

# A Method for Detection of Functional Overlapping Genes

Niv Sabath<sup>1</sup> and Dan Graur<sup>2</sup>

## Introduction

Overlapping genes are widespread in all taxa, but are particularly common in viruses [1]. As far as protein-coding genes are concerned, there is a non-zero probability that at least one of the five possible overlapping sequences of any gene will contain an open-reading frame (ORF) of a length that may be suitable for coding a functional protein. It is, however, very difficult to determine whether or not such an ORF is functional. In non-overlapping genes, the signature of purifying selection is often used as a telltale sign of functionality [2]. Here, we propose an analogous method that assigns functionality to an overlapping reading frame if it can be shown that the sequence is subject to selection. Using a simulation, we test and compare our method with that of Firth and Brown [3].

## Method

Recently, we developed a method for the simultaneous estimation of selection intensities in overlapping genes [4]. This method fits a Markov model of codon substitution to data of two aligned homologous sequences. The model, which accounts for the sequence interdependence imposed by the gene overlap, extends the single-gene model by adding an additional parameter to the estimation of the nonsynonymous/synonymous rate ratio ( $\omega$ ) of the overlapping gene [4]. In order to predict functionality of an ORF that overlaps a known gene, we modified an existing approach that is used to predict functionality of non-overlapping genes [2]. We used two hierarchical models: In model 1, there is no selection operating on the ORF (i.e.,  $\omega$  of the ORF is set to one). In model 2, the ORF is assumed to be under selection (i.e.,  $\omega$  is a free parameter). We calculate the maximum likelihood values under the two models, L1 and L2, and then calculate the likelihood ratio as  $LR = 2(\ln L1 - \ln L2)$ . Finally, LR is compared against a chi-square distribution with one degree of freedom to test whether model 2 fits the data significantly better than model 1 (at p-value  $\leq 0.01$ ), in which case, the ORF is said to be subjected to selection and is most probably functional.

## Results

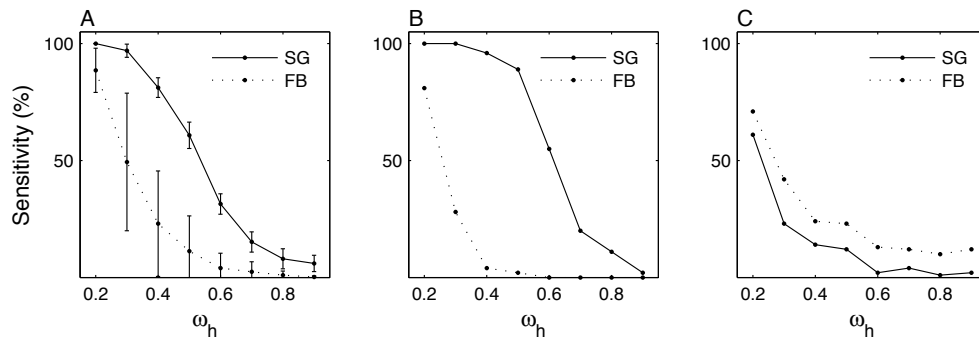
To test the performance of our new method (SG) and compare it to Firth and Brown's (FB)[3] we simulated the evolution of overlapping genes (as described in [4]). In each run of the simulation, one gene was designated as known and the second as hypothetical. We examined the effects of the following factors on the ability of the two methods to detect selection in the hypothetical gene: (1) nonsynonymous/synonymous rate ratio in the hypothetical gene ( $\omega_h$ ), (2) overlap types (same-strand phase-1 and phase-2 overlaps and opposite-strand phase-0, phase-1, and phase-2 overlaps [4]), (3) sequence divergence ( $t$ ), and (4) sequence length. We initially set the sequence length to 300 codons and  $t = 0.4$ , which corresponds to a sequence divergence of  $\sim 12\%$ . We set the nonsynonymous/synonymous rate ratio of the known gene to 0.2 and varied  $\omega_h$  between 0.2 (strong purifying selection) and 0.9 (weak purifying selection). For each set of parameters, we generated 100 random pairs of overlapping genes. We defined sensitivity as the percent of hypothetical genes under selection that were identified correctly by the method.

In Figure 1A, we present the sensitivity of the methods against  $\omega_h$ , averaged over the five overlap types. As expected, the sensitivity of both methods decreased with increase in  $\omega_h$ . In all overlap types, SG exhibits a higher sensitivity than FB, up to 50% for  $\omega_h = 0.4$ .

To test the performance of the methods at high sequence divergence levels, we set  $t = 1$  (corresponding to a sequence divergence of  $\sim 24\%$ ) and used opposite-strand phase-0 overlaps (Figure 1B). For both methods, there is no significant difference in the sensitivity between low and high sequence divergence (chi-square test of independence,  $p = 0.16$  and  $p = 0.43$ , for SG and FB, respectively), suggesting that both methods are not affected by sequence divergence.

In the next step, we tested the performance of the two methods on short sequences. We set sequence length = 50 codons and  $t = 0.4$  (Figure 1C). For SG, the sensitivity is significantly lower for short sequences (chi-square test of independence,  $p < 0.001$ ). In contrast, the sensitivity of FB is significantly higher for short sequences (chi-square test of independence,  $p < 0.005$ ). The sensitivity of FB is not significantly higher than that of SG (chi-square test of independence,  $p = 0.35$ ).

Finally, we tested the specificity of the methods, i.e., the percent of hypothetical genes under no selection that were correctly identified as such by the method. We set  $\omega_h = 1$  (no selection on the hypothetical gene),  $t = 0.4$ , and generated 1000 random pairs of overlapping genes for each of short and long sequences (50 and 300 codons, respectively). Table 1 shows that the specificity of SG is fairly robust as far as short sequence lengths are concerned. In contrast, the specificity of FB decreases with sequence length.



**Figure 1:** Sensitivity versus selection intensity on the hypothetical gene ( $\omega_h$ ) for the SG (solid line) and FB (dotted line) methods. A. Average over the five types of overlap. B. High sequence divergence (opposite-strand phase-0 overlaps). C. Short sequence length (opposite-strand phase-0 overlaps).

	Sequence Length (codons)	
	50	300
SG	98.0%	99.2%
FB	93.5%	100%

**Table 1:** Specificity of the SG and FB methods for short and long sequences

## Conclusions

We developed a new method that can identify functionality in overlapping reading frames. Through simulation, we tested this method under several conditions and compared it with that of Firth and Brown [3]. Under most conditions, our method predicts functionality with higher sensitivity while maintaining near-perfect specificity. Our method has the potential to identify functional overlapping genes that have been overlooked by current annotation methods.

## References

- [1] Belshaw, R., O.G. Pybus, and A. Rambaut, *The evolution of genome compression and genomic novelty in RNA viruses*. *Genome Res*, 2007. **17**(10): p. 1496-504.
- [2] Firth, A.E. and C.M. Brown, *Detecting overlapping coding sequences in virus genomes*. *BMC Bioinformatics*, 2006. **7**: p. 75.
- [3] Nekrutenko, A., K.D. Makova, and W.H. Li, *The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study*. *Genome Res*, 2002. **12**(1): p. 198-202.
- [4] Sabath, N., G. Landan, and D. Graur, *A method for the simultaneous estimation of selection intensities in overlapping genes*. *PLoS ONE*, 2008. **3**(12): p. e3996.