The Ribosome Scanning Model for Translation Initiation: Implications for Gene Prediction and Full-Length cDNA Detection

Pankaj Agarwal and **Vineet Bafna**
SmithKline Beecham Pharmaceuticals R&D
UW2230, 709 Swedeland Road
P.O. Box 1539, King of Prussia, PA 19406-0939
{*agarwal,bafnav1*}*@mh.us.sbphrd.com*

Abstract

Biological signals, such as the start of protein translation in eukaryotic mRNA, are stretches of nucleotides recognized by cellular machinery. There are a variety of techniques for modeling and identifying them. Most of these techniques either assume that the base pairs at each position of the signal are independently distributed, or they allow for limited dependencies among different positions. In previous work, we provided a statistical model that generalizes earlier methods and captures all significant high-order dependencies among different base positions.

In this paper, we use a set of experimentally verified translation initiation (TI) sites (provided by Amos Bairoch) from eukaryotic sequences to train a range of methods, and then compare these methods. None of the methods is effective in predicting TI sites. We take advantage of the *ribosome scanning model* (Cigan et al., 1988) to significantly improve the prediction accuracy for full-length mRNAs. The ribosome scanning model suggests scanning from the 5' end of the capped mRNA and initiating translation at the first AUG in good context. This reduces the search space dramatically and accounts for its effectiveness. The success of this approach illustrates how biological ideas can illuminate and help solve challenging problems in computational biology.

Introduction

DNA and mRNA sequences contain *signals*, which are stretches of nucleotides that are recognized by cellular apparatus such as ribosomes, sequence-specific binding proteins, and complementary nucleic acids. For instance, translation initiation signals identify the start of protein translation, and are marked with a strongly conserved AUG (Pain, 1996). Identification of signals in biological sequences is a central problem in computational biology, especially for gene prediction (Fickett, 1996). Gelfand (1995) has written a comprehensive review of signal prediction techniques and the functional sites that have been predicted using these techniques.

The usual starting point for modeling signals is an ungapped alignment of sequences known to contain these signals. Earliest approaches used a representative signal sequence, usually the *consensus* sequence from the alignment. *Profiles* extend these methods to estimate the distribution of nucleotides at each position in the alignment (Staden, 1984; Gribskov et al., 1988; Stormo, 1990). Hidden Markov models (Reese et al., 1997) and the Gibbs sampler (Lawrence et al., 1993) estimate both the alignment and the signal but, for the most part, they ignore all dependencies between bases at different positions.

Recent techniques attempt to capture some of the dependencies, but still have to focus on low-order models (Zhang and Marr, 1993; Salzberg, 1997; Agarwal and Bafna, 1998). In general, the data is insufficient to estimate high-order models reliably. Burge and Karlin (1997) describe an alternative approach (termed *maximal dependence decomposition*) for capture some distant and higher order dependencies. Neural networks are an alternative technique to capture the correlations, but it is difficult to get explicit dependencies (Pedersen and Nielsen, 1997).

In earlier work (Agarwal and Bafna, 1998), we described a technique (now termed *GSP: Generalized Second-Order Profile*) to estimate the dependencies of each base position conditional upon every other base position. Given all the significant dependencies, we utilized an *arborescence* algorithm to choose a set of dependencies that maximize the information content of the signal.

For TI prediction, the information content for GSP (and most other models) is less than 9 bits and is expected to lead to a false positive once every 512 bases. Even if the GSP score only provides a weak approximation to the true information content of the DNA binding site, we do not see any evidence of the signal carrying enough information that the ribosome could bind specifically to it. In this paper, we investigate the *ribosome scanning model*, which resolves this paradox by severely limiting the region in which a ribosome must search for the translation initiation site.

The Ribosome Scanning Model (RSM)

Kozak (1996) has performed an extensive analysis of the translation initiation sites in eukaryotic mRNAs and has proposed the consensus GCACCatgG as the optimal context for initiation. The A corresponding to the ATG/AUG is numbered as +1. Within this consensus motif, nucleotides in two highly conserved positions exert the strongest effect: a G residue following the AUG codon (at position +4), and a purine (A/G), preferably A, at position[2] −3. The importance of these positions has been validated in experimental studies (Kozak, 1986). In addition, mutation studies have also indicated a correlation between the positions −3 and +4 (Kozak, 1986). This context is not specific enough to pinpoint a TI site, and a typical mRNA would contain multiple AUGs in good context. The fact that ribosomes appear to bind specifically to the TI site may be explained by the ribosome scanning model.

