BIOINFORMATICS

Vol. 14 no. 5 1998
Pages 384–390

Assessing protein coding region integrity in cDNA sequencing projects

A.A. Salamov, T. Nishikawa and M.B. Swindells

Helix Research Institute, 1532-3 Yana, Kisarazu-shi, Chiba-ken 292, Japan

Received on November 13, 1997; revised on March 13, 1998; accepted on March 16, 1998

Abstract

Motivation: *In cDNA sequencing projects, it is vital to know whether the protein coding region of a sequence is complete, or whether errors have occurred during library construction. Here we present a linear discriminant approach that predicts this completeness by estimating the probability of each ATG being the initiation codon.*

Results: *Because of the current shortage of full-length cDNA data on which to base this work, tests were performed on a non-redundant set of 660 initiation codon-containing DNA sequences that had been conceptually spliced into mRNA/cDNA. We also used an edited set of the same sequences that only contained the region following the initiation codon as a negative control. Using the criterion that only a single prediction is allowed for each sequence, a cut-off was selected at which discrimination of both positive and negative sets was equal. At this cut-off, 67% of each set could be correctly distinguished, with the correct ATG codon also being identified in the positive set. Reliability could be increased further by raising the cut-off or including homologues, the relative merits of which are discussed.*

Availability: *The prediction program, called ATGpr, and other data are available at http://www.hri.co.jp/atgpr*

Contact: *swintech@hri.co.jp*

Introduction

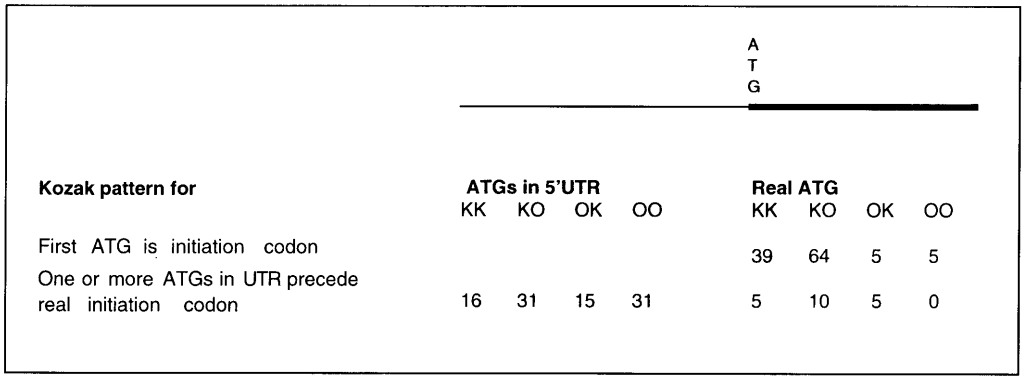
During the past decade, two major types of large-scale sequencing project have been initiated. The first approach is complete genome sequencing (Fleischmann *et al.*, 1995) which has the obvious advantage of mapping all the DNA for a particular organism. However, because of the amount of time required to sequence even a small eukaryotic genome (Goffeau *et al.*, 1997), the subsequent difficulties in predicting open reading frames (ORFs), as well as the absence of expression information, researchers have also used cDNA libraries as a more direct route to identifying novel gene products and expression data (Adams *et al.*, 1995).

In cDNA projects, the selection of clones to be sequenced is essentially a random approach, so for efficiency and cost, most high-throughput methods have only initially sequenced short regions, known as expressed sequence tags or ESTs (Adams *et al.*, 1991). By comparing with public databases, the ESTs can be divided into those which match known sequences and those which appear to be novel. Full-length sequencing can then be limited to the novel genes, provided that the clone from which the EST was derived is complete.

Our meaning of complete requires all of the sequence starting from the cap site and moving through the 5′UTR, initiation codon, protein coding region, termination codon, 3′UTR and poly(A) tail. This relates directly to the full-length sequencing project recently set up at our institute (Barker, 1996) which is establishing techniques (Maruyama and Sugano, 1994) whereby full-length clones can be reliably constructed and sequenced. The main problem in this field is that the fragility of mRNA makes it difficult to guarantee that the cDNA generated is really complete (Gubler and Hoffman, 1983). Therefore, approaches for analysing such data will not only need to identify different regions of the cDNA, but also be able to assess whether the sequenced clone was really complete. Creating a program that successfully balances these different aims is the challenging part of this work.

