BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i418–i423
doi:10.1093/bioinformatics/btm177

Translation initiation site prediction on a genomic scale: beauty in simplicity

Yvan Saeys[1,2,*], Thomas Abeel[1,2], Sven Degroeve[3] and Yves Van de Peer[1,2]

[1]Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium, [2]Department of Molecular Genetics, Ghent University, Ghent, Belgium and [3]Pronota, Technologiepark - Zwijnaarde 927, B-9052 Ghent, Belgium

ABSTRACT

Motivation: The correct identification of translation initiation sites (TIS) remains a challenging problem for computational methods that automatically try to solve this problem. Furthermore, the lion's share of these computational techniques focuses on the identification of TIS in transcript data. However, in the gene prediction context the identification of TIS occurs on the genomic level, which makes things even harder because at the genome level many more pseudo-TIS occur, resulting in models that achieve a higher number of false positive predictions.

Results: In this article, we evaluate the performance of several 'simple' TIS recognition methods at the genomic level, and compare them to state-of-the-art models for TIS prediction in transcript data. We conclude that the simple methods largely outperform the complex ones at the genomic scale, and we propose a new model for TIS recognition at the genome level that combines the strengths of these simple models. The new model obtains a false positive rate of 0.125 at a sensitivity of 0.80 on a well annotated human chromosome (chromosome 21). Detailed analyses show that the model is useful, both on its own and in a simple gene prediction setting.

Availability: Datafiles and a web interface for the StartScan program are available at http://bioinformatics.psb.ugent.be/supplementary_data/

Contact: yvan.saeys@psb.ugent.be

1 INTRODUCTION

The computational identification of translation initiation sites (TIS) is a major component of every gene prediction system, and is thus of major importance in genome annotation projects. In the literature, a large number of machine-learning methods have been described to identify TIS in transcripts such as mRNA, EST and cDNA sequences. These methods are all based on the scanning model (Kozak, 1989), which states that in eukaryotes the first AUG occurring at the 5′ end of the mRNA transcript is typically the correct TIS. However, exceptions can occur through mechanisms, such as leaky scanning, reinitiation and internal initiation of translation (Kozak 1999), resulting in another AUG being the true TIS.

The first method to identify TIS was proposed by Kozak (1987), who described a weight matrix to model the more or less conserved context around the TIS. However, the first real automated system for TIS prediction was the NetStart system, introduced in Pedersen and Nielsen (1997), who used an artificial neural network (ANN) to classify TIS in mRNA sequences, based on a window of 100 bp upstream and 100 bp downstream of the AUG. Around the same time, Salzberg used a conditional probability (CP) matrix (an extension of the weight matrix and equivalent to a first order Markov model) to model TIS (Salzberg, 1997). These ideas were further combined in Zien *et al.* (2000), who combined the use of support vector machines with specially developed kernels based on Salzberg's CP matrices. This work was continued by Li and Jiang (2004) who developed a new Edit–Kernel approach called TIShunter.

Another method based on ANN was proposed by Hatzigeorgiou (2002), who used a multi-step integrated neural network. ATGpr, developed by Salamov *et al.* (1998), is a program that uses a linear discriminant approach for the recognition of TIS. An improved version of ATGpr, named ATGpr_sim, also includes similarity information, and achieves better performance in terms of sensitivity and specificity (Nishikawa *et al.* 2000). Other methods for identifying TIS include the use of Gaussian mixture models (Li *et al.*, 2005) and the expectation-maximization (EM) algorithm (Wang *et al.*, 2003).

A new path of designing TIS models was opened by Zeng *et al.*, (2002), who explored a large number of potentially discriminating and biologically motivated features (*k*-mer frequencies). Feature selection methods were then used to find the most interesting features, relevant to the TIS prediction task, and a large amount of classifiers and meta-classifier schemes were evaluated. In later work, the authors modified their feature set by extracting amino acid patterns instead of *k*-mers (Liu *et al.*, 2004). Their system, referred to as TisMiner, has shown state-of-the-art TIS prediction performance on the dataset originally proposed in Pedersen and Nielsen (1997). Another advantage of the TisMiner system appears to be that, apart from being applied to transcript sequences, it can also be applied to genomic sequences. A more in-depth study of the different techniques involved in TIS recognition can be found in Li *et al.* (2004).

