Neural Network Prediction of Translation Initiation Sites in Eukaryotes: Perspectives for EST and Genome analysis.

Anders Gorm Pedersen* **and** **Henrik Nielsen**†

Center for Biological Sequence Analysis
The Technical University of Denmark, Building 206
DK-2800 Lyngby, Denmark

Abstract

Translation in eukaryotes does not always start at the first AUG in an mRNA, implying that context information also plays a role. This makes prediction of translation initiation sites a non-trivial task, especially when analysing EST and genome data where the entire mature mRNA sequence is not known. In this paper, we employ artificial neural networks to predict which AUG triplet in an mRNA sequence is the start codon. The trained networks correctly classified 88 % of Arabidopsis and 85 % of vertebrate AUG triplets. We find that our trained neural networks use a combination of local start codon context and global sequence information. Furthermore, analysis of false predictions shows that AUGs in frame with the actual start codon are more frequently selected than out-of-frame AUGs, suggesting that our networks use reading frame detection. A number of conflicts between neural network predictions and database annotations are analysed in detail, leading to identification of possible database errors.

keywords: translation initiation, start codon, kozak box, neural networks, signal peptides

Introduction

The choice of start codon in eukaryotes depends on position as well as on context. Usually, translational initiation takes place at the first occurrence of the triplet AUG in an mRNA, but in some cases an AUG further downstream is selected. This is explained by the so-called scanning hypothesis, which states that the small subunit of the ribosome binds at the capped 5'-end of the mRNA and subsequently scans the sequence until the first start codon in a suitable context is found (Kozak 1983; 1984; Cigan & Donahue 1987; Joshi 1987; Kozak 1989). It has been reported that downstream AUGs are used as start codons in less than 10 % of investigated eukaryotic mRNAs (Kozak 1989; Yoon & Donahue 1992). Previous analyses of start codon contexts found the consensus of eukaryotic translation initiation sites to be GCCACCaugG (Kozak 1984; 1987), but further analyses has demonstrated that the pattern varies between different groups of eukaryotes (Cavener 1987; Lütcke *et al.* 1987; Joshi 1987; Cigan & Donahue 1987; Yamauchi 1991; Cavener & Ray 1991) and that these differences are statistically significant (Pedersen & Nielsen 1997). Specifically, all vertebrates that have been investigated have similar start codon contexts, as do the two monocots rice and corn, while several other eukaryotic species have significantly different signals (Pedersen & Nielsen 1997).

Since less than 10 % of all eukaryotic mRNAs reportedly utilize downstream AUGs as start codons, it should be possible to perform prediction of translation initiation sites at more than 90 % accuracy simply by selecting the first AUG, given that complete and error-free mRNA sequences are available. This, however, is very rarely the case in sequence analysis. Thus, we find that even when great care is taken to extract GenBank nucleotide data that is annotated as being equivalent to mature mRNA, almost 40 % of the sequences contain upstream AUGs. This problem is enhanced when using unannotated genome data, and when analysing expressed sequence tags (ESTs). ESTs are partial, single-pass, cDNA sequences, that generally represent the complement of mRNAs in the cell, but that due to the very nature of the technology usually contain more errors (Boguski, Lowe, & Tolstoshev 1993; Boguski & Tolstoshev 1994; Cooke *et al.* 1996; Benson *et al.* 1997). Thus, uncertainties can exist regarding which end of an mRNA the EST corresponds to, it is not always known whether the entire 5' (or 3') end is represented in the EST, the sequence can potentially be contaminated with vector sequence, and the automated single pass sequencing results in a higher error rate than is found in normal genome data.

