Clustering and Multiple Alignment for a Large Number of cDNA Sequences Including Splicing Variants

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

Keywords: clustering, multiple alignment, splicing variants, suffix tree

1 Introduction

Tens of thousands of full-length cDNA sequences has been accumulated in public databases. They include lots of splicing variants. To analyze these sequences, we have to know which sequences are derived from the same genes and to analyze the differences among splicing variants of the same gene. Therefore, a method that conducts clustering and multiple alignment for a large number of cDNA sequences is required. Although aligning cDNA sequences to genome sequences is a widely accepted method for analysis of cDNA sequences, we should have another method that does not need genome sequences because: (i) both of genome sequences and alignment of cDNA sequences to genome sequences may not be perfect and include some errors, and (ii) genome sequences are not yet available for many organisms. However, it is computationally hard to compare a large number of cDNA sequences with each other to obtain their multiple alignment.

There have already been techniques proposed to quickly compare multiple sequences, although they aim not at comparing cDNA sequences but at aligning several closely related genome sequences. Delcher developed MUMer to compare two genome sequences [1]. At first, MUMer searches for substrings called maximal unique matches (MUMs), which are substrings (i) shared by two provided sequences, (ii) appearing at most once in each of provided sequences, and (iii) unable to be extended remaining to be a substring of both of provided sequences. Then, MUMer aligns bases that does not belong to MUMs. Höhl developed Multiple Genome Aligner (MGA for short) that can compare more than two sequences [2]. As MUMer searches for MUMs, MGA searches for maximal multiple exact matches (multiMEMs), which are substrings (i) shared by all provided sequences, (ii) unable to be extended remaining to be a substring of all provided sequences. Then, MGA aligns bases between multiMEMs. A problem common to both of them is that the substrings to be searched for in the first step must be shared by all provided sequences, which means they cannot be applied to cDNA sequences since alternative exons are not directly recognized.

In this study, we developed a method that can quickly conducts clustering and multiple alignment for thousands of cDNA sequences.

2 Methods

Our method consists of the following three steps.

Identification of Unique Exon Blocks (UEBs). First, our method identifies substrings called unique exons blocks (UEBs) [3]. They approximate sequences of exons or concatenations of exons on the assumption that there are no SNPs, sequencing errors, or other sequence variations. It is
proved that UEBs can be identified in $O(N)$-time with suffix trees, therefore our method can quickly identify UEBs.

Integration of UEBs into Fragments. Second, our method integrates any UEBs into a single fragment if and only if they always appear next to each other in the same order. We developed this step so that our method is available to sequences with variations such as SNPs. Since UEBs have to be exactly the same sequences at any of their appearances by their definition, they cannot include any sequence variations. However, this step allows us to recognize a sequence of an exon or a concatenation of them as a single fragment even if they are split into multiple UEBs because of sequence variations. Our method also constructs fragment sequences, each of which is a sequence of fragments detected in a provided cDNA sequence and arranged in the same order as the fragments appear in the cDNA sequence.

Clustering and Multiple Alignment of Fragment Sequences. Finally, our method conducts clustering and multiple alignment of fragment sequences constructed in the previous step. This step can be completed in short time in practice since the number of fragments is much smaller than that of bases.

3 Results

We chose 6357 full-length cDNA sequences (19.6Mbp) in the RefSeq database [4], and applied our method to them. All of chosen sequences had a key phrase “transcript variant” in their DEFINITION field in the GenBank flat-file format. It took only 226 seconds to complete overall computation, which witnesses our method is quite fast and scalable. We also examined the result of clustering and multiple alignment. The locus information in the LocusLink database [4] was converted to form a set of reference clusters $C_0$, which was compared with a set of clusters $C_1$ obtained with our method. We found that a measure which is so-called Jaccard Index defined as $J(C_0, C_1) = |C_0 \cap C_1|/|C_0 \cup C_1|$ was 0.88. Therefore many clusters in our result were consistent with the LocusLink database.

Acknowledgment

This work was partly supported by the New Energy and Industrial Technology Development Organization (NEDO), Japan.

References


