Rapid similarity search algorithm of nucleic acid database

S. Hiraoka
hiraoka@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

Smith and Waterman dynamic programming algorithm is too computer intensive to be used for database search. Most of the common database search programs use approximate scores to reduce search time in stead of Smith and Waterman score. They do not consider all possible alignments and may miss some sequence similarities. We present an algorithm for database search which uses score table of k-tuples. The method calculates the upper limit of Smith and Waterman score for each database sequence in nearly interactive time. Calculated each upper limit of Smith and Waterman score prevents us from missing high sequence similarities.

1 Introduction

With rapid development of nucleic acid database, the need for efficient methods of searching database is increasing. There is the method of Smith and Waterman [1], which is the most accepted method of finding local similarities. The method considers all possible alignments by using dynamic programming algorithm ,and will find optimal alignments which give the highest score. The time required for computation is the order of NM, where N and M are the lengths of the sequences compared. The method is too computer intensive to be used for database search.

Currently, there are several other methods for database search, such as FASTA [2] and BLAST [3]. These methods use approximate scores to reduce search time in stead of Smith and Waterman score. They do not consider all possible alignments and may miss some sequence similarities.

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

2 Method

We present an algorithm for database search which uses score table of k-tuples. We here describe how the method is applied for global comparison for simplicity. The method is also applicable for local comparison which Smith and Waterman algorithm computes. To compare a query with a sequence in a database, we form a two-dimensional array whose rows and columns are the bases of the query and the database sequence respectively. Every possible alignment can be represented by pathways through the array. The best alignment is the pathway for which the score is largest. Dynamic programming algorithm is computer intensive because it calculates scores for every pathways.

The first step in the algorithm is the division of the query into k-tuples. This makes the problem into smaller problems, aligning each k-tuple with the database sequence. The optimal alignment of the query with the database sequence is the connection of the alignments of k-tuples with the database sequence. The sum of the scores between each k-tuple and the database sequence give the upper limit of the score between the query and the database sequence. We prepare the score table of k-tuples which consists of the scores between every possible k-tuple and every database sequence before database search. Thus we can calculate the upper limit of the score between the query and each database sequence only referring part of the score table of k-tuples. Because our algorithm calculates the upper limit of dynamic programming score, we can select all sequences whose dynamic programming scores are more than a certain value without missing.

3 Result

We have implemented the algorithm on a workstation(SPARC station 10). It contains the primate database whose size is one fifth the size of total GenBank database. We are able to search the primate database with a 200-residue query in a few seconds. The system will be expanded to include the complete GenBank database.

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)