

# Evolutionary Conservation and Selection of Alternatively Spliced Human Genes

Steffen Heber<sup>1</sup>, Soma Saha<sup>1</sup>

**Keywords:** alternative splicing, evolutionary conservation, selection

## 1 Introduction.

Alternative splicing significantly increases the diversity of the transcriptome – over half of all known human genes might be alternatively spliced, and some genes create a vast assortment of different transcripts [4, 3]. Often, alternative splicing has important implications for physiology, development and disease [4]. In previous work [2], we developed splicing graphs, a novel data structure, which integrates all transcripts derived from a gene in a single graph representation. Splicing graphs allow us to reliably recover all splice variants represented in the input data, and to reconstruct an exhaustive transcript catalog. The Alternative Splicing Gallery (ASG) [3] collects splicing graphs for human genes based on transcript information from various major sources (Ensembl, RefSeq, STACK, TIGR, and UniGene).

In this study we use ASG to investigate whether human genes with multiple transcripts significantly differ from human genes without alternative splicing with respect to evolutionary conservation, and rates of evolutionary change between human and mouse.

## 2 Material and Methods.

We reconstructed human transcripts as described in [3], and we identified orthologous gene pairs in human and mouse using ENSEMBL’s reciprocal best BLAST hit strategy [5]. To quantify and investigate any relationships between alternative splicing and natural selection, we analysed codon based maximum likelihood estimates generated by the PAML software package [6]. The codon-based evolutionary models implemented in PAML distinguish between the nucleotide substitutions in protein-coding DNA that affect the amino acid sequence (nonsynonymous changes) and those that do not alter the amino acid sequence (synonymous changes). To differentiate between synonymous and nonsynonymous rates of evolution, a parameter  $\omega$  is introduced. This parameter  $\omega$  can be interpreted as the ratio between the rate of a particular nonsynonymous change and the rate the change would have were it actually a synonymous change. When  $\omega = 1$  natural selection does not distinguish between synonymous and nonsynonymous change. Values of  $\omega$  close to 0 indicate stronger intensities of (purifying) natural selection than values that are close to 1.

## 3 Results.

Starting with 22127 human genes stored in ASG we found 17278 (78%) orthologs in mouse. We found a significantly higher percentage (86%) of orthologs among genes with multiple transcripts than among genes with only one transcript (63%). We divided the ASG genes with respect to the number of reported transcripts into 6 categories. The  $\omega$  estimates rely on the F3x4 codon model [1] and human/mouse ortholog pairs. The average  $\omega$  estimate

---

<sup>1</sup>Department of Computer Science, College of Engineering, NC State University, Raleigh, USA.  
E-mail: sheber@ncsu.edu

within each category is shown in Table 1. We have performed various statistical analyses that strongly support (e.g., P-value =  $9.38 \times 10^{-6}$  for a log-linear regression) a negative correlation between the number of transcripts per gene and the value of  $\omega$ , and we found similar results when human ASG genes are compared with rat orthologs (data not shown).

transcripts per gene	mean( $\omega$ ) (human/mouse)	var( $\omega$ )
1	0.289	8.666
2	0.226	3.621
3-5	0.176	0.032
6-10	0.143	0.017
11-50	0.141	0.020
>50	0.131	0.014

Table 1: Tabulation of mean  $\omega$  values derived from orthologous gene pairs in human and mouse. Binned with respect to the reported number of transcripts in ASG.

## References

- [1] Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* 11:725–736.
- [2] Heber, S. and Alekseyev, M. and Sze, S. H. and Tang, H. and Pevzner, P. A. 2002. Splicing graphs and EST assembly problem. *Bioinformatics* 18 Suppl. 1:181–188.
- [3] Leipzig, J. and Pevzner, P. and Heber, S. 2004. The Alternative Splicing Gallery (ASG): Bridging the Gap Between Genome and Transcriptome. *Nucleic Acids Research* 32:3977–3983.
- [4] Modrek, B. and Lee, C. 2001. A genomic view of alternative splicing. *Nature Genetics* 30:13–19.
- [5] Remm, M. and Storm, C. E. and Sonnhammer, E. L. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology* 314:1041–1052.
- [6] Yang, Z. 1977. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* 13:555–556.