Feature Selection for the Prediction of Translation Initiation Sites

Guo-Liang Li* and Tze-Yun Leong

Medical Computing Laboratory, School of Computing, National University of Singapore, Singapore 117543.

Translation initiation sites (TISs) are important signals in cDNA sequences. In many previous attempts to predict TISs in cDNA sequences, three major factors affect the prediction performance: the nature of the cDNA sequence sets, the relevant features selected, and the classification methods used. In this paper, we examine different approaches to select and integrate relevant features for TIS prediction. The top selected significant features include the features from the position weight matrix and the propensity matrix, the number of nucleotide C in the sequence downstream ATG, the number of downstream stop codons, the number of upstream ATGs, and the number of some amino acids, such as amino acids A and D. With the numerical data generated from these features, different classification methods, including decision tree, naïve Bayes, and support vector machine, were applied to three independent sequence sets. The identified significant features were found to be biologically meaningful, while the experiments showed promising results.

Key words: translation initiation site prediction, classification, feature selection

Introduction

One of the main objectives in bioinformatics is to identify important biological markers in genome and cDNA sequences. The translation initiation site (TIS) is one of such markers that has attracted a lot of research interests. TIS is the position in a cDNA sequence where the translation process in a cell begins to construct proteins. In the ideal condition with a full-length error-free cDNA sequence, the translation from the cDNA sequence to the corresponding protein starts from TIS and ends at the first in-frame stop codon downstream from TIS. This means if we can identify TIS in a cDNA sequence, we will gain deep insights into the gene structure and the corresponding protein. Moreover, identifying biological markers such as TIS is an important step to understand the entire biological process in a cell.

Identifying TIS in cDNA sequences has long been an active research area in biological sciences (*1–7*). In most cases, TIS is a tri-nucleotide ATG codon in DNA or an AUG codon in mRNA (There are rare cases that other codons, such as ACG and CTG, are served as TISs. These will not be considered in this paper). There are, however, a lot of ATGs in cDNA sequences and only very few of them act as TISs.

In the full-length cDNA sequences with known structures, the first occurrences of ATGs are frequently TISs. This phenomenon has inspired the scanning model hypothesis (sometimes called the first-ATG rule; ref. *1*, *2*, *8*), which postulates that the small (40S) subunit of eukaryotic ribosome initially binds at the $5'$ end of mRNA, migrates linearly downstream of the sequence, stops at the first AUG codon (*2*), and then starts the translation process. Unfortunately, the new cDNA sequences obtained are often neither in full length nor error-free (*5*). Hence, we cannot apply the first-ATG rule directly to all the new cDNA sequences. Even if the cDNA sequences are complete and error-free, there are some notable exceptions to the first-ATG rule. Pedersen and Nielsen showed (*9*) that only 60% vertebrate mRNAs they extracted from GenBank follow the first-ATG rule. The first ATGs are not TISs due to several reasons (*5*, *6*):

1. Leaky scanning. The ribosome bypasses the first ATG codon, the putative start site, due to the very weak context, and translation starts from a downstream ATG with context more similar to the optimal one.

2. Reinitiation. Translation starts from an ATG near the $5'$ end of mRNA and a small open reading frame (ORF) will be translated, but the ribosome con-

*** Corresponding author.**
E-mail: ligl@comp.nus.edu.sg

tinues scanning until the authentic ATG is reached to construct the protein.

3. Internal initiation. The ribosome binds near the real ATG codon directly without scanning, which is reported for several viral mRNAs.

Many TISs have been verified experimentally along with technology advancement. However, these experimental processes are very costly and time-consuming. Effective and efficient computational approaches, therefore, would greatly facilitate TIS prediction.

The TIS prediction problem can be treated as a classification problem with two classes: positive ATGs, which act as TISs, and negative (or non-start) ATGs, which do not act as TISs. The common computational approach to addressing the TIS prediction problem is to generate the numerical data from the cDNA sequences first, and then apply computational methods to predict TISs from all occurrences of ATGs. The main factors that affect the prediction performance in this approach include the nature of the cDNA sequence sets, the selected features to generate numerical data, and the computational methods used. There are three independent cDNA sequence sets (*9–11*) available in the literature for TIS prediction. Many different computational methods have been applied to the TIS prediction problem, including neural networks (*9*, *10*, *12*, *13*), linear discriminant analysis (*14*, *15*), support vector machines (SVMs; ref. *16*, *17*), mixture Gaussian models (*18*), and the Expectation-Maximization (EM) algorithm (*19*). The relevant features for TIS prediction include direct coding (*9*, *10*, *12*, *16*), *k*-gram usage measure (*20*), position weight matrix (PWM; ref. *14*, *16*, *21*, *22*), the generalized second-order profiles (*23*), coding difference between the regions before and after ATGs (*10*), and a few others.

