Extracting System of Accurate ORFs

Tetsuo Nishikawa, Katsuhiko Murakami, Koji Hayashi, Hiroyuki Sato, Tetsuji Otsuki, Naoko Kasahara, Tomohiro Yasuda, Kouichi Kimura, Keiichi Nagai, Ryotaro Irie, Tomoyasu Sugiyama, Takao Isogai.
An extracting system of accurate ORFs from cDNA sequences.
Genome Informatics, 13: 545-547 (2002).

1. Introduction

The development of the Human Genome Project has brought about a rapid expansion of the range of databases concerning gene sequences. EST sequences registered in high volume [1] are sequences that are collected with an emphasis on throughput, and therefore the precision of these sequences is not very high (It is reported that approximately 3% of EST sequences is error [2]). The frame-shift errors make it difficult to extract reliable amino acid sequences from DNA sequences. Furthermore, reverse transcription is known to be an error prone process [3]. It is therefore required that amino acid sequence with high precision is extracted from these sequences by identifying these artifacts.

Conventionally, an amino acid frame display has generally been used for the extraction of amino acid sequence from a cDNA sequence. In the frame each position of initiation and termination codon is displayed and a segment that starts at an initiation codon and terminates at a termination codon is identified; the obtained segments are identified as possible open reading frames (ORF), and among them, the longest ORF is identified as an amino acid sequence extracted from the cDNA. In the case where a frame shift error exists on a cDNA sequence, an ORF is split and displayed over 2 frames. Further, since the border of the split ORF is not clear, an amino acid sequence is, in general, identified with an error of tens of bases. It has been reported that statistical information included in a DNA sequence, such as coding potential, can be used to identify cloning errors including frame shifts [3]. Dr. Hirosawa showed that the application of a modified GeneMark program for detection of artifacts in cDNA clones. This program serves to provide a warning when any spurious split of protein-coding regions is detected. Though this method is effective for detecting the split of protein-coding regions, it is difficult to detect the strict location of the frame-shift, because of the limitation of the statistical analysis. The most reliable method to identify the frame-shift errors in a DNA sequence is to use similarity information to known amino acid sequences. Methods of comparing a cDNA sequence with amino acid sequences in consideration of the occurrence of frame-shift errors in the DNA sequence have been developed including FASTY [4] and TRANS series developed by our laboratory [5]. Using these methods, even where a frame shift error exists, a single alignment can be obtained and it is possible to identify the frame shift site.

Thus, to extract an amino acid sequence from a DNA sequence, there are three methods, a method of using amino acid frames, a method of using statistical information, and a method of using similarity information to known amino acid sequences. However, in order to extract a highly reliable amino acid sequence even where a frame shift error exists on a cDNA sequence, the application of either one of these methods is not sufficient. Here, we propose a new method, mRNA check system that integrates these three types of information, amino acid frames, statistical information, and similarity information with known amino acid sequences, in order to obtain a highly reliable amino acid sequence from a DNA sequence.

