Sequence Conservation and Variation in Alternative Splicing in Relation to Evolutionary Change

Liliana Florea, George Washington University
Xiaoyue Zhao, University of California at Berkeley

24th Summer Symposium in Molecular Biology - Comparative and Functional Genomics (CFG 2005), *July 2005, Penn State University, University Park PA.*

Abstract

Alternative splicing of eukaryotic genes is an important regulatory process with the potential to explain how a large repertoire of proteins can be achieved from a relatively small number of genes. With more than 60% of the human genes believed to undergo alternative splicing to produce multiple mRNA and protein isoforms[1,2], questions abound about the role of alternative splicing in effecting or reflecting evolutionary change and in increasing the complexity of organisms during evolution. The availability of genomic and cDNA sequence data for several eukaryotic genomes and their whole-genome alignments[3] offer a tremendous resource to study sequence conservation and divergence patterns across the genomes to infer relationships between evolution and splicing variation.

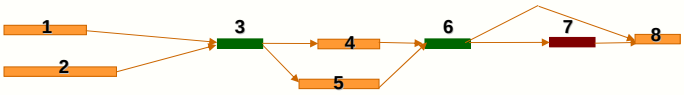
We explore patterns of sequence conservation and point mutation in six genomes (human, chimp, mouse, rat, dog, chicken) in different categories of exons and the introns surrounding them. We show that *skipped exons from minor-form transcripts[4] form a category that has resulted mainly from recent exon creation events*, and the exon and surrounding intron sequences exhibit features consistent with this hypothesis and with positive selection.

Introduction

- What is alternative splicing?
  - Tissue/ developmental stage/ disease-specific selection of different combinations of a gene's exons to produce different mRNA and protein isoforms, with potentially different biological functions.
- What is the extent of alternative splicing?
  - ~60% of the human genes are estimated to undergo alternative splicing[1,2]
  - Some genes can have as many as 40,000 splice variants (DSCAM[5]), but most genes have <100 variations
- Why is it important?
  - Alternative splicing can contribute to organism complexity
  - Defects in alternative splicing are associated with diseases[6] (BRCA1, FGFR2)
- We study alternative splicing in relation to evolutionary change (exon insertions and deletions), and address three issues:
  - Evolutionary analysis of exon creation
  - Sequence conservation in introns
  - Sequence variation in exons

Materials and Methods

- We annotated alternative splicing events in the entire human genome (Assembly version *hg17)* based on genomic alignments of EST and mRNA sequences from NCBI RefSeq, dbEST and MGC repositories, produced with the Sim4/ESTmapper cDNA-to-genome alignment software[7,8].
  - Resources: dbEST: 6.05M ESTs (2.65M unspliced), RefSeq: 23000, MGC: 17952
  - Sim4/ESTmapper cDNA-to-genome alignments
  - Cluster overlapping alignments that share a splice junction, separately by strand, into **splice graphs = genes,** using the gene annotation pipeline AIR[8]



  - Within each cluster, identify *constitutively* (▬) and *alternatively spliced* (▬) exons by comparing pairs of cDNA-genomic alignments
    - Only internal exons are allowed
    - Exons should not overlap with any other exon

- Three main classes of exons are considered:

  i. **Constitutive** (nonAlt)
  ii. Alternatively spliced exons included in the major transcript form (**'major-form' alternative exons**) (AltD): *'on'/'off' ≥ 2.0*
  iii. Alternatively spliced exons included in the minor transcript form (**'minor-form' alternative exons**) (AltI): *'on'/'off' ≤ 0.5*

- For alternatively spliced exons, the major transcript form (exon 'on' or 'off') is determined based on the expression level, estimated from the amount of cDNA evidence (var. 4).

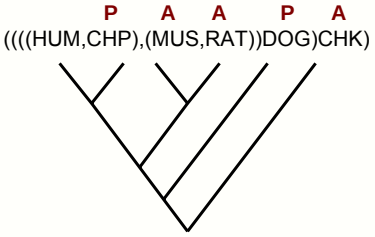
  Minor-form exons were earlier associated with increased frequency of exon creation and/or loss based on human-mouse-rat comparisons[4]. With the addition of the dog and chicken genomes, we are able to further clarify the nature of the evolutionary events.

