A cDNA Evaluation System for Highly Efficient Sequencing of Splicing Variant cDNAs

Jun-ichi Yamamoto[1]
yamamoto-j@reprori.jp

Naoto Hatano[1]
hatano-n@reprori.jp

Kenji Araki[1]
araki-k@reprori.jp

Hiroshi Makita[1]
makita-h@reprori.jp

Kouichi Kimura[2]
kokimura@crl.hitachi.co.jp

Ai Wakamatsu[1]
wakamatsu-a@reprori.jp

Shizuko Ishii[1]
ishii-s@reprori.jp

Tetsuo Nishikawa[1,2]
nisikawa-t@reprori.jp

Takao Isogai[1]
isogai-t@reprori.jp

[1] Reverse Proteomics Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan
[2] Biosystems Research Department, Central Research Laboratory, Hitachi, Ltd., 1-280 Higashi-Koigakubo, Kokubunji, Tokyo 185-8601, Japan

Keywords: cDNA, splicing variant, alignment, genome, mapping

1 Introduction

In the full-length human cDNA sequencing project supported by New Energy and Industrial Technology Developmental Organization (NEDO), about 30,000 full-length and more than one million 5' one-pass cDNA sequences of oligo-capping cDNAs were determined. Consequently, the NEDO human splicing variant cDNA project started in 2002, taking advantage of these oligo-capping cDNAs which are regarded as a valuable source of splicing variant cDNAs. In the project, splicing variant cDNAs are selected from more than one million oligo-capping cDNAs based on the 5' one-pass sequences of those cDNAs.

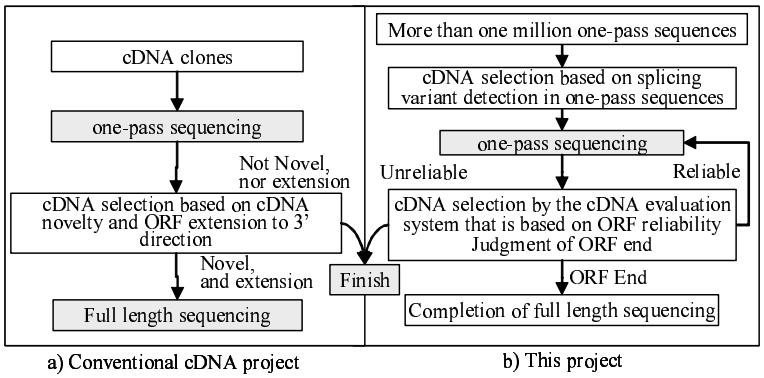

Figure 1: Comparison of flow of splicing variant sequencing project with conventional cDNA projects.

In conventional cDNA sequencing projects as shown in Figure 1 a), firstly one-pass sequences of many cDNAs are determined, secondly, those sequences are filtered based on novelty and quality of ORFs in the sequences. Though this method sequences novel cDNAs rather efficiently, it often mistakes to select cDNAs when there are problems in the region not covered by the 5' end one-pass sequences, because the selection is based only on the 5' end one-pass sequences. In fact, it has been clarified that there are frame-shifts, immature cDNAs, truncated cDNAs, and point mutations, etc in cDNA libraries. This means that it is necessary to evaluate the cDNAs during their sequencing, to obtain full-length sequences of the problem-free cDNAs efficiently.

Thus, in the splicing variant sequencing project, we choose the sequencing strategy shown in Figure 1 b) where, first, we take an initial 5' end one-pass sequence, then we repeatedly evaluate and determine extended one-pass sequence to obtain a full-length sequence. We developed a cDNA evaluation system that finds various types of cDNA problems and detects splicing variants.

2 Methods and Results

2.1 An Overview of the Clone Evaluation System

An overview of the cDNA evaluation system is shown in Figure 2. For a given cDNA sequence, the following three steps for checking the cDNA are performed. 1) check on the reliability of the cDNA. 2) detection of the splicing variants in the cDNA. 3) check on the transcription start site. Based on the checking results, the judgment whether or not the successive sequencing will be performed for extension of the sequence. The cDNA evaluation system consists of a graphical check and a sequence level check. We developed the following three graphical checking tools. 1) Translated Region Inspector (TRins) [2]. 2) cDNA sequence check system [3]. 3) Intris [1]. TRins helps us to check the sequence from a point of cDNA problems, such as frame-shifts, immature cDNAs, truncated cDNAs, and point mutations by displaying coding potential, and similarity with proteins, ESTs, and genome sequences. The cDNA sequence check system has similar functions as TRins, but has a unique function to display the similarity information on three open reading frames. Intris helps us to view splicing variants easily by displaying the alignments of cDNA onto genome sequences with intron regions compressed. The problems that are detected by using the graphical tools are further analyzed by the sequence level checking tool. The sequence level checking is also necessary to detect point mutations.

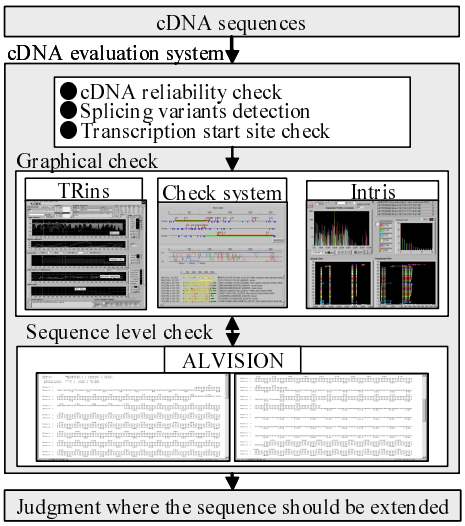

Figure 2: An overview of the clone evaluation system.

A newly developed sequence level checking tool, ALVISION, aligns two cDNA sequences that are splicing variants to each other, allowing large gaps. Figure 3 shows an overview of ALVISION. For given two DNAs or two amino acids sequences, ALVISION executes BLAST (bl2seq). After removing redundancy based on a simple rule from local alignments obtained by a BLAST calculation, a global alignment is built by assembling those local alignments. These graphical and sequence level checking tools help us to find the clone problems and splicing variants efficiently.

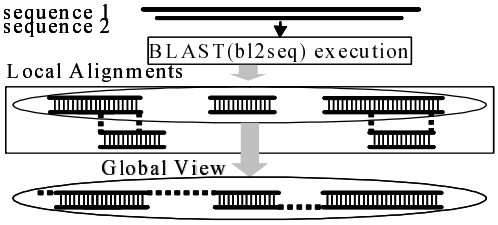

Figure 3: An overview of ALVISION.

3 Future Developments

In the current cDNA evaluation system, a user has to make a judgment based on the graphical data. We plan to automate the judgments in a future release. Although a complete automation is difficult to achieve, we expect that listing up of problematic clones can improve the sequencing efficiency as a whole. This work was supported by a grant from NEDO Project of the Ministry of Economy, Trade and Industry of Japan.

References

[1] Kimura, K. *et al.*, Intris:a viewer for cDNA-genome alignments enabling efficient detection of splicing variants and expression profiles, *Genome Informatics*, 13:548–550, 2002.

[2] Kimura, K. *et al.*, Translated region inspector for cDNA sequences, *The 5th International Workshop on Advanced Genomics*, 5:91, 2003.

[3] Nishikawa, T. *et al.*, An extracting system of accurate ORFs from cDNA sequences, *Genome Informatics*, 13:545–547, 2002.

[4] Wakamatsu, A. *et al.*, Analysis of splicing patterns of human full-length cDNAs, *The 5th International Workshop on Advanced Genomics*, 5:87, 2003.