The genomes of *Toxoplasma gondii*, *Aspergillus fumigatus*, *Arabidopsis thaliana*, *Plasmodium falciparum* (malaria), *Mus musculus* (mouse), and *Homo sapiens* (human) were all subjected to a feature-extraction procedure similar to that used by GlimmerM (section 10.6.1), in which four features were extracted for each CDS in a randomly selected subset of the annotated genes for the organism:

1. The WAM score for the signal at the 5′ end of the ORF
2. The WAM score for the signal at the 3′ end of the ORF
3. The histogram-based length probability of the ORF
4. The *hexamer score* S_{hex} of the ORF, which was computed via Equation (10.36):

$$S_{hex} = \sum_{\substack{hexamers \\ H}} \log \frac{P(H \mid coding)}{P(H)} \tag{10.36}$$

A roughly equal number of noncoding ORF's were also selected at random and used to produce negative training and test cases.

Files:

*.data : training data
*.test : hold-out set for testing model accuracy
*.names : list of attributes.