Pertea and Salzberg achieved accuracy of 84% on both *Arabidopsis thaliana* (*A.t.*) and human genomic sequences (Pertea and Salzberg, 2002). Sparks and Brendel developed MetWAMer tools, achieving an accuracy of 85% on *A.t.* open reading frame sequences (Sparks and Brendel, 2008). In this study, using *A.t.* information, we developed a model for predicting TIS motifs within plant genomic DNA sequences, and we generated a number of features to characterize the genomic surroundings of these motifs. Some of these features have already been used for related tasks (Li and Leong, 2005; Liu and Wong, 2003), but we introduced a number of new features not previously used for TIS predictions. Out of all the features initially considered, we selected 47 as the best set of features for the TIS motif recognition task. Our feature selection is based on a wrapper method that uses a genetic algorithm (GA) and an Artificial Neural Network (ANN). There are other studies that deal with the generation and selection of features for the TIS recognition, for example (Zeng *et al.*, 2002). To the best of our knowledge, our TIS predictor is the only publicly available one for plants. The sensitivity/specificity of our model for *A.t.* is 90.75%/92.2% and is the highest compared with those reported in the literature. The accuracy tests on chromosomes of other plant genomes show sensitivity/specificity for *Vitis vinifera* of 66.8%/94.4%, and *Populus trichocarpa* of 81.6% / 94.4%. The web-based tool that implements our algorithm and our datasets are freely accessible at http://cbrc.kaust.edu.sa/dts.

2 METHODS

2.1 Datasets

TIS data for *A.t.* was extracted from *A.t.* genome and the corresponding annotations file obtained from the TAIR database, version 10 (http://www.arabidopsis.org). We extracted a total of 27 388 genuine TIS samples for positive dataset that correspond to database entries annotated as 'protein coding gene'. The same number of false TIS samples was generated from *A.t.* chromosomes 1 to 5, ensuring that any such sequence is not present in the TIS-positive set. Positive and negative TIS sequences are 300 bp in length with the TIS covering positions 150–152 counted in 5′–3′ direction. The number of negative samples taken from each of the chromosomes was proportional to the chromosome size. Even though negative cases are far more prevalent in the genome, we used equal-sized positive and negative datasets for training because we believe that these

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

data sets contain sufficiently rich distinguishing features to separate genuine TIS motifs from the false ones.

2.2 Feature generation

For TIS prediction, many useful features in sequences surrounding ATG signals are reported. Prominent amongst these can be found, for example, in (Li et al., 2004; Liu and Wong, 2003; Liu *et al*., 2004; Ma *et al*., 2006; Saeys *et al*., 2007; Tzanis and Vlahavas, 2006). Many of the reported features are local in the sense that they primarily characterize properties of the sequences immediately surrounding a candidate TIS. We extended this set of features with some that are affected by nucleotides up to 150 bp from the ATG motif site. Since selection of the optimal combination of candidate features is a combinatorial problem, we first reduce the size of the search space by defining a predetermined subset of features used in all feature-selection iterations. This fixed subset consists of features that we selected based on the previously reported results for which we believe to play a significant role in TIS recognition. We expanded the considered features with a number of new ones. The feature selection method enlarged the fixed feature set. The core step in our feature selection process is the application of genetic algorithm (GA) in search of an optimal features combination. Briefly, the process stipulates that all candidate features are numbered and assigned a value of 0 (not selected as a member of a feature set) or 1 (selected). In this way we form a 'chromosome' in the GA terms. We use a single point crossover together with mutation where each bit in a chromosome is subjected to 15% chance of having its value altered. Finally, we define evaluation function as the accuracy of model based on a 3-fold cross-validation on the training data. Description of major features and more details on feature selection, training and testing, are given in Supplementary Material 1. A full list of the used features is given in Supplementary Material 2.

2.3 Main classifier

Our prediction model is an ANN-based classifier. ANNs were used before for TIS prediction (Pedersen and Nielsen, 1997; Rajapakse and Ho, 2005; Tikole and Sankararamakrishnan, 2008). We used a 31-node single hidden-layer ANNs and the backpropagation algorithm for weights optimization. After selecting features using GA, we train the ANN. Available data, 27 388 positive (real TISs) and 27 388 negative (false TISs) samples, are split into the training and testing sets. From each of these two sets, 65% (18 802) are reserved for model training and the remaining 35% (9586) for testing. The training data (18 802 positive and 18 802 negative samples) were further divided into three parts. The first one, containing 5000 positive and 5000 negative samples were exclusively used to generate feature values. The second set containing 10 882 positive and 10 882 negative samples is used for ANN training. To avoid overfitting, the early stopping with validation method (Prechelt, 1998) is used on the remaining 2920 positive and 2920 negative samples as a validation set.