> Recall that eukaryotic ribosomes do not engage the mRNA directly at the ATG (AUG) start codon. Rather the small (40S) ribosomal subunit enters at the capped, 5' end of the mRNA and then migrates or "scans" linearly until it encounters the first ATG codon, which is then recognized by base-pairing with the anti-codon in Met-tRNA$_i$ (Cigan et al., 1988; Kozak, 1989; Kozak, 1992). The sequence flanking an ATG codon—the "context" seems to affect the efficiency with which a particular ATG triplet stops the scanning 40S subunit. When the 40S ribosomal subunit stops, the 60S subunit joins and the ribosome is ready to form the first peptide bond. ...(Kozak, 1996)

In computational terms, if this hypothesis were true, the correct method should scan the mRNA sequence and pick the first AUG in "good" context, rather than the AUG in the best scoring context. The hypothesis also suggests that there could be other high scoring sites downstream of the true initiation site, making prediction difficult on partial sequences, such as ESTs.

An extension to this hypothesis has also been proposed (Kozak, 1996). Within certain limits, eukaryotic ribosomes can hold on to the mRNA at a terminator codon, resume scanning, and reinitiate downstream at another AUG codon, albeit inefficiently. An mRNA in which the first AUG in a good context is followed by a terminator codon after a short distance is a candidate for reinitiation. In other words, three elements are necessary for reinitiation: a good TI site, an in-frame stop codon shortly thereafter, and another good start shortly after the stop codon.

There is evidence in viruses and prokaryotes (Futterer et al., 1993; Sonenberg, 1994) that the ribosome can bypass a segment of the 5' UTR (possibly due to secondary structure) or may bind to an internal (rather than the 5' cap) site. These mechanisms, if present in eukaryotic cells, argue against the ribosome scanning model. McBratney and Sarnow (1996) have discovered transacting factors that may play a role in AUG start site selection during translational initiation; thus the scanning model as described in this paper is not the entire story. However, for most eukaryotic mRNAs the ribosome scanning model is believed to be correct (Kozak, 1989; Pain, 1996).

Clearly, the ribosome scanning model resolves the paradox of the ribosome being able to recognize and bind to the correct TI site in spite of the low information in the signal. If this model is correct, then computational approaches that simply seek to find the signal with the maximum information content are doomed. Indeed, there is some anecdotal evidence to suggest that methods that just look for the first AUG in a full-length mRNA seem to do as well as the more sophisticated methods. In this paper, we take the logical next step and compare different models of the translation initiation signal with and without the ribosome scanning idea.

Data and Methods

A subset of proteins (from the SwissProt database) with experimentally verified translation initiation sites was provided by Amos Bairoch (personal communication). The start of the mature proteins coincided with the TI signal for 239 proteins. The protein sequences were searched against GenBank (release 104.0, December 1997, with no ESTs), using tblastn from the WU-BLAST 2.0a17 suite (Altschul et al., 1990; Altschul and Gish, 1996), to find all nucleotide sequences with 100% identity over the entire length of protein sequence (including the initiating methionine). For every protein sequence, we selected at most 1 nucleotide sequence if it had enough upstream context (at least 10 bases 5' of the correct TI site), and this reduced the set to 189 sequences.

An additional 357 secreted proteins were also provided by Amos Bairoch and searched similarly against GenBank, except for an additional step. For these proteins, the mature peptide has been experimentally verified but the TI is not exactly known. However, it may be inferred reliably, if there is a unique AUG upstream of the start of the mature protein that is consistent with the length of the signal peptide (14-41 aa). This additional step reduced this set to 225 sequences.

A redundancy check was performed on these 189 + 225 = 414 sequences to eliminate biases in the data set. We created two data sets. The nr-75 data set retained 372 sequences with less than 75% nucleotide identity to each other. The nr-50 data set with 298 sequences, retains only those sequences which have less than 50% identity to each other.

A 14 base pair region surrounding the true initiation sites (from −9 to +5) was extracted from the sequences. This was done separately for both nr-50 and nr-75 data sets. Ungapped alignments for each of these sets were used to train all the models listed later in this section.

[2]In this numbering scheme (as often in molecular biology), there is no base numbered zero.

In addition to testing the techniques on the training data set, a much larger test data set of 4523 full-length human mRNA sequences was created from the Unigene database (Boguski and Schuler, 1995). The annotations in this test data set are not as reliable as in the custom data set provided by Amos Bairoch, so some manual filtering was used to remove obvious sources of noise, such as GenBank entries that were inconsistent. In addition, sequences lacking enough 5' context or containing multiple CDS entries were eliminated to yield a clean data set of 2993 sequences. We note that this data set overlaps with the training set; 165 sequences had the same GenBank identifier, and presumably some additional ones had sequence homology. The training set with at most 372 sequences is, however, small enough so as to not significantly bias the results on the much larger test set.

