International Journal of Bioinformatics Research
ISSN: 0975–3087, E-ISSN: 0975–9115, Vol. 3, Issue 1, 2011, pp-161-166
Available online at http://www.bioinfo.in/contents.php?id=21

A FAST SEARCH METHOD FOR DNA SEQUENCE DATABASE USING HISTOGRAM INFORMATION

QIU CHEN[1*], KOJI KOTANI[2], FEIFEI LEE[1], and TADAHIRO OHMI[1]
[1]New Industry Creation Hatchery Center, Tohoku University, Japan
[2]Department of Electronics, Graduate School of Engineering, Tohoku University, Japan
*Corresponding Author: Email- qiu@fff.niche.tohoku.ac.jp

Received: February 01, 2011; Accepted: February 17, 2011

Abstract- DNA sequence search is a fundamental topic in bioinformatics. The Smith-Waterman algorithm achieved highest accuracy among various sequence alignment tools, but it usually spends much computational time to search on large DNA sequence database. On the contrary, BLAST and FASTA have improved the search speed by using heuristic approaches, by there is a possibility of missing an alignment or giving inaccurate output. This paper presents an efficient hierarchical method to improve the search speed while the accurate is being kept constant. For a given query sequence, firstly, a fast histogram based method is used to scan the sequences in the database. A large number of DNA sequences with low similarity will be excluded for latter searching. The Smith-Waterman algorithm is then applied to each remainder sequences. Experimental results show the proposed method combining histogram information and Smith-Waterman algorithm is a more efficient algorithm for DNA sequence search.
Key words - Fast search, DNA sequence, Histogram information, Smith-Waterman algorithm

Introduction

Comparison of genome sequences (DNA, mRNA and protein) is the most important task in the life science area. There are 4 types of the DNA nucleotides, namely, A (adenine), C (cytosine), G (guanine) and T (thymine), which are utilized to encode DNA. If gene A and gene B have high homology, it is surmisable that the function of gene A is similar to that of gene B.

Normally, when a new DNA or protein sequence is determined, it would be compared to all known sequences in the annotated databases such as GenBank [7], EMBL [1], and DDBJ [2], etc. Because the database is very large, a lot of algorithms are studied and used for the speeding-up of data search. Needleman and Wunsch presented the Needleman-Wunsch algorithm [3], which calculates similarities between sequences by the dynamic programming, and Smith-Waterman algorithm is the improved approach [4].

However, it takes much time to search data with these algorithms because they require too many amounts of calculation. Furthermore, the enormous quantity of data has been accumulated in the database like GenBank, EMBL, and DDBJ, etc., and the volume of data of Genome Database still increases in exponential as shown in Figure 1 [8]. Blast [5], FASTA [6] and PatternHunter [9][10] are three rapid heuristic algorithms are regularly used for searching protein and DNA sequence databases. The idea in these tools is to find subsequences that share some patterns called as filtration techniques.

However, while BLAST and FASTA have improved the search speed by using heuristic approaches, there is a possibility of missing an alignment or giving inaccurate output. Thus, many researches have been trying to improve both the search time and the precision.

In this paper, we propose an efficient method combining histogram features and Smith-Waterman dynamic programming algorithms [4] in order to improve speed and precision. The effects will be demonstrated by using GenBank sequence data.

This paper is organized as follows. At first, we will introduce the proposed algorithm in detail. Then experimental results will be discussed. Conclusions of our research will be given in the last section.

Proposed method

When using classical Smith-Waterman algorithm [4] to align two sequences, searching and comparing a query sequence with the databases with large size of sequences is complicated and requires for more time and spaces complexity. Therefore, the need of mechanism to discard the unrelated or irrelevant sequences compared to a query is highly demanded. In this paper, we present a new search method for DNA sequence matching in a large size

of DNA sequence databases. Histogram features of sequences are firstly used to compare the query sequence with the sequences in database and similarity scores would be obtained. Only the sequences whose similarities exceeded a given threshold are then aligned using exhaustive Smith-Waterman dynamic programming algorithm (4).

