ELSEVIER

Journal of Biotechnology 69 (1999) 19–26

JOURNAL OF Biotechnology

DNA sequence comparison considering both amino acid and nucleotide insertions/deletions because of evolution and experimental error

R. Irie *, S. Hiraoka, N. Kasahara, K. Nagai

Hitachi Ltd., Central Research Laboratory, 1-280, Higashi-koigakubo, Kokubunji-shi, Tokyo, 185-8601, Japan

Received 7 July 1998; received in revised form 5 October 1998; accepted 3 December 1998

Abstract

Amino acid similarity often needs to be considered in DNA sequence comparison to elucidate gene functions. We propose a Smith–Waterman-like algorithm which considers amino acid similarity and insertions/deletions in sequences at the DNA level and at the protein level in a hybrid manner. The algorithm is applied to cDNA sequences of *Oryza sativa* and those of *Arabidopsis thaliana*. The results are compared with the results of application of NCBI's tblastx program (which compares the sequences in the BLAST manner after translation). It is shown that the present algorithm is very helpful in discovering nucleotide insertions/deletions originating from experimental errors as well as amino acid insertions/deletions due to evolutionary reasons.

Keywords: DNA; Sequence; Similarity; Translation; Insertion; Deletion

1. Introduction

The rapidly growing DNA sequence databases are being utilized to elucidate the biological function of DNA sequences of interest with various approaches. One of the most effective approaches is homology search in which the DNA sequence of interest is compared with DNA sequences in the database and the similarities between the sequence of interest and the database sequences are evaluated and sorted. The traditional general method of identifying the maximally similar subsequence among sets of long sequences is the Smith–Waterman algorithm (Smith and Waterman, 1981), which is derived by extending the original homology algorithm of Needleman and Wunsch (1970). This algorithm can compare any pair of DNA sequences considering any possible insertions/deletions (indels) in sequences. However, the application of the Smith–Waterman algorithm to the rapidly growing large-scale DNA

* Corresponding author. Fax: +81-423-27-7751.
E-mail address: r-irie@crl.hitachi.co.jp (R. Irie)

0168-1656/99/$ - see front matter
PII: S0168-1656(98)00201-6

database is too laborious for computers with average performance. To meet such a practical requirement, more simplified but more rapid homology search algorithms (e.g. FASTA (Pearson and Lipman, 1988) and BLAST (Altschul et al., 1990)) have been proposed. The BLAST algorithm in particular is widely used for searching the similar sequences in the large database while indels in sequences are no longer considered in the initial search.

If DNA sequences are compared after being translated into amino acid sequences, the sensitivity of detecting the similarity becomes higher because of the degeneracy of the genetic code (many amino acids correspond to more than one triplet of nucleotides), and therefore the similarity evaluation is much more appropriate to prediction of the gene function. There are a few heuristic methods which were produced by this motivation. The most popular one is the tblastx program based on the BLAST algorithm (e-mail server: blast@ncbi.nlm.nih.gov). There is another method which considers indels but which needs informations about the reading frames (Hein and Stovlbaek, 1996). We propose here a Smith–Waterman-like algorithm which compares a pair of DNA sequences by considering amino acid similarity in the optimal reading frames and aligns not only the translated sequences but also the original DNA sequences by treating nucleotide indels explicitly.

In this paper, the present algorithm is described in detail, and the results of its application to comparing cDNA sequences of *Arabidopsis thaliana* with those of *Oryza sativa* are demonstrated. The benefits of the explicit treatment of indels are discussed by referring to the results of the tblastx application.

2. Algorithm

The two DNA sequences will be $\mathbf{A} = \{a_1a_2...a_n\}$ and $\mathbf{B} = \{b_1b_2...b_m\}$. Sets of contiguous three nucleotides $\{a_ia_{i+1}a_{i+2}\}$ and $\{b_jb_{j+1}b_{j+2}\}$ will be translated and represented as A_i and B_j, respectively. To find pairs of segments with high degrees of similarity, one sets up a matrix **H**. The values of **H** have the interpretation that H_{ij} is the maximum similarity of two segments ending in sets of contiguous three nucleotides corresponding to amino acids A_i and B_j, respectively. First set

$$H_{kl} = 0 \quad \text{for} \quad -6 \le k \le n-2 \text{ and } -6 \le l \le 0 \tag{1}$$

$$H_{kl} = 0 \quad \text{for} \quad -6 \le k \le 0 \text{ and } 1 \le l \le m-2 \tag{2}$$