As mentioned earlier, all the previously described methods are focused on recognizing TIS in transcripts, except for the TisMiner system. However, recognizing TIS in mRNA, cDNA or EST transcripts is different from recognizing TIS in genomic sequences, mainly because of the following reasons: (1) scanning models cannot be applied to genomic sequences unless transcription start sites (TSS) are known, (2) transcripts typically contain zero or one TIS, which facilitates recognition significantly, (3) genomic data contains introns, which disrupt

*To whom correspondence should be addressed.

Table 1. Dataset characteristics

Name	Type	Number of Positives	Number of Negatives	Positive/Negative ratio
Pedersen–Nielsen	mRNA	3312	10 191	1/3
CCDS	Genomic	13 917	350 578	1/25
Human chromosome 21	Genomic	258	1 267 443	1/4912

the coding structure downstream of the TIS and (4) eukaryotic genomes contain millions of candidate TIS, which requires the implementation of the TIS prediction system to be computationally efficient. In this article, we focus on the identification of TIS at the *genomic* level. We investigate some simple, but extremely fast TIS prediction techniques, based on the essential characteristics of TIS recognition, and compare them to state-of-the-art models. Based on these results, we formulate some new insights for TIS prediction at the genomic scale, and we propose some guidelines for future research.

2 METHODS

2.1 Dataset construction

In our experiments, three datasets were used (see Table 1). The first dataset was constructed by Pedersen and Nielsen, and already dates back to 1997 (Pedersen and Nielsen, 1997). Although being rather old, it is included here for comparison purposes, as many TIS prediction techniques still use this dataset to validate their method. The dataset was originally extracted from Genbank and checked for suspicious annotations. Subsequently, all sequences were spliced by removing introns and joining the exon parts to mimic mRNA. The parts of the sequence upstream of the TIS are limited, the average length of the 5′ UTR being 96 bp. For model building, pseudo TIS were defined as all ATGs in the sequences that were not annotated as TIS.

The second dataset was compiled from the consensus CDS (CCDS) database.[1] The CCDS project is a collaborative effort to identify a core set of human protein coding regions that are consistently annotated and of high quality. Annotation updates represent genes that are defined by a mixture of manual curation and automated computational processing. The quality tests performed include consistency in cross-species analysis, analyses to identify putative pseudogenes, retrotransposed genes, consensus splice sites, supporting transcripts and protein homology. We downloaded the CCDS database (release date: 2 March 2005), which contains 14 802 genes. The pseudo TIS were defined as all ATG trinucleotides appearing in a window of 1000 bp upstream and 1000 bp downstream of the actual TIS of each gene. Furthermore, all genes from chromosome 21 were discarded, to ensure a non-biased validation on this chromosome (see further).

The third dataset consists of human chromosome 21, for which the sequence and annotation were downloaded from Ensembl[2] (based on NCBI build 36 version 1). This dataset contains 294 genes, 258 of which have a consensus TIS (i.e. the triplet ATG). These ones were chosen as the positive examples of the dataset, while the remaining ATGs were included as negative examples. This dataset was only used for testing purposes at the genomic scale. Some more detailed characteristics of the three datasets are shown in Table 1.

[1]http://www.ncbi.nlm.nih.gov/CCDS/
[2]http://www.ensembl.org

2.2 Features for TIS identification

In order to build a successful classifier for the recognition of TIS, the model should include sequence features that help as much as possible in discriminating between true TIS and pseudo TIS. One of the earliest observations was made by Kozak (1987), who defined a consensus context around the TIS, representing some binding mechanism to the TIS. Apart from this short context around the TIS, other features such as presence of signaling peptides (Pedersen and Nielsen, 1997) and the presence/absence of coding potential downstream/upstream of the TIS have been investigated in the literature (Liu *et al.* 2004). From previous research on identifying the most relevant features for TIS prediction (Li *et al.*, 2004; Saeys, 2004), it is known that the most prominent feature for TIS identification is the transition from a non-coding region to a coding region in the first reading frame (i.e. coinciding with the codon structure). It is clear that future TIS predictors should thus focus on these characteristics.