Figure 2 shows the processing steps of our proposed method. When an unknown query base sequence is input, it will firstly be divided into *n* parts. It is thought that more robust features can be extracted if order information of the base sequence is added. For each separate partial sequence, it will be divided into small sequence, for instance, ACT and CGG, etc. A small sequence can be considered as a three dimensional vector. This processing overlaps over all the sequence. After that, the histogram feature is calculated. There are only 4 types of DNA bases, so the number of combination of 3-dimensional vector is 64. A reference table with the size of 64 is shown in Figure 3, by which the index number of the 3-dimensional vector is very easy and fast to be determined. The number of vectors with same index number in each separate partial sequence is counted and feature vector histogram is easily generated, and it is used as histogram feature of the separate partial sequence. As the input query base sequence is divided into *n* parts, the histograms of *n* parts are generated. On the other hand, the histogram features are also extracted from the DNA sequence in the database using the same method respectively.

The histogram generated from each partial sequence is compared with the histograms from the same partial sequence in the database by calculating similarities (s_i) between them (as shown in formula (2)). Then the integrated similarities (S) are obtained by averaging as shown in the following formula (1).

$$S = \frac{\sum s_i}{n}, i = 1,...i,... \quad n \qquad (1)$$

$$s_i = 1 - \frac{\sum_{j=1}^{64} \left| (freq_j^{in(i)} - freq_j^{db(i)}) \right|}{2N} \qquad (2)$$

$freq_j^{in(i)}$, $freq_j^{db(i)}$ are the frequencies of 3-dimensional vectors that belong to a separate partial sequence of an input query sequence and that belong to the same separate partial sequence of full length sequences in the database, respectively. N is number of vectors in the separate partial sequence.

The integrated similarities (S) are then compared with a given threshold (T), only the sequences whose similarities exceeded the given threshold are then aligned using exhaustive Smith-Waterman dynamic programming algorithm (4).

Experiments and Discussions

GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences There are approximately 106,533,156,756 bases in 108,431,692 sequence records in the traditional GenBank divisions and 148,165,117,763 bases in 48,443,067 sequence records in the WGS division as of August 2009 [8].

We have downloaded plant GenBank DNA sequence database which contain approximately 1,432,314 sequences. Figure 4 shows the distribution of base lengths (Plant sub database). From this database, 162,021 DNA sequences with the sequence length within 400-500 have been selected to be used in experiments. The performance and reliability of the developed algorithm was evaluated. The query sequences have been chosen randomly from the 162,021 sequences.

We performed all of the experiments on a conventional PC@3.2GHz (2G memory). The algorithm was implemented in ANSI C.

We select 50 results with highest scores among the whole results of the entire DNA sequences which given by the Smith-Waterman algorithm [4], and perform the same search by using histogram information algorithm, and calculating the recall and the precision. Recall indicates the proportion of results yielded from histogram information algorithm to the highest 50 scores, and precision indicates the proportion of correct scores included in the results from histogram information algorithm. The recall and precision are defined as follows.

$$\text{Recall} = \frac{\text{number of correct - match}}{\text{total number of positives}} \qquad (3)$$

$$1\text{- precision} = \frac{\text{number of false - match}}{\text{total number of matches}} \qquad (4)$$

Table 1 shows the comparison between the recall and precision in the whole search range and the search range for the histogram information algorithm. The average search domain for the recall of 1.00 is 583.6, which is about 0.36% of the whole range 162,021. The comparison result of required search time for the experiment in Table 2. The time spending of the same search with histogram information algorithm is about 28.6 seconds, which is 0.39% of about 20 minutes (7207.8 sec) of exhaustive search by Smith-Waterman algorithm. We can obtain the same results in both cases.

Conclusion

In this paper, we proposed a novel search method that improves both the speed and the precision of search by combining histogram features and Smith-Waterman dynamic programming algorithms in the

search of DNA sequences. Experimental results shows histogram information algorithm is efficient in both the precision and the speed of search.

Acknowledgment

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan, Grant-in-Aid for Young Scientists (B), No.21710207, 2009-2011.

