



Translation initiation start prediction in human cDNAs with high accuracy

Artemis G. Hatzigeorgiou^{1,2}

¹Metagen GmbH, Ihnestr.63, 14195 Berlin-Dahlem, Germany and Synaptic Ltd., Science and Technology Park of Crete, PO Box 1447, Voures Herakleion, 71110 Greece

Received on March 2, 2001; revised and accepted on October 24, 2001

ABSTRACT

Motivation: Correct identification of the Translation Initiation Start (TIS) in cDNA sequences is an important issue for genome annotation. The aim of this work is to improve upon current methods and provide a performance guaranteed prediction.

Methods: This is achieved by using two modules, one sensitive to the conserved motif and the other sensitive to the coding/non-coding potential around the start codon. Both modules are based on Artificial Neural Networks (ANNs). By applying the simplified method of the ribosome scanning model, the algorithm starts a linear search at the beginning of the coding ORF and stops once the combination of the two modules predicts a positive score.

Results: According to the results of the test group, 94% of the TIS were correctly predicted. A confident decision is obtained through the use of the Las Vegas algorithm idea. The incorporation of this algorithm leads to a highly accurate recognition of the TIS in human cDNAs for 60% of the cases.

Availability: The program is available upon request from the author.

Contact: agh@pcbi.upenn.edu

1 INTRODUCTION

In February 2001, the International Human Genome Sequencing Consortium published its results about the human genome (Consortium, 2001). On the matter of human genes, the consortium states '... human genes tend to have small exons (encoding an average of only 50 codons) separated by long introns (some exceeding 10 kb). This creates a signal-to-noise problem, with the result that computer programs for direct gene prediction have only limited accuracy. Instead, computational prediction of human genes must rely largely on the availability of cDNA sequences or on sequence conservation with genes

and proteins from other organisms.' This paper discusses the final stage of computational annotation of human genes at the DNA level. The central focus of the paper is the definition of the coding part of a gene, the information that leads to the protein sequence. Coding regions in cDNA sequences are surrounded by non-coding regions, also called 3' and 5' UnTranslated Regions (UTRs). The Translation Initiation Start (TIS) is defined as the start of the coding region. Following the coding frame, the coding region ends with the first stop-codon in frame, assuming of course that no frame-shifts are included in the sequence.

The original work for the identification of the TIS in a cDNA sequence dates back to 1987, when Kozak developed the first weight matrix from an extended collection of data (Kozak, 1987). The consensus motif derived from this matrix is *GCCACC*atg*G*, where a *G* residue following the ATG codon, and a purine, preferably *A*, three nucleotides upstream, are the two highly conserved positions that exert the strongest effect. Attempting to describe what really happens in the cell, Kozak developed the ribosome-scanning model. According to this model (Kozak, 1996), the ribosome first attaches to the specific cap region in the 5' end of the mRNA and then *scans* the sequence until it finds the first ATG that is in an optimal nucleotide context. This is where translation of codons to amino acids begins. Although this is true for most mRNA's studied, there are some notable exceptions (Kozak, 1996 and Pain, 1996) (Figure 1):

- *Leaky scanning*, in which case the first ATG codon has a less than optimal nucleotide context and therefore can be bypassed by the ribosome, which then initiates translation from a start codon in a more optimal nucleotide context further downstream.
- *Reinitiation*, where the translation starts from an ATG codon upstream of the coding region, that is in optimal nucleotide context in the 5' UTR and ends at the first stop codon, normally a short distance away. Scanning then continues until the authentic ATG codon (start codon) is reached.

²Present address: Department of Genetics, University of Pennsylvania, School of Medicine, 1407 Blockley Hall, 418 Guardian Drive, Philadelphia, PA 19104-6021, USA.