All internal exons	Other	nonAlt	Alt	AltD	AltI
242,069	92,896	114,744	34,429	22,212	8,932

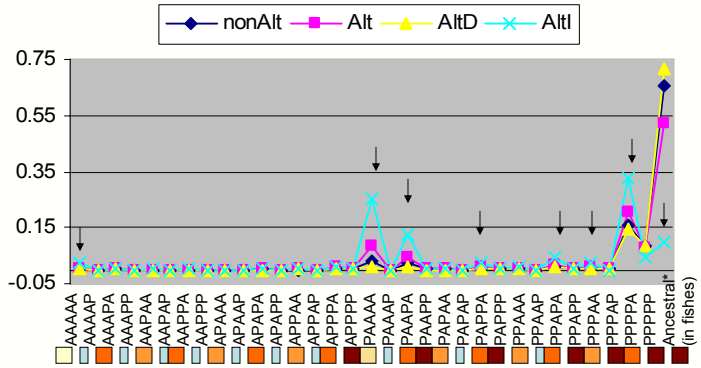
- 'Multiz' whole-genome alignments[3] (http://genome.ucsc.edu)
  – hg17-panTro1-mm5-rn3-canFam1-galGal2-fr1-danRer2 (human-chimp-mouse-rat-dog-chicken-fugu-zebrafish)

I. Evolutionary analysis

- We detected the presence (P)/ absence (A) of human exons in each of the other species (>50% presence in 'multiz' alignments)
  – PAAPA = present in chimp and dog, but absent from rodents and chicken



- Use the P/A status in fishes to resolve some of the ambiguities associated with multiple insertion / deletion events, e.g.:
  – **PAAPA**AA – 'insertion' in primate-rodent-dog lineage,
  – **PAAPA**PA – 'ancestral', and 'deleted' from rodents, chicken and zebrafish



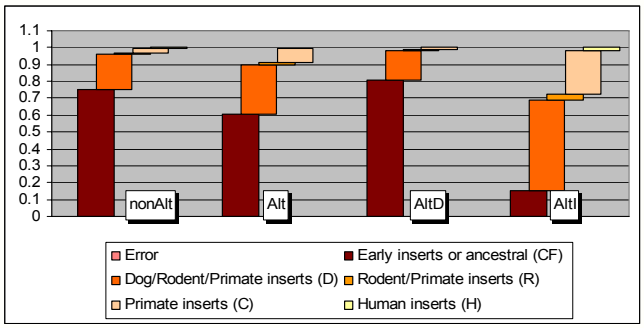
We further group categories into event classes corresponding to exon creation (insertion) stages in the phylogenetic tree:

- (H) Insertion in Human
- (C) Insertion in Primates
- (R) Insertion in Primates+Rodent
- (D) Insertion in Primates+Rodent+Dog
- (CF) Early insertion or ancestral
- (E) Potential error

Number and fraction of exons in each category that are presumed to have been inserted at the various branch points of the phylogenetic tree:

	242069 All		**114744** nonAlt		**34429** Alt		**22212** AltD		**8932** AltI	
H	870	(0.004)	390	(0.003)	240	(0.003)	14	(0.001)	191	(0.021)
C	8605	(0.036)	3317	(0.029)	2826	(0.082)	222	(0.010)	2259	(0.253)
R	2570	(0.011)	1069	(0.009)	540	(0.016)	119	(0.005)	354	(0.040)
D	57430	(0.237)	24004	(0.209)	10053	(0.292)	3926	(0.177)	4779	(0.535)
CF	172594	(0.713)	85964	(0.749)	20770	(0.603)	17931	(0.807)	1349	(0.151)
E	286	(0.001)	146	(0.001)	27	(0.001)	18	(0.001)	7	(0.001)

H Insertion in Human
C Insertion in Primates
R Insertion in Primates+Rodent
D Insertion in Primates+Rodent+Dog
CF Early insertion or ancestral
E Potential error