References

[1] http://www.embl.org/
[2] http://www.ddbj.nig.ac.jp/
[3] Needlman S.B. and Wunsch C.D. (1970) *Journal of Molecular Biology*, 48, 443–453.
[4] Smith T. F. and Waterman M. S. (1981) *Journal of Molecular Biology*, 47, 195–197.
[5] Altscgul S. F., Gish W., Miller W., Myers E. W. and Lipman D. J. (1990) *Journal of Molecular Biology*, 215, 403–410.
[6] Lipman D. and Pearson W. R. (1985) *Scienc*e, 227, 1435-1441.
[7] Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. (2011) *Nucleic Acids Res*. 39(Database issue):D32-7.
[8] http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html.
[9] Li M. and Ma B. (2003) *Genome Informatics*, 14, 164-175.
[10] Ma B., Tromp J. and Li M. (2002) *Bioinformatics*, 18(3), 440- 445.

Table-1 - Comparison between the recall and precision:

Query	Recall	Precision	Search range	Rate
Q1	1.0	0.109	458	0.28%
Q2	1.0	0.16	305	0.19%
Q3	1.0	0.062	798	0.49%
Q4	1.0	0.099	504	0.31%
Q5	1.0	0.125	399	0.25%
Q6	1.0	0.044	1123	0.69%
Q7	1.0	0.074	677	0.42%
Q8	1.0	0.18	267	0.16%
Q9	1.0	0.055	915	0.56%
Q10	1.0	0.13	390	0.24%
Ave.	1.0	0.105	583.6	0.36%

Table-2 - Comparison between the Smith-Waterman and proposed method.

Query	Smith-Waterman(s)	Proposed method(s)	Rate
Q1	7,527	25	0.33%
Q2	7,125	22	0.31%
Q3	6,942	33	0.48%
Q4	7,296	26	0.36%
Q5	7,369	24	0.33%
Q6	7,064	44	0.62%
Q7	7,198	30	0.42%
Q8	7,271	21	0.29%
Q9	7,469	37	0.53%
Q10	7,357	24	0.33%
Ave.	7,207.8	28.6	0.39%