These problems make the prediction of translation initiation sites a non-trivial task. In this paper we

* Phone: (+45) 45 25 24 84; Fax: (+45) 45 93 48 08, email: gorm@cbs.dtu.dk

† Phone: (+45) 45 25 24 70, email: hnielsen@cbs.dtu.dk

present a method for prediction of start codons, that is based on the use of artificial neural networks. The results presented here are preliminary and we are still in the process of developing the method, but we find the current performance to be convincing. The method does not require any knowledge of the position of AUGs in relation to mRNA 5' ends, and we believe it can be useful in connection with analysis of EST data and incompletely annotated genome sequences.

Methods

Data

Extraction All data were extracted from GenBank, release 95 (Benson *et al.* 1997). We extracted a vertebrate group consisting of sequences from *Bos taurus* (cow), *Gallus gallus* (chicken), *Homo sapiens* (man), *Mus musculus* (mouse), *Oryctolagus cuniculus* (rabbit), *Rattus norvegicus* (rat), *Sus scrofa* (pig), and *Xenopus laevis* (African clawed frog). We have previously shown that these vertebrates have similar start codon contexts (Pedersen & Nielsen 1997). Additionally, we have chosen a data set showing large deviation from vertebrates, *Arabidopsis thaliana* (thale cress, a dicot plant).

Nuclear genes with an annotated start codon were selected. The sequences were processed in the following way: all sequences were "spliced" by removing possible introns, and joining the remaining exon parts. From the resulting data set, sequences containing at least 10 nucleotides upstream of the initiation point and at least 150 nucleotides downstream (relative to the A in AUG) were selected. All sequences containing non-nucleotide symbols in the interval mentioned above (typically due to incomplete sequencing) were excluded.

Redundancy All sequence databases are redundant due to the presence of genes belonging to gene families, homologous genes from different organisms, and sequences submitted to the database more than once. Unless this redundancy is reduced before performing statistical analysis, the result will be biased for the over-represented sequences, and the performance of prediction methods will be overestimated (Sander & Schneider 1991; Hobohm *et al.* 1992). We performed very thorough reduction of redundancy using algorithm 2 from (Hobohm *et al.* 1992) and a novel method for finding a similarity cut-off, that we have described elsewhere (Pedersen & Nielsen 1997). Briefly, this method is based on performing all pairwise alignments for a data set, fitting the resulting Smith-Waterman scores to an extreme value distribution (Altschul *et al.* 1994), and choosing a value above which there are more observations than expected from the distribution.

The sizes of the redundancy reduced data sets were: 3312 vertebrate sequences, and 523 *Arabidopsis thaliana* sequences. These data sets are available from the authors upon request.

Neural Networks

The neural networks used in this study were of the feed-forward type, and had three layers of neurons (Hertz, Krogh, & Palmer 1991). They were written in the FORTRAN programming language by Søren Brunak, and has previously been used for several other prediction purposes [*e.g.*, (Brunak, Engelbrecht, & Knudsen 1990; 1991; Hansen *et al.* 1995)]. Inputs were presented to the networks by encoding the DNA sequence into a binary string, using a coding scheme where each nucleotide is represented by 4 binary digits: A=0001, C=0010, G=0100, and T=1000 (sparse encoding). The output layer consisted of two neurons — one predicting whether the central position in the window was the A in a start codon AUG, the other predicting whether it was the A in a non-start codon AUG. The output of the network was interpreted by believing the output neuron with the highest score (the "winner takes all" approach). Neural network performance was estimated using the Mathews correlation coefficient (Mathews 1975).

Prediction of Signal Peptides

In order to test our method for prediction of start codons, we have combined it with a method for prediction of signal peptides in amino acid sequences: The SignalP server (Nielsen *et al.* 1997). This method uses a combination of neural networks to predict the presence of signal peptides and the location of their cleavage sites.

SignalP returns three scores from every position in the sequence: a cleavage site score (C-score) from networks trained to recognise cleavage sites, a signal peptide score (S-score) from networks trained to distinguish between signal peptide and non-signal peptide positions, and a combined cleavage site score (Y-score), which optimises the prediction of cleavage site location by combining the C-score with the derivative of the S-score. Discrimination between signal peptides and N-terminals of non-secretory proteins is performed using the maximal value of one of the three scores or the mean value of the S-score (from the N-terminus to the position with maximal Y-score). Each network ensemble has a specific threshold value for each of these measures.