Many researchers have noticed that different features affect the TIS prediction performance significantly (*14*, *16*). The experiments in Salamov *et al* (*14*) showed that the positional triplet weight matrix around an ATG and the ORF hexanucleotide characteristics are of the most importance in the encoding measures they used. The experiments in Zien *et al* (*16*) showed that the combination of the aggregated local information with direct coding could improve the prediction performance significantly while the codon usage measure would lower the performance.

However, there has been no significant work to compare the effects of different features on TIS prediction or to integrate the numerical data generated from different features to improve TIS prediction. In most previous efforts, the authors only tried a few data-encoding measures and features to generate numerical data in their work. An exception is the work by Zeng *et al* (*20*), in which four data-encoding measures were integrated together to generate numerical features; the important features were selected with the correlation-based feature selection method. However, they only considered a small set of the available numerical features.

In this paper, we extend our earlier investigations (*24*) on feature integration for TIS prediction with a substantial set of relevant features, classification methods, and independent cDNA sequence sets. The three independent cDNA sequence sets include one from Pedersen and Nielsen (*9*), one from Hatzigeorgiou (*10*), and one from Nadershahi *et al* (*11*). The three different feature selection methods applied include the Relief (*25*), the chi2-based (*26*), and the information-gain-based methods. The results show that the good candidate features for TIS prediction are the features from PWM and the propensity matrix, the number of nucleotide C in the sequence downstream ATG, the number of the downstream stop codons, the number of the upstream ATGs, and some others. In terms of prediction accuracy and Matthews correlation coefficient (MCC), our method is comparable to other state-of-the-art methods; the selected features can also be used as potential references for future work on TIS prediction.

Results

We generated numerical data from the three sequence sets based on the set of selected features (see Materials and Methods). The sequence set from Pedersen and Nielsen was used to build the computational models as it has more TISs than the other two sequence sets; the other two sequence sets were used as validation sequence sets. Specifically, Pedersen and Nielsen's sequence set was randomly split into six equally-sized subsets for the cross validation, in which each subset was respectively used as the testing sequence set and the other five subsets were used to build a model accordingly. The Hatzigeorgiou's sequence set and the Nadershahi *et al*'s sequence set were tested on each of the models and the average performances were reported.

Selected features from different methods

We used six training sequence sets and selected the features from each training set for the cross validation purpose. In most of the cases, the feature rankings did not totally agree with each other. However, for a specific feature selection method, the significant features from different training sets, with possibly different orderings, overlapped with each other. The rankings of the features from different training sets were summed together. Table 1 summarizes the fifteen top-ranked features from the three feature selection methods. The complete rankings of the features are available at http://www.comp.nus.edu.sg/∼ligl/publications/TIS/feature_rankings.htm.

Table 1 Top Selected Features from Three Different Feature Selection Methods

Rank	Relief method	Chi2-based method	Information gain method
1	2-gram PWM	1-gram PWM	2-gram PWM
2	# G upstream	# C in the region of [−36, −7]	# C in the region of [−36, −7]
3	3-gram PWM	# G at upstream codon position 1	3-gram PWM
4	1-gram propensity matrix	2-gram PWM	# G downstream
5	# C upstream	3-gram PWM	# G at upstream codon position 1
6	# T upstream	# G downstream	# C downstream
7	# amino acids AA downstream	# C at upstream codon position 3	1-gram propensity matrix
8	# ATG downstream	# downstream stop codon	# T downstream
9	# A upstream	C at position 139	# stop codon downstream
10	# G at downstream codon position 1	# C downstream	# amino acid A downstream
11	# C downstream	# downstream in-frame stop codon	# downstream in-frame stop codon
12	G at position 127	# 2-gram amino acids AG downstream	# A downstream
13	C at position 3 of potential downstream codons	T at position 149	# C at upstream codon position 3
14	# 2-gram amino acids GC upstream	C at position 148	2-gram propensity matrix
15	# amino acid A upstream	# amino acid D downstream	# ATG upstream

Note: # represents the number of the items followed. For example, "# G upstream" means the number of nucleotide G in the upstream sequence relative to the corresponding ATG. "# 2-gram amino acids GC upstream" means the number of 2-gram amino acids GC that are possibly encoded by the upstream sequence relative to the corresponding ATG.

Table 1 shows that the three feature selection methods favor different properties of the features and rank them in different orders. However, we can observe some commonalities in the top-ranked features. In Table 1, the common top features from all the three methods are those generated from PWM, which means that features from PWM are good ones for TIS prediction, as consistent with the results described in Salamov *et al* (*14*). As the characteristics of the propensity matrix are similar to those of PWM, the features generated from the propensity matrix are also highly ranked. Another common feature is the number of nucleotide C downstream in the sliding window. This means that the content of nucleotide C in the coding sequence is quite different from that in the non-coding sequence. Other common features are the number of downstream stop codons (both in-frame and non-in-frame), the number of upstream ATGs, the number of nucleotide C in the region [−36, −7], the number of nucleotide G downstream, and the nucleotide frequencies at the three codon positions. Observations on the downstream amino acids A and D are consistent with the results reported in Liu *et al* (*27*).