The values of H are obtained from the relationship

$$H_{ij} = \max\{\max\{G_{ij}(K) \mid 0 \le K \le 10\}, 0\} \text{ for } 1 \le i \le n\text{-}2 \text{ and } 1 \le j \le m\text{-}2 \tag{3}$$

Here $G_{ij}(K)$ are obtained as follows.

$$G_{ij}(0) = H_{i-3,j-3} + s(A_i,B_j) \tag{4}$$

$$G_{ij}(1) = H_{i,j-3} + W_{\mathrm{a}} \tag{5}$$

$$G_{ij}(2) = H_{i-3,j} + W_{\mathrm{a}} \tag{6}$$

$$G_{ij}(3) = H_{i-5,j-6} + W_{\mathrm{n}} + s(A_i,B_j) \tag{7}$$

$$G_{ij}(4) = H_{i-6,j-5} + W_{\mathrm{n}} + s(A_i,B_j) \tag{8}$$

$$G_{ij}(5) = H_{i-4,j-3} + W_{\mathrm{n}} + s(A_i,B_j) \tag{9}$$

$$G_{ij}(6) = H_{i-3,j-4} + W_{\mathrm{n}} + s(A_i,B_j) \tag{10}$$

$$G_{ij}(7) = H_{i\text{-}6,j\text{-}7} + s(A_{i\text{-}3},\{b_{j-4}b_{j\text{-}3}b_{j\text{-}1}\}) + W_{\mathrm{n}} + s(A_i,Bj) \tag{11}$$

$$G_{ij}(8) = H_{i-6,j-7} + s(A_{i\text{-}3},\{b_{j-4}b_{j-2}b_{j-1}\}) + W_{\mathrm{n}} + s(A_i,B_j) \tag{12}$$

$$G_{ij}(9) = H_{i-7,j-6} + s(\{a_{i-4}a_{i-3}a_{i-1}\},B_{j-3}) + W_{\mathrm{n}} + s(A_i,B_j) \tag{13}$$

$$G_{ij}(10) = H_{i-7,j-6} + s(\{a_{i-4}a_{i-2}a_{i-1}\},B_{j-3}) + W_{\mathrm{n}} + s(A_i,B_j) \tag{14}$$

Here the function $s(A,B)$ is a similarity score between amino acids (or codons) A and B. The similarity score $s(A,B)$ is set equal to zero when A

or B can not be obtained. W_a denotes a penalty for a gap produced by an indel of an amino acid and takes two values as follows.

$W_a = w_o$ for a gap immediately after an amino acid,

$W_a = w_e$ for a gap after another gap (an extension). (15)

Thus the gap penalty function used at the protein level is the linear function of the length of indels which is most popular (Gotoh, 1982). W_n denotes a penalty for a gap produced by indels of nucleotides in a codon.

The pair of segments with maximum similarity and the corresponding alignment is determined with the trace-back procedure in the Smith–Waterman algorithm (Smith and Waterman, 1981).