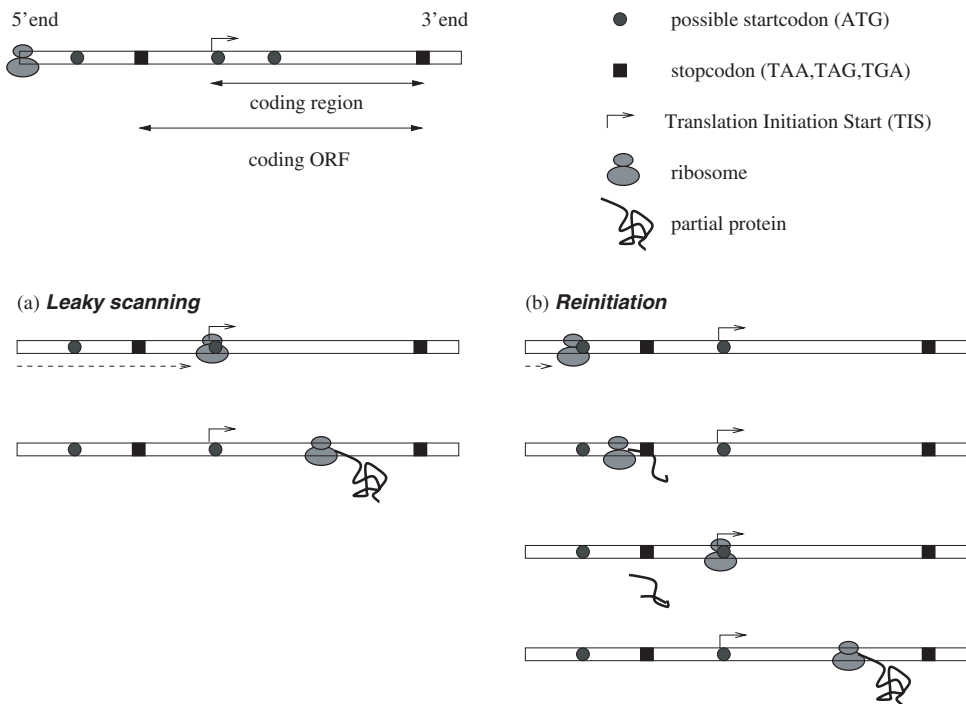


Fig. 1. Description of the ribosome scanning model on the cDNA. See also text.

- *Internal initiation*, where the ribosome, without any scanning, directly binds near the authentic ATG codon. This is only characteristic of some viral mRNAs of peculiar structure.

In the following, only the prediction of TIS in non-viral sequences will be discussed. The importance of defining the TIS within a cDNA can also be highlighted by the fact that, in recent years, several approaches have been developed to improve this prediction. Some of these methods take into account only the nucleotide context in the vicinity of the TIS. These include, for example, positional conditional probability matrix (Salzberg, 1997) and generalized second-order profiles (Agarwal and Bafna, 1998a). According to an evaluation of these methods, there were no significant differences between the performance of these methods and a weight matrix. However, these approaches cannot be used separately because of their inherent high rate of false positives. There is a biological explanation for the high rate of false positives that is observed. According to the ribosome scanning model, the start of translation can only be influenced by the ATG codons that are positioned on the 5' UTR. The existence of ATGs in functional context downstream of the first ATG in the coding region will not affect the protein translation. Therefore we can assume that ATGs in good functional context exist downstream of the first ATG in the coding region without any effect on the translation of the protein. From the point of

software development this means that a TIS software prediction program should include more information than the local context around an ATG.

In Pedersen and Nielsen (1997) there is a description involving the analysis of a larger region—100 bases before and 100 bases after a putative start codon. Predictions are made using an Artificial Neural Network (ANN), which recognizes the local context around the TIS, as well as the statistical properties before and after the TIS. In Salamov *et al.* (1998), a method is presented in which as many as six characteristics are applied to analyze the area around putative TIS. Linear discriminate analysis is used for the final scoring. According to their investigation, the most important components for the correct prediction are the positional triplet weight matrix around ATG and the hexanucleotide difference before and after the ATG in a 50 nucleotide long window. In Agarwal and Bafna (1998b) an algorithmic idea of the ribosome scanning model is implemented. The search starts from the 5' end of the mRNA. A putative start codon is defined as any ATG followed by an ORF longer than 200 nucleotides. The procedure stops once a putative ATG in a good local binding motif is reached.

