Codon-sensitive comparison of DNA sequences containing insertions/deletions and statistical significance of the smilarity scores

Ryotaro Irie [1]
r-irie@crl.hitachi.co.jp
Naoko Kasahara [1]
kasahara@crl.hitachi.co.jp
Susumu Hiraoka [1]
hiraoka@crl.hitachi.co.jp
Keiichi Nagai [1]
k-nagai@crl.hitachi.co.jp

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

Abstract

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. We also developed a procedure to evaluate the statistical significance of the similarity scores of sequence alignments with gaps by using Karlin-Altschul statistics. The present DNA sequence comparison algorithm is applied to cDNA sequences of oryza sativa and those of arabidopsis thaliana. The possibilities, where a random (or unrelated) sequence has the obtained similarity scores against a query, are evaluated by the present statistical procedure. The results are compared with those of BLAST (tblastx) and the usefulness of the present algorithm and procedure is discussed.

1 Introduction

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, 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. However the introduction of insertions/deletions (indels) into alignments of sequences is desired in order to deal with the sequencing errors or the evolutional indels. We propose here an 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 addition to it, to avoid the apparently similar but unrelated sequences and to select the substatially similar and significant sequences, we developed a program to evaluate the statistical significance of the similarity scores of sequence alignments with indels .

2 Algorithm and Procedure

The following is the present algorithm for DNA sequences. The two DNA sequences will be $\mathbf{A} = \{a_1 a_2 ... a_n\}$ and $\mathbf{B} = \{b_1 b_2 ... b_m\}$. Sets of contiguous three nucleotides $\{a_i a_{i+1} a_{i+2}\}$ and $\{b_j b_{j+1} b_{j+2}\}$ 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 $\mathbf{H}$. The values of $\mathbf{H}$ have the interpretation that H_{ij} is the maximum similarity of two segments ending in sets of contiguous three nucleotides A_i and B_j, respectively. First set

$H_{kl} = 0$ for -6œB!eœ(B k œB!e$œ(B$n-2$ and -6œB!eœ(B l œB!e$œ(B$0$, (1)
$H_{kl} = 0$ for -6œB!eœ(B k œB!e$œ(B$0$ and 1œB!eœ(B l œB!e$œ(B$m-2$. (2)

The values of **H** are obtained from the relationship

$Hij = max\{G_{ij}(K)|0$œB!eœ(B K œB!e$œ(B$10\}$ for 1œB!eœ(B i œB!e$œ(B$n-2$ and 1œB!eœ(B j œB!e$œ(B$m-2$. (3)

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

$G_{ij}(0) = H_{i-3,j-3} + s(A_i, B_j)$ (4)
$G_{ij}(1) = H_{i,j-3} + W_a$ (5)
$G_{ij}(2) = H_{i-3,j} + W_a$ (6)
$G_{ij}(3) = H_{i-5,j-6} + W_n + s(A_i, B_j)$ (7)
$G_{ij}(4) = H_{i-6,j-5} + W_n + s(A_i, B_j)$ (8)
$G_{ij}(5) = H_{i-3,j-4} + W_n + s(A_i, B_j)$ (9)
$G_{ij}(6) = H_{i-4,j-3} + W_n + s(A_i, B_j)$ (10)
$G_{ij}(7) = H_{i-6,j-7} + s(A_{i-3}, \{b_{j-4}b_{j-3}b_{j-1}\}) + W_n + s(A_i, B_j)$ (11)
$G_{ij}(8) = H_{i-6,j-7} + s(A_{i-3}, \{b_{j-4}b_{j-2}b_{j-1}\}) + W_n + s(A_i, B_j)$ (12)
$G_{ij}(9) = H_{i-7,j-6} + s(\{a_{i-4}a_{i-3}a_{i-1}\}, B_{j-3}) + W_n + s(A_i, B_j)$ (13)
$G_{ij}(10) = H_{i-7,j-6} + s(\{a_{i-4}a_{i-2}a_{i-1}\}, B_{j-3}) + W_n + s(A_i, B_j)$ (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 (w_o for a gap immediately after an amino acid, w_e for a gap after another gap (an extension)). The pair of segments with maximum similarity and the corresponding alignment is determined with the traceback procedure in the Smith-Waterman algorithm[1].

The present procedure for statistical significance evaluation is similar to that described by the previous work for comparison of protein sequences[2] except for the following features. In our program, the extreme-value distribution function is fitted to the distribution of the regression-scaled scores for each query by using the top-populated score and the population around the top.

3 Results

The algorithm is applied to cDNA sequences of *oryza sativa* and those of *arabidopsis thaliana*. The results (alignments) are compared with the results of tblastx program. It can be shown that the present algorithm is very powerful in detecting nucleotide indels originating from the errors. As for the present statistical procedure, the fitting of the statistical function was found to be successful enough to evaluate the statistical siginificance of the top scores.

References

[1] Smith, T. F. and Waterman, M. S., "Identification of Common Molecular Subsequences ," *J. Mol. Biol.*, 147:195-197, 1981.

[2] Pearson, W. R., "Effective Protein Sequence Comparison ," *Methods in Enzymology*, 266:227-258, 1996.