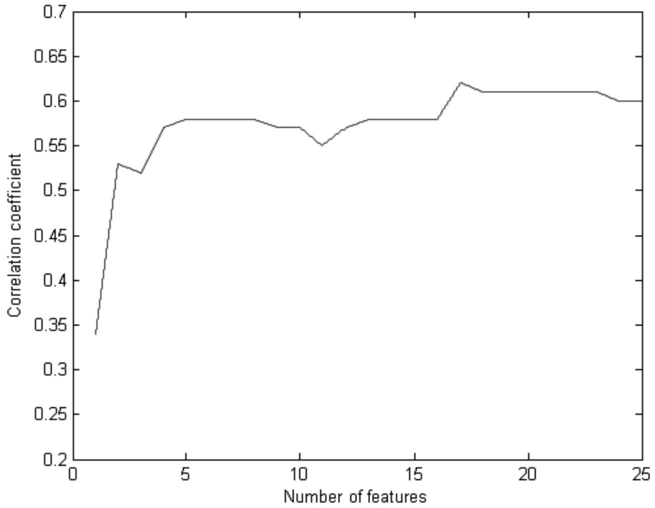

Fig. 1 The Matthews correlation coefficients for different numbers of features with information-gain-ranked features and the decision tree method.

Number of features for computational methods

An information-gain-based approach was used to decide on the number of features for use with the respective computational methods. Information gain measures the feature's discriminant power for classification; this approach has been successfully applied to the decision tree model building. After calculating the information gain of each feature, all the features are ranked in descending order. The top feature is first added to the selected feature set. A decision tree model is built based on the current selected features; the MCC value on the testing set is then calculated. Then the second feature is added to the selected feature set and the above process is repeated. After selecting the top 25 features, the MCC values are shown in Figure 1, in which the set with 17 features achieves the highest MCC values, and then the MCC values remain relatively constant. It shows that the top 17 features are the best with respect to the MCC values for decision tree methods. To consider the variation of the number for other feature selection methods, we selected 20 (3 more than 17) as the number of the selected features for further analysis.

The ranked features from the Relief method were also used to determine the number of features for the computational methods. We tried different numbers of features (10, 20, 40, 60, 80, 100, 120, 150, 180, and 200). The results with top 100 features are comparable with the results with more features (data not shown). Therefore we did experiments with 100 features, too.

Classification results

With 20 top-ranked features from three different feature selection methods, we fed the data into three different classification methods (Table 2). The results with 100 top-ranked features are shown in Table 3, in which the best results were obtained with the Relief method and the classification methods of decision tree and SVM. The accuracy is 97% and the MCC value is 0.91, which are among the best results available in the literature, as shown in Table 4. Li and Jiang (*17*) achieved good results on TIS prediction with the edit kernel and the SVM method. In our experiment, the kernel method is the linear kernel for SVM and the parameter C in SVM is equal to 1. The best result from the decision tree method with 100 features is easier to interpret.

With the models built on the sequence set from Pedersen and Nielsen, we also tested them with sequence sets from Hatzigeorgiou and Nadershahi *et al* (Table 5). The results in Table 5 show that the best model from the Pedersen and Nielsen's sequence set can achieve good performance on these two independent sequence sets, although the results are not as promising as those on Pedersen and Nielsen's sequence set. We observed that the models can always achieve better sensitivity than specificity—the models can identify TIS properly but always perform badly on the non-start ATGs. The possible reasons are that the models are sequence-set-dependent and/or the negative ATGs in Pedersen and Nielsen's sequence set are not representative enough.

Table 2 Classification Results with 20 Top-ranked Features for Three Different Feature Selection Methods and Three Different Classification Methods

Feature selection	Classification	Training				Testing			
method	method	Se	Sp	Acc	MCC	Se	Sp	Acc	MCC
Relief	Decision tree	83%	92%	90%	0.74	81%	91%	89%	0.71
	Naïve Bayes	98%	75%	81%	0.63	97%	75%	81%	0.63
	SVM	75%	93%	89%	0.69	75%	93%	89%	0.69
Chi2	Decision tree	70%	92%	86%	0.63	66%	90%	84%	0.56
	Naïve Bayes	86%	80%	81%	0.59	86%	80%	81%	0.59
	SVM	57%	90%	82%	0.50	57%	90%	82%	0.50
Information gain	Decision tree	66%	92%	85%	0.59	65%	91%	85%	0.58
	Naïve Bayes	97%	70%	77%	0.58	97%	70%	77%	0.58
	SVM	68%	90%	85%	0.58	68%	90%	85%	0.58

Table 3 Classification Results with 100 Top-ranked Features for Three Different Feature Selection Methods and Three Different Classification Methods

