Human Full-Length Pre-mRNA Sequence Dataset for Computational Gene Prediction and Alternative Splicing Analysis

Masahiko Mizuno[1] **Osamu Gotoh**[1,2] **Makiko Suwa**[1]
mizuno@cbrc.jp o.gotoh@aist.go.jp m-suwa@aist.go.jp

[1] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-43 Aomi, Koto-ku, Tokyo 135-0064, Japan
[2] Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Keywords: human full-length pre-mRNA sequence, gene prediction, alternative splicing, splice site, exon-intron structure, GenBank annotation, experimental evidence, EST-genome alignment

1 Introduction

Computational gene prediction has recently become essential to identify all the genes from enormous genome sequences and to define their functions. However, gene prediction methods still show low specificity (45% at the exon level) [7]. Although computationally predicted data by automatic annotation system are growing rapidly, experimentally verified data provide the cornerstone for improvements of gene prediction. Because there is no guarantee that the predicted events occur *in vivo*. It has become important to discriminate experimentally verified data from computationally predicted data [5].

Aligning expressed sequence tags (ESTs)/mRNAs to the genomic sequences has been a practical approach to detect gene regions and to identify alternative splicing on a genomic scale. In previous studies EST-genome alignments were made using 90-93% sequence identity as threshold [2, 3, 4]. However, their low thresholds allow ESTs to incorrectly align with paralogous genes or pseudogenes. Moreover, EST sequences have low-quality sequencing region or contaminated region [6].

We propose to collect true data in the view of quality rather than quantity for further improvement of gene prediction and functional analysis. In this study, we retrieved human full-length pre-mRNA (FL-pre-mRNA) sequences by parsing experimentally verified annotation in GenBank records, and then confirmed complete length and exon-intron boundaries (splice sites) in the sequences by multiple EST alignment. Additionally, we identified and classified constitutive and alternative splice sites on the sequences. Moreover, we demonstrate novel findings about human gene structure and splice sites based on this high-quality dataset.

2 Materials and Methods

By carefully checking keywords in GenBank annotation, we retrieved human protein-coding FL-pre-mRNA sequences from primate division of GenBank database (release 126). Pseudogenes, mitochondrion-encoded genes, partial sequences, and sequences with computationally predicted annotation were discarded. All-against-all sequence comparison in the selected FL-pre-mRNAs was performed by gapped BLASTN with $\geq$97% identity and $\geq$50% overlap length to remove redundancy, and to retain splice variants that have identical sequences but distinct exon-intron structures. We searched ESTs identical to the remained FL-pre-mRNA sequences by a homology search program, megablast [8] with $\geq$97% identity and $\geq$50 nt alignment length. The detected ESTs were precisely re-aligned to the corresponding FL-pre-mRNAs by a dynamic programming alignment tool, ALN [1] with $\geq$97% identity, $\geq$100 nt of total alignment length, and $\geq$75% of the EST length.

3 Results and Discussion

We retrieved 1,104 human FL-pre-mRNA sequences by carefully examining GenBank keywords. After comparing these sequences and exon-intron structures in their annotations, 830 non-redundant

genes consisting of 979 splice variants were obtained. The 830 genes were aligned with 221,649 EST sequences. Since additional exons were found at upstream/downstream of the annotated pre-mRNA regions by EST mapping, 80 genes were discarded as partial. After purging the remained 750 complete genes with stringent sequence similarity (≥50%), 723 non-homologous genes were constructed as the final dataset.

In order to reduce incorrect alignments of sequencing errors, paralogous sequences or pseudogenes, we refined a few cutoff values of EST alignment as described above. Against the 723 genes, 2,732 ESTs with <100 nt of total alignment length were removed at the stage of EST alignment by ALN. In addition, 19,310 ESTs were discarded as the total alignment length was shorter than 75% of the EST length. Consequently, 159,381 identical ESTs were aligned to the 723 genes (220.4 ESTs per gene). A sufficient number of ESTs (107.9 ESTs per splice site) confirmed each splice site. A total of 11,982 splice sites were identified in the 723 genes. Of these, 2,245 and 4,167 sites were defined as constitutive and alternative splice sites, respectively. Constitutive splice sites were used in all the splice variants through GenBank annotation and EST mapping. Alternative splice sites were used in at least one annotation or more than two ESTs, while not in another annotation or more than two other ESTs. Therefore, both types of splice sites are highly reliable.

11,039 and 8,950 splice sites were identified from GenBank annotation and EST mapping, respectively. 3,032 splice sites (25.3%) were detected by GenBank annotation alone. These sites were not detected by EST mapping. This result suggests our dataset retains low expressed genes or rare splice variants. On the other hand, EST mapping found 943 novel sites (7.9%) which were not annotated in GenBank records. Therefore, our approach detected more splice sites than those in previous studies which used single data resource. Furthermore, 8,007 splice sites (66.8%) were confirmed doubly by GenBank annotation and EST mapping.

Sequence comparison between the 723 genes and Refseq mRNAs showed that 70.3% of our genes had the extended or equal 5'-end sequences, or were not found in the Refseq mRNAs. This result suggests our genes are well confirmed about complete exon-intron structure. A high-quality dataset of the FL-pre-mRNA sequences could be applied to more accurate gene prediction from the human genome sequences and to alternative splicing analysis.

4 Acknowledgments

This work was supported in part by The New Energy and Industrial Technology Development Organization (NEDO) industrial technology fellowship program. The computation was performed partly on the MAssively parallel computer for Genome Informatics (Magi) Cluster at Computational Biology Research Center.

References

[1] Gotoh, O., Homology-based gene structure prediction: simplified matching algorithm using a translated codon (tron) and improved accuracy by allowing for long gaps, *Bioinformatics*, 16(3):190–202, 2000.

[2] Honkura, T., Ogasawara, J., Yamada, T., and Morishita, S., The gene resource locator: gene locus maps for transcriptome analysis, *Nucleic Acids Res.*, 30(1):221–225, 2002.

[3] Kan, Z., Rouchka, E.C., Gish, W.R., and States, D.J., Gene structure prediction and alternative splicing analysis using genomically aligned ESTs, *Genome Res.*, 11(5):889–900, 2001.

[4] Kan, Z., States, D.J., and Gish, W.R., Selecting for functional alternative splices in ESTs, *Genome Res.* 12(12):1837–1845, 2002.

[5] Mathé, C., Sagot, M.F., Schiex, T., and Rouzé, P., Current methods of gene prediction, their strengths and weaknesses,*Nucleic Acids Res.*, 30(19):4103–4117, 2002.

[6] Wolfsberg, T.G. and Landsman, D., A comparison of expressed sequence tags (ESTs) to human genomic sequences, *Nucleic Acids Res.*, 25(8):1626–32, 1997.

[7] Zhang, M.Q., Computational prediction of eukaryotic protein-coding genes, *Nat. Rev. Genet.*, 3(9):698–709, 2002.

[8] Zhang, Z., Schwartz, S., Wagner, L., and Miller, W., A greedy algorithm for aligning DNA sequences, *J. Comp. Biol.*, 7(1-2):203–214, 2000.