2.3 Simple classifiers for TIS recognition

When designing new algorithms for a given modeling task, it is standard practice to start with the most simple models, evaluate them, and then gradually build further upon these models to improve modeling performance. However, one should be cautious not to end up with overly complex models, as a simpler hypothesis should always be preferred over a complex one when performance is about the same. From a machine learning point of view, simple methods are preferred over complex ones because they have less risk of overfitting, i.e. they better generalize to unseen examples.

Based on these design principles, we here define three simple models for TIS recognition: a model based on a position weight matrix, a model based on interpolated context models (ICM), and a model based on stop codon frequencies. The first of these models has already been applied to TIS recognition, while the latter two are new in the field of TIS identification. We will now describe each of these TIS prediction models in more detail.

2.3.1 Position weight matrices Position weight matrices (PWM) are one of the simplest, and most widely used techniques to model sequence motifs. As many of these motifs (including TIS) can be linked to biological binding mechanisms, some parts of the motif may be more conserved than others. Recognition of motifs using PWM is done by observing the frequencies of nucleotides (in the case of DNA or RNA) at each position in a set of example occurrences of the motif. A new motif is then scored by observing the nucleotides at each position, and multiplying their probability of occurrence, as was estimated from the example occurrences.

For the problem of TIS prediction, we define this more formally as follows. Given a local context of u nucleotides upstream, and d nucleotides downstream of the TIS, the motif length is $w = u + d$ (the ATG triplet is invariant and thus we do not include it in the context). A training set T consists of a number of positive examples (true TIS) and negative examples (pseudo TIS), and for each of these

two classes (true/pseudo) we calculate the frequency of observing nucleotide i at position j: $p_i^j|\text{class}$ with $i \in \{A, T, C, G\}$, $1 \leq j \leq w$ and class $\in$ {true, pseudo}. A putative TIS occurrence $z = z_1 \cdots z_w$ can then be scored by calculating its log odds score, and a threshold value can be chosen for which all examples having a score higher than the threshold are classified as true TIS, while the other ones are predicted as pseudo TIS. The score is calculated as

$$\text{pred}_{\text{pwm}}(z) = \ln \frac{\prod_{i=1}^{w} p_{z_i}^i|\text{true}}{\prod_{i=1}^{w} p_{z_i}^i|\text{pseudo}} \quad (1)$$

when using the PWM to identify TIS, only the parameters u and d need to be tuned. In our experiments, tuning these parameters was done using a 3-fold cross-validation of the training set. In the case of testing the PWM on the genomic scale, u and d were tuned using a 3-fold cross-validation on the CCDS set, exploring all possible combinations of $u, d \in \{1, \ldots, 40\}$. The optimal values on the CCDS dataset were $u = 6$ and $d = 39$, and these settings were evaluated on the chrom21 dataset (see Results Section).

2.3.2 Interpolated context models As mentioned earlier, TIS are characterized by a transition from a non-coding, untranslated region (UTR) to a coding (translated) region in the first reading frame. In order to better model this transition, a new model is proposed, based on a method that was specifically designed to make the difference between coding and non-coding regions. In gene prediction, such techniques to identify coding potential lie at the heart of the gene prediction system (Borodovsky and McIninch, 1993), and a wealth of markov based models has been designed to cope with this problem. All of these methods have in common that they look for certain k-mers that are highly specific to some region, the so-called sequence biases. Based on these biases, unseen sequences can be scored as being either more likely to code for proteins, or not (Fickett and Tung, 1992).

In this work, we chose the ICM (Delcher *et al.*, 1999) as a submodel to identify coding and non-coding sequences. The ICM has the advantage of both being able to identify correlations between nucleotides and being computationally efficient. The order k of an ICM determines the size of the window in which to look for nucleotide correlations, and these need not necessarily be adjacent. In this way, the ICM extends the traditional markov models to the notion of a Bayesian decision tree, i.e. a sparse probability distribution expressed as a tree. A more extensive description of this algorithm falls outside the scope of this article, and we refer the interested reader to Delcher *et al.*, 1999 and Salzberg *et al.*, 1999 for more details.