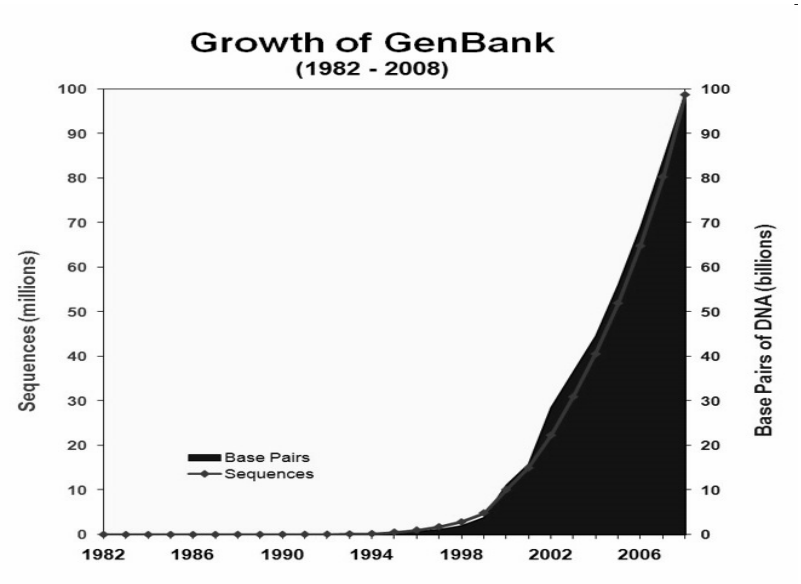

Fig. 1- The growth of DNA sequences in GenBank

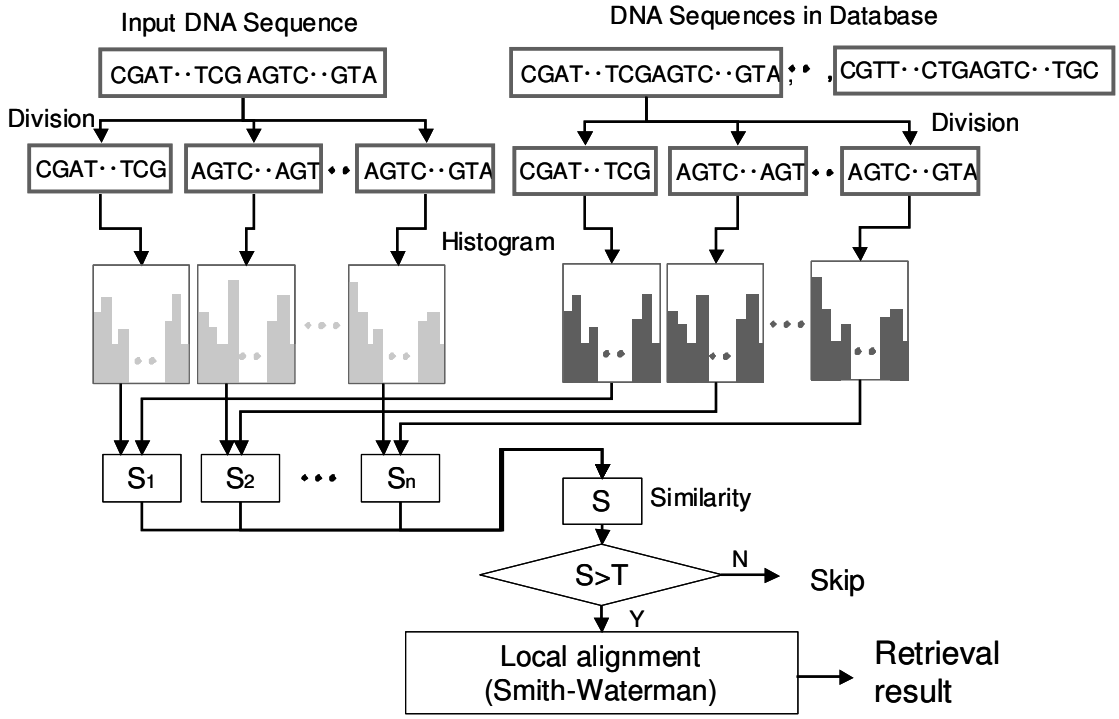

Fig. 2- Processing steps of proposed method

CCC	CCT	CCG	CCA	CTC	CTT	CTG	CTA
0	1	2	3	4	5	6	7
CGC	CGT	CGG	CGA	CAC	CAT	CAG	CAA
8	9	10	11	12	13	14	15
TCC	TCT	TCG	TCA	TTC	TTT	TTG	TTA
16	17	18	19	20	21	22	23
TGC	TGT	TGG	TGA	TAC	TAT	TAG	TAA
24	25	26	27	28	29	30	31
GCC	GCT	GCG	GCA	GTC	GTT	GTG	GTA
32	33	34	35	36	37	38	39
GGC	GGT	GGG	GGA	GAC	GAT	GAG	GAA
40	41	42	43	44	45	46	47
ACC	ACT	ACG	ACA	ATC	ATT	ATG	ATA
48	49	50	51	52	53	54	55
AGC	AGT	AGG	AGA	AAC	AAT	AAG	AAA
56	57	58	59	60	61	62	63

Fig. 3- Reference table

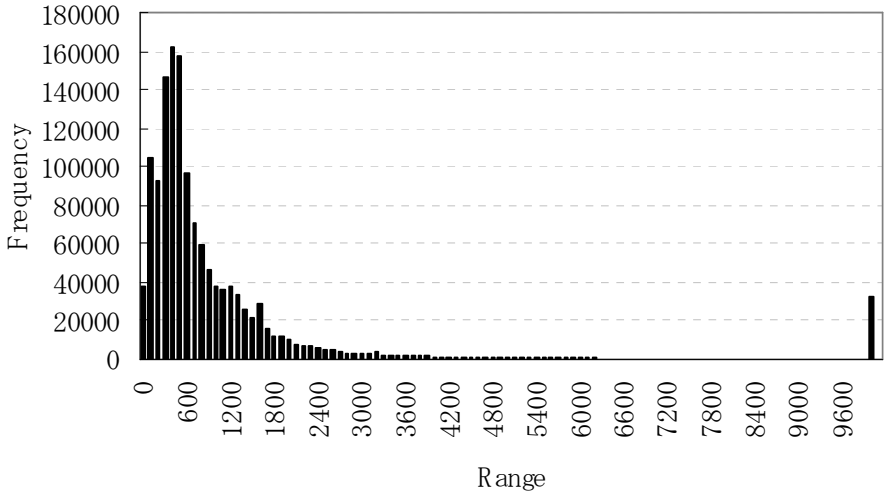

Fig. 4- Distribution of base lengths (Plant sub database)