Results and Discussion

As mentioned, it should be possible to predict translation initiation start sites at better than 90 % accuracy, if one has access to entire error-free mRNA sequences. However, when we analysed our data sets with the purpose of extracting sequences corresponding to mature mRNAs, we found that only about 10 % (387 out of 3312) of the sequences in the vertebrate set had sufficient annotation for this purpose. (In the remaining cases the exact *in vivo* transcriptional startpoints and upstream splice sites have not been determined). Further analysis of the resulting vertebrate mRNAs demonstrated that almost 40 % (150 out of 387) contained one or more upstream AUGs. Thus, it was only possible to use the simple "first-AUG" rule in the remaining 237 sequences, corresponding to approximately 7 % of the entire vertebrate set. In the case of genome or EST data the situation is going to be even worse.

Hence, we thought it could be interesting to construct a method for prediction of translation initiation sites that was not dependent on knowledge of the position of an AUG in relation to mRNA 5' ends. To this end, we trained artificial neural networks on the entire non-redundant data sets. Specifically, we trained on 80 % of a sequence set and tested the predictive performance on the remaining 20 %. Only AUGs were examined and predicted to be either start AUGs or non-start AUGs. The vertebrate data set contained 2684 AUGs of which 660 (25 %) are start codons. The *A. thaliana* set consisted of 412 AUGs, 105 (25%) of which are start codons. Several different network architectures were investigated in order to find one that performed well on the problem at hand. Specifically, we examined all combinations of the following ranges of parameter values:

- Number of neurons in the hidden layer: 0, 1, 2, 5, 10, 20, 30, 50
- Input window size (nucleotides): 13, 33, 53, 73, 93, 113, 133, 153, 173, 203

Among these architectures the optimal performance was observed with a 203 nucleotide input window and 30 hidden units. Generally, we found that performance varied only slightly with different numbers of hidden units, while the size of the input window was very important. Thus, the predictive performance was found to get significantly better with bigger input windows (data not shown), suggesting that the network is using global information. This could be either the nucleotide composition, which is known to be different between coding and non-coding regions, or reading frame characteristics such as the presence of stop codons, and nucleotide frequencies at the three codon positions.

The best networks that were obtained on the vertebrate set showed a Mathews correlation coefficient of 0.6208 and were able to predict 85 % of the AUGs correctly (corresponding to 78 % of start codons, and 87 % of non-start AUGs). The best *Arabidopsis thaliana* network reached a Mathews correlation coefficient of 0.7122, meaning that they were able to predict 88 % of all AUGs correctly (90 % of start codons and 87 % of non-start AUGs). For the purpose of analysing the predictive performance of these preliminary networks we used only one partition of the data into training and test sets. It is, however, important to note that the very low redundancy of our data sets, means that the predictive performance we observe is unlikely to be overestimated.

In order to investigate whether the networks utilized local sequence information (such as the start codon context signals) we performed the following experiment: neural networks were presented with input windows which covered the aforementioned 203 nucleotides, except for one position (a "hole") from which the input was disregarded. In a series of runs, the position of the hole was shifted along the input window, and the maximum correlation coefficient-values (C_{max}) for the test set were determined. In this way it should be possible to detect areas of the input window that are particularly important by looking for positions of the hole that causes the prediction ability of the network to be partly destroyed. Specifically, these positions can be seen as local minima in the plot of C_{max} *vs.* position of the hole (figure 1). This method has previously been used to analyse *Escherichia coli* promoters with regard to localisation of regions important for transcriptional initiation (Pedersen & Engelbrecht 1995). As it can be seen, there is a clear effect on the prediction ability of the network when the hole covers positions close to the start codon, indicating that local information is indeed also important for prediction of translational initiation sites. Furthermore it can be seen that by far the most important position is 3 nucleotides upstream of the start codon. This is in accordance with previous sequence compilations and systematic mutagenesis showing that the presence of a purine (A or G) at position −3 is very important for efficient initiation of translation (Kozak 1986; 1987).