We consider that the two main ‘junctions’ in a piece of cDNA lie at the initiation codon, which separates the 5′UTR from the beginning of the coding region, and the termination codon which separates the end of the coding region from the 3′UTR. If the presence or absence of these two features could be reliably identified (together with their location when present), most of the regions of a cDNA would be reliably identified, as the poly(A) tail is trivial to determine. In fact, this would only leave completeness of the 5′UTR at the cap site undetermined. We have therefore considered ways to discriminate these features.

Our approach is to concentrate on the initiation codon, because successful identification will enable us to determine nearly all other data. For example, in sequences which contain the initiation codon, identification will lead to the partition of the 5′UTR/coding region. This, in turn, facilitates the detection of the termination codon when present, and hence the coding region/3′UTR boundary. Failure to locate the initiation site would imply that the coding region was incomplete and the 5′UTR was missing. Absence of the poly(A) tail would give similar information about the 3′ end of the



Kozak pattern for	ATGs in 5'UTR				Real ATG			
	KK	KO	OK	OO	KK	KO	OK	OO
First ATG is initiation codon					39	64	5	5
One or more ATGs in UTR precede real initiation codon	16	31	15	31	5	10	5	0

Fig. 1. Occurrence of ATG and Kozak-type patterns in our 133 sequence dataset. KK is complete [AG]ATGG pattern, whereas KO is just {AG]ATG, OK is ATGG and OO is ATG.

3′UTR. The only characteristic that would escape consideration is whether the 5′UTR is complete to the cap site.

Ideally, we would like to work with a large non-redundant set of full-length cDNA sequences in which the coding region, UTRs and other features have been fully documented. Unfortunately, such data are limited; indeed, this is the reason why our institute has undertaken such a sequencing project. Of the mRNA/cDNA deposited in Genbank, little of it is truly full length and certainly not enough to make a suitable test set. Looking at cases where both the mRNA and genomic DNA have been deposited, we find that most mRNA sequences cover the protein coding region, but have a truncated 5′UTR. Use of these mRNA data for analysis would incorrectly imply that the initiation codon was easy to detect, as 5′UTR truncation frequently results in the initiation codon being the first ATG in the sequence. Given this situation, we have opted to start with high-quality, annotated, genomic DNA and then splice our own mRNA/cDNA from this information.

The essential question for our work is how to identify the initiation codon correctly, when present, while also rejecting incomplete sequences. Researchers have previously used many 'rules of thumb' to determine whether the initiation codon is present. One of the best known trends is the preference for an [AG]xx**ATG**G pattern around the initiating ATG codon (Kozak, 1986). However, while the absence of such a pattern will usually exclude an ATG from being the initiation site, the pattern is so general that it will match many other ATG triplets in each sequence. A recent review (Kozak, 1996) has suggested that the best way of identifying the initiation site would be to find the ATG closest to the 5′ end having a pattern that matched the above. As we stated above, if one used the current mRNA data from sequence databases, this approach is likely to be quite successful, but in truly full-length clones the 5′UTR will be longer and the likelihood of a non-initiating ATG preceding the initiation codon will be higher. Even this discussion negates a further complication of introns being retained in the 5′UTR. In cases where a 5′UTR intron has not been spliced out during pre-mRNA processing, the region preceding the initiation codon will be even larger, making detection even more difficult.

In order to quantify the problems of an approach that merely chooses the most 5′ consensus matching ATG, we made conceptually spliced mRNA from 133 high-quality human gene entries from Genbank, in which information regarding the complete 5′UTR as well as 5′UTR introns was clearly documented (Figure 1). We found that even when all 5′UTR intronic regions had been removed from consideration (i.e. making the likelihood of success higher), there remained a significant number of ATG triplets in the 5′UTR that also corresponded to the most prevalent [AG]xx**ATG**G and [AG]xx**ATG** patterns.

In fact, there would be three distinct scenarios if such an approach were used. If the 5′UTR were complete, the method would frequently fail, but if the 5′UTR became truncated, while still retaining the initiation codon, the approach would become more successful. However, if the sequence were so incomplete that even the initiation codon were missing, success would fall to zero as the method would continue to predict the most 5′ Kozak consensus as the initiation codon.