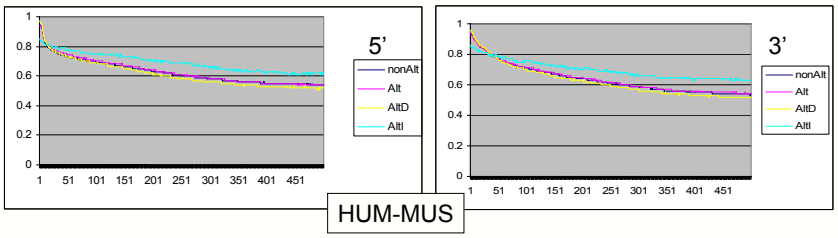


• **O1: AltI** exons are **more frequently associated with exon creation** (insertion) events than the other categories.

• **O2: AltI** exons have resulted mostly by **recent insertions** (~15% occurred before the chicken split, compared to ~80% for AltD and ~75% for nonAlt).

II. Sequence conservation in introns

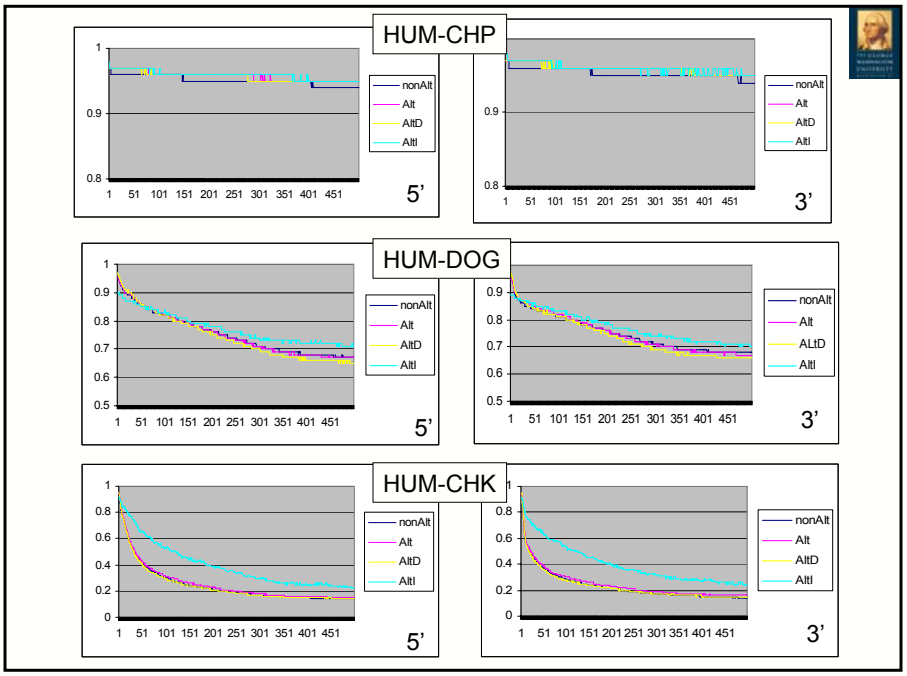
- We plotted the frequency of aligned positions between human and each of (CHP, MUS, RAT, DOG, CHK) in two sets of windows:
  - within the 500 bp upstream of the exons' 5' end
  - within the 500 bp downstream of the exons' 3' end



O1: Sequence surrounding **AltI** exons is **more frequently conserved** than for the other exon categories.

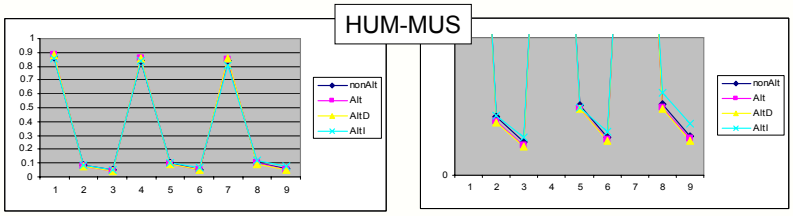
O2: Sequence surrounding **AltD** exons is **less frequently conserved** than for nonAlt and AltI exons.

O3: These tendencies are valid for all HUM-X pairs, and become **stronger as the evolutionary distance increases**.

