A Local Search Method Using Histogram Features for Fast Retrieval of DNA Sequences

Qiu Chen, Koji Kotani, Feifei Lee, and Tadahiro Ohmi

***Abstract*—DNA sequence retrieval is a very important topic in bioinformatics algorithm development. However, this task usually spends much computational time to search on large DNA sequence database. In this paper, an efficient hierarchical method is proposed to improve the search speed while the accuracy is being kept constant. For a given query sequence, firstly, a fast local search method using histogram features is used as a filtering mechanism before scanning 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 using GenBank sequence data show the proposed method combining histogram information and Smith-Waterman algorithm is a more efficient algorithm for DNA sequence retrieval.**

***Index Terms*—Fast retrieval, DNA sequence, Histogram information, Smith-Waterman algorithm, Local search**

I. Introduction

The decipherment of 3-billion-base human genome sequence which was called Apollo project of life sciences [1], [2], was finally completed by the international cooperation in April 2003. Since this achievement of human genome project, researchers around the world are now having a very keen competition on clarification of the structure and performance analysis of the protein, genes and protein networks, and new gene sequences are clarified every day. The enormous quantity of data has been accumulated in the database like GenBank [7], EMBL, and DDBJ, etc. Moreover, the volume of data of Genome Database still increases in exponential as shown in Figure 1 [8].

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, EMBL, and DDBJ, 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].

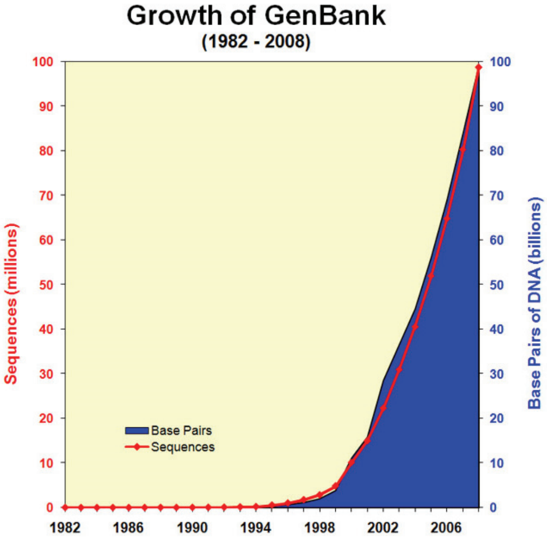

Figure 1. The growth of DNA sequences in GenBank [8].

However, it takes much time to retrieve data with these algorithms because they require too many amounts of calculation. 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. While BLAST and FASTA have improved the retrieving speed with heuristic algorithms, there is a possibility of missing an alignment or giving inaccurate output. Thus, many researches have been trying to improve both the retrieval time and the precision.

In our last paper [11], we proposed an efficient method combining histogram features and Smith-Waterman dynamic programming algorithms [4] in order to improve both speed and precision. The effects have been demonstrated by using GenBank sequence data. For sequences which range of length variation is not very large, the experimental results show the proposed algorithm is very efficient, but the efficiency decreases with variation in sequence length.

In this paper, we propose a local search method in order to improve both efficiency and speed even the sequence length changes largely. The effects will be demonstrated by using GenBank sequence data.

This paper is organized as follows. Section II introduces

This work 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.

Qiu Chen, Feifei Lee, and Tadahiro Ohmi are with New Industry Creation Hatchery Center, Tohoku University, Sendai, 980-8579 Japan (phone: +81 -22-795-3977; fax: +81-22-795-3986; e-mail: qiu@fff.niche.tohoku.ac.jp).

Koji Kotani is with Department of Electronics, Graduate School of Engineering, Tohoku University, Sendai, 980-8579 Japan..