Clearly, a more comprehensive method is required, and in this paper we describe an approach that uses linear discriminant analysis to combine a number of empirical observations reported in the literature. Discriminant analysis has previously been used in various aspects of gene recognition (Solovyev *et al.*, 1994) and independent tests have shown the approach to be as effective as other more complex methods (Burset and Guigo, 1996). The following discussion concentrates on sequences from human cell lines, but the method is general and could be applied to any other species, given that sufficient training data were available.

System and methods

133 sequence data set

All human sequences having the annotation 'prim-transcript' in the feature table were selected from Version 100 of GenBank, and those having any the following problems were removed: (i) proposed start codon was not ATG; (ii) initiation codon was not located in the exon stated to be the first coding exon; (iii) coding region did not end with a stop codon; (iv) an in-frame stop codon was identified in the proposed coding region; (v) the number of bases constituting the proposed coding region was not a multiple of three; (vi) donor and acceptor splice sites lacked highly conserved GT and AG dinucleotides, respectively. Then, coding regions of the remaining sequences were subjected to all-against-all global sequence alignment (Myers and Miller, 1988) and 133 entries with pairwise sequence identities of <50% were selected. Finally, from each entry, the 5′UTR introns were removed according to the annotation information.

660 sequence data set

To construct our test set, we started with GenBank sequences in Version 100 that had not been deposited by large-scale sequencing projects, as these were likely to have been verified experimentally (but see more detailed check later on). From these data, we first extracted 2432 human entries containing CDS information and having the 'DNA' label in the LOCUS field. Although our aim is to predict the initiation codon from a cDNA/mRNA sequence, we 'constructed' our own cDNA data by splicing out introns from the coding regions of complete genes. This was because only half of the human mRNA/cDNA entries available had >50 bp upstream of the initiation codon. In other words, most data do not start at the cap site and do not contain a significant portion of the 5′UTR. As the deposited mRNA/cDNA entries are significantly different to the data currently being sequenced from full-length cDNA libraries (Maruyama and Sugano, 1994), our approach of generating conceptually spliced cDNA from gene data is the most appropriate.

This initial set was cleaned to remove entries in which any of the following observations were made: (i) proposed start codon was not ATG; (ii) initiation codon was not located in the exon stated to be the first coding exon; (iii) coding region did not end with a stop codon; (iv) an in-frame stop codon was identified in the proposed coding region; (v) the number of bases constituting the proposed coding region was not a multiple of three; (vi) donor and acceptor splice sites lacked highly conserved GT and AG dinucleotides, respectively; (vii) more than one gene was present in the entry; (viii) 5′UTR was <50 bp; (ix) protein coding region was <100 bp. The remaining sequences were then checked in detail to ensure that they were experimentally verified genes. Entries containing the names of well-known prediction programs were eliminated by a computer search, and then the remaining entries were subjected to a manual analysis.

This left 1110 sequences. Introns occurring within the protein coding region were removed in order to simulate mRNA data, but in this test set (unlike our 133 sequence set described in the Introduction) 5′UTR introns were retained, as mRNA sequences sometimes have unspliced 5′UTR introns (Kozak, 1996). This has the effect of making our test set as difficult as possible.

Finally, to obtain a non-redundant dataset, we conducted all-against-all global alignments of the predicted protein products (Myers and Miller, 1988) and selected 660 sequences, with pairwise sequence identities of <50%. A total of 660 sequences contained 660 real ('true') ATG start codons and 35 668 other ('false') ATG trinucleotides. We refer to these sequences as the positive set. However, we also generated a negative set which only includes the regions which occur after the initiation codon. This is an important control because in real predictions we will not know whether the cDNA has the initiation codon present or not.

Cross-validation procedure

Predictions were tested by a 20-fold cross-validation approach in which 577 sequences were used to calculate statistical preferences, 50 were used to estimate coefficients for the linear discriminant function and 33 sequences for set aside for testing. Once this procedure had been repeated 20 times, all entries had been included once in the test set.

Constituents of the linear discriminant function

Discrimination of initiation codons from all other ATG trinucleotides was performed by linear discriminant analysis which has been successfully applied to gene recognition (Solovyev *et al.*, 1994). Discriminant analysis allows the statistical significance of various characteristics to be assessed (Afifi and Azen, 1979) and in our case we used contextual nucleotide preferences to discriminate real ATG start codons (positive class) from all other ATGs which might occur in the 5′UTR, coding and 3′UTR regions (negative class). We found the following characteristics to be the most useful for this task.