More recently, a new method for TIS prediction based on support vector machines has been introduced (Zien *et al.*, 2000).

The methods mentioned above are flexible and in part

can also be used on Expressed Sequenced Tag (EST) data. However, according to their published results, none of these methods succeed in giving results better than 85% on the test set.

According to a simple statistic on the set of genes used here (475), the first ATG in the frame of the coding ORF results in correct prediction in 90% of the cases.

The aim of this work was to develop a program that performs better than existing methods. This was achieved by using two modules, one sensitive to the conserved motif and one sensitive to the coding/non-coding potential around the start codon (this idea was first used for splice site prediction in Brunak *et al.* (1991)). Both modules are based on ANN. By applying the simplified method of the ribosome scanning model, the algorithm starts a linear search at the beginning of the coding ORF and stops once the combination of the two modules gives a positive score.

2 METHODS

2.1 Construction of the dataset

The verification of a TIS experimentally is quite a time consuming wetlab experiment and as a result of this most of the annotated TIS in the databases are not experimentally verified.

In order to obtain a validated dataset, the first step of data collection was made on the protein database Swissprot, rather than on the genomic databases. All the human proteins whose N-terminal sites are sequenced at the amino acid level were collected and manually checked (by Amos Bairoch, personal communication). The next step was to retrieve the full-length mRNAs for these proteins whose TIS had been indirectly experimentally verified. 475 corresponding human cDNAs, completely sequenced and annotated, were found. Out of these 475 cDNAs, three-quarters were used for the extraction of the training data, here called the training gene pool and one quarter was used for the extraction of the test data, here called the test gene pool. For the full design of the modular structure of the algorithm, several datasets were constructed. These datasets were used for:

- the training of the ANNs, called ANN-training set,
- the evaluation of the generalization performance of the ANNs during the training, called ANN-evaluation set, and
- the final testing of the performance of a trained ANN, called ANN-test set.

ANN-training and evaluation sets were extracted from the training gene pool, while the ANN-test set was extracted from the test gene pool. In the last step of the algorithm, different modules were integrated into one approach. Again, parameter estimation was performed on genes

retrieved from the training gene pool, this time on full length sequences retrieved out of the training gene pool. Final testing of the algorithm was performed on whole sequences included in the test gene pool.

2.2 Consensus ANN

For the consensus-ANN a window 12 nucleotides long was used. This sequence includes the positions from -7 to $+5$, where $+1$ is the position of the first nucleotide of an ATG triplet (see Figure 2). Every cDNA sequence provides only one positive data example for a TIS. Consequently, only a relatively small amount of data was used to train this ANN (325 positive and 325 negative examples). The input is presented to the network through the universal encoding system, where each nucleotide is transformed into a binary 4-digit string (Figure 3). A number of different feedforward ANN architectures were tested during the training procedure:

- feed forward nets without hidden units (perceptron);
- feed forward nets with hidden units; and
- feed forward nets with hidden units and short cut connections (direct connections from the input to the output units).

In the second and third case, a number of different hidden units were tested (results not shown).

The best performance was achieved by an ANN with short cut connections and two hidden units that was trained with the cascade correlation algorithm (Fahlman and Lebiere, 1990). In cascade correlation, training starts with a perceptron, which is an ANN with weights only between the input and the output units. After some iterations, part of the weights freeze (stop changing) and a hidden unit is added. In the remaining iteration, the new weights are trained to learn examples which had not been successfully learned in the first steps of the training.

2.3 Coding—ANN

In the second step, an ANN is trained for the recognition of the coding region. In this case the windows of the sequences that are used are 54 nucleotides in length. For the training of this second ANN it is possible to extract more positive data from every gene. Here, a possible homology between training and test data could influence the result. For this reason, such homologies between the training and test genes were eliminated through pair-wise alignment with the full Smith–Waterman algorithm. Only genes from the training pool with less than 70% homology to the genes of the test pool were used for extracting the training data—a total of 282 genes. From these genes, 700 positive and 700 negative sequence regions were extracted. An additional 500 regions (half positive–half