Feature selection	Classification	Training				Testing			
method	method	Se	Sp	Acc	MCC	Se	Sp	Acc	MCC
Relief	Decision tree	98%	98%	98%	0.95	95%	97%	97%	0.91
	Naïve Bayes	99%	80%	85%	0.70	99%	80%	85%	0.70
	SVM	96%	97%	97%	0.92	95%	97%	97%	0.91
Chi2	Decision tree	96%	98%	97%	0.93	72%	90%	86%	0.61
	Naïve Bayes	100%	76%	82%	0.66	100%	76%	82%	0.66
	SVM	84%	93%	91%	0.75	82%	92%	90%	0.73
Information gain	Decision tree	94%	98%	97%	0.91	76%	92%	88%	0.68
	Naïve Bayes	99%	77%	83%	0.67	99%	77%	83%	0.67
	SVM	84%	93%	91%	0.76	83%	93%	90%	0.75

Table 4 Comparison of the Results from Different Methods

Method	Se	Sp	Acc	MCC
Neural network*	82.4%	64.5%	84.6%	0.627
Salzberg method*	68.1%	73.7%	86.2%	0.619
SVM Salzberg kernel*	78.4%	76.0%	88.6%	0.696
SVM edit kernel III ASCM250#	99.8%	99.9%	99.9%	0.997
Decision tree (20 features from Relief)	81%	91%	89%	0.71
Decision tree and SVM (100 features from Relief)	95%	97%	97%	0.91

*The results from Zien *et al* (*16*). #The results from Li and Jiang (*17*).

Table 5 Results on Sequence Sets 2 and 3 with 100 Top-ranked Features

Feature selection	Classification	Sequence set 2				Sequence set 3			
method	method	Se	Sp	Acc	MCC	Se	Sp	Acc	MCC
Relief	Decision tree	97%	76%	77%	0.30	80%	78%	78%	0.30
	Naïve Bayes	100%	46%	47%	0.17	95%	45%	48%	0.18
	SVM	98%	76%	76%	0.30	81%	78%	78%	0.30
Chi2	Decision tree	82%	47%	48%	0.11	71%	54%	55%	0.12
	Naïve Bayes	99%	8%	11%	0.05	90%	23%	26%	0.07
	SVM	91%	52%	53%	0.16	81%	57%	58%	0.17
Information gain	Decision tree	88%	59%	60%	0.17	80%	64%	65%	0.20
	Naïve Bayes	100%	10%	13%	0.06	90%	24%	27%	0.08
	SVM	92%	63%	64%	0.20	86%	68%	69%	0.25

Discussion

In this paper, we have examined different measures for recognizing TISs. With three feature selection methods, we have identified some significant features for TIS prediction, such as features from k-gram PWMs, the propensity matrix, the number of downstream stop codons, the number of the upstream ATGs, downstream amino acids A and D, and the content of nucleotide C downstream of an ATG. With the selected features, the results show that our proposed methodology can achieve good performance for TIS prediction, compatible with a well-known sequence-similarity-based method. Moreover, we postulate that the top-ranked features are biologically meaningful, which can be used as the potential biological markers for future experiments. And we discuss the top-ranked features as follows.

First, among the top-ranked features, PWM is an aggregated measure, which may take most of the local information around TIS into consideration. Second, amino acid usage measure can capture the coding difference information in the sequences, since the sequences after TIS are used for constructing proteins while the sequences before TIS are not. Some researchers suggested that the coding difference is a good measure for TIS recognition and they have tried other measures to catch the coding difference, such as neural networks (*10*). Third, many features from other categories are top ranked, such as the number of downstream stop codons, the number of upstream ATGs, the number of nucleotide C in the region [−36, −7], and the nucleotide frequencies in codon positions. This means that the other ATGs and stop codons around the potential TISs are important to distinguish the TISs from the non-start ATGs. Fourth and interestingly, a well-known feature, a conserved purine at position −3, is not in the top-ranked features. A possible reason is that many non-start ATGs follow such a pattern, too. Furthermore, few features from direct coding and k-gram usage measures are top ranked. The possible reason is that numerous features are generated from direct coding and k-gram usage measures, the discriminative power is scattered to many features and then each feature has less information for TIS prediction. The 3-gram usage measure is a codon usage measure, which is very similar to the amino acid measure but with less generalization power. This can explain why features from direct coding and k-gram usage measures are dominated by features from other measures in any of the three feature selection methods.

In our experiment, the model with the highest MCC also reached the best accuracy. But we did not compare the accuracy of our results with those in previous studies (*10*, *17*, *18*, *28*). In the studies of Hatzigeorgiou (*10*) and Li *et al* (*18*), the complete and error-free cDNA sequences were used for TIS prediction, while our system does not have such requirement. In the studies of Li and Jiang (*17*) and Nishikawa *et al* (*28*), the sequence similarity was used for TIS prediction. Li and Jiang combined sequence similarity and the state-of-the-art computational methods—kernel methods, and achieved the very good result for TIS prediction. Nishikawa *et al* required protein databases for the sequence similarity search. Although we know that sequence similarity implies the similar function and is a good way to predict sequence function, it is not easy to obtain the specific biological markers for experiments.

