2 Method

We present an algorithm for DNA sequence comparison which translates the sequences most reliably and compares the translated sequences. The method enables us to find protein sequence similarity in DNA sequences even if we do not know the protein sequences which are coded in the DNA sequences. The algorithm produces temporal DNA sequence for each original DNA sequence. They are divided into codons from the beginnings and translated into protein sequences. Each temporal and original DNA sequence are compared using DNA scoring system. Translated sequences are compared using protein scoring system. The sum of the scores are calculated for all possible temporal DNA sequences. The temporal DNA sequences are optimized to make the sum of the scores maximum. The translation of the optimized temporal DNA sequences are the most reliable protein sequences. The production of the temporal DNA sequences and the maximization of the sum of the scores are achieved by the dynamic programming method. We modified the algorithm of Smith and Waterman to be used to compare translated sequences [2].

3 Result

We present an algorithm for DNA sequence comparison which selects the most reliable translation and compares the translated sequences. We are able to find protein sequence similarity in DNA sequences even if we do not know the protein sequences which are coded in the DNA sequences. We are also able to find gaps in DNA sequences which alter reading frames. The method enables us to detect distant relations between DNA sequences, while the comparison of the same sequences by conventional methods give us only false positives.

References

[1] M. Dayhoff, *"Atlas of Protein Sequence and Structure"*, Vol. 5, Suppl.3, 345(1978)

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

DNA sequence comparison based on amino acid similarity

S. Hiraoka
hiraoka@crl.hitachi.co.jp

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

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

Abstract

DNA databases are growing exponentially. Sequence similarities are often the most valuable information we can get from DNA databases. Especially for protein-coding sequences, comparison of translated sequences give us clues to protein function. However gaps in DNA sequences prevent us from translation and force us to compare them as they are. We present an algorithm for DNA sequence comparison which translates the sequences most reliably and compares the translated sequences. The method enables us to find protein sequence similarity in DNA sequences even if we do not know the protein sequences which are coded in the DNA sequences.

1 Introduction

Most of protein sequences are determined from DNA sequences. DNA database size is increasing exponentially. If we can translate all protein-coding sequences in DNA databases it will make the most up-to-date protein databases. However translation is impossible for some of the sequences because of difficulty of predicting coding regions and alteration between reading frames caused by gaps. Sequence comparison plays an important role in protein function analysis. They may provide the information about the structure, function and evolution of the protein. It is preferable to compare protein sequences rather than comparing protein-coding DNA sequences. The reasons are the degeneracy of genetic code and the scoring system. The degeneracy of genetic code can cause partial mismatches between DNA sequences coding an identical protein. Furthermore, the scoring system for protein is more sophisticated than the scoring system for DNA that has only two scores, matches and mismatches [1].

平岡進、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