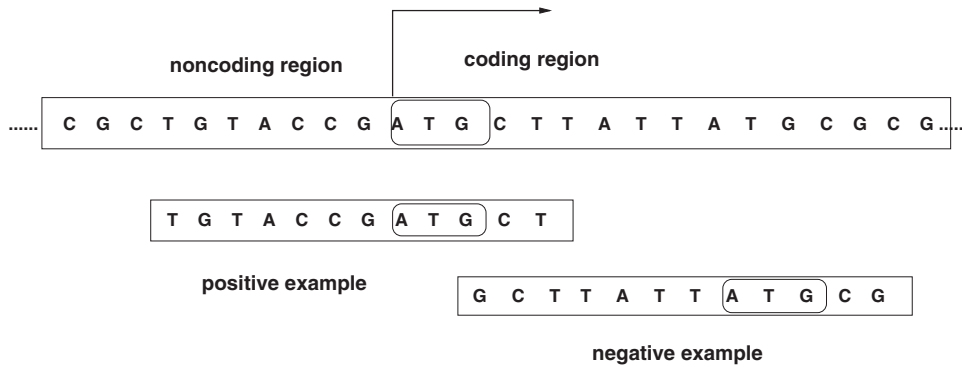


Fig. 2. Construction of the positive and negative pattern using 12 bases long windows with an ATG starting at the 8th position.

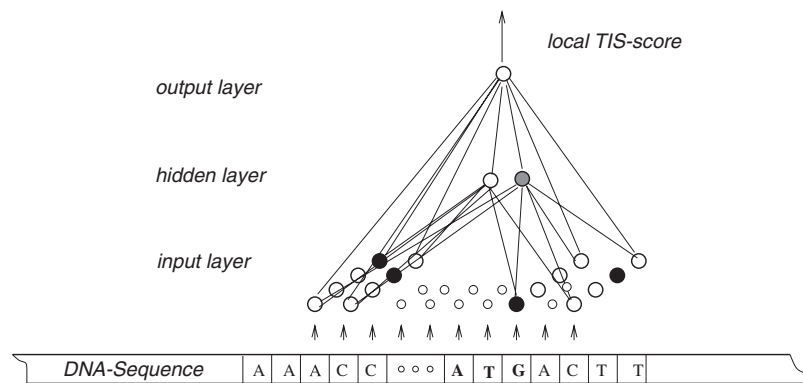


Fig. 3. The architecture of the module for the recognition of the consensus motif around TIS.

negative) were extracted from the test gene pool for testing the performance of the coding ANN.

Previous investigation has shown that preprocessing the data through a coding measure can significantly improve the performance of the ANN (Hatzigeorgiou *et al.*, 1999). Through a variety of existing coding measure methods the best results were obtained by applying the codon usage statistic to the sequence window. The counting starts with the first nucleotide of the window, and counts all non-overlapping codons.

This leads to a transformation of the sequence window to a vector of 64 units. Every unit gives the normalized frequency of the corresponding codon appearing in the window.

If the window starts with the first nucleotide of a codon then the window is in coding-frame and the ANN has a high score (close to 1), otherwise the score is low (close to 0).

The training of the ANN is done with the algorithm Resilient Backpropagation (RPROP; Riedmiller and Braun, 1993, an improved version of the *classical* Backpropagation; Riedmiller and Braun, 1993), applied to a feedforward ANN.

The main differences between RPROP and Backpropagation are that:

- the change of the weights depends only on the sign of the potential derivation of the error and not on its size;
- the weight update phase incorporates the current gradient and the gradient of the previous step; and
- every weight has its own learning parameter for the changing of the weight value.

The experiments show that RPROP yields better results than Backpropagation (Hatzigeorgiou and Reczko, 1999). Also it was experimentally verified that a neural network with 2 hidden units gives the best generalization results.

2.4 Training procedure and generalization performance of the ANNs

The performance of TIS prediction on the test set has an accuracy of 76.4%, where the accuracy is taken to be the average of the prediction on the positive and the negative data.

One of the critical issues in the training of an ANN is determining the appropriate moment to stop training.