The ribosome scanning with reinitiation method (RSM) may be described procedurally: select the first AUG that scores above a fixed threshold. If this AUG has a short ORF (< 200 bp),[3] we assume reinitiation and resume scanning after the terminator codon. If it has a long ORF, stop and report the site. Note that the idea can be used with different scoring methods, and we use it with First-AUG, Profiles, WAM, and GSP, all described below. For each method a threshold is computed that maximizes its prediction accuracy. The following methods for scoring sites are compared:

Profiles (PWM) (Staden, 1984; Gribskov et al., 1988; Stormo, 1990): A profile or a position weight matrix is constructed from the position-specific frequencies of the bases from an alignment of true signals. All positions are assumed to be distributed independently.

Weight array matrices (WAM) (Stormo, 1990; Zhang and Marr, 1993; Salzberg, 1997): This is a second-order profile that assumes that each base is conditional upon the previous base.

Generalized Second-order Profile (GSP) (Agarwal and Bafna, 1998): This method includes a selection of most informative correlations that include both adjoining and non-adjoining ones. This set of correlation is selected using an arborescence algorithm. All the correlations included in the model are guaranteed to be statistically significant (p=0.05). This statistical significance is based on a sampling technique for the background distribution. This is a very conservative significance estimate because we account for every hypothesis tested.

GSP with χ^2 significance (GSP_{χ^2}): Instead of sampling from the background distribution for correlations (Agarwal and Bafna, 1998), we compute the χ^2 significance for each correlation. All the correlations above a certain p-value are included in the second-order model. Though each correlation is significant at the stated p-value, it is quite likely that one or more of them is insignificant because of multiple hypothesis testing. Burge and Karlin (1997) have previously used the χ^2 test for signal prediction. We use two different p-values (0.05 and 0.001) for the χ^2 test. For a 4×4 contingency table[4] (9 degrees of freedom), a p-value of 0.05 corresponds to a $\chi^2 = 16.72$ (Snedecor and Cochran, 1989). This p-value may be somewhat optimistic considering the number of hypotheses tested; thus we also repeated the experiments with a p-value of 0.001, which corresponds to $\chi^2 = 27.88$ (df=9).

First-AUG: In this method, each AUG gets the same score. This method is only useful in the context of RSM, a model in which the first AUG from the 5' end of the mRNA is often the correct TI signal. This null model mimics a seemingly naive consensus prediction based on the RSM.

In evaluating methods for predicting translation initiation, we distinguish between testing on full-length mRNA sequences and other data, namely, partial mRNA, EST, and genomic sequences. Methods that do not use RSM can be applied to all of the above data. These methods often do not explicitly predict sites, but simply report all sites scoring above a certain threshold. So if a computational decision is required, the highest scoring site is the best choice, but often other information, such as coding potential or open-reading frames, is used by the decision-making process or the scientist. In contrast, our method, using the RSM, explicitly predicts a TI site for full-length mRNA.

When comparing these computational methods, we require each method to pick a site explicitly. Each test sequence (of full-length mRNA) results in either one correct or one false prediction. Thus, the fraction of sequences for which the TI site is correctly predicted is a useful measure of the performance of a method. This simple statistic is equivalent to the minimum error rate very often used in classification (Duda and Hart, 1973) and it has also been used in sequence analysis (Agarwal and States, 1998). Note that this statistic does not involve choosing a threshold score; it minimizes the number of errors over all choices of thresholds.

Results and Discussion

In computing the model for GSP, we evaluated all possible correlations, but the final model involved mostly adjoining correlations (data not shown). One interesting experimentally validated correlation (Kozak, 1986) that we examined carefully is between positions -3 and $+4$. The experimental evidence suggests that the absence of a purine (A/G) at -3 necessitates a G at $+4$. The χ^2 value for this $(-3, +4)$ pair is 16.2 with p=0.063. Thus, it is not statistically significant at the traditional significance level of p=0.05. This may illustrate the problems of using statistical significance to

[3]The choice of 200 bp as the short ORF length is based on paucity of proteins with length less than 66aa.

[4]A 4x4 contingency table is used to capture the joint frequencies of every pair of bases in the two positions.

