Dataset Construction for Gene Structure Prediction and Alternative Splicing Analysis

Masahiko Mizuno  Hideki Nagasaki  Makiko Suwa
mizuno@cbrc.jp  h-nagasaki@aist.go.jp  m-suwa@aist.go.jp
Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-41-6 Aomi, Koutou-ku, Tokyo 135-0064, Japan

Keywords: gene prediction, alternative splicing, splice sites, ORF, pre-mRNA, human genome, Genbank database

1 Introduction

The performance of gene finding from genome sequences strongly depends on the accuracy of splice site prediction. Recent gene finding programs, however, still do not reach enough levels. To improve the accuracy of splice site prediction, it is required to understand the splicing mechanism and to make a model from clear experimental evidences.

For this purpose, genomic full-length precursor mRNA sequences (FL-pre-mRNAs), together with expression information are indispensable. The FL-pre-mRNAs have entire gene structure such as the 5' and 3' end of mRNA, initiation codon, splice sites, stop codon, and polyadenylation signals, etc. They also contain all the alternative splice sites except the first or last exons in alternative transcripts. However, databases of FL-pre-mRNAs are still not reported in previous works.

Aligning expressed sequence tags (ESTs) to the genomic sequences has been a common method for gene prediction or splice site analysis (1, 3). However, ESTs are not suitable for collecting FL-pre-mRNAs because ESTs are partial sequences and the 5' ends of mRNAs are unknown in most cases, and even EST contigs clustered in UniGene (2) or RefSeq database (4) are not evident to be full-length. It is because ESTs are single sequencing reads that contain mutations, insertions, or deletions (5).

Growing genomic and EST sequence data, computational approach has become one of methods to annotate the sequences as putative genes or ORFs. Whereas, Genbank database has accumulated the entries in which genomic complete protein-coding sequences or full-length mRNA sequences are characterized by experimental evidence. The sequences and the annotation (the positions of gene boundaries and functional signals) with the information more reliable than that determined by in silico prediction are expected to be high quality. Thus, we constructed datasets with experimental annotation from Genbank database for gene structure prediction and splice site analysis. Moreover, the analysis for constitutive and alternative splice sites with the correlation with several biological descriptors will be discussed.

2 Materials and Methods

Two sequence datasets were constructed from human entries in Genbank database (release 125); genomic full-length pre-mRNA (FL-pre-mRNA) and full-length mRNA sequences, respectively. Genes encoded by mitochondrion were also excluded. Although annotation in Genbank entries includes experimental evidence or computational prediction, only sequences with experimental annotation were selected.
Sequences were compared between the two datasets. If mRNAs matched pre-mRNA sequences with 99.5% and more identity, they were considered to be identical. The alignment of these identical sequences, with the position of splice sites in "FEATURE" table of Genbank, demonstrated two classes; constitutive and alternative splice sites. The splice sites aligned in all the identical sequences were categorized as constitutive splice site. Otherwise, the sites were as alternative splice site. The four classes of the aligned splice sites constructed the consensus sequences, for the 5' and 3' splice sites, respectively, and for constitutive and alternative splice sites, respectively.

3 Results and Discussions

1733 candidates of genomic FL-pre-mRNAs were retrieved from human entries in Genbank database, and 936 full-length mRNA sequences were selected. The data quality was examined in detail, and also the list of biological functions of the genes was constructed.

Consensus sequences at splice sites have been constructed from single sequence dataset which includes both constitutive and alternative splice sites. Constitutive splice sites are selected in any condition of cells, while alternative splice sites are selected depending on the cell types or conditions; tissues or developmental stages. We compared these two classes of consensus sequences to demonstrate the similarity and difference of the sequence patterns and profiles. In addition, the correlation between the consensus patterns and the tissue distribution will be discussed. Recent studies reported that alternative splice sites were weak in comparison with known consensus sequences. That suggests additional sequences around splice sites might function as splicing regulatory element. We will discuss these possible novel sequence elements that involve in the splicing.

4 Future works

We will apply the profile of consensus sequences at splice sites and the other splicing sequence elements to develop algorithms to predict splice sites and identify gene sequence structures in human genomic sequences. The algorithms for gene prediction, named OpenGenes, will be developed. Furthermore, we will focus on the order, combination, and balance of the recognition of the sequence signals to develop the hypothetical models of splicing machinery. The models will be tested by the accuracy of gene prediction and the correlation between splicing sequence elements in the datasets constructed in this study. These results could elucidate the mechanism of splicing.

5 Acknowledgments

This work was supported in part by The New Energy and Industrial Technology Development Organization (NEDO) industrial technology fellowship program.

References