2. Methods and Results

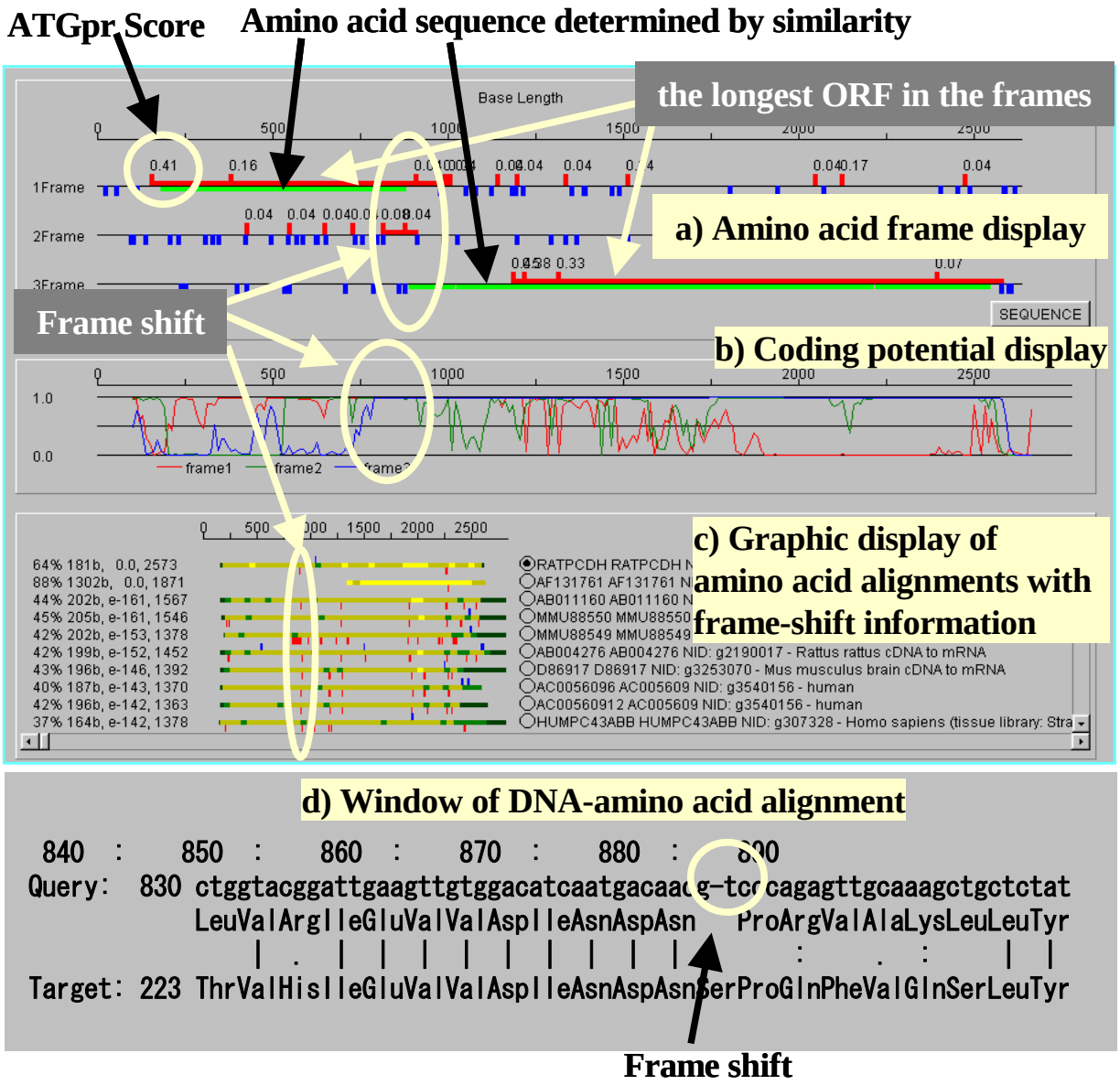

Figure 1 Window of analysis results and window of DNA-amino acid alignment. Window of analysis results consists of a) Amino acid frame display, b) Coding potential display, c) Graphic display of amino acid alignments, and d) Window of DNA-amino acid alignment.

The mRNA check system performs the following analyses, ORFs extraction, similarity analysis to amino acid sequences by using BLASTX and TRANSQ [5], and statistical analyses, ATGpr and coding potential analysis, for a given cDNA sequence. The analyzed results are displayed on the window shown in Figure 1. This window consists of the following four sub-windows, a) amino acid frame sub-window, b) coding potential sub-window, c) sub-window for graphic display of amino acid alignments, and d) window of DNA-amino acid alignment. In the amino acid frame sub-window, initiation codons, termination codons, and the longest ORF in each frame are displayed on three frames. On the initiation codon mark, ATGpr score that is a measure for likelihood of

initiation codon is displayed. In the coding potential sub-window, coding potentials of three frames are plotted along the given DNA sequence. In the sub-window for graphic display of amino acid alignments, the multiple alignments results with amino acid sequences are displayed graphically. The location of the frame-shifts detected by TRANSQ are indicated in the graphical alignment lines and local identity in the alignment discriminated by colors provides the reliability of the detected frame-shifts. For the selected amino acid sequence, alignment between the given cDNA sequence and amino acid sequence is shown in the sub-window of DNA-amino acid alignment. The amino acid sequence determined by this alignment is displayed in the frames in amino acid frame sub-window. Users can inspect the frame-shifts by observing ORFs, amino acid sequences obtained by similarity, coding potential, and alignment with multiple amino acid sequences, graphically at the same time. This observation makes users judge the reliability of the frame-shifts effectively. The mRNA check system enables the effective detection of incomplete DNA sequences such as truncated cDNAs, frame-shifts in database amino acid sequences and chimera.

The next target is to develop automatic detection method for frame-shifts with higher reliability. This work was supported by a Grant from NEDO Project of Ministry of Economy Trade and Industry of Japan.

[1] Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M., dbEST--database for "expressed sequence tags", Nat Genet, 4, 332-323, 1993.

[2] Nishikawa T., Nagai, K. EST Error Analysis in a Large-scale GenBank Search of ESTs using Rapid-idenyity-searching Progrm, Genome Sequencing and Mapping, 1996.

[3] Hirosawa, M., et al., Detection of Spurious Interruptions of Protein-Coding Regions in Cloned cDNA Sequences by GeneMark Analysis, Genome Res., 10, 1333-1341, 2000

[4] Pearson W.R., Wood T., Zhang Z., Miller W., Comparison of DNA sequences with protein sequences, Genomics, 46, 24-36, 1997.

[5] Kasahara, N., et al., Highly Sensitive Homology Search Methods on Parallel Computer, GIW ’97, 294-295, 1997.