Furthermore, the sample sequences used in our work are segments around ATGs, and the features considered are local features. Some previous efforts to predict TISs from full-length cDNA sequences with global features are shown in the literature (*10*, *14*, *18*, *28*) with significant results. Since EST (expressed sequence tag) sequences, which are partial cDNA sequences, are shorter and easier to get under the current sequencing technology, our work focuses on the local features.

Materials and Methods

Data sets

Three independent cDNA sequence sets were used in our work for TIS prediction.

1. Sequence set 1: Pedersen and Nielsen's sequence set has been used in a series of earlier investigations (*9*, *16*, *17*, *20*). This data set consists of a selected set of vertebrate genome sequences extracted from GenBank. The possible introns are spliced from all sequences. The sequences satisfied with the following conditions are kept for the similarity test: (1) ATG as annotated TIS; (2) without non-nucleotide symbols; (3) with at least 10 nucleotides upstream of TIS and at least 150 nucleotides downstream of TIS. After the similarity test, 3,312 vertebrate sequences are left as the training sequence set. There are 13,503 ATGs in this sequence set, and 3,312 (24.5%) of them

are TISs.

2. Sequence set 2: Hatzigeorgiou's sequence set was originally extracted from Swissprot. The steps to extract the sequence set are as follows: (1) collect human protein sequences whose N-terminal sites are sequenced at the amino acid level (sequences manually checked by Amos Bairoch); (2) retrieve the full-length mRNAs for these proteins whose TISs have been indirectly experimentally verified. A total of 480 completely-sequenced and annotated human cDNAs are found. There are 14,108 ATGs in this sequence set, and 480 (3.4%) are TISs.

3. Sequence set 3: Nadershahi *et al*'s sequence set was collected for comparison of computational methods for identifying TISs in EST data. It was extracted from UniGene (*29*) Build #160. A total of 50 UniGene clusters were randomly selected from 371 UniGene clusters with complete CDS (coding sequences) annotation, and then 50 EST sequences with TIS and 50 EST sequences without TIS were randomly selected from the selected 50 UniGene clusters. There are 942 ATGs in this sequence set and 50 (5.3%) are TISs.

With the available sequence sets, we use a sliding window with length 204 to generate samples—there are 54 nucleotides upstream of an ATG and 150 nucleotides downstream of an ATG. For the positions with missing values, we pad with "N". The samples from the real TISs are positive samples and those from non-start ATGs are negative samples.

Feature generation

We partition the features into eight categories, mainly based on the different data-encoding measures. The details are as follows.

1. Direct coding. This is a simple way to generate numerical data from DNA sequences for TIS prediction (*9*, *10*, *12*, *16*). Generally, each nucleotide in a DNA sequence is encoded by four bits under direct coding: 0001 for A, 0010 for C, 0100 for G, 1000 for T, and 0000 for others. The sliding window in our work for direct coding is in the range [-54, 150] with respect to ATG's position (The nucleotides in cDNA sequences are numbered relative to ATG. The "A" in ATG is numbered as +1, and the numbers increase downstream of the cDNA sequences. The upstream nucleotide adjacent to ATG is numbered as -1 and the numbers decrease upstream of the cDNA sequences). This measure generates $4\times204 = 816$ features.

2. Consensus motif. The consensus motif GCCACCatgG around TIS was derived by Kozak (*30*), which states that these nucleotides GCCACCatgG frequently appear at the corresponding positions to TIS. Especially, a purine, preferably A, at position -3 is a significantly conserved signal for TIS. If each nucleotide in positions [-6, -1] and +4 of the sequence is the same as that in the consensus, a feature is encoded as 1; otherwise as 0. Moreover, if there is a purine at position -3, a feature is encoded as 1; otherwise as 0. The total number of the nucleotides that are the same as the consensus is counted. This measure generates 9 features.

3. K-gram usage measure ($k = 1\sim3$). A k-gram is a segment of sequence with k continuous nucleotides together. There are 4^k entries in k-gram for a specific k. The k-gram usage measure counts the frequency of each k-gram in the sliding window. In our work, the value of k is from 1 to 3. Although the measures with k greater than 3 can also be used, in our experiment, we found that the numerical data generated from the k-gram usage measures with k greater than 3 will overfit the training data easily. A possible reason is that if k is greater than 3, the total entries in k-gram will be very large and may memorize the training patterns. Therefore we do not consider the k-gram usage measure with k greater than 3 in this paper. Here we generate numerical data from the sequences before and after ATG separately. Then this measure generates $(4^1+ 4^2+ 4^3)\times2 = 168$ features.