An extended training can decrease the global error of the training set, but on the other hand may also lead to over-training the ANN. In the latter case, the ANN learns the characteristic of individual examples and not their global characteristics.

In order to avoid this, only 2/3 of the examples of the training set are used for training; the remaining 1/3 are used for evaluating the performance of the ANN following every iteration. Once the performance of the *evaluation* group of examples starts to decrease, the ANN is stopped. As mentioned in the dataset section, the extraction of the evaluation examples is produced from genes belonging to the training pool rather than from the test pool.

The prediction accuracy of the consensus ANN on the examples of the test set was 76.4%, while the prediction of the coding ANN on the examples of the test set was 82.5%. The resulting accuracy is taken to be an average of the correct prediction rate on the positive (true TIS) and negative examples.

All the training of the ANNs was performed by the Stuttgarter Neural Network Simulator (SNNS), publicly available from the University of Stuttgart, Germany (Zell *et al.*, 1993).

2.5 Integrated method

The final algorithm is designed for analysis of full-length mRNA sequences.

In the first step of the analysis, a coding score was calculated for every nucleotide of the mRNA sequence. To do this, a window of 54 nucleotides in length slides along the sequence; for every window, the codon usage statistics were calculated and then used as input to the coding ANN. The ANN output ranges between 0 and 1. In the coding regions a high score is expected for every third position relative to every third window starting in frame.

In the second stage of analysis, the coding evidence of the putative coding region included in the longest ORF of the sequence was calculated. The putative coding region starts with the first in-frame ATG and ends with the first in-frame stop codon. For the calculation of the coding potential, the average coding score of every third position starting from the first nucleotide (the A of the ATG) was calculated.

If the coding score is very low, the user can analyze another ORF within the sequence. In all of the examples used here, this occurred in only two genes with very short coding regions (smaller than 200 nucleotides).

In the third step, for every in-frame ATG a consensus score was calculated. To do this, a window of 12 nucleotides was extracted (as described for the training dataset) and used as input to the consensus ANN. The output of the consensus ANN was again a score between 0 and 1.

In the fourth step, for the same in-frame ATG, a coding

difference score was calculated by building the difference between all coding scores (calculated by the coding ANN) of in-frame 60 positions before the ATG and all coding scores of in-frame 60 positions after it.

The final score for every putative TIS was obtained by combining the output of the consensus ANN and the coding difference. Among many potential ranking strategies, a simple multiplication of the two scores was chosen. The first suitable ATG starting from the 5' end of the investigated ORF with a score greater than 0.2 was examined as the correct TIS. This method provides only one prediction for every ORF. According to the results of the test group, 94% of the TIS were correctly predicted; 6% of the predictions were false positive. In a simple test where the first ATG from the 5'-prime was chosen as the startcodon, without taking into account the coding scores, the prediction was 92%. Figure 4 gives the prediction of the two modules along the first part of a cDNA sequence.

For a correct annotation, a prediction with 100% confidence is required. This is possible using a new generation of algorithms, the *Las Vegas algorithms* (Brassard and Bratley, 1996). Such algorithms provide a correct prediction in some cases and have a *no answer* option in the remaining cases. The term *Las Vegas* was introduced to distinguish algorithms that reply correctly when they reply at all from *Monte Carlo* algorithms, which occasionally make mistakes (Sze and Pevzner, 1997).

A simple statistic on the training set shows that all TIS starts at one of the three first ATGs of the coding ORF. If S is defined to be the set of possible solutions for the TIS prediction, then S is defined through these ATGs. For a given parameter q , the set of Competing Solutions $CS(q)$ is defined to be all solutions in S with a score higher than q . Intuitively, a parameter q needs to be found such that $CS(q)$ is a non-empty set and the true TIS is guaranteed to be in the $CS(q)$. If the simple condition that $CS(q)$ contains only one solution is added, a 100% accurate prediction is achieved. In this case q is defined to be 0.2, and yields only one solution (the correct one) in 60% of cases for CS. This means that for the other 40% either no or more than one answer is given.

In other words: 'if only one out of the three suitable ATGs has a high score and only one correct candidate is expected then a 100% correct prediction is guaranteed.'