(1) *Positional triplet weight matrix around an ATG*. This is a simple extension of a singlet weight matrix approach, which has previously been used for the prediction of 5′-exons (Solovyev *et al.*, 1995). Our tests (data not shown) suggested that the triplet weight matrix gave better discrimination of start codons than singlet or doublet weight matrices. For each triplet i (i = 1, 64) and position j = (–14, +5), we first calculate the frequency for the subset of initiation ATGs and then repeat the calculation for all ATGs. Dividing

these two gives the propensity for a particular triplet to be in a specific position relative to the initiation codon.

$$P_{triplet}(i,j) = f_{initiation\ ATG}(i,j)/f_{total\ ATG}(i,j)$$

To apply these propensities, the total score around each ATG region is added together for the window –14 to +5. We also tried the more mathematically correct approach of summing logarithms of these values, but found that once the linear discriminant function had taken all characteristics into account, the results were effectively identical. We suspect that the 'noise' in sequence data plays a larger role than variations which result from specific manipulations of the propensities.

$$P_{hexamer}(k) = f_{coding}(k)/f_{noncoding}(k)$$

(2) When applying these to the test set, the propensities for each in-frame hexanucleotide downstream of an ATG is totalled, up to a maximum of 300 nucleotides (100 amino acids) or the end of the cDNA, which ever is shorter. Adding the propensities gives a preference for longer reading frames with suitable hexanucleotide compositions, while limiting the calculation to 300 nucleotides ensures that full-length cDNA is not required. However, as >85% of the sequences analysed also contain at least one false ORF of >300 nucleotides, this parameter is only removing short ORFs that are unlikely to encode proteins, rather than pointing the method towards the longest candidate. Tests on cDNA from our own libraries (data not shown) suggest that there is almost always 300 nucleotides of the ORF present when a cDNA is sequenced from its 5′ end.

(3) *5' UTR-ORF hexanucleotide difference.* The difference between the average hexamer coding propensities in the potential 5′UTR region [–1, –50] and potential coding region [+1, +50] for a given ATG was calculated. For real start codons, this characteristic has a higher value. Although related to the calculation in (2), this procedure suggests whether the initiation codon is actually present, whereas the data in (2) merely suggests the presence of a coding region. In cases where, for some reason, a cDNA fragment starts after the initiation codon, (2) will score highly whereas (3) will be low.

(4) *Signal peptide characteristic.* The most hydrophobic 8-residue peptide found within a 30 amino acid window, downstream of each ATG, was identified. This characteristic approximates the likelihood of a signal peptide being present (McGeoch, 1985). Hydrophobicity was calculated using the hydropathy scale of Kyte and Doolittle (1982).

(5) *Presence of another upstream in-frame ATG.* This is a simple binary characteristic with values 0 or 1. If an extra ATG is found upstream of the ATG under analysis (without encountering an in-frame stop codon), the likelihood of the ATG under analysis being the initiation codon is down-weighted.

(6) *Upstream cytosine nucleotide characteristic.* The frequency of cytosine in the region [–7, –36] upstream of a given ATG is included, as it has been observed that 5′UTRs of human genes are often cytosine rich (Louis and Ganoza, 1988).

These characteristics were then combined into a single linear discriminant function (LDF). The techniques for computing optimal coefficients for the LDF and estimating the statistical significance of potential discriminatory characteristics are the same as those described previously (Solovyev *et al.*, 1994; Salamov and Solovyev, 1997).

Simple position weight matrix

In order to show the improvements obtained by our approach, we compared the results to a simple position weight matrix of the form:

$$\sum_{i=-14}^{i=5} P(i,j) = f_{initiation}ATG(i,j)/f_{totat}ATG(i,j)$$

where f is the frequency of each nucleotide i at position j for the initiation codon-containing ATG sites (numerator) and all ATGs (denominator).

Measures of accuracy

Sensitivity and specificity for the positive, initiation codon-containing dataset are defined as follows, where $N_{correct}$ is the number of correct predictions at a particular threshold, N_{total} is the total number of sequences in the dataset and $N_{AboveThreshold}$ is the number of sequences with a prediction that is above the threshold.

$$sensitivity\ (\%) = \left(\frac{N_{correct}}{N_{total}}\right)100$$

$$specificity\ (\%) = \left(\frac{N_{correct}}{N_{AboveThreshold}}\right)100$$

For the negative set, it is best to consider the sensitivity in terms of the percentage of sequences correctly rejected at a particular threshold.