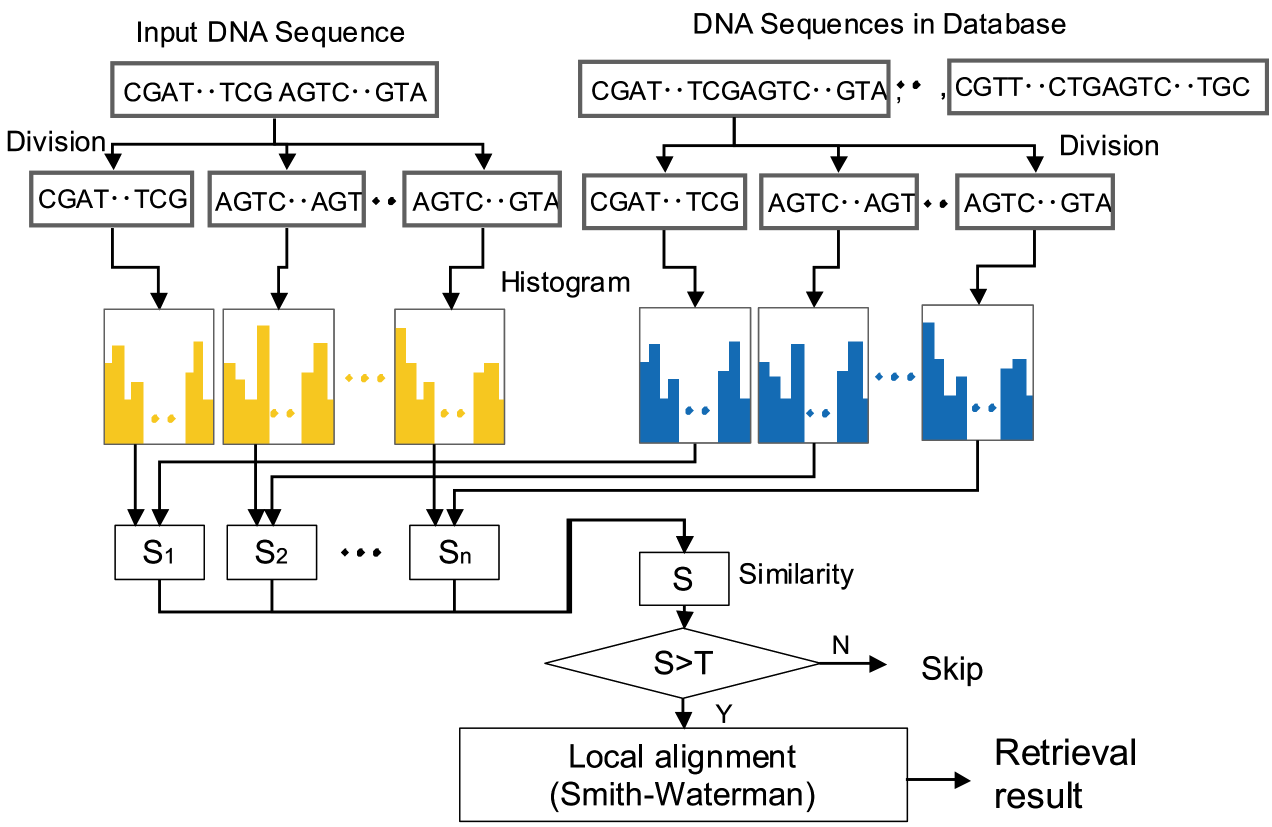

Figure 2. Processing steps of proposed method.

the proposed algorithm in detail. Experimental results will be discussed in section III. Finally, conclusions are given in section IV.

II. 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 can also be extracted from the DNA sequence in the database using the same method respectively. In our previous method, the histogram generated from each partial sequence is then compared with the histogram from the same partial sequence in the database by calculating similarities between them. The shortcoming of this approach is, when the difference of sequence length between the input

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

Figure 3. Reference table.

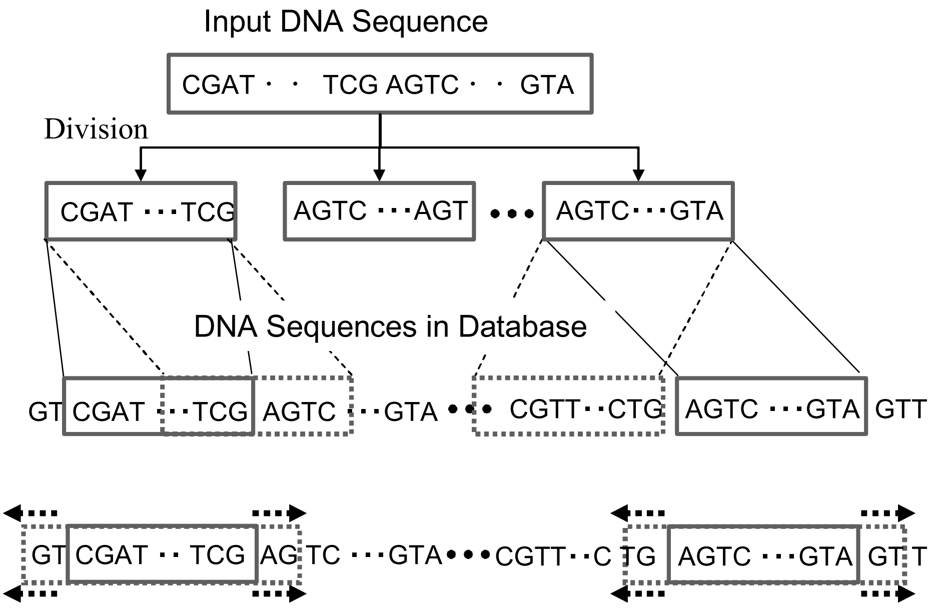
Input DNA Sequence
CGAT · · TCG AGTC · · GTA
Division
CGAT ···TCG
AGTC ···AGT
AGTC···GTA
DNA Sequences in Database
GT CGAT ···TCG AGTC ···GTA ••• CGTT··CTG AGTC ···GTA GTT
GT CGAT ·· TCG AG TC ···GTA•••CGTT··C TG AGTC ···GTA GT T

Figure 4. Local search approach.

base sequence and that in the database is large, the error of the normalization of histogram can not be ignored.