3 RESULTS AND DISCUSSION

In general, the rate of 94% correct TIS prediction compares favorably with the 85% success rate reported in Agarwal and Bafna (1998b), a method which also allows only one prediction per gene. For the comparison we should mention that the training data of the above mentioned method was the same as the training and test data reported here. The prediction reported in Salamov *et al.* (1998) is 79% and that of Pedersen and Nielsen (1997)

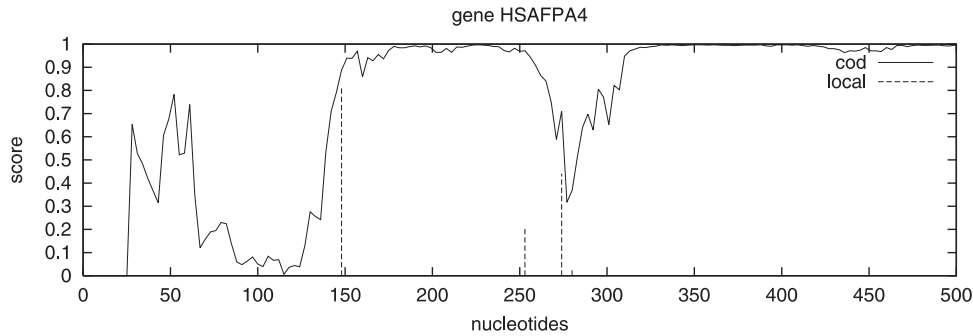


Fig. 4. The prediction of the two ANNs on a cDNA sequence (part of the gene). The cod-line gives the score of the coding ANN for the coding frame. The local-line gives the position and the score of the consensus ANN for all ATGs in coding frame. The correct TIS is in position 148.

78%. However, it should be taken into account that these last two methods allow more than one prediction per gene. In addition, it should be mentioned that these numbers only give indications about the different performances; they should not be directly compared, since the results come from different datasets.

Some analytical results of the method described here (named DNA Intelligent ANALYSIS—DIANA for TIS) are documented in Table 1. In addition, the prediction for the same genes by ATGPred (Salamov *et al.*, 1998) and NetStart (Pedersen and Nielsen, 1997) is given. The comparison could only be made with programs available over the internet. The examples given in this table are all full-length mRNAs with TIS at the second or third in-frame ATG. For ATGPred and NetStart the three TIS predictions with the highest scores are documented. For DIANA-TIS, the scores for the first three suitable ATGs are given.

Documented in Table 2 are the predictions of the nucleotide sequence of four proteins. These proteins are known to have an unusually long signal peptide with suitable ATGs before their cleavage sites (Pedersen and Nielsen, 1997). After applying their method, the authors suggest that there could be an incorrect annotation of the start site of these proteins in Swissprot and proposed the entry MCP3_HUMAN in Swissprot to be replaced by the entry MCPT_HUMAN with the start codon in position 329.

The application of the method DIANA-TIS on the same proteins yields the same conclusion. Both programs NetStart and DIANA-TIS are based on ANN. The main difference between the two programs is that DIANA-TIS uses two ANNs (one sensitive to the consensus sequence and one sensitive to the coding potential), whereas NetStart uses only one (with a 200 nucleotide window). This difference allows DIANA-TIS to distinguish better than NetStart between the TIS and other ATGs. For example,

Table 1. Performance of three programs for TIS prediction along mRNAs

Gene–protein	TIS	ATGPred	NetStart	DIANA-TIS
HSCLMF35 P29459		68 0.31	68 0.52	68 0.43 –0.21
	170	269 0.23	170 <u>0.82</u>	170 0.68 <u>0.55</u>
		320 0.22	320 0.72	269 0.25 –0.05
HSFCERI P12319		68 0.31	68 0.52	68 0.07 0.04
	107	98 0.65	98 0.82	98 0.43 0.24
		107 <u>0.20</u>	107 <u>0.72</u>	107 0.65 <u>0.27</u>
HSG6PDR P11413		117 0.61	255 0.62	381 0.11 0.01
	471	193 0.56	381 0.69	471 0.93 <u>0.75</u>
		381 0.90	471 <u>0.67</u>	579 0.93 0.02
HSINSR P06213		79 0.69	79 0.73	79 0.48 0.07
	139	139 <u>0.80</u>	139 <u>0.80</u>	139 0.95 <u>0.69</u>
		385 0.42	250 0.80	250 0.77 –0.02