Implementation

The code for this algorithm, called ATGpr, has been written in C and has been implemented on standard Sun and Silicon Graphics platforms.

Results

We have applied our method to the prediction of initiation codons in a dataset of 660 initiation codon-containing sequences, as well as an edited set which only contain the sequence after the initiation codon. The reason for this ap-

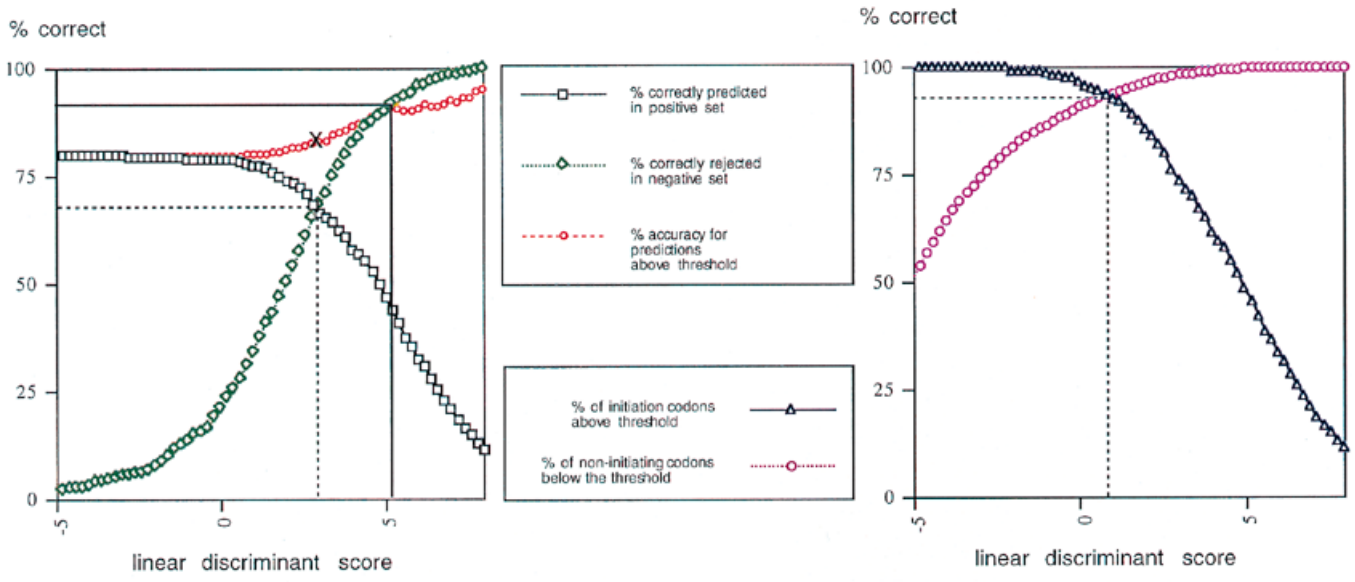


Fig. 2. (**a**) (Left) Results for the positive and negative set when only one prediction is allowed per entry. Black squares show the percentage of initiation codons correctly predicted in the positive set, and green diamonds shows the per cent correctly rejected in the negative set. The red circles show the increase in specificity for the positive set (i.e. when only predictions above the threshold are considered). The dotted line shows the point at which identification in the positive set is equal to rejection in the negative set and the x identifies the specificity for this threshold. The solid line corresponds to the threshold at which specificity for the positive set is equal to the sensitivity in the negative set. (**b**) (Right). The effect of threshold when all ATGs lying above the threshold are considered. The blue triangles describe the percentage of initiation codons lying above a particular threshold, whereas the pink circles show the percentage of non-initiating ATG triplets that lie below. The point at which these two percentages are equal (93%) is shown by a dotted line.

proach is that not only do we need to find the initiation codon when present, but we also need the ability to reject incomplete sequences. There are only 660 initiation codons in the complete sequences, yet there are 35 668 other ATG nucleotides which must be rejected. In contrast with methods for predicting protein secondary structure, where even a random assignment of helix, strand or coil to each residue will lead to a reasonably high percentage of correct predictions, the aim of finding the correct initiation codon is much harder.

