Proceedings of the 7th WSEAS International Conference on Mathematics & Computers in Biology & Chemistry, Cavtat, Croatia, June 13-15, 2006 (pp34-37)

ANALYSIS OF EXON SKIPPING EVENTS IN HUMAN AND MOUSE *

Chia-Hung Liu
Institute of Information Science, Academia Sinica, Taiwan

Fang Rong Hsu†
Bioinformatics Research Center, Department of Information Engineering and Computer Science, Feng Chia University, No. 100 Wenhwa Rd., Seatwen, 40724 Taichung, Taiwan

Hwan-You Chang
Department of life science
National Tsing-Hua University, Hsin-chu, Taiwan

Wei-Chung Shia
Department of Information Engineering and Computer Science,
Feng Chia University, Taiwan

HUI-LING PENG, YING TSONG CHEN
Department of Biological Science and Technology,
National Chiao-Tung, University

Yaw-Lin Lin, Yin-Te Tsai
Department of Computer Science and Information Management
Providence University, Taichung, Taiwan

Abstract: -Comparison of human and mouse genomics revealed a similar long range sequences organization, and many genes that are orthologous between human and mouse. Alternative splicing is a very important post-transcriptional event leading to an increase in the transcriptome diversity. Recently, genomics studies estimate that 40–60% of human genes undergo alternative splicing. It is interesting to appraise the level of conservation of alternative splicing. To address this question, we developed a bioinformatics method to identify alternative splicing events that are conserved and alternative splicing events that are not conserved between human and mouse. Here we report an analysis of 3,613 orthologous genes in human and mouse, which discover 2,628 alternative splicing events are conserved in both transcriptome and 2,783 alternative splicing events are not conserved. Further classification of these conserved alternative splicing events reveals that 239 events are due to exon skipping, 34 events are due to mutually exclusive, 1,540 events are due to 3' splice sites, 815 events are due to 5' splice sites, and no events are due to intron retention. By comparison, non-conserved alternative splicing events reveals that 609 events are due to exon skipping, 47 events are due to mutually exclusive, 1,048 events are due to 3' splice sites, 960 events are due to 5' splice sites, and 370 events are due to intron retention. We combined with the human exons and mouse exons to discover the cross-species alternative splicing events. The cross-species alternative splicing event means that in orthologous genes, there is no alternative splicing event in both species and their splicing forms are distinct. We discovered 235 such events.

* This work is supported in part by NCS Taiwan, grant no. NSC-92-3112-B-468-001, NSC94-2218-E-305-012.
† Corresponding author, e-mail: *frhsu@fcu.edu.tw*

1 Introduction

Eukaryotic genes contain stretches of DNA that usually do not code for protein products. These non-coding stretches, called introns, must be spliced out of pre-mRNA transcribed from the gene before the RNA can be translated into the protein. The protein coding parts of this splicing process accurately joins genes, called exons, segments into a continuous stretch of protein-coding mRNA. Alternative splicing (AS) the result is occurred when the splicing machinery (a spliceosome) produces different mature mRNAs from the transcript of a gene [1-3].

Comparative analysis of ESTs and cDNAs with genomic DNA predict a high frequency of alternative splicing in human and mouse genes. However, the extent of alternative splicing conservation among mammals is still being discussed [4-6]. If an alternative splicing event is conserved in both human and mouse, and thus more likely to be biologically significant and provides some advantage to the organisam. However, if an alternative splicing event in one species is not conserved in the other species, it can also be functional, representing exons created after the divergence of the human and the mouse lineages. It is interesting to assess the level of conservation of alternative splicing.

The conservation of alternative splicing events in human and mouse can be classified according to comparison result. We would like to discover three types of comparison result. If an alternative splicing event in one species and it also exists in other species, the event is called a conserved alternative splicing event (CAS events). If there is an alternative splicing event in one species and there is only one form in other species, the event is called a non-conserved alternative splicing event (NCAS events). We combined with the human exons and mouse exons to discover the cross-species alternative splicing events (CSA events). The cross-species alternative splicing event means event that in orthologous genes, there is no alternative splicing event in both species and their splicing forms are distinct. On the other words, if we combine exons and splicing events of two species, it looks like there is one alternative splicing event in this combination.

2 Materials

The all EST sequences were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/repository/dbEST/gzipped/dbEST.reports.date.no.gz). As of FEB 14, 2004; the dbEST database has nearly 5.1 million human ESTs and 3.7 million mouse ESTs. Human and mouse genomic sequences retrieved from NCBI contig build 34(genbank format) and build 31, respectively (ftp://ftp.ncbi.nih.gov/genbank/genomes). Gene information and mRNA sequences were fetched from NCBI RefSeq project [7].