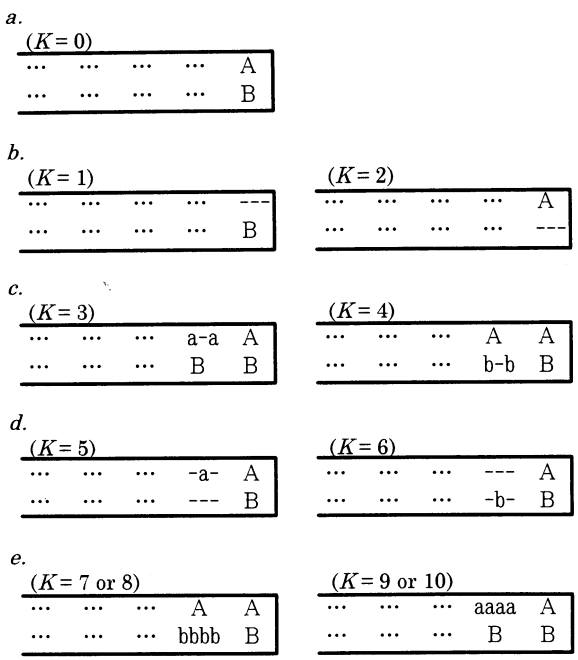

Fig. 1. Probable cases for ending segments at any A_i and any B_j, which are considered in the present algorithm. Here A and B denotes sets of contiguous three nucleotides (amino acids or codons) of DNA sequences **A** and **B**, respectively. a and b denote nucleotides of sequences **A** and **B**, respectively. (-) denotes a deletion at the DNA level and (- - -) denotes a deletion at the protein level. (· · ·) denotes an amino acid or a deletion at the protein level.

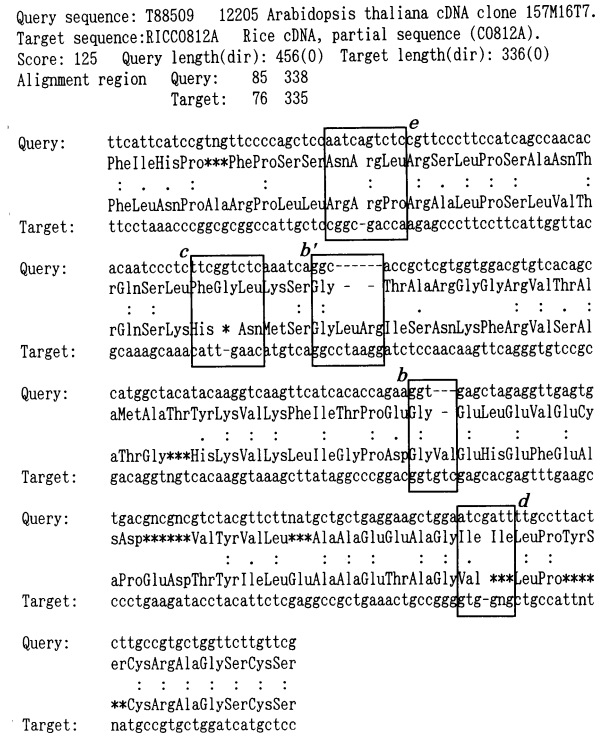

Fig. 2. Example of the cDNA sequence alignment obtained by the present algorithm. The length of a sequence (a query or a target) is measured by counting the recorded nucleotides. The numbers in parentheses denote the direction (dir) of reading (0: forward, 1: reverse). The alignment region is described with four numbers, (q_1,q_2) for a query and (t_1,t_2) for a target. In this case, the query sequence region between the q_1-th nucleotide and the q_2-th nucleotide is aligned with the target sequence region between the t_1-th nucleotide and the t_2-th nucleotide. Hyphens (-) in sequences are used to denote the deletions (of nucleotides and amino acids) generated by the present algorithm. Asterisks (*) are used to denote the cases where there is no translation for the nucleotide triplet. Rectangles enclose examples for introduction of gaps. Symbols placed near the rectangles are consistent with symbols in Fig. 1 except for b′, which denotes the case where the length of the gap at the protein level is extended.

The above formulas for H_{ij} are produced by considering the probable cases for ending segments at any A_i and B_j, which are illustrated in Fig. 1. The illustration *a* in Fig. 1 represents the

case when A_i and B_j are associated, where Eq. (4) is applied. The illustration *b* represents the cases when A_i/B_j is at a deletion of an amino acid, where Eq. (5) or Eq. (6) is applied. The illustration *c* represents the cases when immediately before A_i/B_j (associated with B_j/A_i) exists a gap which is produced by the deletion of one nucleotide in an amino acid (codon), where Eq. (7) or Eq. (8) is applied. In Eqs. (7) and (8), the similarity score between $\{a_{i-2}a_{i-1}\}$ and $\{b_{j-3}b_{j-2}b_{j-1}\}$ and that between $\{a_{i-3}a_{i-2}a_{i-1}\}$ and $\{b_{j-2}b_{j-1}\}$ are set neutral or zero because the amino acid corresponding to $\{a_{i-2}a_{i-1}\}/\{b_{j-2}b_{j-1}\}$ is unclear. The illustration *d* represents the cases when immediately before A_i/B_j (associated with B_j/A_i) exists a gap which is produced by the insertion of one nucleotide between codons, where Eq. (9) or Eq. (10) is applied. The illustration *e* represents the cases when immediately before A_i/B_j (associated with B_j/A_i) exists a gap which is produced by the insertion of one nucleotide into an amino acid (codon), where Eqs. (11)–(13), or Eq. (14) is applied.