Our first test is the most simple, in which only the top scoring ATG from each sequence in the positive test set (see methods) is predicted to be the initiation codon. Using this approach, we are able to predict 79% of the initiation codons correctly (Figure 2a). This is the upper limit for our method when only a single prediction is permitted per sequence and every sequence must be predicted, because the remaining 21% will always have an ATG triplet which scores more highly than the real initiation codon. The discriminant function that achieves this result has six components (see Methods), of which the most important two are the positional triplet weight matrix around an ATG (which performs better than a simple position weight matrix) and the ORF hexanucleotide characteristic (Table 1).

Table 1. Variation of accuracy with different components of discriminant function and comparison with a simple position weight matrix

	% of correct initiation codons identified in positive set	% of correct rejections in negative set at same threshold	% correct when positive and negative sets are equal
All six characteristics	79	12	67
Only ORF length	35	5	27
Only triplet weight matrix	29	2	27
ORF length and triple weight matrix	57	6	42
Simple position weight matrix	23	4	21

This would be quite a good result if it was certain that all the sequences analysed had an initiation codon. Unfortunately, we do not have this guarantee and, at the threshold which enables 79% of the initiation codons to be predicted in the positive test set, only 12% of the sequences in the negative test would be correctly identified as having no initiation codon. This result is not nearly as good and so it is necessary to find a balance between these conflicting aims.

To do this, we first defined a stricter threshold at which the number of initiation codons identified in the positive set was equal to the number of rejected sequences from the negative set (Figure 1). At this threshold, 67% of each set of 660 sequences could be identified correctly. However, while this adequately describes the negative set, the situation in the positive set has become more complex as it now includes sequences that are above the threshold and correctly predicted, others that are above the threshold but with the wrong initiation codon predicted, and finally some that are below the threshold and not predicted at all.

It is advantageous, therefore, to use two distinct measures when considering the positive dataset. Sensitivity is expressed in terms of the total number of sequences in the dataset and is the measure that has been used so far in this paper. Specificity, however, is calculated in terms of the number of predictions that lie above the threshold (see Methods). In a real situation, we may have automatically decided that predictions below a certain threshold were unsuitable and would therefore be most concerned about the reliability of those which lie above the threshold. This is the measure that specificity addresses.

If we consider the specificity for the positive set (see the red circles in Figure 2a), we see that this steadily increases with cut-off because it is limiting consideration to only the most confident predictions. In the case described above where the sensitivity is 67%, only 528 predictions are actually above the threshold, and thus the specificity for the positive set is 83% (as represented by the cross in Figure 2a). By selecting higher thresholds, specificity can be increased further as only the most confident predictions will be included. For instance, at the point where the percentage of sequences correctly rejected in the negative set is equal to the specificity for the positive set (the point where the green and red points conveniently cross in Figure 2a), we see that 89% specificity is obtained for the positive set, though with the disadvantage that sensitivity is now only 52%.

From a more academic viewpoint, it is also interesting to look at how we might differentiate between initiation codons and other ATG triplets in any piece of DNA. This aim is somewhat different to what we described above, as now we only need to consider the initiation codon-containing set. It is immediately clear from Figure 2b that the threshold which balances the percentages of initiation codons detected and other ATGs rejected is quite high (93%). Although this suggests that the procedure is rather effective, enthusiasm must be tempered by the sheer number of non-initiating ATGs in 7% of the false data. In fact, this will result in about four false positives being predicted for every initiation codon identified. By choosing a higher threshold, we can reduce the number of false positives, and at the cut-off where there is only one false positive per initiation codon identified, we are able to detect 61% of all initiation codons.

Discussion

In our tests, we have used a dataset where bias from sequence similarity has been minimized and where sequences have been further jackknifed into training and testing sets. Developing methods for predicting aspects of gene structure is now a popular area of research, and for obvious reasons there is competition to find procedures which have the best sensitivity and specificity. However, there are two common errors in papers which report significant improvements. The first problem is that there is almost always significant redundancy between the training and testing sets. The second problem is that direct comparison between results published in different years will be biased towards the most recent procedure, as that is the one that had the largest dataset to work from.