We further analysed the erroneously predicted positions in order to see whether the prediction was dependent on the reading frame and the position relative to start codon. The error rate (percent incorrectly predicted AUGs of each class) are shown in table 1, together with the average neural network output score for the AUGs in each class. For all the non-start AUGs. these two measures are correlated; when the average score is closer to that of the start codons, the possib-

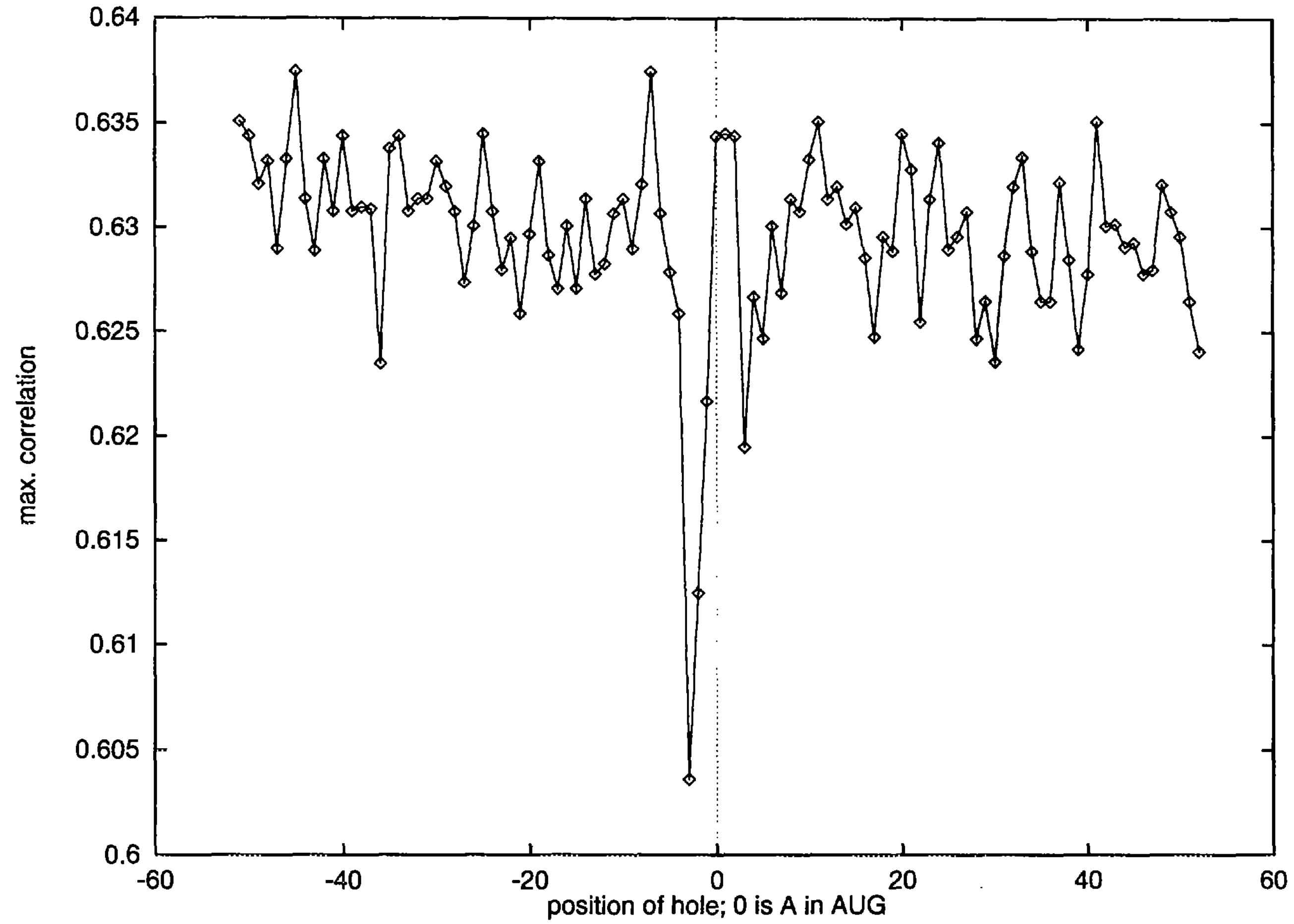

Figure 1: Scanning the input window of the neural network with a one-nucleotide hole. The maximum correlation coefficient C_{max} is shown as a function of the position of the hole (translation initiation site = 0). A network with a window size of 203 and 30 hidden units was used. Notice the local minimum around position −3 which indicates that local information is important for the prediction performance of the network, and that this nucleotide is one of the most important.

Category	Vertebrates (total: 2684)			*Arabidopsis* (total: 412)		
	number	error rate	score mean	number	error rate	score mean
Start codons	660	22%	0.609	105	10%	0.703
Upstream, in frame	105	34%	0.426	14	50%	0.510
Upstream, out of frame	301	15%	0.350	28	18%	0.340
Downstream, in frame	612	20%	0.263	87	24%	0.294
Downstream, out of frame	1006	5%	0.169	178	4%	0.157

Table 1: The error rates and neural network scores for five classes of AUG codons: start codons (correct initiation sites), and 4 distinct classes of non-start AUGs. The non-start AUGs are divided into 4 classes based on whether they are upstream (5') or downstream (3') of the start codon and whether they are in or out of the reading frame defined by the start codon. The total number in each set and each class is also shown.