'Gene–protein' column contains the names and protein accession numbers of the analyzed genes, 'TIS' column—positions of correct TIS as annotated in the database. Within 'ATGPred' and 'NetStart' first column gives predicted position of the TIS and the second column the final score. For 'DIANA-TIS' first column gives predicted position of the TIS, second-TIS motif score in this position and third- final score. The correct predictions are underlined.

in the prediction of the HSBPIAA genes two suitable ATGs are located only 12 nucleotides away. For this reason, the coding/non-coding information will not change significantly between the two ATGs but the consensus motif is completely different (0.10 versus 0.82), thus leading to a significantly different score.

A favorable prediction does not work for all examples. In INIP_HUMAN, the TIS motif score for the next ATG is also high, although the combined score is much lower. This can be explained by the observation that in some signal peptides sequences, the coding potential score is relatively low, and can thus affect the combined score.

The method described here is designed to work for full-

Table 2. Performance of the programs for TIS prediction along the mRNA with signal peptide sequences

Gene-protein	TIS	SP	ATGPred	NetStart	DIANA-TIS
HSANG	40	33	40	0.26	40 0.40 40 0.05 0.02
ANGT_HUMAN	67	24	-	-	67 0.70 67 0.22 0.11
HSBPIAA	31	31	31	0.41	31 0.61 31 0.10 0.05
BPLHUMAN	43	27	-	-	43 0.72 43 0.82 0.51
HSIIP	41	37	41	0.65	41 0.75 41 0.14 0.02
INIP_HUMAN	74	26	74	0.39	74 0.77 74 0.52 0.06
HSMCP3A	299	33	299	0.32	299 0.51 299 0.01 0.01
MCP3_HUMAN	329	23	329	0.27	329 0.80 329 0.83 0.48

All designations as in Table 1, 'SP'—length of signal peptide.

length (or almost) mRNAs. For ESTs some adjustments are needed. Since ESTs are characterized as partial cDNA sequences and frequently have sequencing errors, the coding ORF information can not be used. Instead, it is possible to make a coding prediction tolerant of sequencing errors. This prediction is then combined with the consensus ANN score (Hatzigeorgiou and Reczko, 1999).

Reliable annotation methods based on similarity seem to be the most promising solution for defining gene structure. However, several examples show that even very closely related proteins can have different start sites which are not revealed by a simple comparison (Stroemstedt *et al.*, 1996; Croop *et al.*, 1997). Therefore, the author regards the TIS prediction to be one of the problems still to be solved, basically via statistical methods. Recently a further development of the ATGpred program was published (Nishikawa *et al.*, 2000) that combines statistical information and similarity with protein sequences. The program described in this paper can be very useful as a further step in improvement of sensitivity and reliability of the TIS prediction programs.

The problem of gene identification is one of the main tasks of bioinformatics. There has been a great deal of progress in gene identification methods in the last few years. The older coding region identification methods have given way to methods that can suggest the overall structure of genes.

Mapping sequences from cDNAs is still the most direct way to characterize the coding parts and provide reliable information for the structural annotation of genes in genomic sequences. This task can not be achieved without proper annotation of cDNAs. A successful method for annotation of cDNA sequences has been demonstrated in this paper.

After the large sequencing projects, such as the Human Genome Project, are completed the much larger part of the

analysis of these sequences starts and the computer simulated prediction systems play a major role. Algorithms using sophisticated ANN, fuzzy logic, integrated methods and hybrid systems are in great demand.

ACKNOWLEDGEMENTS

I am grateful to Martin Reczko, who participated in the early stages of this project. I also wish to thank Amos Bairoch for providing the dataset, James W. Fickett for helpful comments on the TIS problem, and Slava Bolshakov for comments on the manuscript. This work was supported by the Greek Secretary of Research & Development.