In the area of gene recognition, Burset and Guigo (1996) tried to reduce training and test set bias by creating two large test sets: one which contained homologous entries and a second which was non-redundant. Their results revealed that when tested independently, all the methods gave rather similar results and that those for the non-homologous test were noticeably worse than the redundant test set. Furthermore, all of the methods performed worse under this independent assessment than in the original reports. Owing to the popularity of their work, their datasets have since been used as a benchmark for testing newer algorithms (Kulp *et al.*, 1996; Burge and Karlin, 1997; Zhang, 1997). However, subsequent tests are difficult to perform in a fair manner because unless one rigorously cross-validates the process using only the non-homologous dataset, the database bias will always be a contributing factor. Even if cross-validation, as a method, is performed correctly, a dataset with redundancy will almost certainly result in an artificial level of accuracy, because homologous sequences will have the opportunity to appear in both the training and test sets simultaneously.

To give some idea of the improvements that can be expected by adopting a less rigorous attitude, we applied our algorithm to the Guigo dataset which contains homologues (Burset and Guigo, 1996) taking care to cross-validate all the predictions in the manner described in Methods. If we only allow one prediction per sequence, 86% of the top predictions are now correct. This is an apparent increase of 7% and is purely due to sequence bias. If we search for the mid-point where the accuracy in both positive and negative datasets is

equal, 75% of the data are correctly discriminated. This is an apparent increase of 8%, but is again purely an effect of database bias. Such increases suggest that improvements resulting from less rigorous validation techniques should be treated with caution.

As different types of sequencing projects come on-line, innovative solutions will be required to deal with the specific problems they encounter. The program we have produced here offers a general solution to the detection of 5 full-length cDNA when no homology to a known gene exists. It is expected that this method, like gene recognition methods, will be improved by including database searches as a pre-filter and, with this aim, an integrated approach is currently being developed.

References

Adams,M.D. *et al.* (1991) Complementary DNA sequencing: Expressed Sequence Tags and the human genome project. *Science*, **252**, 1651–1656.

Adams,M.D., Kerlavage,A.R., Fleischmann,R.D., Fuldner,R.A., Bult,C.J., Lee,N.H. and Kirkness,E.F. (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, **377(Suppl.)**, 3–174.

Afifi,A.A. and Azen,S.P. (1979) *Statistical Analysis. A Computer Oriented Approach.* Academic Press, New York.

Barker,S. (1996) Japanese genomics combines state and industry backing. *Nature*, **380**, 375.

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

Burset,M. and Guigo,R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.

Fleischmann,R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd. Science*, **269**, 496–512.

Goffeau,A., Aert,R., Agostini-Carbone,L.M., Ahmed,A., Aigle,M., Alberghina,K. and Albermann,K. (1997) The Yeast Genome Directory. *Nature*, **387(Suppl.)**.

Gubler,U. and Hoffman,B.J. (1983) A simple and very efficient method for generating cDNA libraries. *Gene*, **25**, 263–269.

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

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

Kulp,D., Haussler,D., Reese,M.G. and Eeckman,F.H. (1996), A *Generalized Hidden Markov Model for the Recognition of Human Genes in DNA*, **4**, 134–142, ISMB, AAAI/MIT Press.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Louis,B.G. and Ganoza,M.C. (1988) Signals determining translational start-site recognition in eukaryotes and their role in prediction of genetic reading frames. *Mol. Biol. Rep.*, **13**, 103–115.

Maruyama,K. and Sugano,S. (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligonucleotides. *Gene*, **138**, 171–174.

McGeoch,D.J. (1985) On the predictive recognition of signal peptide sequences. *Virus Res.*, **3**, 271–286.

Myers,E.W. and Miller,W. (1988) Optimal alignments in linear space. *Comput. Applic. Biosci.*, **4**, 11–17.

Salamov,A.A. and Solovyev,V.V. (1997) Recognition of 3-processing sites of human mRNA precursors. *Comput. Applic. Biosci.*, **13**, 23–28.

Solovyev,V.V., Salamov,A.A. and Lawrence,C.B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.*, **22**, 5156–5163.

Solovyev,V.V., Salamov,A.A. and Lawrence,C.B. (1995) Prediction of human gene structure using linear discriminant functions and dynamic programming. *ISMB*, **3**, 367–375.

Zhang,M.Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl Acad. Sci. USA*, **94**, 565–568.