Some similar approaches for aligning a DNA sequence and a protein sequence are discussed (Zhang et al., 1997). The ideas may be generalized for aligning a pair of DNA sequences. The way of introducing indels for aligning a DNA sequence and a protein sequence is more systematic than ours for aligning a pair of DNA sequences. However our algorithm is built up under the guidance of the observations including the following one. The probability of the nucle-

Table 1
EST sequence pairs between *A. thaliana* and *O. sativa* the similarity scores of which are enhanced by introducing gaps

Locus names of pairs[a]	Scores[b]	Introduced gaps[c]
T88476: RICC0682A	114/69	0 *ntg*+2 *aag*
T88476: RICS5087A	101/59	4 *ntg*+0 *aag*
T88476: RICR2082A	96/60	0 *ntg*+2 *aag*
T88476: RICR2802A	92/54	0 *ntg*+2 *aag*
T88478: RICS16345A	97/65	4 *ntg*+2 *aag*
T88489: RICS2250A	110/62	0 *ntg*+2 *aag*
T88498: RICS3281A	91 /—	3 *ntg*+2 *aag*
T88509: RICC0812A	125 /—	3 *ntg*+3 *aag*
T88538: RICS14994A	114/59	2 *ntg*+5 *aag*
T88538: RICS12037A	104/59	1 *ntg*+5 *aag*
T88538: RICS12974A	100/58	1 *ntg*+5 *aag*
T88551: RICR0260A	165/58	1 *ntg*+4 *aag*
T88551: RICS1706A	161/65	2 *ntg*+4 *aag*
T88551: RICS4216A	160/65	2 *ntg*+4 *aag*
T88551: RICS15107A	151/60	2 *ntg*+4 *aag*
T88551: RICS2252A	145/58	2 *ntg*+4 *aag*
T88551: RICR0034A	137/58	1 *ntg*+4 *aag*
T88551: RICR2521A	137/69	2 *ntg*+4 *aag*
T88551: RICS5055A	134/69	1 *ntg*+4 *aag*
T88551: RICS2694A	129/58	2 *ntg*+4 *aag*
T88551: RICS5071A	118/69	1 *ntg*+4 *aag*
T88551: RICS0787A	117/58	2 *ntg*+4 *aag*
T88551: RICS2729A	112 /—	2 *ntg*+4 *aag*
T88557: RICC2857A	119/57	2 *ntg*+9 *aag*
T88563: RICC10346A	114/69	1 *ntg*+3 *aag*
T88626: RICS14478A	90/61	2 *ntg*+0 *aag*
T88643: RICC10428A	107/69	0 *ntg*+3 *aag*
T88645: RICS14373A	100/66	2 *ntg*+2 *aag*
T88661: RICS2213A	91/65	3 *ntg*+0 *aag*

[a] Locus names of EST sequence pairs (*A. thaliana*: *O. sativa*).

[b] Maximum similarity scores of the sequences' alignments obtained using two algorithms (the present algorithm/the BLAST algorithm (tblastx)). Hyphens mean that the tblatsx program does not output the similarity data for the sequence pairs under the default condition. However, except for the hyphen cases and several other cases, tblastx outputs the *P*-values that are less than 0.05 (or imply significant similarity).

[c] Gaps in the alignment which are introduced by the present algorithm. An *ntg* denotes a gap produced by a nucleotide insertion/deletion (the gap types *c*, *d*, or *e* in Fig. 1). An *aag* denotes a gap produced by an amino acid insertion/insertion (the gap type *b* in Fig. 1).