The simple ICM-based TIS prediction algorithm that we suggest here finds its rationale in the fact that a 'good' TIS candidate should have a clear UTR region upstream, combined with a coding region in the first reading frame downstream. Pseudo TIS will typically not have these characteristics, and thus a scoring function can be constructed to measure the 'goodness' of a putative TIS. This score is defined as

$$\text{pred}_{\text{icm}}(z) = \text{ICM}^{\text{utr}}(z_{\text{upstream}}) + \text{ICM}^{\text{cod1}}(z_{\text{downstream}}) \quad (2)$$

where $\text{ICM}^{\text{utr}}(x)$ denotes the score of a homogenous (i.e. independent of the reading frame) ICM for sequence x, and ICM^{cod1} denotes the score of being in the first reading frame of a coding region for a frame-dependent ICM. The sequence z_{upstream} denotes the upstream part of the TIS context $z_1, \ldots, z_u$, while $z_{\text{downstream}}$ denotes the downstream part of the TIS context, i.e. $z_{u+1}, \ldots, z_w$.

To obtain these scores, two separate ICM models were trained: one for detecting UTR, and one for detecting coding sequences. For the UTR model, windows of 100 bp upstream of each true TIS were extracted from the training set, and these sequences were used to train the ICM. To train the coding model, first exon sequences were extracted for each true TIS in the training set.

For testing purposes however, the parameters u and d still have to be defined to identify unseen TIS. The ICM-based method thus depends on three parameters: the upstream and downstream context length u and d, and the order k of the ICM. In our experiments, all ICM models were evaluated using order $k = 8$, which enables to include dependencies between up to 9 nt. The context size parameters u and d were again tuned using a 3-fold cross-validation of the training set, thereby exploring all possible combinations of $u, d \in \{10, 15, 20 \cdots, 200\}$. A 3-fold cross-validation on the CCDS dataset revealed the optimal parameters to be $u = 60$ and $d = 140$, and these values were used for genome wide testing on the chrom21 dataset (see Results Section).

2.3.3 Stop codon frequencies Another simple measure to score putative TIS consists of looking at the stop codon frequencies downstream of the TIS. The rationale for this approach is the following. TIS are characterized by the fact that they represent the start of the first exon, so we know the reading frame of the first exon. In general, the first exon will have a minimal length, and thus there will be a minimal amount of sequence downstream of the TIS that does not contain an in-frame stop codon. On the other hand, pseudo TIS will not have this constraint, so the presence of in-frame stop codons can be used to discriminate between true and pseudo TIS. A very simple predictor can now be constructed that looks at the region following a putative TIS for the occurrence of in-frame stop codons. The earlier in this region an in-frame stop codon occurs, the less likely it is that the putative TIS is a true TIS. To obtain a simple scoring function that constructs a classifier out of this observation, we calculate the (cumulative) probability of observing an in-frame stop codon for the positive, respectively negative examples in the training set. It turns out that there is a significant difference in the cumulative distributions of the in-frame stop codons in both datasets.

Then, for each testing example, we scan the downstream part of the sequence until we find an in-frame stop codon. For this first occurrence of an in-frame stop codon, we record the position x, and we check the model to find the probability of having a first in-frame stop codon at position x following a true TIS. More formally, the score of the stop codon model is defined as

$$\text{pred}_{\text{stop}}(z) = \text{In}\left(p\left(z_{\text{downstream}}^{\text{in-frame stop}}\right)\right) \quad (3)$$

where $p(j)$ denotes the probability of finding the first in-frame stop codon at position j downstream of the TIS. The notation $z_{\text{downstream}}^{\text{in-frame stop}}$ denotes the first position in $z_{\text{downstream}}$ where an in-frame stop codon occurs. As a technicality, we would like to mention that one could also use information from the negative examples, by subtracting $\text{pred}_{\text{stop}}(z)$ by the corresponding score of finding in-frame stop codons in pseudo TIS. However, experiments showed that this never led to significantly better results, so we did not include information of the negative examples in the stop codon score.

3 RESULTS AND DISCUSSION

In this section, we discuss the results of our experiments. The most important evaluation of the methods discussed earlier consists of an evaluation of TIS predictions on human chromosome 21. The large number of putative TIS, and the limited number of true TIS among those renders this a hard test for computational TIS prediction techniques. Furthermore, we provide an analysis of our methods in a very rudimentary gene prediction setting, for which we discuss some case studies on human chromosome 21.

