Detection of Frame-Shift Error in the *Yeast* Genome Sequences

Naoko Kasahara kasahara@crl.hitachi.co.jp
Naoyuki Harada n-harada@crl.hitachi.co.jp
Keiichi Nagai k-nagai@crl.hitachi.co.jp
Central Research Laboratory, Hitachi, Ltd.
1-280 Higashi-koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan

1 Introduction

Recently the number of organisms, whose whole genome sequences have been determined, is increasing. They are utilized to various analyses, such as comparative genomics. In case of protein sequence analyses, frame-shift errors are crucial. Even in the 99.99 % accuracy, there might be about 1,000 errors including insertions and deletions in the YEAST genome sequence. We have developed two kinds of homology search methods based on Smith-Waterman-like algorithm considering nucleotide and amino acid gaps simultaneously. One compares a translated DNA sequence and a protein sequence (*transq*) [1]. The other compares two translated DNA sequences (*transw*) [2]. We also developed their parallel computation programs to realize practical computation time for database search [3]. We utilized them to detect frame-shift errors in the YEAST genome sequence.

2 Method and Results

We obtained the YEAST genome sequence from SGD
(http://genome-www.stanford.edu/Saccharomyces/, June, 1998). We tried to find frame-shift errors in the regions where short ORFs are located. We compared the sequences in those regions with the all amino acid sequences in SwissProt (rel. 34.0) using *transq*. An example of the results is shown in Fig. 1. The alignment obtained by *transw* is also shown in Fig. 2. From the alignments with a YEAST hypothetical protein including an intron, there might be two frame-shift errors in this region. As a result, the originally assigned three ORFs [4] could be connected to one ORF as shown in Fig. 3. The potential substitution errors are also shown in Fig. 1 and Fig. 2. We consider our methods is efficient to detect frame-shift errors as well as substitution errors in the genome sequences.

References

[1] Kasahara, N., Hiraoka, S., and Nagai, K., Direct comparison between DNA and amino acid sequences based on a dynamic programming method, *Genome Informatics 1996*, Universal Academy Press, 202–203, 1996.

[2] Irie, R., Kasahara, N., Hiraoka, S., and Nagai, K., Codon-sensitive comparison of DNA sequences contains insertions/deletions and statistical significance of the similarity scores, *Genome Informatics 1997*, Universal Academy Press, 286–287, 1997.

[3] Kasahara, N., Hiraoka, S., Irie, R., and Nagai, K., Highly sensitive homology search methods on parallel computer, *Genome Informatics 1997*, Universal Academy Press, 294–295, 1997.

[4] Mewes, H.W., Albermann, K., Bahr, M., Frishman, D., Gkeissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S.G., Pfeiffer, F., and Zollner, A., Overview of the yeast genome, *Nature*, 387:7–65, 1997.

```
Query sequence:Yeast Chromosome VI ( posision 1500 )
Target sequence:>sp:YH18_YEAST HYPOTHETICAL 68.9 KD PROTEIN IN PUR5 3'REGION.
Score: 6929 - strand
Alignment region  Query :   3709 .. 1802
                             Target:      1 ..  603
                                  ⋮

Query: 3469 tataatgagttgagtttccgtgtcctggaacgttgtcacgaaatagcgagtgccaggccg
            TyrAsnGluLeuSer PheArgValLeuGluArgCysHisGluIleAlaSerAlaArgPro
             |  |  |  |  |   |  |  |  |  |  |  |  :  |  |  |  |  |  |  |
Target:  81 TyrAsnGluLeuSer PheArgValLeuGluArgCysTyrGluIleAlaSerAlaArgPro


Query: 3408 aacgacagctctacgatgcgtactttcactgactttgtttctggcgcacctattgtaagg
            AsnAspSerSerThrMetArgThrPheThrAspPheValSerGlyAlaProIleValArg
             |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  .  |  |  |  |
Target: 102 AsnAspSerSerThrMetArgThrPheThrAspPheValSerGlyThrProIleValArg
```

□ :detected frame-shift errors   ⬚ :detected substitutions

Fig. 1 Alignment result obtained by *transq*

```
Query sequence:Yeast Chromosome VI ( posision 1500 )
Target sequence:>YSCH9117 Saccharomyces cerevisiae chromosome VIII cosmid 9177.

Query:  3469 tataatgagttgagtttccgtgtcctggaacgttgtcacgaaatagcgagtgccaggccg
             TyrAsnGluLeuSer PheArgValLeuGluArgCysHisGluIleAlaSerAlaArgPro
              |  |  |  |  |   |  |  |  |  |  |  |  :  |  |  |  |  |  |  |
             TyrAsnGluLeuSer PheArgValLeuGluArgCysTyrGluIleAlaSerAlaArgPro
Target:51705 tataatgagttgagt-ttccgtgtcctggaacgttgttacgaaatagcgagtgccaggccg

Query:  3408 aacgacagctctacgatgcgtactttcactgactttgtttctggcgcacctattgtaagg
             AsnAspSerSerThrMetArgThrPheThrAspPheValSerGlyAlaProIleValArg
              |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  .  |  |  |  |
             AsnAspSerSerThrMetArgThrPheThrAspPheValSerGlyThrProIleValArg
Target:51765 aacgacagctctacgatgcgtactttcactgactttgtttctggcacacctattgtaagg
```

□ :detected frame-shift errors   ⬚ :detected substitutions

Fig. 2 Alignment result obtaind by *transw*

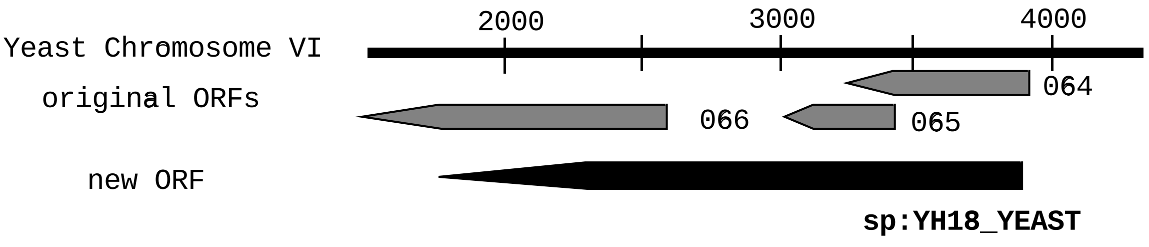

Fig. 3   Comparison original and new ORFs