Technique	nr-50		nr-75	
	% correct	Threshold	% correct	Threshold
Profile	30.0%	-	30.0%	-
WAM	26.8%	-	27.0%	-
with *ribosome scanning model*				
First AUG	65.5%	-	65.5%	-
Profile	83.5%	125	83.9%	109
WAM	83.1%	29	84.3%	8
GSP	80.0%	156	81.8%	119
GSP_{χ^2} (p=0.05)	80.0%	156	82.2%	108
GSP_{χ^2} (p=0.001)	82.0%	117	82.8%	121

Table 1: Percentage of full-length mRNAs (of 2,993) from Unigene (test set) for which the TI site was correctly predicted using the various techniques. The thresholds are in units of 0.01 bits, so the thresholds are rather low.

Technique	nr-50		nr-75	
	% correct	Threshold	% correct	Threshold
Profile	48.0%	-	45.5%	-
WAM	44.6%	-	44.6%	-
with *ribosome scanning model*				
First AUG	80.1%	-	81.3%	-
Profile	92.7%	177	93.0%	182
WAM	91.9%	197	91.7%	184
GSP	91.5%	169	91.7%	216
GSP_{χ^2} (p=0.05)	91.5%	169	92.0%	203
GSP_{χ^2} (p=0.001)	90.0%	182	90.8%	168

Table 2: Percentage of *training* set mRNAs for which the TI site was correctly predicted. The thresholds are in units of 0.01 bits.

imply biological significance, especially in view of small data sets. Alternatively, the mutation experiments that established the correlations may only be a valid for the consensus AUG sequence used, and may not generalize.

Table 1 presents the results of different methods applied to the test data set. Methods based purely on signal content, such as profiles and WAM, can predict the correct TI sites for only 27%–30% of the full-length mRNAs. The success rate jumps to 65% with First-AUG, which is the simplest model that incorporates ribosome scanning. Other techniques, which are based on a better description of the context around the initiating AUG, improve the prediction accuracy to 80-84%. As our test depends upon GenBank annotations, it could be argued that the annotations are not based upon experimental evidence and are perhaps even biased towards picking the first AUG.

We tested the methods on the training set as well, so as to help confirm or deny the above hypothesis (see table 2). Not surprisingly, on this data set, all methods perform better than on the test set. However, the conclusions regarding RSM remain unchanged. Without RSM, the performance is no more than 48%. With RSM, the prediction accuracy becomes as good as 93%.

It is evident that the success rate of methods that do not use RSM is low (irrespective of the test data set). These results indicate that these methods are not very accurate in predicting TI sites on genomic and partial mRNA (EST) data. The ribosome scanning idea is crucial to picking up the weak translation initiation signal and is applicable only to full length mRNA. A popular and misguided use of these programs is to check if a 5' EST fragment contains the TI site, implying that the corresponding cDNA clone is full length. Our results indicate that these methods are of little or no use in making such predictions. As a very rough illustration, consider an extreme example in which we have 1,000 genes each with 5 non-overlapping ESTs, each EST about 256 bases long. A TI prediction method, based on an information content of 8 bits in the signal (see table 3), would predict a site once every 256 bases. This would imply one prediction every EST for a total of 5,000 predictions. But, only 1 in 5 ESTs has a TI site, thus resulting in 4,000 errors. Additionally, of the 1,000 ESTs with TI sites the prediction accuracy is below 50%, resulting in another 500 errors. In summary, we will make 5,000 predictions, of which only 500 will be correct. Thus, predicting TI sites in isolation is not useful; however, it is possible to combine it with other measures (such as coding potential and ORF lengths) to get reasonable predictions. This is the strategy employed by most successful gene pre-

diction programs (Fickett, 1996). On the other hand, for partial mRNAs (or ESTs), it may be possible to restrict possible TI sites to a few AUG codons by eliminating some AUGs because of poor contexts, secondary structure, or good upstream AUGs (Kozak, 1996). Pedersen and Nielsen (Pedersen and Nielsen, 1997) have employed neural networks trained on rather large windows around the AUG codon to detect initiating AUG codons with 85% accuracy in vertebrate mRNA. Their technique exploits the context around the AUG plus frame detection, and possibly the hexamer (or other coding potential) differences between 5' UTR and coding sequence.

Interestingly, the first-order profile does at least as well if not better than all the higher order methods. In particular, GSP captures all the information in a profile, and some extra information from statistically significant correlations. It may, therefore, be considered surprising that the profile performs better than GSP when tested on the *training* set. A possible explanation is that the RSM was used to test these models while it was not used in the training. Alternatively, Lapedes et al. (1997) have observed that at least for protein sequences, phylogenetic similarities in the training set may often provide correlations that are statistically significant but have no biological significance. This may indeed be the case with some of the correlations in the GSP model.

Without reading too much into the actual values, we can conclude that the increase in power due to the higher order models does not improve the prediction and may even adversely affect it because of overtraining by some of the methods. On a pessimistic note, efficient translation may not even be the goal of many genes, and some upstream AUGs in good context may be present simply to inhibit translation of the gene (Kozak, 1986). Therefore, pure computational discrimination of the TI signal may not be feasible.

