Rapid identity searching program for DNA sequences and its applications to cDNA grouping

T. Nishikawa S. Hiraoka
N. Kasahara K. Nagai
nisikawa@crl.hitachi.co.jp hiraoka@crl.hitachi.co.jp
kasahara@crl.hitachi.co.jp k-nagai@crl.hitachi.co.jp

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

Abstract

We developed a program that determines whether or not a query sequence is included in a database within a permitted matching error rate. It consists of two steps: bit-table filtration and dynamic programming matching. The bit table filtration quickly excludes many sequences that have no relation to the query sequence and identifies the sequences without missing that match the query sequence within the given error rate. The application of this program to large-scale human cDNA grouping showed that it took only one tenth the time required by FASTA for grouping all human cDNA.

1 Introduction

The need for rapid searching for identical DNA sequences allowing for a certain sequencing error rate, is increasing as DNA sequencing projects increase in scale. We show two examples in which rapid identity searching is important. First, the identity searching in GenBank that determines whether or not the query sequence is included in the database is the most basic procedure when new sequences are determined. Second, the assembling problem, especially EST assembling requires the rapid comparing between every pair of sequences to find correct overlaps. The rapid identity searching is also necessary for this purpose. Conventional tools for DNA similarity searching, Smith Waterman(S-W) method [1], FASTA [2], BLAST [3], are optimized for homology searching, and not for the identity searching. The speed and the searching conditions therefore is not enough or not appropriate using these conventional tools for identity searching. So, the target of our study is to develop a tool for rapid identity searching allowing several percentages of error rate, which can be applicable for these applications.

西川哲夫，平岡進，笠原直子，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

It consists of two steps: filtration and dynamic programming matching. An algorithm using a bit table of the k-tuples in the database sequences is introduced for the filtering. The bit table of the k-tuples has the information of whether or not each k-tuple is included in each database sequence. Given- length sub-sequences in the query sequence are compared with the bit table to calculate the score for each database sequence. The way of picking up sub-sequences can be chosen adequately depending on the applications. The score is obtained by counting the number of k-tuples commonly included in both the sub-sequence and each database sequence. The database sequences that have scores greater than a threshold value are sent to DP matching. The threshold values are defined as the minimum hit k-tuple numbers with a given allowed error rate. By this definition, the searching of sequences within a permitted error rate can be performed without missing. The bit table search therefore quickly excludes many sequences that have no relation to the query sequence and identifies the sequences without missing that match the query sequence within the given error rate.

3 Results and Discussion

We implemented the algorithm on SPARC station 10. In the first application to GenBank identity searching, we picked up 45 base sub-sequences at every certain distance D in a query to deal with partial overlapping of sequences. Using optimized parameters, size of k-tuple=9 and D=9, we could search the sequences within 8 percent error rate in the Primate devision in GenBank in 2.5 seconds, which is one-tenth the time obtained using FASTA. Second, this program was applied to large-scale cDNA assembling. Human cDNA sequences in the dbEST in GenBank were grouped by searching the EST sequences with taking every sequence as a query. We picked up 45 base sub-sequences at both ends in a query. It took 10 hours to group fifteen thousand sequences. This is one-tenth the time required by FASTA. Furthermore, all group members were identified, unlike BLAST that tends to miss group members.

4 Acknowledgement

We would like to thank Dr. Okubo in Osaka University and Dr. Ogiwara in Tokyo University for the helpful discussion with them about the application of our program.

References

[1] T. F. Smith and M. S. Waterman, Identification of common molecular subsequences. *J. Mol. Bol.*,147,195(1981)

[2] W. R. Pearson and D. J. Lipman, Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*,85,2444(1988)

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.*,215,403(1990)