An Approach to Alternative Gene Transcripts Discovery through Assembly of Fragmental cDNA Sequences

Ryotaro Irie irie@hri.co.jp
Yasuhiko Masuho masuho@hri.co.jp
Keiichi Nagai k-nagai@hri.co.jp
Helix Research Institute, Inc., 1532-3 Yana, Kisarazu-shi, Chiba 292-0812, Japan

Keywords: cDNA, alternative, transcript, assembly

1 Introduction

Genome sequencing efforts for Homo sapiens and other model organisms, which will be completed in a few years or have been completed, have opened the post-genomic age where the function of genes should be explored. The coding regions of genes on the genome sequences are identified or predicted by comparison with cDNA sequences and by *ab initio* methods.

However, each gene often has several gene transcripts (mRNAs). Thus, it is helpful or might be necessary to know the transcription multiplicity and alternative transcripts of each gene before one proceeds to the detailed study of the gene, because alternative transcripts from a gene (locus) can provoke different biological functions. Besides, the transcript-oriented classification of fragmental cDNA sequences, which are usually accompanied by expression-related information (as in dbEST of GenBank), could supply us with the additional expression information of each transcripts (tissue, developmental state, disease, etc.) [2]. Even in our full-length cDNA sequencing project [1], sequencing parts of cDNAs (including 5' end sequence with complete 5'UTR) is the first step to be executed, and the transcript-oriented classification of fragmental sequences as well as the gene-oriented one is important to obtain a cDNA sequencing strategy.

The previous works on transcript-oriented classification of the fragmental cDNA sequences derived from a gene or associated with a gene index are represented by a work using a multiple-alignment software [2] and a work using an assembling software which is tailored for assembling sequences to create a genome sequence [4, 3]. We developed a DNA sequence assembly program which is tailored for assembling fragmental cDNA sequences to create all contigs (sets of consistently aligned fragments) each of which could correspond to a transcript.

2 Method and Results

Major steps of our DNA sequence assembly algorithm (called MakeAllContigs) are described as follows.

1. Computation of alignment between any pair of members of a given set of DNA sequences.

   We used the popular and reliable pairwise alignment software, blastn (gapped alignment software for DNA sequences in WU-BLAST 2.0 package). We assume that the directions of sequences are given appropriately (e.g., by clustering or homologue collection).

2. Selection of a *nucleus* sequence from the members which are not assigned to any contigs.

3. Construction of a new contig starting from the *nucleus* selected in Step 2.

The following construction cycle is repeated until the current contig grows as large as any more tried member of the given set cannot be incorporated in the current contig. *Construction cycle* : Each member of the given set is tried to *major* contig members. (Initially, the *major* contig member is the *nucleus* only.) The trial starts from the most upstream *major* contig member. If inconsistency with any *major* contig member is detected in the course of the trial, the tried sequence is rejected by the contig. If the tried sequence is virtually included by any *major* contig member, the tried sequence is incorporated as a subordinate of the *major* contig member. If the tried sequence virtually includes one or more *major* contig members and the overall consistency is found, the included *major* contig members become the subordinates of the tried sequence (a new *major* contig member in place of the former *major* contig members). If overall consistency is found in other cases, the tried sequence is simply incorporated as a *major* contig members.

4. If any member of the given set is not assigned to any contig, go to Steps 2. and 3. If all members of the given set are assigned to any contig, the whole process for the initially given sequence set are finished.

The contigs produced by MakeAllContigs sometimes overlap each other as sets of sequences because all consistent sequences are incorporated in a contig. It is different from the splitting strategy in the previous assembling softwares. According to our algorithm, each contig is represented by *major* contig members, each of which is longer and often more reliable than the subordinates. For example, a contig including a full-length cDNA sequence has only one *major* contig member, that is, the full-length one. We applied MakeAllContigs to some UniGene clusters (including clusters to which a previous transcript-oriented classification technique [2] was applied). As a result of the application, it was confirmed that the known alternative transcripts are assigned to different contigs. In addition to it, the predictability of a tissue-specificity of a transcript which is derived from a UniGene cluster was confirmed.

References

[1] Barker, S., Japanese genomics combines state and industry backing [news], *Nature*, 380(6573):375, 1996.

[2] Burke, J., Wang, H., Hide, W. and Davison, D.B., Alternative gene form discovery and candidate gene selection from gene indexing projects, *Genome Res.*, 8(3):276–290, 1998.

[3] Huang, X. and Madan, A., CAP3, *Genome Res.*, 9(9):868–877, 1999.

[4] Quackenbush, J., Liang, F., Holt, I., Pertea, G. and Upton, J., The TIGR gene indices, *Nucleic Acids Res.*, 28(1):141–145, 2000.