We used the information about alternative splicing from Avatar (a value added transcriptome database) [8]. Briefly, the database filters the expressed sequences that may be contaminated, so only ESTs alignments with similarity scores greater than 94% and contained canonical introns were used for further alternative splicing analysis. In Avatar, authors used the alignment result from expressed sequences aligning to the genome generated by MUGUP [9] to detect the alternative splicing sites.

The five classes are including exon skipping, mutually exclusive, alternative 3' splice site, and alternative 5' splice site. Using these data, there were 4,124 and 1,351 alternative exon skipping sites; 536 and 131 mutually exclusive sites; 11,479 and 6,260 alternative 3' splice site; 8,010 and 3,682 alternative 5' splice site and 16,267 and 3,518 retained intron identified in human and mouse orthologous pairs, respectively.

We used the orthologous pairs between human and mouse genes from a reciprocal best match relationship according to annotation in the Homology Map database [10]. There are 9,136 orthologous pairs in the Homology Map database.

3 Methods

Our method is carried out in three main steps: First we choose carefully the orthologous genes pairs for further analysis. Second we identified orthologous exons between human and mouse by the results of the first step. Then we discovered CAS events NCAS events and CSA events by the results of the second step. Due to page limitation, we only describe our method for exon skipping events. Methods of other types of alternative splicing events are similar and omitted.

3.1.Preprocessing phase

In the preprocessing phase, we would show how to choose the orthologous genes pairs for further analysis. Based on the Homology Map catalog of gene pairs that are orthologous between human and mouse, we searched for these 9,136 gene names in Avatar (fetched from NCBI RefSeq project) and 7,078 human genes can be found.

For those genes were not found in Avatar, 324 matched with 18,729 HGNC gene aliases file [12] and were thus taken as additional genes sets. The genes corresponding to these 7,402 (= 7,078 + 324) pairs were examined for occurrence in mouse gene name sets (obtained from NCBI RefSeq project). We searched 7,402 human gene names in Avatar and 6807 human genes can be found in mouse gene lists.

3.2.Mapping phase

To identify orthologous exons between human and mouse, we used pairs of genes reported as reciprocal best matches in the Homology Map. For all target orthologous gene pairs, we first identified their corresponding exons in Avatar.

To match the orthologous exons between the homology gene

pairs, we used Sim4 [10] to align exons of each human gene with exons of the orthologous mouse genes. The acceptable match had to cover at least 80% of each exon tags and the sequence identity in this region had to exceed 80%. Here we identified a orthologous exon map between each orthologous gene. Among these human exon tags, 69,069 exons (3,613 genes) were matched successfully onto mouse exon tags. There are several reasons for the phenomenon of exons missing (corresponding orthologous exons haven't been found). It may cause by only few ESTs covering the corresponding orthologous genes. Or, the huge mutation rate for specific exons may cause the lower similarity which under our criterion.

3.3. Discovering conserved exon skipping events (CES events) and non-conserved exon skipping events (NCES events)

An exon skipping event must contain both the internal exon and the two flanking exons. We denominate the form including the middle exon as Form I. Besides, one sequence that contained the two flanking exons but skipped the middle one, and we denominate the form skipping the middle exon as Form II. We can use the exon information of orthologous exon map and human and mouse exon skipping lists to discover the conserved exon skipping (CES) events and the non-conserved exon skipping (NCES) events. An example is shown in Figure 1. The non-conserved exon skipping event means that there is an exon skipping event in one species and there is only one form in the other species. How we know there is only one form? There must be enough ESTs supporting one form and no any ESTs supporting the other one.

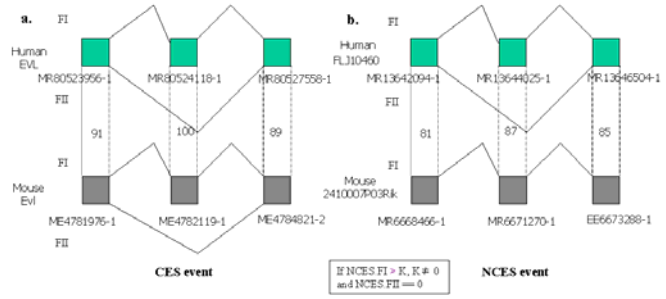

Figure 1. (a) A CES event in the EVL-Evl gene pair. (b) A NCES event in the FLJ10460 -2410007P03Rik gene pair.

We detected 239 CES events. If the quantity of ESTs was not considered, we detected 1,501 NCES events. If only the cases at least ten ESTs support one form and no EST support the other form is considered as non-conservation, we detected 609 NCES events. Consider the cases that we could not classify them as CES or NCES events. They can also be functional. The missing forms are potentially occurred in another species. The predictions can be confirmed by using RT-PCR or microarray in the future.