4. Position weight matrix. PWM is the frequencies of each singlet, doublet, or triplet at every position in the sequence (*22*, *31*, *32*). For each k-gram i and position $j = (-54, +150)$ [$i = (1, 4)$ when $k = 1$, $i = (1, 16)$ when $k = 2$, and $i = (1, 64)$ when $k = 3$], let $f(i, j)$ be the frequency of k-gram i at position j. Then $f_{TIS}(i, j)$ is the frequency of k-gram i at position j from the positive training set, $f_{Total}(i, j)$ is the frequency of k-gram i at position j from the total training set, and $f(i)$ is the total k-gram i in the training data. The PWM entry $pwm(i, j) = \log\left(f_{TIS}(i, j)/f(i)\right)$. By this definition, we can generate three PWMs for $k = 1$, 2, and 3, respectively. The PWM score of one sequence is the sum of the individual scores at each position for each k. This measure generates 3 features.

5. Propensity matrix. The frequencies are defined as above. The entry in a propensity matrix $pm(i, j) = \log\left(f_{TIS}(i, j)/f_{Total}(i, j)\right)$. Here the value

of k is from 1 to 3. The reason is the same as mentioned above for PWM. This measure generates 3 features.

6. Amino acid k-gram usage measure. It is about the frequencies of k-grams of 20 amino acids and stop codons that can be expressed by the cDNA sequences in the sliding window, which is used in Li and Leong (*24*) and Liu *et al* (*27*). There are 21^k entries in amino acid k-gram for a specific k. In our work, we chose k to be 1 and 2. With the sequences upstream and downstream of ATG considered separately, this measure generates $(21 + 21\times21) \times 2 = 924$ features.

7. Signal peptide characteristic. This measure approximates the likelihood of a signal peptide being present (*33*), which was used in Salamov *et al* (*14*). Within a 30-amino-acid window downstream of each ATG, the hydrophobicity of the amino acids in a sliding window with length 8 is summed up. Also the most hydrophobic 8-residue peptide and the sum of the total hydrophobicity of the 30 amino acids are identified. There are 25 features in this measure.

8. Other features. The measures in this category include: the number of upstream and downstream ATGs (in-frame and non-in-frame), the number of upstream and downstream stop codons (in-frame and non-in-frame), the number of C and G (upstream and downstream), the number of C in the region of $[-36, -7]$ relative to TIS's position (*14*) and nucleotide occurrences at the three codon positions (considering the ORF, and the downstream and upstream of ATG). There are 35 features under this measure.

As described above, there are 1,983 features in total and most of them are binary. The continuous features are discretized with a method based on the minimum description length (MDL; ref. *34*) for the feature selection and classification methods.

In the literature, some other measures have also been used to help TIS identification, such as the sequence similarity (*28*), the length of the $5'$ UTR (*35*), and global features (*18*). The first measure (*28*) assumes that extra sequence database, other than the training set and testing set, are available. The second one (*35*) assumes that the sequences are complete and error-free at the side of $5'$ UTR of the cDNA sequences. The third one (*18*) requires the full-length cDNA sequences. However, in this work, most of the sequences used are short and not in full-length. Therefore the features reported in these previous efforts (*18*, *28*, *35*) for TIS prediction were not used in this work.

Feature selection

The numerical data from different features can measure the difference between TISs and negative ATGs to some extent. When more features are used to generate the numerical data, the difference between TISs and negative ATGs will be measured more completely. This will increase the chance to predict TISs accurately. At the same time, however, when the number of the features increases, more noise and redundant information will be introduced in the data set, which may deteriorate the prediction performance and slow down the computation process.

Generally, the different features do not have the same discriminative power for TIS prediction. In order to reduce the noise and redundant information and take advantage of the benefits from more features, we can apply a set of feature selection methods to keep the most significant features for TIS prediction. Then, the computation process could be speeded up with only the significant features. The selected significant features may also be biologically meaningful, which can give biologists clues to choose features to examine in their biological experiments.

In our work, three different feature selection methods were adopted: the Relief method, the chi2-based method, and the information-gain-based method. These three feature selection methods are based on different criteria and represent different categories. We compare their effectiveness for TIS prediction as follows.

1. Relief. This is a filter method to select the relevant features based on statistical methods. It works by randomly sampling an instance and identifying its nearest neighbors in the same class and the other classes respectively. It sets the initial weight of all features to 0 and adjusts the weight of the features based on the sample's nearHit and nearMiss in the data set. The nearHit of instance i is the nearest instance in the data set with the same class label as instance i. The nearMiss of instance i is the nearest instance in the data set with a different class label compared with instance i. With one sample's nearHit and nearMiss in the data set, the weight of each feature decreases by the square of the difference of values of this feature in this sample and its nearHit, and increases by the square of the difference of values of this feature in this sample and its nearMiss. After enough samples are drawn from the original data, Relief selects the feature whose weight is greater than a specified

threshold. An extension of Relief, Relief-F (*36*), was used in this work.