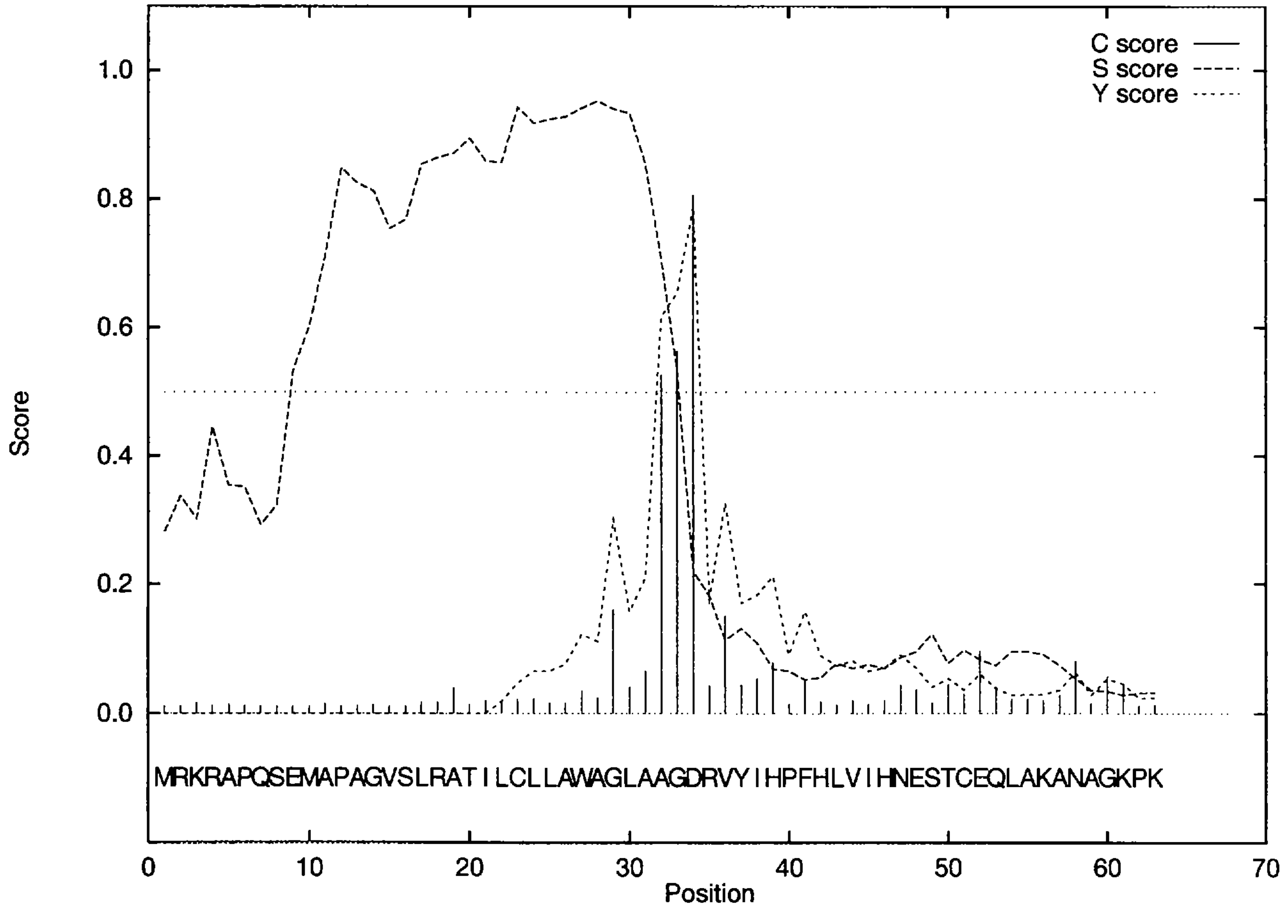

Figure 2: SignalP prediction of the sequence of **ANGT_HUMAN**, Human angiotensinogen. The S score (signal peptide score) has a high value for residues within signal peptides, while the C and Y scores (cleavage site scores) are high at position +1 immediately after possible cleavage sites. Note that the S score is comparatively low for the region between the first Met and the second Met.

lity of errors increases.

It is clearly seen that the network is far more likely to predict AUGs as start codons if they are in the correct reading frame, both upstream and downstream of the start codon. This suggests that the trained network is able to recognize reading frames, and is in accordance with the observation that global information is used. Furthermore, errors are more frequent upstream than downstream. This is somewhat surprising in light of the scanning hypothesis: if the first good translation initiation site from the 5' end of the mRNA is used, there should be no selective pressure against downstream AUGs being similar to initiation sites. However, this consideration applies to the local information only; the reading frame detection, on the other hand, is likely to be easier when a larger part of the input window consists of coding sequence.

In order to further investigate the networks ability to recognize start codons, we set out to investigate a set of atypical signal peptides from the SWISS-PROT database, version 29 (Bairoch & Boeckmann 1994): the average length of eukaryotic signal peptides is 22.6 residues, and only approximately 6% are longer than 30 residues. We wanted to investigate the longer signal peptides in order to see whether any of them could be explained by mis-assigned start codons. Hence, we selected the 32 vertebrate signal peptides longer than 30 residues that had at least one Met between the initiator Met and the proposed cleavage site. We then applied SignalP to these sequences using both the upstream annotated and the alternative downstream methionines. In 21 cases, a downstream Met gave a better average S score than the one indicated in the SWISS-PROT annotation. The S score was typically low around the upstream start codon, and changed to a high value around the alternative start codon (figure 2). We then extracted the corresponding GenBank nucleotide sequences for these 21 signal peptides and

