**GENE RECOGNITION USING EST DATA: UNEXPECTEDLY FREQUENT ALTERNATIVE SPLICING OF HUMAN GENES.**

MIRONOV A.A.1, GELFAND M.S.2+

1 State Center of Biotechnology NIIGenetika, Moscow, 113545, Russia;

2 Institute of Protein Research, Russian Acad. Sci., Pushchino, 142292, Russia;

+Corresponding author: e-mail: misha@imb.imb.ac.ru

Keywords: human genes, alternative splicing, gene recognition, contigs

**Introduction**

Procrustes-EST was used to predict exon-intron structures of human genes using EST contigs from TIGR Human Gene Index. It turned out that 35% genes are spliced alternatively, and the majority of splicing event occured in 5’ untranslated regions. Most of the alternative splices of coding regions generate additional protein domains rather than alternating domains.

The total size of human genomic DNA sequences in GenBank exceeds 100 million bases and is rising exponentially. It is also estimated that at least half of the human genes are represented in the existing EST collections [1]. However, the majority of human genomic sequences are uncharacterized or characterized incompletely.

Several groups (most notably, GRAIL [2]) attempted to use ESTs for genomic DNA annotation. However, the problem is not a trivial one. Simple matching of ESTs to genomic sequences by BLAST-like programs is not sufficient, since the most informative ESTs are those that correspond to several exons, whereas BLAST does not map exactly the exon-intron boundaries. Besides a considerable number of ESTs map to intergenic or intronic regions, or could be products of aberrant or incomplete splicing. It is likely that these matches constitute at least one fifth of the existing EST databases [3].

We have developed a program for prediction of exon-intron structure of genomic DNA fragments using EST data. The program Procrustes-EST is based on modified spliced alignment algorithm [4]. When applied to known human genes and TIGR EST assemblies [5], the program found a large number of alternatively spliced genes (about 40%). Most of the alternative splicing events occured in 5’ untranslated regions. In many cases the use of the program allowed to link and merge the exising assemblies into single contigs.

**1. Data and Methods**

Human genomic DNA fragments containing complete multi-exon genes were compiled by merging samples from [4] and [6]. Genes were considered to be duplicates if their described exon-intron structures were identical (minor differences in intron lengths were allowed), and the longest representative from each group of duplicates was selected. The final sample consisted of 392 genes.

Repeats were filtered from the genomic sequences by RepeatMasker [7]. EST assemblies corresponding to a gene (*targets*) were selected from the TIGR Human Gene Index [5] using BLASTN [8]. The E-value threshold was set to 10-50; at most 10 highest scoring assemblies per gene were retained. Genes having at least one common target were grouped into clusters. The target sets for each cluster were merged and ascribed to each member of the cluster. Finally, the sequences complementary to the targets were added to the target sets.

Exon-intron structures were predicted using Procrustes-EST. This program predictes candidate splicing sites with a very weak threshold and then finds a chain of exons with the highest similarity to the target using the spliced alignment algorithm [4]. The threshold for accepting the prediction was set to 80% relative similarity [9] so as to avoid interference of members of multigene families. Local drops of similarity between the predicted exon chains and the targets were analyzed manually and several prediction errors caused by loss of sites were corrected.

At the postprocessing stage predictions corresponding to one gene were merged if there were no contradictions. More exactly, two chains of exons were merged if they had at least one common splicing site together with the adjacent part of an exon. All possible superassemblies were formed. EST contigs that could not be merge with other contigs ("orphans") were ignored.

**2. Results**

Relationships between targets, superassemblies and genes are presented in Tables 1, 2, and 3

The alternatives at 5’ end (5’ forks) occurred in 73 genes (18.5%), internal alternatives (loops) in 41 genes (10.4%), and 3’ forks in 64 genes (16.2%). 23% of loops were generated by alternative acceptors, 16%, by alternative donors, 27% were exons that were present in one of the two variants, there were rate instances of retained introns, alternative introns and alternative exons, and 25% were complex cases that could not be classified. Further, 22% of 5’ forks were alternative 5’ exons, 18% had different start points and an additional intron in one of the variants, and the rest were complex cases. Finally, 11% of 3’ forks were alternative terminal exons, 35% had different end points and an additional intron in one variant, and the rest were complex cases.

Column 0 gives the number of chimaeric superassemblies generated because the procedure for merging of exon chains does not take retain long distance information.

Table 1. Distribution of the number of EST contigs merged during formation of superassemblies.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| # of contigs | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| % of superassemblies | 10.6% | 48.3% | 19.3% | 10.8% | 6.6% | 1.8% | 1.4% | 0.8% | 0.4% |

Table 2. Distribution of the number of superassemblies per contig. Column 0 gives the number of orphan contigs.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| # of superassemblies | 0 | 1 | 2 | 3 | 4 | 5 | 6 | >6 |
| % of contigs | 27.1% | 55.5% | 12.2% | 2.0% | 1.6% | 0.2% | 0.7% | 0.7% |

Table 3. Distribution of the number of superassemblies per gene. Column 1 gives the number of genes without alternative splicing.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| # of superassemblies | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | >9 |
| % of contigs | 65.6% | 18.6% | 4.6% | 5.4% | 1.3% | 1.5% | 0.5% | 1.0% | 1.5% |

**3. Discussion**

The results of this study demonstrate that the alternative splicing is much more frequent than one could expect judging from existing database annotations. Functionally, 80% of alternatively spliced genes had an alternative in 5’ untranslated region, whereas only 20% of them had alternatives in the coding region, and 19% in 3’ untranslated region (the total exceeds 100% since alternatives may occur in two or all three of these regions). The alternative splicing at 5’ end is probably a mechanism allowing the cell to use different promoters (with different regulation) for the same gene.

Another result is that the use of genomic data allows one to merge EST contigs in the situations where the EST overlaps alone provide insufficient evidence for contig construction. Indeed, 40% of superassemblies were produced by more than one contig. On the other hand, the fact that 10% of the superassemblies were chimaeric demonstrates that the alternative splicing of different exons is not independent.

More detailed analysis of complex cases is required before something can be said about the prevalence of particular types of alternative splicing. It seems that the majority of alternative splicing events within the coding regions produce additional protein domains rather than alternating domains.

**Acknowledgements**

This work was partially supported by the Russian State Scientific Program "Human Genome", Russia Fund of Fundamental Research, and the USA Department of Energy. We are grateful to Jim Fickett for running BLAST comparisons, to T.V.Astakhova for the help with the data, and to Pavel Pevzner and Mikhail Roytberg for useful discussions.

**References**

1. G.D. Schuler et al. Science 274, 540-546 (1996)
2. E.D. Uberbacher, Y. Xu. J. Comput. Biol. 4, 325-338 (1997)
3. T.G. Wolfsberg, D. Landsman. Nucl. Acids Res. 25, 1626-1632 (1997)
4. M.S. Gelfand, A.A. Mironov, P.A. Pevzner. Proc. Natl. Acad. Sci. USA 93, 9061-9066 (1996)
5. M.D. Adams et al. Nature 377 (Suppl.), 3-17 (1995)
6. D. Kulp, D. Haussler, M.G. Reese, F.H. Eckerman. 4th ISMB. AAAI Press, Menlo Park CA, 134-142 (1996)
7. A. Smit, P. Green (unpublished)
8. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman. J. Mol. Biol. 215, 403-410 (1990)
9. A.A. Mironov, M.A. Roytberg, P.A. Pevzner, M.S. Gelfand. Genomics (in press)