In addition to the three simple TIS predictors that were introduced in the previous section, two additional techniques

were evaluated. Each of these two predictors combines two or more of the simple prediction functions: $\text{pred}_{\text{pwm}+\text{icm}}$ combines pred_{pwm} and pred_{icm} by summing their score functions:

$$\text{pred}_{\text{pwm}+\text{icm}} = \text{pred}_{\text{pwm}} + \text{pred}_{\text{icm}} \quad (4)$$

and $\text{pred}_{\text{pwm}+\text{icm}+\text{stop}}$ combines all three simple prediction functions:

$$\text{pred}_{\text{pwm}+\text{icm}+\text{stop}} = \text{pred}_{\text{pwm}} + \text{pred}_{\text{icm}} + \text{pred}_{\text{stop}} \quad (5)$$

For simplicity, we will further refer to $\text{pred}_{\text{pwm}+\text{icm}+\text{stop}}$ as the StartScan system.

3.1 Evaluation on the genomic level

For this evaluation, the CCDS dataset was used for training, and the chrom21 dataset was used for testing. To provide a fair comparison between all methods, all of them were compared at the same level of sensitivity. Sensitivity (or true positive rate) is defined as the portion of all predicted positives that turn out to be true TIS, i.e. sensitivity = TP/(TP + FN). For all methods, the sensitivity was fixed at a value of 0.80, and the false positive rate (= FP/(FP + TN)) for this level of sensitivity was compared.[3] We will denote this measure as Se80. Obviously, the lower the Se80 measure, the better the classification model.

A comparison of all Se80 measures on the chrom21 dataset is shown in Table 2. As a baseline method, the simple PWM model with a context of 10 bp upstream and 10 bp downstream was chosen. This method obtains a value of Se80 = 0.27, which is not too bad for such a simple model. If we optimize the context of the PWM (as explained in the previous section), we obtain a value of Se80 = 0.19, which heavily reduces the number of false positives, compared to the non-optimized context.

The next method that was evaluated was based on the (ICM). Again, we optimized the context parameters u and d, which resulted in a score of Se80 = 0.237, which is a little worse than the optimized PWM method.

Next, the predictor based on stop codon frequencies was analyzed. Surprisingly, this method clearly outperformed both the optimized PWM and ICM methods, resulting in an Se80 measure of 0.147 and thereby drastically reducing the percentage of false positive predictions compared to the simple PWM. Although being extremely simple, the stop codon model thus performs extremely well on a genomic scale, emphasizing the importance of in-frame stop codon frequencies for a correct identification of the TIS.

Subsequently, we evaluated the results of the combined methods $\text{pred}_{\text{pwm}+\text{icm}}$ and StartScan. To find the optimal parameters for $\text{pred}_{\text{pwm}+\text{icm}}$, we adopted a greedy approach. First, the optimal parameters of pred_{icm} were chosen ($u = 60$, $d = 140$), and these were combined with all possible u and d values for the PWM submodel, $u, d \in \{1, \ldots, 40\}$. This resulted in the PWM parameters $u = 6$ and $d = 37$. We note that $\text{pred}_{\text{pwm}+\text{icm}}$ obtains an Se80 measure of 0.148, which significantly outperforms both the single pred_{pwm} and pred_{icm} models. The StartScan system, which extends $\text{pred}_{\text{pwm}+\text{icm}}$ with the stop codon model, adopts the same parameter settings, and performs best (Se80 = 0.125), obtaining a reduction of more than 50% in FP rate compared to the simple PWM model.