			Annotated / Predicted		initiation	
Protein	SWISSPROT	GenBank	SP	mean S	start	init.
name	ID	Locus	length	score	codon	score
Human angiotensinogen	ANGT_HUMAN	HUMANG	33	0.707	40	0.400
			24	0.909	*67	0.700
Human acid sphingomyelinase	ASM_HUMAN	HUMASM	46	0.483	101	0.578
			[a]14	0.956	197	0.671
Bactericidal permeability-increasing protein	BPI_HUMAN	HUMBPIAA	31	0.738	31	0.613
			27	0.852	43	0.725
Human C4b-binding protein, α chain	C4BP_HUMAN	HUMPRPC4B	48	0.512	139	0.358
			13	0.897	244	0.755
Mouse C4b-binding protein, α chain	C4BP_MOUSE	MUSBPC4B	56	0.440	203	0.494
			13	0.889	*332	0.748
Trout pro-opiomelanocortin B	COLJ_ONCMY	IMPOMCB	36	0.750	122	0.310
			22	0.843	164	0.561
Bovine adrenal dopamine β-monooxygenase	DOPO_BOVIN	BOVADBM	[b]32	0.624	8	0.738
			[b]19	0.884	47	0.822
Human glucocerebrosidase	GLCM_HUMAN	HUMGCBPRC	39	0.491	123	0.476
			19	0.912	**183	0.477
Human γ-interferon-inducible protein IP-30	INIP_HUMAN	HUMIIP	37	0.817	41	0.757
			26	0.865	74	0.775
Human integrin α-4 subunit	ITA4_HUMAN	HSINTAL4	39	0.290	25	0.410
			13	0.571	103	0.528
Human monocyte chemotactic protein 3	MCP3_HUMAN†	HSMCP3A	33	0.754	299	0.518
			23	0.863	*329	0.803
Mouse meprin A α-subunit	MEPA_MOUSE	MUSMEPRINA	33	0.860	16	0.192
			20	0.906	55	0.605
Human platelet factor 4, variant 1	PLFV_HUMAN	HUMPF4V1A	[c]34	0.767	281	0.657
			[c]18	0.882	329	0.746
Mouse BDNF / NT-3 growth factors receptor	TRKB_MOUSE	MSTRKB	31	0.919	512	0.401
			20	0.973	545	0.490

Table 2: Prediction of alternative start codons in connection with prediction of signal peptides.

[a] When the predicted initiation is used, SignalP predicts a cleavage site two positions downstream of the annoated, resulting in a signal peptide of 16 residues.

[b] SignalP predicts a cleavage site four positions downstream.

[c] SignalP predicts a cleavage site two positions downstream.

* The possibility that the downstream initiation site may be the correct one is acknowledged in database entry remarks.

** Both initiation sites are used.

† In newer versions of SWISS-PROT, MCP3_HUMAN has been replaced by MCPT_HUMAN with the initiation site suggested by our prediction.

tested them with our inititation prediction network. For 14 of the 21 signal peptides, the downstream Met yielded a higher start codon score, indicating that these unusually long signal peptides are likely to be caused by errors in the database asignment of start codon (table 2).

Concluding Remarks

Upstream initiation codons are not infrequent in GenBank sequences. Regardless of whether these exist *in vivo* or only in the database annotations, they comprise a concrete problem facing the sequence analyst. This problem will be encountered more frequently in the future, as EST and unannotated genome data become still more abundant.

In the present work, we have shown preliminary results indicating that neural networks are indeed capable of predicting start codons with some confidence. However, our analysis of the trained networks suggest that the local information present around the cleavage site (Pedersen & Nielsen 1997) is not optimally utilized. A combination of networks trained specifically on local and global information might improve the performance, and we are currently in the process of implementing such a procedure and extending the analysis to other eukaryotic groups.

Acknowledgements

Both authors were supported by a grant from the Danish National Research Foundation.

References

Altschul, S.; Boguski, M. S.; Gish, W.; and Wootton, J. C. 1994. Issues in searching molecular sequence databases. *Nature Genetics* 6:119–129.