Model	Information (in bits)	
	nr-50%	nr-75%
Profile	8.31	8.32
WAM	9.07	9.02
GSP	9.07	9.01
GSP_{χ^2} (p=0.05)	9.07	8.98
GSP_{χ^2} (p=0.001)	8.73	8.73

Table 3: The information content of various models.

Acknowledgments

We are grateful to Amos Bairoch for providing the data set, which made this study possible. In addition, we would like to thank James Fickett for a valuable critique and many useful references. Jianmei Fang Duckworth, Istvan Ladunga, Bill Marshall, David Searls, Randall Smith, and Wyeth Wasserman provided valuable comments on the manuscript. Finally, we thank Laureen Treacy for proofreading the manuscript and the Bioinformatics Group at SmithKline Beecham for support and encouragement.

References

Agarwal, P. and Bafna, V. (1998). Detecting non-adjoining correlations within signals in DNA. In *Proceedings, Second Annual International Conference on Computational Molecular Biology, RECOMB98*, pages 1–7. ACM Press.

Agarwal, P. and States, D. (1998). Comparative accuracy of methods for protein-sequence similarity search. *Bioinformatics*, 14(1). in press.

Altschul, S. and Gish, W. (1996). Local alignment statistics. *Methods Enzymol.*, 266:460–480.

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410.

Boguski, M. and Schuler, G. (1995). ESTablishing a human transcript map. *Nature Genet.*, 10:369–371.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268(1):78–94.

Cigan, A., Feng, L., and Donahue, T. (1988). tRNA functions in directing the scanning ribosome to the start site of translation. *Science*, 242:93–97.

Duda, R. and Hart, P. (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons.

Fickett, J. (1996). Finding genes by computer: The state of the art. *Trends in Genetics*, 12(8):316 320.

Futterer, J., Kiss-Laszlo, Z., and Hohn, T. (1993). Nonlinear ribosome migration on cauliflower mosaic virus 35S RNA. *Cell*, 73(4):789–802.

Gelfand, M. (1995). Prediction of function in DNA sequence analysis. *J. Comp. Biol.*, 2(1):87–115.

Gribskov, M., Homyak, M., Edenfield, J., and Eisenberg, D. (1988). Profile scanning for three-dimensional structural patterns in protein sequences. *Comput. Appl. Biosci.*, 4:61–66.

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

Kozak, M. (1992). Regulation of translation in eukaryotic systems. *Ann. Rev. Cell Biol.*, 8:197–225.

Kozak, M. (1996). Interpreting cDNA sequences: Some insights from studies on translation. *Mammalian Genome*, 7:563–574.

Lapedes, A., Giraud, B., Liu, L., and Stormo, G. (1997). Correlated mutations in protein sequences: Phylogenetic and structural effects. In *Proceedings of the AMS/SIAM Conference on Statistics in Molecular Biology*.

Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A., and Wootton, J. (1993). Detecting

subtle sequence signals: A gibbs sampling strategy for multiple alignment. *Science*, 262:208–214.

McBratney, S. and Sarnow, P. (1996). Evidence for involvement of trans-acting factors in selection of the AUG start codon during eukaryotic translational initiation. *Mol. Cell Biol.*, 16(7):3523–3534.

Pain, V. (1996). Initiation of protein synthesis in eukaryotic cells. *Eur. J. Biochem.*, 236:747–771.

Pedersen, A. and Nielsen, J. (1997). Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. In *Proceedings, Fifth International Conference on Intelligent Systems for Molecular Biology, ISMB97*, pages 226–233. AAAI press.

Reese, M., Eeckman, F., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. In *Proceedings, First Annual International Conference on Computational Molecular Biology, RECOMB97*, pages 232–240. ACM press.

Salzberg, S. (1997). A method for identifying splice sites and translation start sites in eukaryotic mRNA. *Comput. Appl. Biosci.*, 266:365–376.

Snedecor, G. and Cochran, W. (1989). *Statistical Methods.* Iowa State University Press/Ames, 8th edition.

Sonenberg, N. (1994). mRNA translation: Influence of the 5' and 3' untranslated regions. *Curr. Opin. Genet. Dev.*, 4:310–315.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, 12(1 Pt 2):505–519.

Stormo, G. (1990). Consensus patterns in DNA. *Methods Enzymol.*, 183:211–221.

Zhang, M. and Marr, T. (1993). A weighted array method for splicing signal analysis. *Comput. Appl. Biosci.*, 9:499–509.