4 Result

In this study, we have performed genome-wide alternative splicing analysis for both mouse and human conservation and non-conservation.

The data set (Avatar) of transcript-confirmed constitutively and alternatively spliced introns and exons from human and mouse was done through careful spliced alignment of genome and transcript sequences (EST and mRNA). In Table 1, we list the number of CAS events with different AS class. In Table 2, we list the number of NCAS events with different AS class and different numbers of supported ESTs. In Table 3 and Table 4 we list the number of CSA events with different forms and different numbers of support ESTs.

Table 1 .The number of CAS events with different AS class.

Class	No. of supported ESTs	Exon Skipping	3' splice site	5' splice site	Mutually exclusive	Retained intron	Total
CAS	1	239	1,540	815	34	0	2,628

Table 2. The number of NCAS events with different AS class and different numbers of supporting ESTs.

Class	No. of supported ESTs	Exon Skipping	3' splice site	5' splice site	Mutually exclusive	Retained intron	Total
NCAS	1	1,501	2,961	2,433	47	370	7,312
NCAS	10	609	1,048	960	21	145	2,783

Table 3. The number of CSA events (Human Form I + Mouse Form II) with different numbers of supporting ESTs.

Class	Form	No. of supporting ESTs	No.
CSA	Human Form I + Mouse Form II	1	2,357
CSA	Human Form I + Mouse Form II	5	354
CSA	Human Form I + Mouse Form II	10	122

Table 4. The number of CSA events (Human Form II + Mouse Form I) with different numbers of supporting ESTs.

Class	Form	No. of supporting ESTs	No.
CSA	Human Form II + Mouse Form I	1	2,363
CSA	Human Form II + Mouse Form I	5	269
CSA	Human Form II + Mouse Form I	10	113

5 Conclusions

We have performed genome-wide alternative splicing analysis for both mouse and human. The work presented here shown that the sets of conserved alternative splicing events and non-conserved alternative splicing events between human and mouse. We also have discovered cross-species alternative splicing events that one form from human and the other form from mouse. Such events looks extremely diverse expression, maybe these events are associated with exon creation and/or loss during evolution. In addition, we measured the conserved

alternative splicing events according their level of expression.

References

1. R.E Breitbart, A. Andreadis and B. Nadal-Ginard, "Alternative splicing: a ubiquitous mechanism for generation of multiple protein isoforms from single genes." Annu. Rev. Biochem. 1987, 56, pp. 467-495.
2. Caceres, J. F., and Kornblihtt, A. R. “Alternative splicing: multiple control mechanisms and involvement in human disease” Trends Genet 2002, 18, pp. 186-193.
3. A. Javier Lopez, "Alternative splicing of pre-mRNA: Developmental and Mechanisms of Regulation" Annu. Rev. Genet. 1998. 32, pp.279-305
4. Thanaraj TA, Clark F, Muilu J. “Conservation of human alternative splice events in mouse”. Nucleic Acids Res. 2003, 31, pp. 2544–2552.
5. Modrek B, Lee CJ. “Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss”. Nat Genet. 2003, 34, pp. 177–180.
6. Sorek R, Ast G. “Intronic sequences flanking alternatively spliced exons are conserved between human and mouse”. Genome Res. 2003, 13, pp. 1631–1637.
7. Pruitt KD, Maglott DR. "RefSeq and LocusLink: NCBI gene-centered resources". Nucleic Acids Res. 2001, 29, pp. 137-140
8. F. R. Hsu et al. "Genome-wide alternative splicing events detection through analysis of large scale ESTs" Prof. of the IEEE 4th symposium on bioinformatics and bioengineering (BIBE2004), pp.310-316.
9. F. R. Hsu and J. F. Chen., "Aligning ESTs to Genome Using Multi-Layer Unique Markers", Proc. Of the IEEE Computational Systems Bioinformatics Conference (CSB2003), Aug., 11-14, San Francisco, USA, pp. 564-566.
10. Wheeler, D.L. et al. "Database resources of the National Center for Biotechnology Information: 2002 update". Nucleic Acids Res. 2002, 30, pp. 13-16.
11. http://www.gene.ucl.ac.uk/public-files/nomen/ens4.txt
12. Florea, L, Hartzell, G, Zhang, Z, Rubin, G.M, and Miller, W. "A computer program for aligning a cDNA sequence with a genomic DNA sequence". Genome Res. 1998, 8, 967-974.
13. Hsin Chun Lin, Discovery the Relationship Between Single Nucleotide Polymorphisms and Alternative Splicing Events, Taichung Healthcare and Management University, master thesis 2004.