In a last experiment, we evaluated the predictive performance of the TisMiner method on a genomic scale. TisMiner consists of a support vector machine classifier, combined with an elaborated feature construction and feature selection procedure (Liu *et al.*, 2004). This method was chosen because it showed state-of-the-art performance on the problem of TIS recognition in transcript data, and it can be also easily applied to genomic data without jeopardizing the underlying model. The method was retrained on the CCDS dataset to allow for a fair comparison, and was then tested on the chrom21 dataset. The result is extremely poor, compared to the simple predictor methods, as a high false positive rate of 0.45 was obtained. This means that TisMiner predicts about twice as much false positives, compared to a simple PWM. Figure 1 shows the ROC curves for the simple PWM, StartScan and TisMiner systems. These curves plot the true positive rate (sensitivity) in function

Table 2. Comparative evaluation of all models on human chromosome 21

Method	Context	Se80
PWM	[10,10]	0.27
PWM_optimized	[6,39]	0.19
ICM_optimized	[60,140]	0.237
STOP	-	0.147
PWM+ICM	[6,37]+[60,140]	0.148
StartScan (PWM+ICM+STOP)	**[6,37]+[60,140]**	**0.125**
TisMiner	[100,100]	0.45

[3]TF = true positives, FP = false positives, TN = true negatives and FN = false negatives.

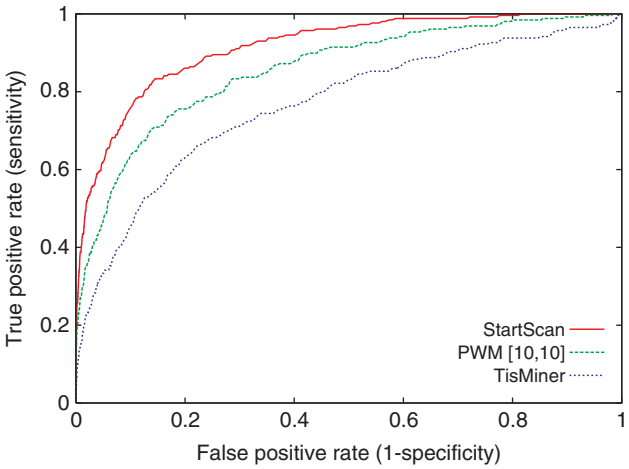

Fig. 1. ROC curves plotting the predictive performance of a simple PWM (with context 10 bp upstream and 10 bp downstream), the TisMiner system and the StartScan system.

of the false positive rate (1-specificity) and are obtained by shifting the classifier's decision threshold.

To see whether the simple methods developed here would outperform the TisMiner method on transcript data, we also explored them on the Pedersen–Nielsen dataset. As this was a single dataset for training/testing, a 3-fold cross-validation was used to asses prediction performance. However, due to the fact that this dataset consists of mRNA, and thus does not contain in-frame stop codons, predictors using $pred_{stop}$ could not be evaluated on the Pedersen–Nielsen dataset. The results of this analysis are shown in Table 3. Clearly, the TisMiner system largely outperforms the simple measures when identifying TIS in transcript data. We can thus conclude that TisMiner is well suited for the task of identifying TIS in transcript data, while the simple measures presented here are to be preferred when identifying TIS in genomic data.

Table 3. Comparative evaluation on the Pedersen–Nielsen dataset

Method	Se80
PWM_optimized	0.348
ICM_optimized	0.421
PWM+ICM	0.396
TisMiner	0.063

3.2 Evaluation in a simple gene prediction setting

To evaluate our methods in a rudimentary gene prediction setting, we combined the StartScan system with a simple sensor for coding regions, based on the Fourier characteristic (Tiwari *et al.* 1997). To evaluate the method on the genomic level, we discuss two case studies on human chromosome 21 (Fig. 2). The first region (part a) is situated around position 32,200,000 and contains the two first exons (shown as black boxes) of gene ENSG00000142149, which is oriented in the sense direction. As can be seen in the figure, both exons are identified by the coding region sensor, and the first exon contains a clear prediction of the TIS by the StartScan system. The prediction is well positioned at the beginning of the exon, and shows that the occurrence of a TIS on the sense strand is the most plausible explanation.