Table 2
Distribution of gap introduction types which are found in an application of the present algorithm[a]

	0 *ntg*	1 *ntg*	2 *ntg*	3 *ntg*	4 *ntg*	5 *ntg*
0 *aag*			1	1	1	
1 *aag*						
2 *aag*	4		1	1	1	
3 *aag*	1	1		1		
4 *aag*		4	8			
5 *aag*		2	1			
6 *aag*						

[a] This table is based on the data in Table 1. The number in each cell denotes the number of sequence pairs into which the present algorithm introduces the corresponding number of *ntg*s in the top row and the corresponding number of *aag*s in the left end column in order to obtain the alignment with the maximum similarity score.

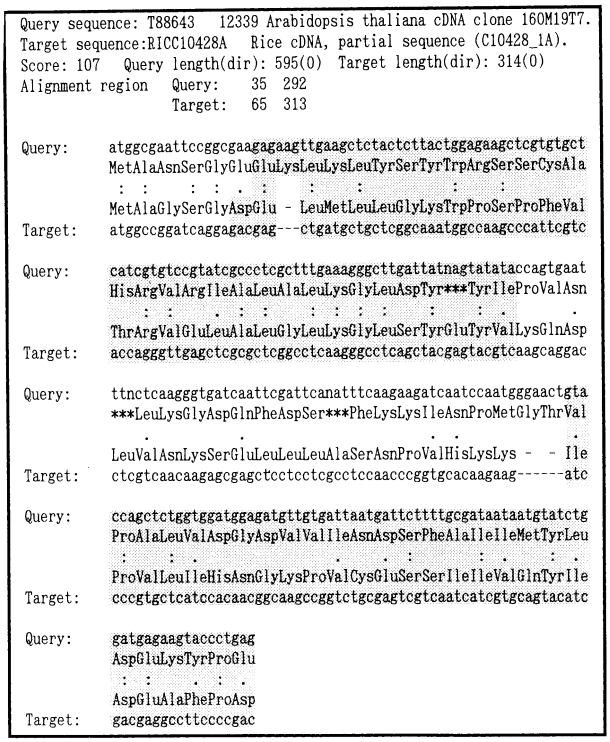

Fig. 3. Example of the cDNA sequence alignment into which the gaps are introduced at the protein level only. The notation concerning the length of a sequence, the direction of reading, the alignment region, and the nucleotide deletions (-) generated by the present algorithm is the same as in Fig. 2. The whole alignment is discovered by the present algorithm. The alignment regions also discovered by the tblastx are hatched.

otide indels caused by the experimental DNA sequencing errors is nearly one percent on the average (Nishikawa and Nagai, 1996). Thus, one cannot expect the consecutive occurrence of indels of nucleotides to be frequent. By eliminating such rare cases, we avoid making the algorithm time-consuming.

As compared with the Smith–Waterman algorithm (simply applied to a comparison of amino acid sequences), Fig. 1 shows that the present algorithm, by flexibly employing the appropriate reading frames, introduces the consideration of insertions of a single nucleotide between amino acids (codons) and indels of a single nucleotide into a codon (an amino acid) adjacent to two other amino acids. If the problem is confined to the nucleotide indels originating from errors of the experimental DNA sequencing, this moderate consideration of indels is practically effective.

3. Application and discussion

The source code for the present algorithm was written in ANSI C language and was compiled and executed on a Solaris UNIX and a Macintosh operating system. The program for parallel computers (e.g. Hitachi SR2201) will be available from Hitachi, Ltd. in the near future.

The present algorithm is applied to the comparison of cDNA sequences of *A. thaliana* with cDNA sequences of *O. sativa* (rice). In this application, we use the standard genetic code and adopt BLOSUM62 (Henikoff and Henikoff, 1992) as a score matrix $s(A,B)$ and set $w_o = -12$, $w_e = -4$, $W_n = -12$.

Fig. 2 shows the representative alignment obtained by the application of the present algorithm. The aligned DNA sequences are a cDNA sequence of *A. thaliana* (T88509) and a cDNA

sequence of *O. sativa* (RICC0812A), both of which are fetched from GenBank (anonymous ftp site: ncbi.nlm.nih.gov/genbank). The results of all types of indel consideration in the present algorithm are found in this example. In this application, it takes 0.8 s to complete the calculation on a Sun workstation (SS20) and about 5 s on an Apple PowerBook 5300c (PowerPC 603e 100MHz). This computation time is reasonable for the interactive investigation of a few sequences of interest.

