A large-scale GenBank search of Expressed Sequence Tags using rapid identity searching program for DNA sequences

T. Nishikawa
nisikawa@crl.hitachi.co.jp

K. Nagai
k-nagai@crl.hitachi.co.jp

Central Research Laboratory, Hitachi, Ltd
1-280 Higashi-koigakubo, Kokubunji-shi, Tokyo 185, Japan

Abstract

We have developed a program for rapid identity-searching of DNA sequences allowing several percentages of sequencing error rates. The program was applied to a large-scale searching of Expressed Sequence Tags (ESTs) against the GenBank sequences, and from this searching results the error information of ESTs was obtained. The 15,666 sequences of human ESTs were searched in the primate division in GenBank release 80 within 23.3 hours that is only one-thirty of the time needed when FASTA is used. The total error rate 2.45 percent was obtained from the alignments between the ESTs and the primate sequences satisfying the identity-conditions.

1 Introduction

The number of registered DNA sequences in GenBank is now doubling every twenty months. Such a rapid growth in the quantity of DNA sequences in databases has given rise to the need for a method by which new sequences can be quickly compared with the sequences already determined. We have developed a program for rapid identity-searching of DNA sequences allowing several percentages of sequencing error rates [1]. We applied this program to a large-scale identity-searching of Expressed Sequence Tags (ESTs) against the GenBank sequences. Although ESTs, that are now being determined in large-scale, are considered to be very useful for mapping or other applications, they are estimated to have many sequencing errors due to single run sequencing. The actual sequencing error included in the ESTs in GenBank, however, has not yet been reported. So, we estimated the error rate in the ESTs in GenBank by using the results of the large-scale searching of the ESTs against gbpri.seq in GenBank.

西川哲夫，永井啓一：（株）日立製作所中央研究所，〒 185 国分寺市東恋ヶ窪 1-280

2 Methods

The 15,666 sequences of human ESTs that contain no Alu repeat were searched in the primate division (29,168 sequences, 28,050,304 bases) in GenBank release 80. The identity condition used was that hit rate is more than 93 percent in 50 bases regions and both ends in the alignments of the sequences are matched allowing up to 5 bases mismatching from ends. The primate sequences are usually determined by runs in both directions and their error-rates are therefore estimated to be much smaller than those of the ESTs sequences. Therefore, we can obtain the sequencing error information of ESTs from the alignments between the query EST sequence and the primate sequence that meets the identity conditions. We identified the differences in the alignments and regarded these differences as errors in ESTs.

3 Results and Discussion

The search of 15,666 ESTs in primate division in Genbank took 23.3 hours. This is only one-thirty of the time needed when FASTA is used. From this search, 1,143 of ESTs were found to have identical sequences in primate division in Genbank. From the alignments between ESTs and the identical sequences, the average total error rate in 1,143 ESTs was obtained to be 2.45 percent. The insertion-deletion error rate, N-error rate, and substitution error rate excluding N-error rate was also respectively obtained 0.67 percent, 0.94 percent, and 0.88 percent. These results are very close to the sequencing error rates obtained from the genome sequencing using auto-sequencer [2]. The dependencies of the error rates on the base length from the 5' end were investigated and the followings were obtained. The insertion error rate of T and G base increased very rapidly as the base length increases in the base length region more than 200 bases. The deletion error rate of G and C base increased as the base length decreases in the base length region less than 100 bases. And the substitution error rate from A to G, from G to A, from C to T, from T to C were found to be greater than the other substitution error rates in the entire base length regions. This error rate information can be useful for the applications of ESTs, for example, for the design of PCR primers from EST sequences.

4 Acknowledgement

We would like to thank Professor Takagi and Dr. Ogiwara at the University of Tokyo for their helpful advice on the program algorithm.

References

[1] T. Nishikawa, S. Hiraoka, N. kasahara and K. Nagai, Rapid identity searching program for DNA sequences and its applications to cDNA grouping. *Genome Informatics Workshop*, 5, 194 (1994)

[2] F. Khurshid, and S. Beck, Error Analysis in Manual and Automated DNA Sequencing. *Analytical Biochemistry*, 208, 138 (1993)