A second case study concerns the region around position 32,075,000 (part b), for which no gene is annotated. However, both the coding region sensor and the antisense TIS prediction

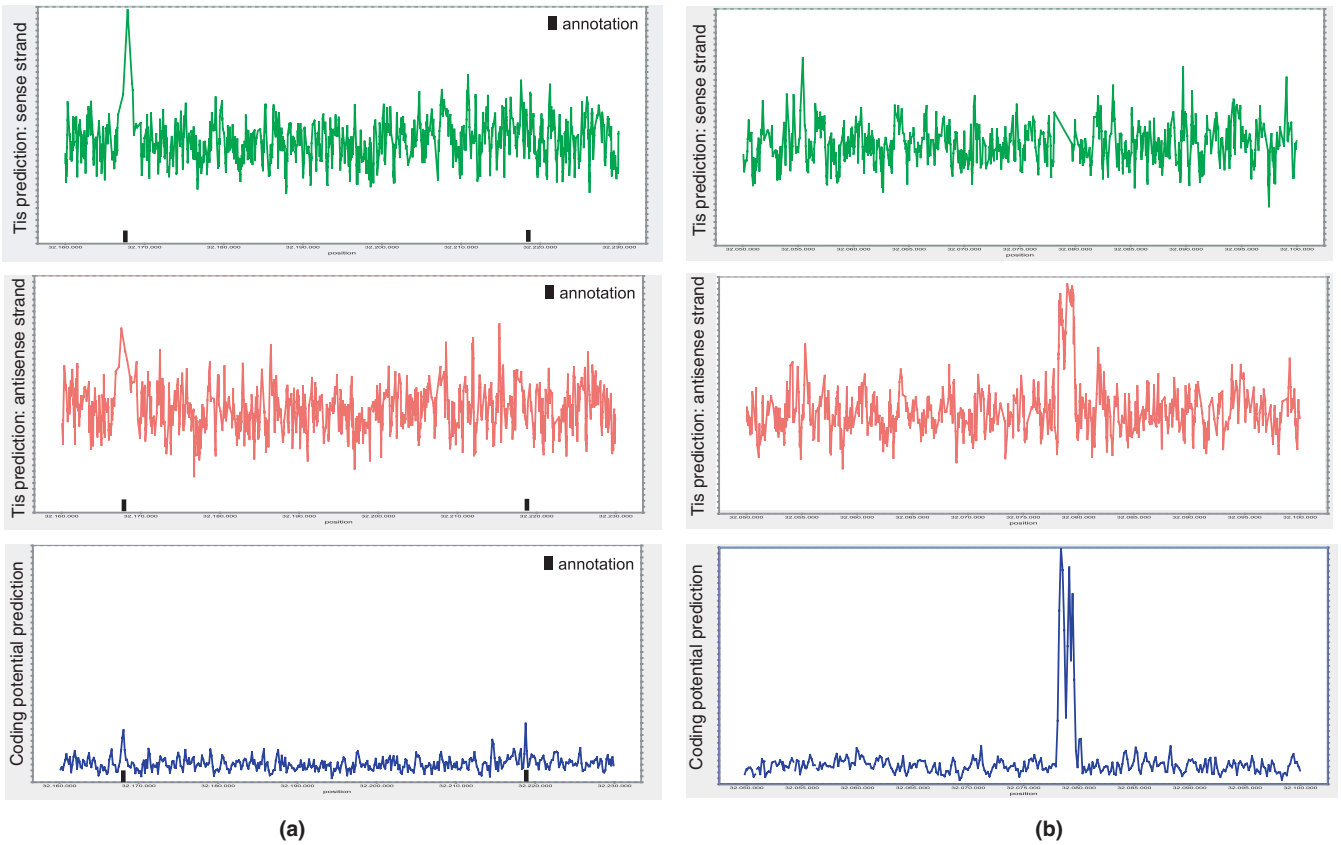

Fig. 2. Two case studies in a simple gene prediction setting on human chromosome 21. Case study (**a**) depicts the region of chromosome 21, extending from position 32,160,000 to position 32,230,000. In this region, the first two exons of the gene ENSG00000142149 can be identified using the coding potential sensor (lower figure), and a clear TIS can be identified on the sense strand (upper figure). Case study (**b**) shows the region of chromosome 21, extending from position 32,050,000 to position 32,100,000. A clear peak in the coding potential sensor (lower figure) and the antisense TIS predictor (middle figure) indicate clear evidence for a gene missed in the annotation.

of StartScan show a clear peak around position 32,080,000, suggesting the presence of a gene that is missed in the annotation. We examined this region in the Ensembl genome browser, and found many mRNAs present in this region, providing clear evidence of a missed gene.

