PSF

Finding pseudogenes in a genomic sequence.
Searching for pseudogenes is performed by aligning set of proteins with the genomic sequence. Protein FASTA-file could contain sequences with unformatted names or (preferably) with specially formatted ones. Proteins with formatted names are produced with a PSF_Pre program (not installed in the current version). This special prot. name format describes nucleotide sequence which translation gives appropriate protein, and number of its exons.

All the alignments containing one of the following are considered pseudogene candidates:
(1) stop-codons/frameshifts in nuc. sequence [for alignment with ANY protein]
(2) PolyA site and/or PolyA signal, if exon is single [for alignment with ANY protein]
(3) Number of exons is much lower than in ancestor gene [for alignment with protein SPECIALLY FORMATTED]
(4) Ka/Ks ratio exceeds 0.5 [for alignment with protein SPECIALLY FORMATTED]

It is recommended to input NR or IPI base as a protein base (better unredundant). In this case only p.(1) and p.(2) will work, but resulting candidates will be more reliable. Note that incorrectly predicted proteins might give a number of false pseudogenes.

Output example:

```
chr @@ chain @@ pos(dir.ch.) @@ len(nt.) @@ identity,@@ coverage,@@ Ka/Ks @@ uali.head @@ uali.tail @@ exons#,lower @@ exons#,upper @@ polyA @@ polyA_signal @@ corr.stops# @@ uncorr.stops# @@ corr.frameshifts# @@ uncorr.frameshifts# @@ prototype_chr @@ prototype_prot_name @@ prototype_exon#,lower @@ prototype_exon#,upper @@ DNA_identity @@ CDS length
ENm009 @@ - @@ 322971 @@ 859 @@ 57.79 @@ 81.61 @@ 0.283 @@ 0 @@ 13 @@ 1 @@ 1 @@ 0 @@ 0 @@ 0 @@ 0 @@ 0 @@ 1 @@ chr11 @@ C11000184 chr11 1 exon (s) 424011 - 423106 ORF: 1 - 900  299 aa, chain - ## BY PROTMAP: gi|21928977|dbj|BAC06074.1| seven transmembrane helix receptor [Homo ## 29 @@ 1 @@ 1 @@ 60.656 @@ 732 @@
ENm009 @@ + @@ 966139 @@ 872 @@ 49.59 @@ 75.63 @@ 0.487 @@ 10 @@ 19 @@ 1 @@ 2 @@ 0 @@ 0 @@ 0 @@ 0 @@ 0 @@ 1 @@ chr11 @@ C11000197 chr11 1 exon (s) 433690 - 432722 ORF: 242 - 1204 orf 4667288 4668250  320 aa, chain - ## gi|13540539|ref|NP_110401.1| (NM_030774) olfactory receptor, family 51, subfamily E, member 2; prostate specific G-protein coupled receptor [Homo sapiens] ## 320 ## orf_perfect ## NM_030774_#_242_#_1204 @@ 1 @@ 1 @@ 60.882 @@ 726 @@
ENm009 @@ + @@ 33573 @@ 928 @@ 62.29 @@ 95.19 @@ 0.284 @@ 3 @@ 1 @@ 1 @@ 1 @@ 0 @@ 0 @@ 0 @@ 0 @@ 0 @@ 1 @@ chr11 @@ C11000202 chr11 1 exon (s) 437411 - 436467 ORF: 1 - 939  312 aa, chain - ## BY PROTMAP: gi|22061831|ref|XP_171424.1| similar to olfactory receptor [Pan trog ## 31 @@ 1 @@ 1 @@ 66.105 @@ 891 @@
....
```

Where:

Fields are separated with '@@' sequence.
First line represent field names.
List of field names: