Towards Splicing Pattern Detection Based on cDNA Sequences

Tomohiro Yasuda tyasuda@crl.hitachi.co.jp
Koichi Kimura kokimura@crl.hitachi.co.jp
Tetsuo Nishikawa nisikawa@crl.hitachi.co.jp
Central Research Laboratory, Hitachi, Ltd., 1-280, Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan

Keywords: alternative splicing, suffix tree

1 Introduction

This study aims at rapid detection of splicing patterns from given cDNA sequences. When given sequences involve sufficient variety of transcript variants, we can identify all splicing patterns by sequence comparison. Our method searches for partial sequences, each of which appears completely from its beginning to its end, or is completely missed, in each of the given cDNA sequences. Since those partial sequences can be regarded as concatenations of exons, searching for them plays a key role in detecting splicing patterns.

2 Methods and Results

In this study, the partial sequences described above are refered as unique exon blocks (UEBs). The following is a formal definition of UEBs. To establish this definition, we extended Delpher's MUMs [1] so that they can be defined for more than two sequences.

Definition 1 *A UEB e is a string that satisfies the conditions below.*
(A1) e is a substring of an input sequence.
(A2) e is longer than u bp, where u is a parameter.
(A3) e appears at most once in each sequence.
(A4) e does not extend beyond any boundaries of MUMs.
(A5) e is not a substring of any other UEBs.

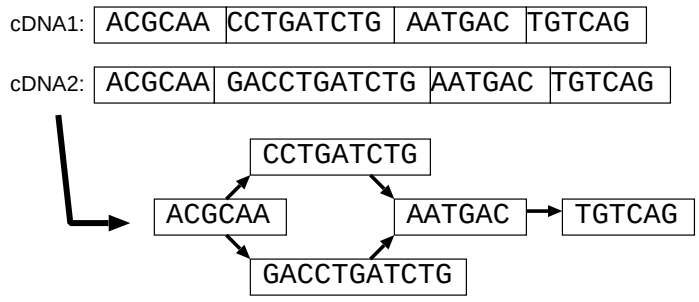

Figure 1: Alternatively spliced cDNA sequences. Difference of splicing patterns can be detected through comparison of cDNA sequences.

We have developed an algorithm that finds all UEBs, and only UEBs, within $O(N)$ time, where N is the number of all bases in given cDNA sequences. To identify UEBs, this algorithm conducts depth-first traversals C times on a *suffix tree* [2] built for all given cDNA sequences, where C is an integer independent of the number and length of given cDNA sequences. At first this algorithm searches for MUMs. It then finds another class of strings which are called right UEB-holders. And finally, it identifies UEBs. In a preliminary experiment, the splicing pattern of the WT1 tumor suppressor gene was reconstructed through the analysis of three transcript variants by our method, as shown in Figure 3. The reconstructed pattern was the same as the one described in the RefSeq database.

This work is supported by a Grant from the NEDO Project of the Ministry of Economy, Trade and Industry of Japan.

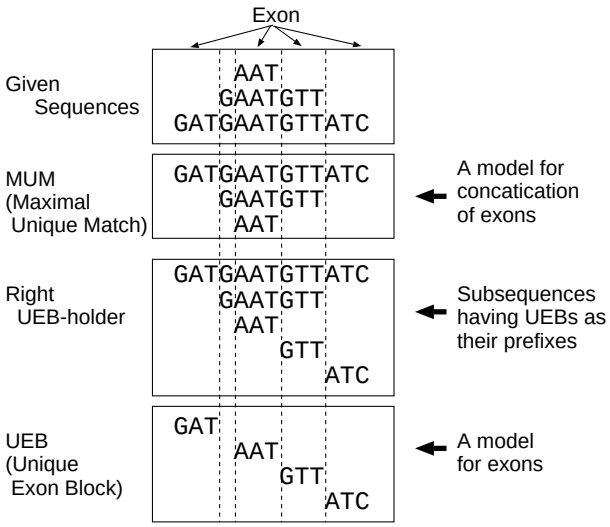

Figure 2: MUMs, right UEB-holders, and UEBs. In order to capture UEBs, our methods detects MUMs and right UEB-holders.

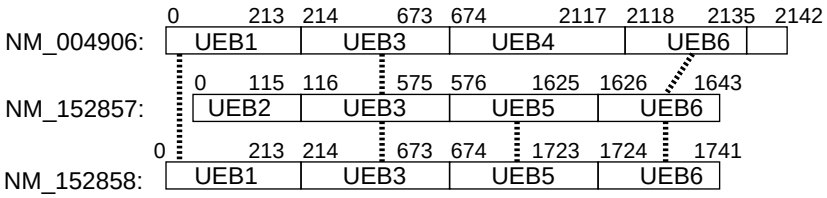

Figure 3: The splicing pattern of WT1 tumor suppressor gene detected by our method. Alternative 5'-ends and 3'-ends were correctly determined. UEB6 was a poly-A signal.

References

[1] Delcher, A., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., and Salzberg, S.L., Alignment of whole genomes, *Nucleic Acids Res.*, 27:2369–2376, 1999.

[2] Gusfield, D., *Algorithms on strings, trees, and sequences. Computer Science and Computational Biology*, Cambridge University Press, 1997.