REFERENCES

- Agarwal, P. and Bafna, V. (1998a) Detecting nonjoining correlations within signals DNA. In *Proceedings of the 2nd Annual International Conference on Computational Molecular Biology, RECOMB98*. ACM Press, pp. 1–7.
- Agarwal, P. and Bafna, V. (1998b) Translation initiation: implications for gene prediction and full-length cDNA. In *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology, ISMB98*. AAI Press, pp. 2–7.
- Brassard, G. and Bratley, P. (1996) *Fundamentals of Algorithmics*. Prentice-Hall.
- Brunak, S., Engelbrecht, J. and Knudsen, S. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, **220**, 2–17.
- Consortium, I.H.G.S. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Croop, J.M., Teller, G.E., Fletcher, J.A., Lux, M.L., Raab, E., Golden-son, D., Son, D., Arciniegas, S. and Wu, R.L. (1997) Isolation and characterization of a mammalian homolog of the *Drosophila* white gene. *Gene*, **185**, 77–85.
- Fahlman, S.E. and Lebiere, C. (1990) The cascade-correlation learning architecture. In Touretzky, D. (ed.), *Advances in Neural Information Processing Systems II, NIPS*. Morgan Kaufmann, pp. 524–532.
- Hatzigeorgiou, A. and Reczko, M. (1999) Feature recognition on expressed sequence tags in human DNA. In *Proceedings of the International Joint Conference on Neural Networks*. CD, INNS Press.
- Hatzigeorgiou, A.G., Papanikolaou, H. and Reczko, M. (1999) Finding the reading frame in protein coding regions on DNA sequences: a combination of statistical and neural network methods. In Mohammadian, M. (ed.), *Computational Intelligence: Neural Networks & Advanced Control Strategies*. pp. 148–158.
- Kozak, M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.*, **15**, 8125–8148.
- Kozak, M. (1996) Interpreting cDNA sequences: some insights from studies on translation. *Mammalian Genome*, **7**, 563–574.
- Nishikawa, T., Ota, T. and Isogai, T. (2000) Prediction whether a human cDNA sequence contains initiation codon by combining statistical information and similarity with protein sequences. *Bioinformatics*, **16**, 960–967.
- Pain, V.M. (1996) Initiation of proteins synthesis in eukaryotic cells. *Eur. J. Biochem.*, **236**, 747–771.

- 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 System for Molecular Biology, ISMB97*. AAAI Press, pp. 226–233.
- Riedmiller,M. and Braun,H. (1993) A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In Ruspini,H. (ed.), *Proceedings of the IEEE International Conference on Neural Networks, (ICNN93)*. IEEE, San Francisco, pp. 586–591.
- Salamov,A.A., Nishikawa,T. and Swindells,M.B. (1998) Assessing protein coding region integrity in cDNA sequencing projects. *Bioinformatics*, **14**, 384–390.
- Salzberg,S.L. (1997) A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput. Appl. Biosci.*, **13**, 365–376.
- Stroemstedt,M., Rozman,D. and Waterman,M. (1996) The ubiquitously expressed human CYP51 encodes lanosterol 14 alpha-demethylase, a cytochrome P450 whose expression is regulated by oxysterols. *Arch. Biochem. Biophys.*, **329**, 73–78.
- Sze,S. and Pevzner,P.A. (1997) Las Vegas algorithms for gene recognition: suboptimal and error tolerant spliced alignment. *J. Comput. Biol.*, **4**, 297–320.
- Zell,A., Mache,N., Hübner,R., Mamier,G., Vogt,M., Herrmann,K.U., Schmalzl,M., Sommer,T., Hatzigeorgiou,A., Döring,S., Posselt,D., Reczko,M. and Riedmiller,M. (1993) SNNS user manual, version 3.0. *Technical Report*. Universität Stuttgart, Fakultät Informatik.
- Zien,A., Raetsch,G., Mika,S., Schoelkopf,B., Lemmen,C., Smola,A., Lengauer,T. and Mueller,K.-R. (2000) Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, **16**, 799–807.