ORFs sequences

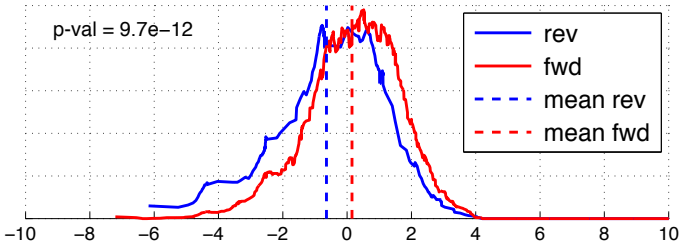
ATGNNNNNNNTGA
ATGNNNNNNNTAG
ATGNNNNNNNTAA
…

Select features to predict expression

Unknown

Already described

Orientation respect to gDNA



Expression

Find sequence features able to distinguish
forward oriented inserts than reverse
(machine learning + sequential feature selection)

>4000 features to dozens

GLM classifier

C base count
GA dimer count
A or G in codon pos 1 (AG1)
T or C in codon pos 3 (TC3)
…

AUC = 0.79

Naive Bayes classifier

A count
CG dimer count
GAT codon count
A in codon pos 2 (A2)
…

AUC = 0.80

Tree Bagger classifier

A count
GA dimer count
A in codon pos 2 (A2)
A or C in codon pos 2 (AC2)
…

AUC = 0.80

Select and reduce the number of features
that are able to predict expression

GLM
(expression ~ classifiers selected features)

Pearson r^2 = 0.11

+

Lasso

Codon bias (tAI)
3'UTR length
GC content
Transcript length

Final predictors

AG1
AC2
A2

Interpreted by
the ribosome

Unknown mechanism
(not used)

GAAAGA
ACGTTA