Prediction of Initiation Codons in cDNA Fragments Using Statistical Information and Similarity with Protein Sequences

Tetsuo Nishikawa nisikawa@hri.co.jp
Toshio Ota ota@hri.co.jp
Takao Isogai isogai@hri.co.jp

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

1 Introduction

More than 100 millions of human cDNA fragment sequences have already been published by several EST projects. These ESTs, however, are often incomplete in the 5'-region of full length cDNA sequences, whereas it is important to obtain the clones including the intact protein coding sequences (complete clones) if one is to analyze gene function. To obtain complete clones efficiently, it is necessary to develop programs which select them efficiently from given cDNA fragments as well as to develop effective methods for generating complete clones [1]. For this purpose, we have previously reported on a computer program, ATGpr [2], which estimates the reliability of the prediction that each ATG is a true initiation codon in a given DNA sequence using statistical information. We evaluated the accuracy of the initiation codon prediction by ATGpr in complete cDNA sequences in that study. In practice, however, it is also of importance to know whether a fragment sequence of cDNA, such as ESTs, contains an initiation codon or not. Since ATGpr uses only statistical information derived from the cDNA sequence, the accuracy can be expected to improve when information on similarity with other proteins are used in the prediction. In this paper, we propose a new method which uses both statistical and similarity information to obtain higher prediction accuracy for fragment sequences of cDNA. We evaluate the accuracy of this method using UniGene data, a source of cDNA fragment sequences.

2 Methods

2.1 Dealing with UniGene

The presence of initiation codons in 5'-EST sequences randomly sampled one from each human UniGene cluster (Build28, 6446 known gene clusters), representative 5'-ESTs, was judged by comparing with the representative mRNA sequences of the clusters. Here, we define a sequence as "complete" when it includes the initiation codon. The BLASTN program was used for this comparison.

2.2 Method for prediction of a fragment with the initiation codon

1) The ATGpr prediction was performed for each representative 5'-EST. The sequence was judged as complete when the maximum score in all ATG codons in the sequence was greater than a score threshold.

2) The prediction using ATGpr combined with information on similarity with known protein sequences including those from other organisms was performed for each representative 5'-EST. Each representative 5'-EST was compared with the OWL database (Ver.30.0) using the BLASTX

program, and alignments with identities of 30-95% and with a consensus length longer than 100 bases were selected. We judged it as complete when the DNA sequence was longer than the protein sequence at the 5'-end in at least one alignment and when the maximum ATGpr score was greater than a given threshold.

2.3 Evaluation Method

The specificity and the sensitivity of the two predictions described above were determined at score thresholds varying from 0 to 1. The specificity and the sensitivity is defined as Nhc/Nh and Nhc/Nc, respectively. Here, Nc is the total number of complete sequences, Nh is the number of the sequences predicted as complete with the score greater than a given threshold, and Nhc is the number of correctly predicted sequences in those sequences predicted as complete. The value where specificity equals sensitivity as function of score threshold is defined as Spn.

3 Results and Discussion

The Spn for fragment sequences when using only ATGpr was 35%. This is significantly lower than that for complete sequences in the previous report; this may be due to the fact that the portion of the "complete" ESTs in UniGene is very small (estimated to be 15%). The Spn when using ATGpr combined with the similarity information was 55%, that is the Spn value increased by 20% by introducing the similarity information. When this method is applied to unknown ESTs the improvement in Spn will of course be less than 20%, because the number of proteins similar to unknown genes are most often fewer than the case with known genes. However, with the rapid increase in the number of protein sequences, one can expect significant improvements in the Spn value.

References

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

[2] Salamov, A.A., Nishikawa, T., and Swindells, M.B., Assessing Protein Coding Region Integrity in cDNA Sequencing Projects, *Bioinformatics*, 14:384–390 , 1998.