In this paper, we propose a local search approach to resolve this problem. As shown in figure 4, when the histograms of *n* parts of input query base sequence are generated, a search processing will be carried out to get a best matched part in the database for each partial sequence. The similarity between these histograms is used and the best match will be located. Next, the partial sequence is then extended from both sides of it until the corresponding similarity between the partial sequence belonging to input query base sequence and that in the database does not increase any more.

The histogram generated from each extended partial sequence is then compared with the histograms from the corresponding matched partial sequence in the database by calculating similarity (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).

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

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

$freq_j^{in(i)}$, $freq_j^{db(i)}$ 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].

III. Experiments and discussions

A. Data sets

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 sub-database of GenBank DNA database which contain approximately 1,432,314 sequences. From this sub-database, 592,570 DNA sequences with the sequence length within 400-1000 have been selected to be used in the experiments. The performance and reliability of the developed algorithm was evaluated. The query sequences have been chosen randomly from the 592,570 sequences.

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

B. Experimental results

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 retrieval 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.

Table 1 shows the comparison between the recall and precision in the whole retrieval range and the retrieval range for the histogram information algorithm. The average retrieval domain for the recall of 1.00 is 2123, which is about 0.358% of the whole range 592,570. The comparison result of required retrieval time for the experiment in Table 2. The time spending of the same retrieval with histogram information algorithm is about 102 seconds, which is 0.386% of about 7.4

hours (443.0minutes) of exhaustive search by Smith-Waterman algorithm. We can obtain the same results in both cases.

IV. CONCLUSIONS

In this paper, we proposed a novel local retrieval method that improves both the speed and the precision of retrieval by combining histogram features and Smith-Waterman dynamic programming algorithms in the retrieval of DNA sequences. Experimental results shows histogram information algorithm is efficient in both the precision and the speed of retrieval.

References

[1] J. C. Venter, M. D. Adams, E. W. Myers, etc., “The sequence of the human genome”, Science, vol. 291, no. 5507, pp. 1304 -1351, 2001.

[2] F. S. Collins, M. Morgan, A. Patrinos, “The human genome project: lessons from Large-Scale Biology”, Science, vol. 300, no. 5617, pp. 286-290, 2003.

[3] S.B.Needlman and C.D.Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins”, Journal of Molecular Biology, vol. 48, pp. 443− 453, 1970.

[4] T. F. Smith and M. S.Waterman, “Identification of common molecular subsequences”, Journal of Molecular Biology, vol. 47, pp. 195− 197, 1981.

[5] S. F. Altscgul, W.Gish, W. Miller, E. W. Myers and D. J. Lipman, “Basic local alignment search tool”, Journal of Molecular Biology, vol. 215, pp. 403− 410, 1990.

[6] D. Lipman and W. R. Pearson, “Rapid and sensitive protein similarity searches”, Science, vol. 227, pp. 1435-1441, 1985.

[7] GenBank, ftp://ftp.ncbi.nih.gov/genbank/

[8] http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html.

[9] M. Li, and B. Ma, “PatternHunter II: highly sensitive and fast homology search”, Genome Informatics, vol. 14, pp. 164-175, 2003.

[10] B. Ma, J. Tromp, and M. Li, “PatternHunter: faster and more sensitive homology search”, Bioinformatics, vol. 18, no. 3, pp. 440- 445, 2002.

[11] Q. Chen, K. Kotani, F. Lee, and T. Ohmi, “A Fast Retrieval of DNA Sequences Using Histogram Information”, 2009 Int’l Conf. on Future Information Technology and Management Engineering (FITME 2009), pp. 529-532, Sanya, China, Dec., 2009.

Table 1. Comparison between the recall and precision.

Query	Recall	Precision	Search range	Rate
Q1	1.0	0.045	1109	0.18%
Q2	1.0	0.027	1833	0.31%
Q3	1.0	0.012	4085	0.69%
Q4	1.0	0.051	971	0.16%
Q5	1.0	0.035	1428	0.24%
Q6	1.0	0.015	3328	0.56%
Q7	1.0	0.020	2462	0.42%
Q8	1.0	0.034	1451	0.24%
Q9	1.0	0.017	2903	0.49%
Q10	1.0	0.03	1666	0.28%
Ave.	1.0	0.029	2123	0.358%

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

Query	Smith-Waterman(m)	Proposed method(s)	Rate
Q1	434.6	77	0.295%
Q2	445.1	92	0.343%
Q3	430.9	157	0.609%
Q4	443.5	73	0.275%
Q5	448.8	84	0.313%
Q6	455.6	132	0.482%
Q7	439.1	110	0.417%
Q8	449.5	88	0.326%
Q9	423.5	121	0.475%
Q10	459.1	92	0.332%
Ave.	443.0	102	0.386%