Mapping Database of cDNA and Genome Designed to Use for Various Applications

Akifumi Yamashita uhmin@gen-info.osaka-u.ac.jp
Ken Kurokawa ken@gen-info.osaka-u.ac.jp
Teruo Yasunaga yasunaga@gen-info.osaka-u.ac.jp
Genome Information Research Center, Osaka University, 3-1 Yamadaoka, Suita city, Osaka 565-0871, Japan

Keywords: genomic position of cDNA, alternative splicing, noncanonical dinucleotides

1 Introduction

There are many databases or tools that handle cDNA-to-genomic alignment on web [3, 2], and they are quite convenient for the aims supposed by the authors. However, for the other purposes which were not supposed by them, or for new cDNAs not registered in the database, they are not effective. That is why we created a procedure making database of cDNA position on genome. The two types of databases were made. One is the database constructed on MySQL relational database management system, which is adequate for searching by cDNA sequence entry name, genomic position, exon length, intron length, or calculating statistics like average exon or intron length. The other is a flat file database which is useful for application programs such as finding sequence motif in the surrounding region of a gene. Using this procedure, we made a cDNA-genome mapping database from FANTOM [2] cDNA clones of mouse and NCBI mouse genome. A database can be made from any other cDNA-genomic sequence pair, of course.

2 Materials and Methods

cDNA sequences were got from the RIKEN full-length cDNA clones of mouse (FANTOM) [2], and contigs of mouse genomic sequence were from NCBI [3].

A cDNA was matched with a genome contig by blastn. A large number of matches were reported by blastn and put into the first filter program (Fig 1-1). At first, the match with highest score was selected as the exon candidate. Then, the match with highest score among the reminder matches was selected as the exon candidate only if the match was not conflict with the already selected exon candidates. Conflictions include match direction, overlapped with already selected exon candidates and mapping position. This process was performed repeatedly while the reminder matches exit. Then a set of exon candidates was obtained for a genome contig.

The first filter program was applied for every genome contigs. In the second filter, sets of exon candidates were compared with each other by the total blast score of matches selected as exon candidates for each genome contig. The exon candidates with the highest total score were chosen as the exons (Fig 1-2). Finally the precise positions of exon-intron boundaries were determined using splicing site GT-AG rule. During the process, there were many cases where splicing site GT-AG could not be found; they were stored in a verifiable manner.

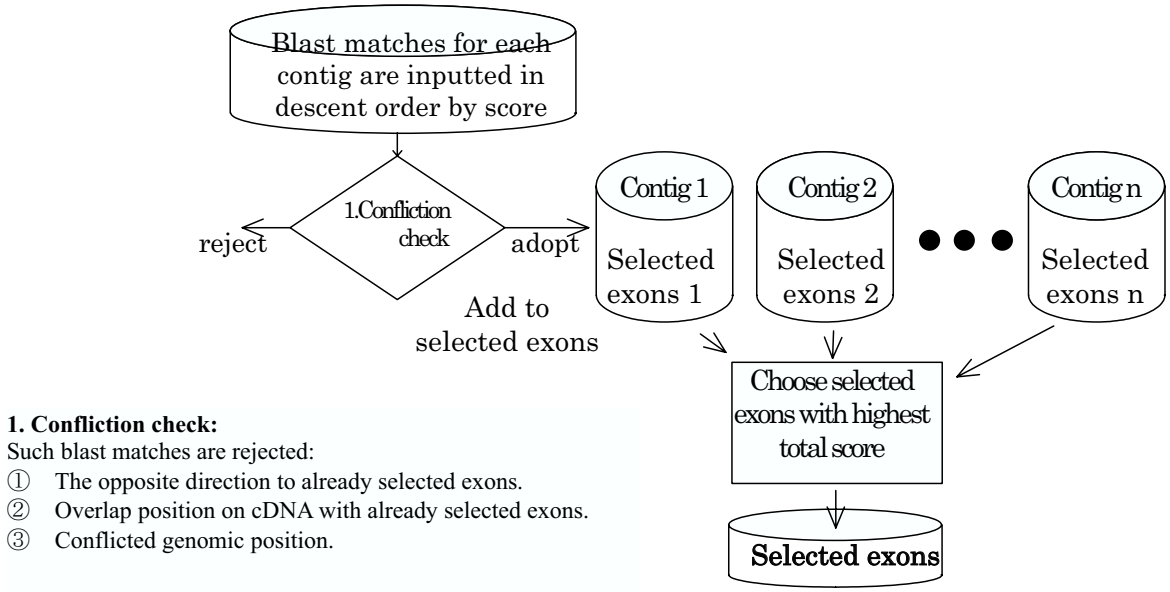

Figure 1: The procedure of mapping cDNA and genome.

3 Results and Discussion

We processed 60770 FANTOM cDNA sequences and predicted 290,668 exons from 60,515 cDNA. There were 15,682 pieces of cDNAs not mapping to the genomic sequence, which had more than 10bp in length. The reasons why there exist such pieces or regions of cDNAs would be the followings; 1) the corresponding genome sequence is not determined yet, 2) some genes are divided into different contigs. The present program does not consider such the case. To solve this problem additional program will be needed, 3) blast search does not report too short match, and 4) sequencing error.

Several statistical values for exon and intron were calculated using MySQL database. The average lengths were 261.6bp for first exons, 150.6bp for internal exons between introns, 1039.5bp for last exons, 1890.6bp for intronless genes and 5624.8bp for introns. It is very interesting that the average length of internal exsons is remarkably shorter than that of first exsons or last exons.

Among 218647 splicing junctions, there were 4,530 pairs (2.07%) which had no canonical dinucleotide (GT-AG). This proportion (2.07%) is consistent with the previous reported value (1.29%) [1]. Although most of them may be due to sequencing error, some of them should be candidates of non-canonical dinucleotide pairs. Further investigation is needed to determine the gene with noncanonical dinucleotides at the splice junctions.

References

[1] Burset, M., Seledtsov, I.A., and Solovyev, V.V., Analysis of canonical and non-canonical splice sites in mammalian genomes, *Nucleic Acids Res.*, 28(21):4364–4375, 2000.

[2] FANTOM DB, `http://fantom.gsc.riken.go.jp/db/search/`

[3] NCBI Map Viewer, `http://www.ncbi.nlm.nih.gov/mapview/static/MVstart.html`