Chapter 5
Recognition of Translation Initiation Sites in *Arabidopsis Thaliana*

Haitham Ashoor
King Abdullah University of Science and Technology, Saudi Arabia

Arturo M. Mora
King Abdullah University of Science and Technology, Saudi Arabia

Karim Awara
King Abdullah University of Science and Technology, Saudi Arabia

Boris R. Jankovic
King Abdullah University of Science and Technology, Saudi Arabia

Rajesh Chowdhary
Biomedical Informatics Research Center, USA

John A.C. Archer
King Abdullah University of Science and Technology, Saudi Arabia

Vladimir B. Bajic
King Abdullah University of Science and Technology, Saudi Arabia

ABSTRACT

Computational identification of translation initiation sites (TISs) has been of great importance in gene discovery and gene loci annotation because it predicts the start of protein coding regions. Many methods have been developed to identify TISs from cDNA and mRNA sequences, but much less work has considered TIS recognition directly from genomic DNA. In addition, to provide an insight into TIS signals conserved between distantly related eukaryotic species, the authors developed a human TIS recognition model that, when applied without modifications to TIS prediction in Arabidopsis thaliana genome, produced an accuracy of over 83 percent. When the model was trained on A. thaliana data, the resulting accuracy increased to 91 percent.

Their results suggest that in spite of the considerable evolutionary distance between Homo sapiens and A. thaliana, our approach successfully recognized deeply conserved genomic signals that characterize TIS. Moreover, they report the highest accuracy of TIS recognition in A. thaliana DNA genomic sequences.

DOI: 10.4018/978-1-61350-435-2.ch005

INTRODUCTION

One of the objectives of bioinformatics is to identify important biological signals in various genomic sequences. The translation initiation site (TIS) is one such signal that denotes the start codon at which translation initiates. Accurate recognition of TIS signals can help in discovery of protein-coding genes and in better annotation of gene loci (Preiss & Hentze, 2003, Do & Choi, 2006). Annotation engines typically assign the TIS to the first ATG codon which generates a maximal Open Reading Frame (ORF), but this by no means is sufficiently accurate.

Canonical TISs consist of the ATG triplet nucleotides, but in rare cases may consist of ACG or CTG triplets. In this study, we focus on the canonical ATG sequences (Preiss & Hentze, 2003). However, an ATG triplet will occur, on average, every 64 nucleotides in random DNA. Thus, in higher eukaryotes with large genomes, there will be a plethora of false TIS signals. For instance, in the 3.3 billion base pairs (bp) human genome with an estimated coding capacity of ~30,000 genes and assuming all are protein coding and with no alternative TISs, there will be ~30,000 real TISs and 103,095,000 false TIS signals, i.e. ~3,436 fold excess of false to true signals. Thus, there is a clear need for accurate prediction of TIS signals contained in the DNA sequence.

The presence of introns within genes, makes the accurate prediction of the TIS signals from genomic DNA sequence much more difficult than from cDNA or mRNA sequences. Extensive research has been carried out to develop computational methods for recognition of TISs mainly in cDNA and mRNA sequences. Perhaps understandably, much less attention has been given to the more difficult problem of identifying computationally these signals within genomic DNA. The associated problem is determination of the best set of features that can be used to discriminate true form false genomic signals (Saeys et al., 2007), in our case TIS signals. In this study, we introduce several new global features to the pool of already studied TIS related features, and we select the set of relevant features using a wrapper method.

Most computational recognition approaches of TIS signals have used mRNA dataset for comparing results (Pedersen and Nielsen, 1997). This dataset contains a mix of mRNA sequences from different vertebrate genomes. They (Pedersen & Nielsen, 1997) implemented an Artificial Neural Network (ANN) to predict TISs and reported an accuracy of 85% on their dataset. Later, (Hatzigeorgiou, 2002) reported an accuracy of 94% on human cDNA sequences that contain complete ORFs. She also employed a combination of two ANNs as a prediction model. Ma and colleagues developed TISKey (Ma et al., 2006), which uses an ensemble of Support Vector Machines (SVMs) and with the Pedersen and Nelsen dataset reported accuracy of 93.7%. Zeng and AlHaj used multiple agent architecture with reinforcement learning and reported 96.72% accuracy(Zeng & AlHaj, 2008). Rajapakse and Ho implemented a hybrid approach of Markov model and ANN on the Pedersen and Nielsen dataset (Rajapakse & Ho, 2005) and reported 93.8% sensitivity and 96.9% specificity using 3-folds cross validation. Li et al. used the Hatzigeorgiou dataset of mRNA sequences with full ORFs, and by using a Gaussian mixture model reported sensitivity of 98.06% and specificity of 92.14% (Li et al., 2004).

Studies based on genomic DNA sequences exhibited lower levels of accuracy. Saeyes et al. reported on human genomic DNA sequences 80% sensitivity, and 87.5% specificity (Saeyes et al., 2007). Sparks and Brendel developed the MetWAMer system which uses a perceptron classification algorithm and clustering of data by the k-medoids algorithm and methionine-weight array matrices to achieve an accuracy of 85% on *A. thaliana* genomic DNA sequences dataset (Sparks & Brendel, 2008). Pertea and Salzberg demonstrated that GlimmerM achieved 84% accuracy on both *A. thaliana* and human genomic sequences (Pertea & Salzberg, 2002).