In order to assess the benefit of the indel consideration, the BLAST algorithm (the tblastx program) is also applied to comparison of *A. thaliana* cDNA sequences with *O. sativa* cDNA sequences. In this work, default values of BLAST parameters were used. (Thus the score matrix used is BLOSUM62.) All the cDNA sequences were collected from the EST sequence entries in GenBank (Release 94.0). The employed *A. thaliana* EST sequences are 199 sequences (having locus names from T88472 through T88671 but missing one locus name). The *O. sativa* database consists of all EST sequences of *O. sativa* in GenBank (11300 sequences).

First of all, 199 EST sequences of *A. thaliana* were compared with all sequences in the *O. sativa* EST database by the present algorithm. One hundred and four sequences of the EST sequences of *A. thaliana* were found to have a relatively high similarity (the score 90) to one or more sequences in the *O. sativa* database. The other sequences of *A. thaliana*, which have a poor similarity (the score <90) to all the sequences in the *O. sativa* database in the application of the present algorithm, were then compared with all sequences in the *O. sativa* database by the tblastx program. In this comparison, the tblastx program could not find the *A. thaliana* sequences with higher similarity (the score ≥ 90), either[1]. This result of score-based comparison could be reasonably understood by the basic difference between the algorithms in that the BLAST algorithm treat with gaps more or less by combining ungapped alignments using statistics for multiple high scoring segments (Karlin and Altschul, 1993) while the present algorithm directly searches gapped alignments using the Smith–Waterman algorithm.

Table 1 shows the list of the *A. thaliana* ESTs, the sequences of which have a higher similarity (the score ≥ 90) against one or more sequences in the *O. sativa* EST database in the application of the present algorithm, but which were found to have a lower similarity (the score <70) against the same *O. sativa* EST sequences in the application of the BLAST algorithm (tblastx). However, except for the sequence pairs listed in Table 1, the present program and the tblastx program output the equivalent similarity scores for each sequence pair between the *A. thaliana* EST database and the *O. sativa* EST database. Table 1 shows that the consideration of indels in the manner of the present algorithm is effective in the enhancement of similarity scores.

There are various kinds of gap introduction in Table 1. The symbol *ntg* denotes a gap produced by an indel of a nucleotide (the type *c*, *d*, or *e* in Fig. 1). The symbol *aag* denotes a gap produced by an indel of an amino acid (the type *b* in Fig. 1). The distribution of these types of gap introduction are two-dimensionally displayed in Table 2. The number in each cell of the table is the number of the sequence pairs into which the present algorithm introduces the corresponding number of *ntg*s in the top row and the corresponding number of *aag*s in the left end column in order to obtain the alignment with the maximum similarity score. The maximum population of the gap introduction is located at the type of introducing 2 *ntg*s and 4 *aag*s. However, the contribution of nucleotide indels can not be neglected on the whole. The number of *aag*s in each alignment (a few *aag*s in an alignment with the length of a few hundred nucleotides) suggests that the corresponding indels of amino acids could be due to evolutionary reasons. Similarly, there are a few *ntg*s in each alignment (with the length of a few hundred nucleotides). According to our recent research on EST databases, the indels originating from the errors during the experimental DNA sequencing are estimated to occur at the rate of about 1 indel per 100 nucleotides (Nishikawa and Nagai, 1996). Thus the indels of nucleotides

[1] Some possibility that tblastx could have found some new alignments remains if the results were assessed using *P*-values.