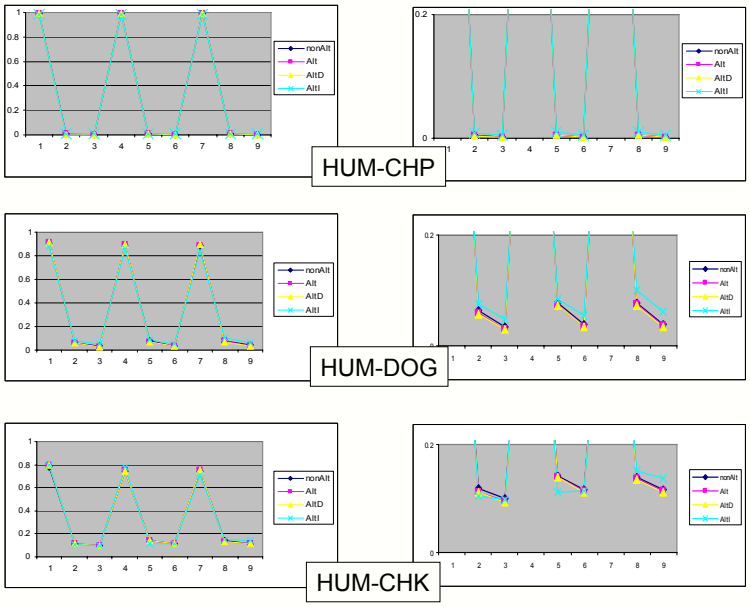


III. Sequence variation in exons

- We determined the frequency of matches (M)/ transitions (I)/ transversions (V) in protein-coding exons for all pairs HUM-X:
  - blastx of exons against the SwissProt database of proteins
  - roughly 60% of the exons had protein matches (≥80% exon coverage, ≥90% sequence identity, no frameshifts or in frame stop codons)



- **O1**: For CHP, MUS, RAT and DOG comparisons, **AltI exons show increased I and V rates** compared to nonAlt and AltD exons, at all 3 codon positions, which may indicate positive selection. The result does not hold true for CHK.
- **O2**: Similarly, **AltD exons show decreased I and V rates** compared to nonAlt and AltI exons, at all 3 codon positions, which may indicate effects of purifying selection.



Conclusions

We determined that 'minor-form' alternative exons form a specific category of exons that is characterized by relatively recent exon creation events. Whereas most constitutive and major-form alternative exons are conserved across the six taxa, exons in the 'minor-form' category show a significantly larger incidence of exon creation events at the various branches of the phylogenetic tree. We also reported on increased sequence conservation in the introns surrounding these exons, as well as higher exonic sequence variation that is consistent with these findings and with the hypothesis of positive selection.

Further studies will be needed to determine the function of the proteins (peptides) encoded in these exons, as well as to deconvolve the effects of evolution and of potential splicing regulatory factors from the observed patterns of sequence conservation and mutation.

References

1. Kan, Z., Rouchka, E.C., Gish, W.R. and States, D.J. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs, *Genome Res.* **11**:889-900.
2. Modrek, B., Resch, A., Grasso, C., and Lee, C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes, *Nucl. Acids Res.* **29**:2850-2859.
3. Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D., and Miller, W. (2004) Aligning multiple genomic sequences with the threaded blockset aligner, *Genome Res.* **14**:708-715.
4. Modrek, B., and Lee, C. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss, *Nature Genetics* **34**:177-180 .
5. Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world, *Trends Genet.***17**:100-107.
6. Faustino, N.A. and Cooper, T.A. (2003) Pre-mRNA splicing and human disease, *Genes Dev.* **17**:419-437.
7. Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M. and Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence, *Genome Res.* **8**:967-974.
8. Florea, L., Di Francesco, V., Miller, J., Turner, R., Yao, A., Harris, M., Walenz, B., Mobarry, C., Merkulov, G.V., Charlab, R., Dew, I., Deng, Z., Istrail, S., Li, P., Sutton, G. (2005) Gene and alternative splicing annotation with AIR, *Genome Res.* **15**:54-66.