

# A Classification Approach to Comparative Gene Finding in Mammals

Michael Lin,<sup>1</sup> Manolis Kellis<sup>2</sup>

**Keywords:** gene finding, comparative genomics, classification

## 1 Introduction

Evolutionary conservation is a powerful signal that can be used to identify protein-coding genes within related genomes. Promising early approaches ([1], [3], [4], [5]) considered conservation between two species, typically human and mouse, to augment existing *ab initio* gene finding approaches. The growing availability of genomes from many different species makes it possible to additionally consider conservation across multiple related species. Siepel and Haussler [6] have designed a complex probabilistic model to measure conservation and identify exons across many different species. Kellis et. al. [2] used whole-genome multiple alignments of four yeast species to complete a sweeping re-annotation of the yeast genome by classifying candidate open reading frames as coding or non-coding, based on evolutionary conservation of the codon reading frame.

We extend this latter technique to mammalian genomes, a considerably more complex task due to the exon-intron structure of most mammalian genes. We identify candidate exons within multiple alignments of human, dog, mouse, and rat genomes by searching for start codons, stop codons, and splice sites. We then classify each candidate exon as coding or non-coding by evolutionary measurements such as reading frame conservation and codon substitution rates, as well as standard protein-coding measures from the literature. Preliminary results suggest that well-chosen conservation measurements have striking power to discriminate coding and non-coding regions in mammals, that will further improve as more mammalian genomes become available.

## 2 Codon Substitution Matrix

A 64-by-64 Codon Substitution Matrix (CSM), specifying the frequency with which each codon of one species aligns to each codon in another species within protein-coding exons, unifies information about the species-specific codon usage bias, selection for silent nucleotide substitutions, and selection for amino acid substitutions with similar chemical properties. We trained CSMs on millions of codons from known protein-coding exons in mammalian species, and, for comparison, comparable numbers of putative codons from noncoding regions. Candidate exons can then be scored by a log-odds ratio measuring their likelihood to have originated from either a coding or non-coding region, based on their codon usage and substitution. Figure 1 shows the CSM scores for human CFTR exons, as well as nearby spurious candidate exons, aligned to their dog and mouse orthologs.

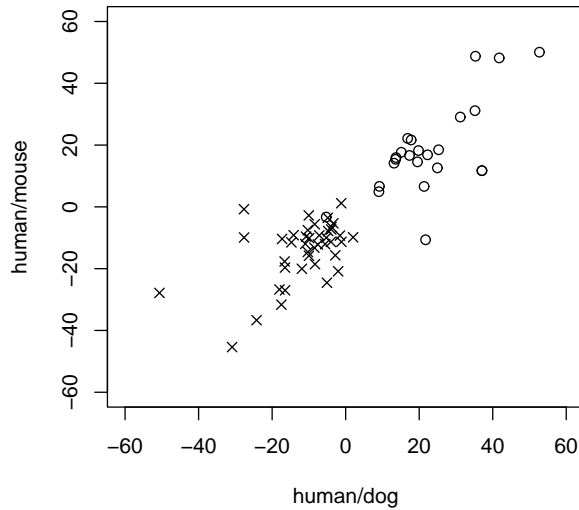
## 3 Reading Frame Conservation

Frameshifting insertions and deletions are negatively selected against in protein-coding exons due to their likely deleterious effect on the biological function of the resulting protein. The Reading Frame

---

<sup>1</sup>Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology. E-mail: mikelin@mit.edu

<sup>2</sup>Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology. E-mail: manoli@mit.edu



**Figure 1:** Coding scores of candidate exons near human CFTR based on Codon Substitution Matrices for human-dog and human-mouse alignments. Circles represent 26 CFTR exons; crosses are 50 randomly selected spurious candidate exons from the same chromosomal region.

Conservation (RFC) test developed by Kellis et. al. [2] single-handedly detects known yeast open reading frames with sensitivity and specificity greater than 99%. The specificity of this test is, however, limited on short sequences of a few hundred nucleotides or less, such as many mammalian exons, simply due to lack of evolutionary distance between the species aligned. Nonetheless, the presence of a frameshifting indel in a candidate exon is strong evidence that it is not, in fact, protein-coding, and preliminary analysis in the CFTR region shows that this eliminates over 75% of spurious candidate exons in a pairwise alignment of human and dog alone.

## References

- [1] Badger, J. H. and Olsen, G. J. 1999. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* 16(4):512–524
- [2] Kellis, M. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241–254
- [3] Korf, I., Flicek, P., Duan, D., and Brent, M. R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* 17:S140–148
- [4] Meyer, I. M. and Durbin, R. 2002. Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics* 18(10):1309–1318
- [5] Parra, G., Agarwal, P., Abril, J. F., Wiehe, T., Fickett, J.W., and Guigo, R. 2003. Comparative gene prediction in human and mouse. *Genome Res* 13(1):108–117
- [6] Siepel, A. and Haussler, D. 2004. Computational identification of evolutionarily conserved exons. *Proc. 8th Int'l Conf. on Research in Computational Molecular Biology (RECOMB '04)*, 177–186.