Highly Sensitive Homology Search Methods on Parallel Computer

Naoko Kasahara
kasahara@crl.hitachi.co.jp
Susumu Hiraoka
hiraoka@crl.hitachi.co.jp
Ryotaro Irie
r-irie@crl.hitachi.co.jp
Keiichi Nagai
k-nagai@crl.hitachi.co.jp
Central Research Lab. Hitachi.Ltd.
1-280 Higashi-koigakubo, Kokubunji-shi, Tokyo 185, Japan

Abstract

We have developed two kinds of highly sensitive homology search methods based on Smith-Waterman-like algorithm. One is direct comparison between DNA sequence and amino acid sequence. The other is comparison between DNA sequences through translated amino acid sequences. Both methods consider gaps in amino acid and nucleotide sequence levels simultaneously. Although these methods attain higher sensitivity and specificity than BLASTX or TBLASTX, they need huge calculation time to perform dynamic programming calculation. We developed parallel computation programs on the parallel computer, Hitachi SR2201, to realize practical computation time.

1 Introduction

DNA database size is increasing exponentially. On the other hand, amino acid sequence comparison plays an important role in protein function analysis. It is preferable to compare amino acid sequences rather than comparing protein coding DNA sequences. BLASTX and TBLASTX were developed for that aim. Both methods translate a DNA sequence into six frame amino acid sequences and compare the sequences with known protein amino acid sequences or six frame amino acid sequences of another DNA sequence. However, since they don't consider any gaps in sequences, the sensitivity and specificity of database search can be degraded in some cases, especially, in case of using EST sequences. So we have developed two kinds of homology search methods based on Smith-Waterman-like algorithm. One is direct comparison between DNA sequence and amino acid sequence [1]. The other is comparison between DNA sequences through translated amino acid sequences [2]. Both methods consider gaps in amino acid and nucleotide sequence levels simultaneously and they can attain higher sensitivity and specificity than BLASTX or TBLASTX. However our methods use dynamic programming calculation, they must take huge calculation time, so we developed parallel computation programs on the parallel computer, Hitachi SR2201, to realize practical computation time.

2 Method and Results

First, we explain the parallel computation program of the first method, that is , the direct comparison between DNA sequence and amino acid sequence. This program is composed by three steps, which are, 1) translation DNA sequence to amino acid sequence nucleotide-by-nucleotide, 2) comparing between translated amino acid sequence and known amino acid sequence, allowing gaps to exist in either sequence, and 3) calculating and displaying the alignment of these sequences. In the second step, we consider seven paths instead of three considered in the conventional Smith-Waterman algorithm [3], in order to allow gaps in amino acid and nucleotide sequence levels simultaneously. We devide these steps for parallel processes. The first and second steps run on many processors and the third step runs on only one processor which controls other processors. One processor, called master, gets the sequences from the database and sends the sequence to other processors, called slave. After receiving the sequences, slave processors calculate the homology score between query sequence and the received sequence from the database and stack the result. When master finishes getting the sequence from the database, slaves send their results to master and the master receives all results and sorts those. Finally, the master calculates the alignments between the query sequence and the homologous sequences, whose homology scores are in the top part of the result.

We developed the parallel computation program on SR2201 and examined the search speed. As a query sequence, we chose the rice EST sequence, RICC2791A, from Genbank (rel.99), whose length is 348 base, and as a database, we used Swissprot (rel.34, including 59,021 sequences, and 21,210,388 residues). Under the above condition and in case of processor number is 256, our method takes 90 seconds to compare the query sequence through all sequences in Swissprot database. Under the same condition, when this method runs on the stand alone workstation SparcStation20, it takes about 5 hours to analyze through all database of Swissprot. The parallel program realizes about 200 times higher calculation speed than the stand alone processor. On the other hand, BLASTX takes about 60 seconds under the same condition. Our method realize almost the same speed of BLASTX in case of 256 processors. As for search quality, our method can detect all the twenty two related sequences, but BLASTX detects fifteen of them.

Next, we describe the second method, that is, comparison between two DNA sequences after translating into amino acid sequences. The parallel computation procedures are almost the same as the first method. In this case, both DNA sequences are translated into amino acid sequences nucleotide by nucleotide. And, in the dynamic programming calculation, the eleven paths are considered to allow gaps in amino acid and nucleotide sequence levels simultaneously. As a query sequence, we chose an *A.thaliana* EST sequence, ATTS0048, whose length is 248 bases and as a database, we made test database which is composed by rice and *A. thaliana* EST sequences and the size is about 8.5 M base. When we use this data set, our method takes one minute on 128 processors. This search speed is reasonable from the difference of the calculation amounts of the first and second methods.

References

[1] N. Kasahara, S. Hiraoka, and K. Nagai; Direct Comparison Between DNA and Amino Acid Sequences based on a Dynamic Programming Method, *GIW '96*, pp.202–203, 1996.

[2] R. Irie, N. Kasahara, S. Hiraoka, and K. Nagai; Codon-sensitive Comparison of DNA sequences contains insertions/deletions and statistical significance of the similarity scores, *GIW '97*, 1997.

[3] T. F. Smith, and M. S. Waterman; Identification of common molecular subsequences, *J. Mol. Biol.*, Vol.147, pp.195–197, 1981.