2. Chi2-based method. This method uses chi-square statistic between features and the class to rank the features. The chi2 value of a feature is defined as

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(A_{ij}-E_{ij})^2}{E_{ij}},$$

where m is the number of values of the current feature, n is the number of classes, A_{ij} is the number of instances with the i^{th} value for the current feature in the j^{th} class, $E_{ij} = R_i * C_j/N$ is the expected frequency of A_{ij}, in which R_i is the total number of instances with the i^{th} value for the current feature, C_j is the total number of instances in the j^{th} class, and N is the number of the total instances. The chi2 value measures the difference of the expected frequencies and the actual frequencies in different categories. The larger the chi2 value is, the more significant the feature is. The features are sorted in descending order by their chi2 values.

3. Information-gain-based method. Information gain is an entropy-based measure of the feature quality, which has been used to build the decision tree (*37*) and other applications. It is the difference between the prior entropy of class C and the posterior entropy given values of a feature F:

$$\begin{aligned} &Information\ gain \\ &\quad = -\sum_{C} P(C)\log_2 P(C) \\ &\quad - \sum_{F}\left(-P(F)\times\sum_{C} P(C|F)\log_2 P(C|F)\right) \end{aligned}$$

The larger the information gain is, the more important the feature is to predict the classes. The features are sorted in descending order by their information gain.

Computational methods

In our work, we chose three representative computational methods—decision tree, naïve Bayes, and SVM—from the available computational methods. The decision tree method (*37*) is a *de facto* classification method to evaluate other classification methods and is built recursively based on the information gain of the features. Naïve Bayes is a probability-based classification method, which is simple but practical; it assumes that all the features are independent of each other given the class. SVM (*38*) selects the fewest instances as the support vectors with the largest margin in the feature space, which is possibly the classification method with the best prediction result up to date, although it sometimes suffers in the presence of noisy data. Based on past results from related work, these three methods can be the representative data mining methods for TIS prediction (*20*). A well-known machine learning package Weka (*39*) has implemented all the three methods. We ran these three methods on the same training and test data under the Weka environment. After feature selection, the three methods were applied to the selected data to predict TISs.

Evaluation measures

Prediction performance is measured by sensitivity (*Se*), specificity (*Sp*), accuracy (*Acc*), and MCC in our work. Let *TP* be the number of the real positive ATGs predicted as positive, *FP* the number of the real negative ATGs predicted as positive, *TN* the number of the real negative ATGs predicted as negative, and *FN* the number of the real positive ATGs predicted as negative. *Se* is defined as $TP/(TP+FN)$, the percentage of the correctly-predicted positives in the total real positives. *Sp* is defined as $TN/(TN+FP)$, the percentage of the correctly-predicted negatives in the total real negatives. *Acc* is defined as $(TP+TN)/(TP+TN+FP+FN)$, the percentage of the total correctly-predicted instances in all the instances. MCC is defined as:

$$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FN)\times(TP+FP)\times(TN+FP)\times(TN+FN)}}$$

In these criteria, *Se* only measures the prediction performance on the positive cases, and *Sp* only measures the prediction performance on the negative cases. With the biased data in our case, they are not good enough to measure the performance of a model (also for the total accuracy). MCC takes into account both positive and negative cases and is also suitable for the biased data. Therefore it was used as the main measure to evaluate our approach in the experiments.

Acknowledgements

This research was supported by Research Grant No. BM/00/007 from the Biomedical Research Council (BMRC) of the Agency for Science, Technology, and Research (A*Star) and the Ministry of Education in Singapore. We thank the anonymous reviewers for some very helpful comments, and thank Artemis Hatzigeorgiou, A. Pedersen, H. Nielsen, A. Nadershahi, S.C. Fahrenkrug, and L.B.M. Ellis, who kindly shared their original cDNA sequence sets.