3.3 Online TIS prediction at the genomic scale

Based on the combination of several simple TIS prediction measure, an online version of the StartScan algorithm was implemented. This program offers the user the opportunity to feed in a part of a genomic sequence, and returns a score for all putative TIS, both in the sense and the antisense direction. The StartScan system is very fast, and allows for genome wide TIS prediction. As an example, training the system on the CCDS dataset, and evaluating it on the chrom21 dataset required less than 10 min on a Pentium 2.4 GHz processor running Linux. The system is available at http://bioinformatics.psb.ugent.be/supplementary_data/

4 CONCLUSIONS AND FUTURE WORK

In this article, we presented several simple prediction methods to identify TIS on a genomic scale. We compared these methods to a state-of-the-art method for identifying TIS in transcript data, and conclude that the simple methods largely outperform it on the genomic scale. This provides evidence that simple methods can deal well with the fact that on a genomic level, many putative TIS are present, while only a very small fraction of them are true TIS. A weakness of current TIS predictors on transcript data appears to be their incapability to deal with such a high class imbalance between true and pseudo TIS, and future research to deal with this problem might render them useful in the context of identifying TIS on a genomic scale.

Further research on the simple methods proposed here includes the extension of PWM to methods that are able to better model nucleotide dependencies in the immediate context of the TIS, and the incorporation of the proposed scores into more complex classification schemes. In addition, we plan to combine the use of TIS prediction with new methods for TSS identification, and initial results indicate that StartScan predictions can be used as a valuable additional component to better annotate the beginning of a gene.

As another line of future research, we plan to incorporate StartScan into a state-of-the-art gene prediction system, where we will also include the TSS prediction methods we are currently developing.

Conflict of Interest: none declared.

REFERENCES

Borodovsky,M. and McIninch,J. (1993) GeneMark: parallel gene recognition for both DNA strands. *Comput. and Chem.*, **17**, 123–133.

Delcher,A.L. *et al.* (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.

Fickett,J.W. and Tung,C.S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, **20**, 6441–6450.

Hatzigeorgiou,A. (2002) Translation initiation start prediction in human cDNAs with high accuracy. *Bioinformatics*, **18**, 343–350.

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

Kozak,M. (1989) The scanning model for translation: an update. *J. Cell Biol.*, **108**, 229–241.

Kozak,M. (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene*, **234**, 187–208.

Li,H. and Jiang,T. (2004) A class of edit kernels for SVMs to predict translation initiation sites in eukaryotic mRNAs. In *Proceedings of the 8th International Conference on Research in Computational Molecular Biology*, 262–271.

Li,G. *et al.* (2005) Translation initiation sites prediction with mixture Gaussian models in human cDNA sequences. *IEEE Trans. Knowl. Data Eng.*, **8**, 1152–1160.

Li,J. *et al.* (2004) Techniques for recognition of translation initiation sites. In *The Practical Bioinformaticion*, World Scientific 2004.

Liu,H. *et al.* (2004) Using amino acid patterns to accurately predict translation initiation sites. *In Silico Biol.*, **4**, 255–69.

Nishikawa,T. *et al.* (2000) Prediction whether a human cDNA sequence contains initiation codon by combining statistical information and similarity with protein sequences. *Bioinformatics*, **16**, 139–168.

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

Saeys,Y. (2004) Feature selection for classification of nucleic acid sequences. PhD thesis, Ghent University, Belgium.

Salamov,A.A. *et al.* (1998) Assessing protein coding region integrity in cDNA sequence projects. *Bioinformatics*, **14**, 384–390.

Salzberg,S. (1997) A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput. Appl. Biosci.*, **13**, 365–376.

Salzberg,S.L. *et al.* (1999) Interpolated Markov models for eukaryotic gene finding. *Genomics*, **59**, 24–31.

Tiwari,S. *et al.* (1997) Prediction of probable genes by Fourier analysis of genomic sequences *Comput. Appl. Biosci.*, **13**, 263–270.

Wang,Y. *et al.* (2003) Recognition of translation initiation sites of eukaryotic genes based on an EM algorithm. *J. Comput. Biol.*, **10**, 699–708.

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

Zien,A. *et al.* (2000) Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, **16**, 799–807.