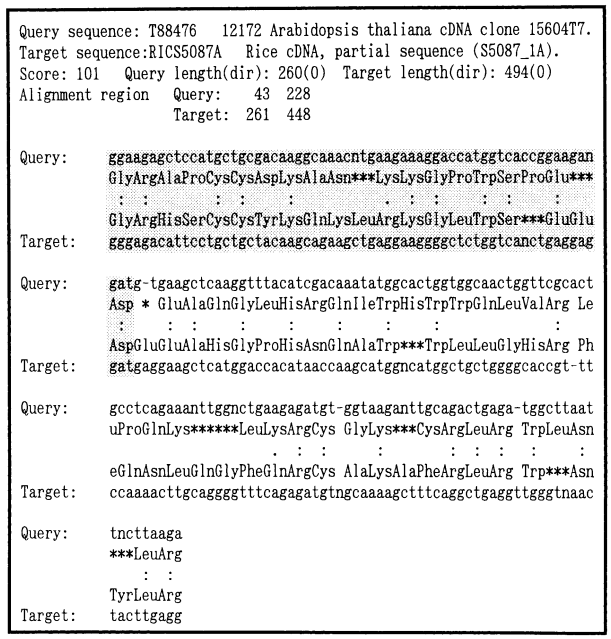

Fig. 4. Example of the cDNA sequence alignment into which the gaps are introduced at the DNA level only. The notation concerning the length of a sequence, the direction of reading, the alignment region, and the nucleotide deletions (-) generated by the present algorithm is the same as in Fig. 2. The whole alignment is discovered by the present algorithm. The alignment regions also discovered by the tblastx are hatched.

shown in Table 1 are expected to have occurred during the experimental DNA sequencing.

Figs. 3 and 4 are the examples of the maximal similarity alignments, which are discovered by the present algorithm, and include only *aag*s and only *ntg*s, respectively. The alignment regions also discovered by the BLAST algorithm are hatched. Here it should be noted that the tblastx program does not output the original DNA sequences. In Figs. 2–4 one can find many examples for the degeneracy of the genetic code, which causes appropriate enhancement of the similarity score by considering the translation. For example, the first associated different nucleotide triplets ({*gga*} and {*ggg*}) in Fig. 4 are treated as identical amino acids (Glys).

When comparing the DNA sequences which are expected to have many nucleotide indels (Figs. 2 and 4), the proper evaluation of the similarity is deemed very difficult for the BLAST homology search after translation (tblastx). However, the present algorithm, which dynamically considers amino acid similarity and appropriately introduces gaps (not only *aag*s but also *ntg*s), can discover the longer alignments of similar putative proteins. These similarities are the starting points for the functional analysis of genes. In addition, the present algorithm is expected to be very powerful in detecting DNA sequencing errors and will therefore be useful in the proper evaluation of the similarity of DNA sequences containing experimental errors.

The computation time which the present algorithm takes to search through a usual DNA database using a conventional computer might be lengthy. Thus one had better use parallel computers to search through databases with the present algorithm. If no parallel computer is available, one may search through databases with tblastx before the detailed investigation of the selected sequence pairs with the present algorithm while several significant alignments could be missed.

4. Conclusions

We have proposed a Smith–Waterman-like algorithm which considers amino acid similarity and indels in sequences at the DNA level and at the protein level in a hybrid manner. The algorithm is applied to cDNA sequences of *O. sativa* and those of *A. thaliana*. The results are compared with those of applying NCBI's tblastx program (which compares the sequences in the BLAST manner after translation). It is shown that the present algorithm is very powerful in detecting nucleotide indels originating from the errors during the experimental DNA sequencing as well as in discovering amino acid indels which are due to the evolutionary reasons. Thus the present algorithm is very useful for evaluating the similarity of DNA sequences containing experimental errors.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Gotoh, O., 1982. An improved algorithm for matching biological sequences. J. Mol. Biol. 162, 705–708.

Hein, J., Stovlbaek, J., 1996. Combined DNA and protein alignment. In: Doolittle, R.F. (Ed.), Methods in Enzymology, Computer Methods for Macromolecular Sequence Analysis, vol. 266. Academic Press, London, pp. 402–418.

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. 89, 10915–10919.

Karlin, S., Altschul, S. F., 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. Proc. Natl. Acad. Sci. U.S.A. 90, 5873–5877.

Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.

Nishikawa, T., Nagai, K., 1996. EST error analysis in a large-scale GenBank search of ESTs using rapid-identity-searching program for DNA sequences. In: Genome Mapping and Sequencing. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, p. 172.

Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A. 85, 2444–2448.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.

Zhang, Z., Pearson, W.R., Miller, W., 1997. Comparing a DNA sequence with a protein sequence. In: Proceedings of First Annual International Conference on Computational Molecular Biology. ACM Press, New York, pp. 337–343.