3 RESULTS

As a representative measure of model performance, we used the model sensitivity defined as $Se = TP/(TP + FN)$ and specificity $Sp = TN/(TN + FP)$, where TP, TN, FP and FN are the numbers of true positive predictions, true negative predictions, false positive predictions and false negative predictions, respectively. When evaluated on the test data only, the performance of our TIS prediction model for *A.t*. resulted in $Se = 90.75\%$ and $Sp = 90.77\%$. When we tested our model on the whole *A.t*. genome excluding the training data, we obtained $Se = 90.75\%$ and $Sp = 92.2\%$. The tests of our TIS prediction in other plant genomes, with the unmodified Arabidopsis model, resulted on *Vitis vinifera* (entire chromosomes 1 and 2) in $Se = 66.8\%$ and $Sp = 94.4\%$, and on *Populus trichocarpa* (entire chromosome 1) in $Se = 81.6\%$ and $Sp = 94.4\%$. Details are in Supplementary Material 3.

4 CONCLUSION

We developed a web tool for the recognition of TIS motifs in plant genomic DNA sequences that is based on an ANN classifier. Model features are selected by a GA as a part of the model optimization process. The model demonstrates not only an improved prediction accuracy over the reported TIS predictors for *A.t*., but also performs well on two other plant species for which it was not specifically trained. We hope that our tool will find good use in studies and annotation of gene properties of plants and may provide a further insight into the mechanisms of translation initiation.

Conflict of Interest: none declared.

REFERENCES

Do,J.H. and Choi,D.K. (2006) Computational approaches to gene prediction. *J. Microbiol*., **44**, 137–144.

Hann,S.R. (1994) Regulation and function of non-AUG-initiated protooncogenes. *Biochimie*, **76**, 880–886.

Kozak,M. (1989) Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. *Mol. Cell. Biol*., **9**, 5073–5080.

Li,G. *et al*. (2004) Translation initiation sites prediction with mixture gaussian models. In: Jonassen,I. and Kim,J. (eds) *WABI* 2004, *LNBI* 3240. Springer-Velrag, Berling Heidelberg, pp. 338–349.

Li,G.L. and Leong,T.Y. (2005) Feature selection for the prediction of translation initiation sites. *Genomics Proteomics Bioinformatics*, **3**, 73–83.

Liu,H. *et al*. (2004) Using amino acid patterns to accurately predict translation initiation sites. *In Silico Biol*., **4**, 255–269.

Liu,H. and Wong,L. (2003) Data mining tools for biological sequences. *J. Bioinform. Comput. Biol*., **1**, 139–167.

Ma,C. *et al*. (2006) Feature mining integration for improving the prediction accuracy of translation initiation sites in eukaryotic mRNAs. In *Fifth International Conference on Grid and Cooperative Computing Workshops*. IEEE Computer Society, Washington DC, pp. 349–356.

Pedersen,A.G. and Nielsen,H. (1997) Neural network prediction of translation initiation sites in eukaryotes perspectives for EST and genome analysis. In *Proceedings 5th International Conference on Intelligent Systems for Molecular Biology*. pp. 226–233.

Pertea,M. and Salzberg,S. (2002) A method to improve the performance of translation start site detection and its application for gene finding. In: Guigo,R. and Gusfield,D. (eds.) *WABI 2002, LNCS 2452*. Springer-Verlag, Berlin Heidelberg, pp. 210–219.

Prechelt,L. (1998) Early stopping–but when?. In: Orr,G.B. and Müller,K.R. (eds.) *Neural Networks: Tricks of the Trade, LNCS 1524* Springer-Verlag, Berlin Heidelberg. pp. 55–69.

Preiss,T. and Hentze,M. (2003) Starting the protein synthesis machine: eukaryotic translation initiation. *Bioessays*, **25**, 1201–1211.

Rajapakse,J.C. and Ho,L.S. (2005) Markov encoding for detecting signals in genomic sequences. *IEEE/ACM Trans. Comput. Biol. Bioinform*., **2**, 131–142.

Saeys,Y. *et al*. (2007) Translation initiation site prediction on a genomic scale: beauty of simplicity. *Bioinformatics*, **23**, i418–i423.

Sparks,M.E. and Brendel,V. (2008) MetWAMer: eukaryotic translation initiation site prediction. *BMC Bioinformatics*, **9**, 381.

Tikole,S. and Sankararamakrishnan,R. (2008) Prediction of translation initiation sites in human mRNA sequences with AUG start codon in weak Kozak context: a neural network approach. *Biochem. Biophys. Res. Commn*., **369**, 1166–1168.

Tzanis,G. and Vlahavas,I. (2006) Prediction of translation initiation sites using classifier selection. Chapter in advances in artificial intelligence. In: *Lecture Notes in Computer Science*. Vol. 3955, Springer-Verlag, Berlin Heidelberg, pp. 367–377.

Zeng,F. *et al*. (2002) Using feature generation and feature selection for accurate prediction of translation initiation sites. *Genome Inform*., **13**, 192–200.