

Study of pseudogenes and splice sites in the annotated *C. elegans* genome

Cyril Chua, Teng Kah Wee, Ng Foong Li, Neo Zhi Yuan, Lim Yong Liang, Yee Pei Shan, Koh Siok Im

Keywords: Data mining, *C. elegans*, pseudogenes, splice sites

Abstract

To enable biological analysis, good gene annotation is required. Gene annotation efforts are hampered by the existence of pseudogenes and inaccurate splice site prediction methods.

We have identified pseudogenes which are wrongly annotated as genes in *C. elegans*. They were identified based on slight dissimilarities with homologs at intron-exon junctions. Using computational methods, we were able to enrich the genomic data 7 times for pseudogenes. Results show that pseudogenes are found more easily in large gene families. In addition, pseudogenes could only be identified when they have highly similar BLAST hits.

Some identified pseudogenes have wrongly chosen splice sites. We created a simple splice site detection program which found a TNNAGIRY consensus greater than 90% at the exon/intron junction and will use this for future studies. We hope to increase our pseudogene prediction using improved splice site prediction as well as other methods.

Software and Files

Data was obtained from ACeDB (www.acedb.org) and mined using Microsoft Excel. Please contact Cyril Chua at cyril@sp.edu.sg for software and files.

Figures and Tables

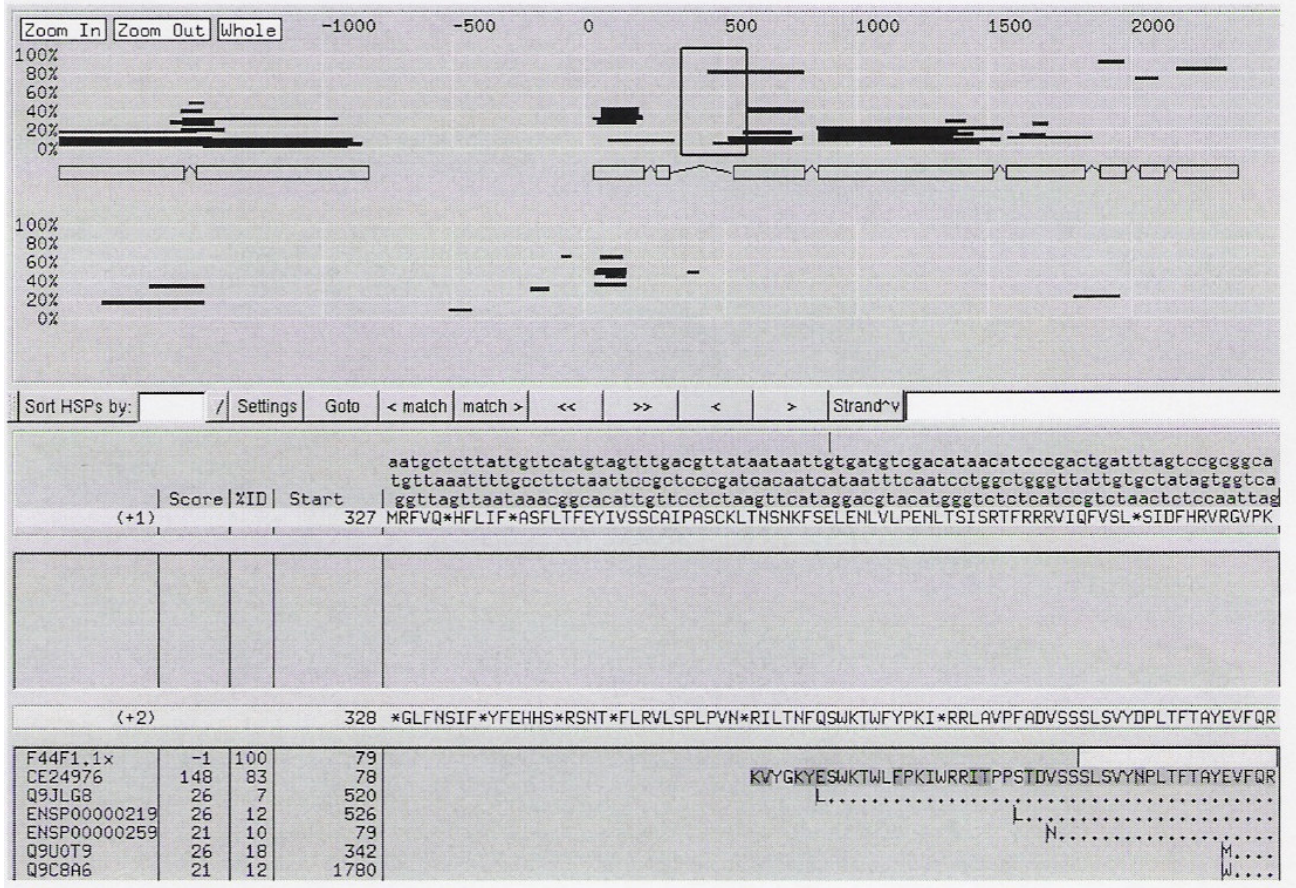


Figure 2. ACeDB Blixem view of F44F1.1, an identified pseudogene showing intron stop codon homology with CE24976 (F44F1.3). Around the stop codon in frame 2 at nucleotide position 480, both have a consensus sequence of PKI(*W)RR where the fault lies.

References

1. Harrison, P.M., Echols, N., & Gerstein, M.B. 2001. *Nucleic Acids Research* **29**:818-830.
2. Mounsey, A., Bauer, P., & Hope, I.A. 2002. *Genome Research* **12**:770-775.