References

1. Cigan, A.M., *et al.* 1988. tRNAi(met) functions in directing the scanning ribosome to the start site of translation. *Science* 242: 93-97.
2. Kozak, M. 1989. The scanning model for translation: an update. *J. Cell Biol.* 108: 229-241.
3. Kozak, M. 1992. A consideration of alternative models for the initiation of translation in eukaryotes. *Crit. Rev. Biochem. Mol. Biol.* 27: 385-402.
4. Kozak, M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 44: 283-292.
5. Kozak, M. 1996. Interpreting cDNA sequences: some insights from studies on translation. *Mamm. Genome* 7: 563-574.
6. Kozak, M. 2002. Pushing the limits of the scanning mechanism for initiation of translation. *Gene* 299: 1-34.
7. Liu, H. and Wong, L. 2003. Data mining tools for biological sequences. *J. Bioinform. Comput. Biol.* 1: 139-167.
8. Kozak, M. 1978. How do eucaryotic ribosomes select initiation regions in messenger RNA? *Cell* 15: 1109-1123.
9. Pedersen, A. and Nielsen, H. 1997. Neural network prediction of translation initiation sites in eukaryotes: prespectives for EST and genome analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 5: 226-233.
10. Hatzigeorgiou, A.G. 2002. Translation initiation start prediction in human cDNAs with high accuracy. *Bioinformatics* 18: 343-350.
11. Nadershahi, A., *et al.* 2004. Comparison of computational methods for identifying translation initiation sites in EST data. *BMC Bioinformatics* 5: 14.
12. Derst, C., *et al.* 2000. Prediction of human translational initiation sites using a multiple neural network approach. *Int. J. Comput. Syst. Signal* 1: 169-179.
13. Stormo, G.D., *et al.* 1982. Use of the “Perceptron” algorithm to distinguish translational initiation sites in *E. coli. Nucleic Acids Res.* 10: 2997-3011.
14. Salamov, A.A., *et al.* 1998. Assessing protein coding region integrity in cDNA sequencing projects. *Bioinformatics* 14: 384-390.
15. Wang, Y., *et al.* 2003. Recognizing translation initiation sites of eukaryotic genes based on the cooperatively scanning model. *Bioinformatics* 19: 1972-1977.
16. Zien, A., *et al.* 2000. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics* 16: 799-807.
17. Li, H. and Jiang, T. 2004. A class of edit kernels for SVMs to predict translation initiation sites in eukaryotic mRNAs. In *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology*, pp. 262-271, San Diego, USA.
18. Li, G., *et al.* 2005. Translation initiation sites prediction with mixture Gaussian models in human cDNA sequences. *IEEE Trans. Knowl. Data Eng.* 17: 1152-1160.
19. Wang, Y., *et al.* 2003. Recognition of translation initiation sites of eukaryotic genes based on an EM algorithm. *J. Comput. Biol.* 10: 699-708.
20. Zeng, F., *et al.* 2002. Using feature generation and feature selection for accurate rrediction of translation initiation sites. In *Proceedings of 13th International Conference on Genome Informatics*, pp. 192-200, Tokyo, Japan.
21. Salzberg, S.L. 1997. A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput. Appl. Biosci.* 13: 365-376.
22. Kozak, M. 1987. An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* 15: 8125-8148.
23. Agarwal, P.K. and Bafna, V. 1998. Detecting nonadjoining correlations within signals in DNA. In *Proceedings of the Second Annual International Conference on Research in Computational Molecular Biology*, pp. 2-8, New York, USA.
24. Li, G. and Leong, T.Y. 2004. A feature-based data mining approach to improve translation initiation site prediction (Abstract). In *Proceedings of the World Congress on Medical Informatics*, San Francisco, USA.
25. Kira, K. and Rendell, L. 1992. The feature selection problem: traditional methods and a new algorithm. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pp. 129-134, San Jose, USA.
26. Liu, H. and Setiono, R. 1995. Chi2: feature selection and discretization of numeric attributes. In *Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*, pp. 388-391, Washington, USA.
27. Liu, H., *et al.* 2004. Using amino acid patterns to accurately predict translation initiation sites. *In Silico Biol.* 4: 255-269.
28. Nishikawa, T., *et al.* 2000. Prediction whether a human cDNA sequence contains initiation codon by com-

bining statistical information and similarity with protein sequences. *Bioinformatics* 16: 960-967.

29. Pontius, J.U., *et al.* 2002. UniGene: a unified view of the transcriptome. In *The NCBI Handbook*. Bethesda, USA.

30. Kozak, M. 1987. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J. Mol. Biol.* 196: 947-950.

31. Fickett, J.W. 1996. The gene identification problem: an overview for developers. *Comput. Chem.* 20: 103-108.

32. Stormo, G.D. 1990. Consensus patterns in DNA. *Methods Enzymol.* 183: 211-221.

33. McGeoch, D.J. 1985. On the predictive recognition of signal peptide sequences. *Virus Res.* 3: 271-286.

34. Fayyad, U.M. and Irani, K.B. 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of 13th International Joint Conference on Artificial Intelligence*, pp. 1022-1027, Chambery, France.

35. Rogozin, I.B., *et al.* 2001. Presence of ATG triplets in $5'$ untranslated regions of eukaryotic cDNAs correlates with a “weak” context of the start codon. *Bioinformatics* 17: 890-900.

36. Kononenko, I. 1994. Estimating attributes: analysis and extensions of RELIEF. In *Proceedings of the European Conference on Machine Learning*, pp. 171-182, Catania, Italy.

37. Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, USA.

38. Burges, C.J.C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining Knowl. Discov.* 2: 121-167.

39. Witten, I.H. and Frank, E. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, San Francisco, USA.