Bairoch, A., and Boeckmann, B. 1994. The SWISS-PROT protein sequence data bank: current status. *Nucleic Acids Res.* 22:3578–3580.

Benson, D.; Boguski, M.; Lipman, D.; and Ostell, J. 1997. Genbank. *Nucl. Acids Res.* 25:1–6.

Boguski, M. S., and Tolstoshev, C. M. 1994. Gene discovery in dbEST. *Science* 265:1993–1994.

Boguski, M. S.; Lowe, T. M. J.; and Tolstoshev, C. M. 1993. dbEST — database for "expressed sequence tags". *Nature genetics* 4:332–333.

Brunak, S.; Engelbrecht, J.; and Knudsen, S. 1990. Neural network detects errors in the assignment of mRNA splice sites. *Nucleic Acids Res.* 18:4797–4801.

Brunak, S.; Engelbrecht, J.; and Knudsen, S. 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* 220:49–65.

Cavener, D. R., and Ray, S. C. 1991. Eukaryotic start and stop translation sites. *Nucl. Acids Res.* 19:3185–3192.

Cavener, D. R. 1987. Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucl. Acids Res.* 15:1353–1361.

Cigan, A. M., and Donahue, T. F. 1987. Sequence and structural features associated with translational initiator regions in yeast — a review. *Gene* 59:1–18.

Cooke, R.; *et al.* 1996. Further progress towards a catalogue of all *Arabidopsis* genes: analysis of a set of 5000 non-redundant ESTs. *The Plant Journal* 9:101–124.

Hansen, J. E.; Lund, O.; Engelbrecht, J.; Bohr, H.; Nielsen, J. O.; Hansen, J.-E. S.; and Brunak, S. 1995. Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAC: polypeptide *N*-acetylgalactosaminyltransferase. *Biochem J.* 308:801 813.

Hertz, J.; Krogh, A.; and Palmer, R. G. 1991. *Introduction to the Theory of Neural Computation.* Studies in the Sciences of Complexity. Addison-Wesley Publishing Company.

Hobohm, U.; Scharf, M.; Schneider, R.; and Sander, C. 1992. Selection of representative data sets. *Prot. Sci.* 1:409–417.

Joshi, C. P. 1987. An inspection of the domain between putative TATA box and translation start site in 79 plant genes. *Nucl. Acids Res.* 15:6643–6653.

Kozak, M. 1983. Translation of insulin-related polypeptides from messenger RNAs with tandemly reiterated copies of the ribosome binding site. *Cell* 34:971–978.

Kozak, M. 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucl. Acids Res.* 12:857–872.

Kozak, M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 44:283–292.

Kozak, M. 1987. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucl. Acids Res.* 15:8125.

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

Lütcke, H. A.; Chow, K. C.; Mickel, F. S.; and Moss, K. A. 1987. Selection of AUG initiation codons differs in plants and animals. *EMBO J.* 6:43–48.

Mathews, B. W. 1975. Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405:442–451.

Nielsen, H.; Engelbrecht, J.; Brunak, S.; and von Heijne, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering* 10:1–6.

Pedersen, A. G., and Engelbrecht, J. 1995. Investigations of *Escherichia coli* promoter sequences with artificial neural networks: new signals discovered upstream of the transcriptional startpoint. *Proceedings, third international conference on intelligent systems for molecular biology (ISMB-95)* 292–299.

Pedersen, A. G., and Nielsen, H. 1997. Statistical analysis of translation initiation sites in eukaryotes: patterns are specific for systematic groups. *In preparation.*

Sander, C., and Schneider, R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *PROTEINS: Structure, Function, and Genetics* 9:56–58.

Yamauchi, K. 1991. The sequence flanking translational initiation site in protozoa. *Nucl. Acids Res.* 19:2715–2720.

Yoon, H., and Donahue, T. F. 1992. Control of translation initiation in *Saccharomyces cerevisiae*. *Mol. Microbiol.* 6:1413–1419.