from a nucleotide database, GenBank, exons typically arise at most every 2,000 bases on average in long genes. However in this sequence an exon arises every 16,000 bases on average. The fact possibly cause software to find false exons. One of our interests is to examine how many false positives these tools will produce for this sequence.

3 Methods

As the sequence is too long to input into the tools, we set a 80,000 base long window starting from the first base, and transferred the sequence in the window. Each time the window shifted half size of itself, we gave the sequence in the window to the tools. Then we repeated the procedure. Therefore, most bases were analyzed twice. This way to analyze keeps the both sides of any fragments long enough except for edges of the whole sequence. The results of each software for the windows were combined later.

4 Results and Discussion

We performed a biochemical experiment about some predicted exons with high scores. The experiments indicated that two exon candidates predicted by GRAIL are certain repetitive sequences. Then we regard these predictions as true. Based on the fact, FEX is more sensitive than GRAIL, but less specific. As for sensitivity, GRAIL predicted 80% (15/19) of the known exons and FEX predicted 95% (18/19). With regard to the specificity, 64% (30/47) of the predicted exons by GRAIL are false, and 82% (81/99) by FEX are false.

The numbers of the predicted exons of both tools in the forward strand are much less than one hundred fifty, which is the expectation estimated from the length of the sequence and a ratio (0.5 exon/kb). The ratio is what we grossly calculated from many human genes stored in the GenBank in advance, as a ratio of the number of exons to the sequence length. The fewness of the candidate exons is praiseworthy but it is still many.

We studied the dependence of specificity on the predicted scores for GRAIL and FEX. We found that the output scores are very correlated with the accuracy. Then we propose the guidelines of a 60 percent correct criterion to make biochemical experiments efficiently. They are score 80 for GRAIL, and score 9 for FEX.

5 Summary

In this study, FEX is more sensitive for both exons and splicing sites, but less specific than GRAIL. The numbers of the predicted exons were much less than our estimations. We propose the guidelines of a 60 percent correct criterion. They are score 80 for GRAIL, and score 9 for FEX. Although both tools are not complete yet, they are practical tools for DNA sequences longer than 100,000 bases, or exon-sparse sequences.

Acknowledgments

This work is partially supported by Grant-in-Aid for Scientific Research on Priority Areas, "Genome informatics" from the Ministry of Education, Science, Sports and Culture, Japan and by Science and Technology Agency, Japan.

References

[1] Xu,Y., Einstein,J.R., Mural, R.J., Shah,M., and Uberbacher,E.C. "An Improved System for Exon Recognition and Gene Modeling in Human DNA Sequences" ISMB, 1994.

[2] Solovyev,V.V., Salamov,A.A., and Lawrence, C. B. "The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames" ISMB, 1994.

Evaluation of exon prediction tools using a long DNA sequence data

Katsuhiko Murakami [1 2] Shiho Tsukuni [1]
Toshihisa Takagi [1] Masahira Hattori [1]

[1] Human Genome Center
Institute of Medical Science, University of Tokyo
{katsu, takagi}@ims.u-tokyo.ac.jp

[2] Central Research Laboratory, Hitachi Ltd.

Abstract

We have evaluated the ability to locate coding regions of two exon prediction software, GRAIL and FEX, using a long (about 301k bases) genomic DNA sequence. We performed an experiment to check the correctness of the exon candidates with high scores. FEX was more sensitive but less specific than GRAIL. The numbers of the exons predicted by both tools were much less than our simple estimation from the sequence length. To reduce more unreliable candidates, we proposed guidelines for users. If one uses the guidelines, both tools would be more practical even for DNA sequences longer than 100,000 bases.

1 Introduction

Locating and identifying genes from a newly sequenced genomic DNA by computing is one of the major challenges in bioinformatics. To date, several systems to predict protein coding regions or modeling gene structure have been developed in this decade. Although the performances of these systems have been compared in the papers which each system was presented in, the objective comparison among different systems has not been reported.

If a predicted exon was not described in the public databases, it has been defined as a false positive. To know the real correctness of the prediction, we must perform experimental confirmations about the predicted exon.

In addition, the relation between the predicted score and the error rate of these systems are not announced in some systems. A user don't know how reliable the prediction is. Accordingly we need to know the relation between the score and the reliability of the prediction.

We evaluated two systems which identify coding regions, GRAIL-2 [1] (hereafter we call GRAIL) and FEX [2]. GRAIL is a system that uses multiple sensors and neural networks. FEX uses an algorithm based on the distribution of oligonucleotides and linear discriminant analysis. The reasons why we selected these systems are that they are the latest systems and available by E-mail.

2 Data

Recently we have determined the complete sequence of amyloid precursor protein (APP) gene. The sequence is about 301,000 bases in length. We used it as an input for the following reason. In our simple calculation

[1]村上 勝彦、津國 志保、高木 利久、服部 正平：東京大学医科学研究所ヒトゲノム解析センター，〒 108 東京都港区白金台 4-6-1
[2]村上 勝彦：（株）日立製作所中央研究所，〒 185 国分寺市東恋ヶ窪 1-280