pORFect predictor

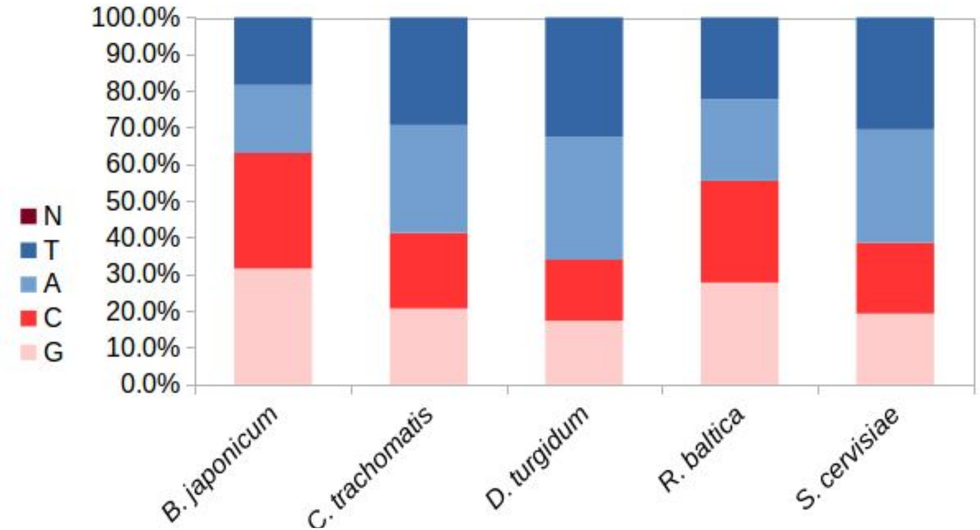
A Magnum Opus by
Maria Hesselman and Maryia Ropat
Group 4

Genomes

Bradyrhizobium japonicum
Chlamydia trachomatis
Dictyoglomus turgidum
Rhodopirellula baltica
Saccharomyces cerevisiae

Statistics

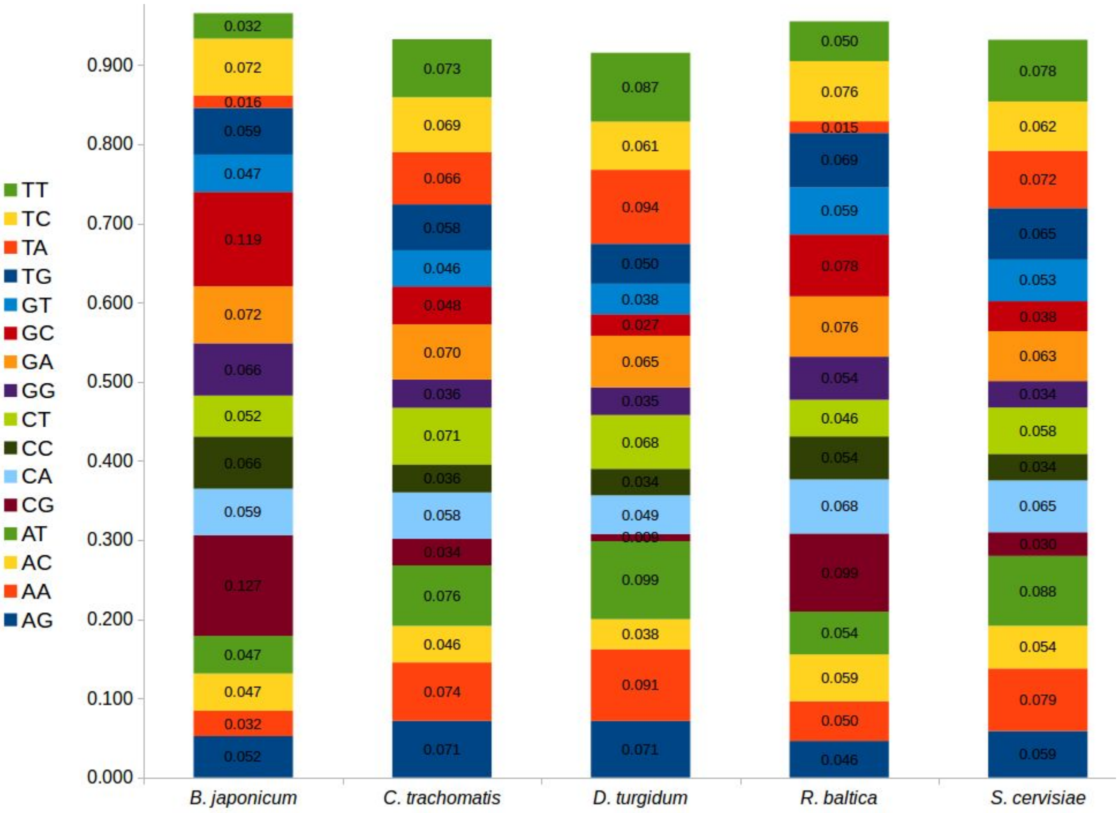
Nucleotide frequencies



GC content =((G+C)(A+T+C+G))100

Dinucleotide frequencies

Di-nucleotide frequency = XX/(total X-1)



Prediction

Frame retrieval

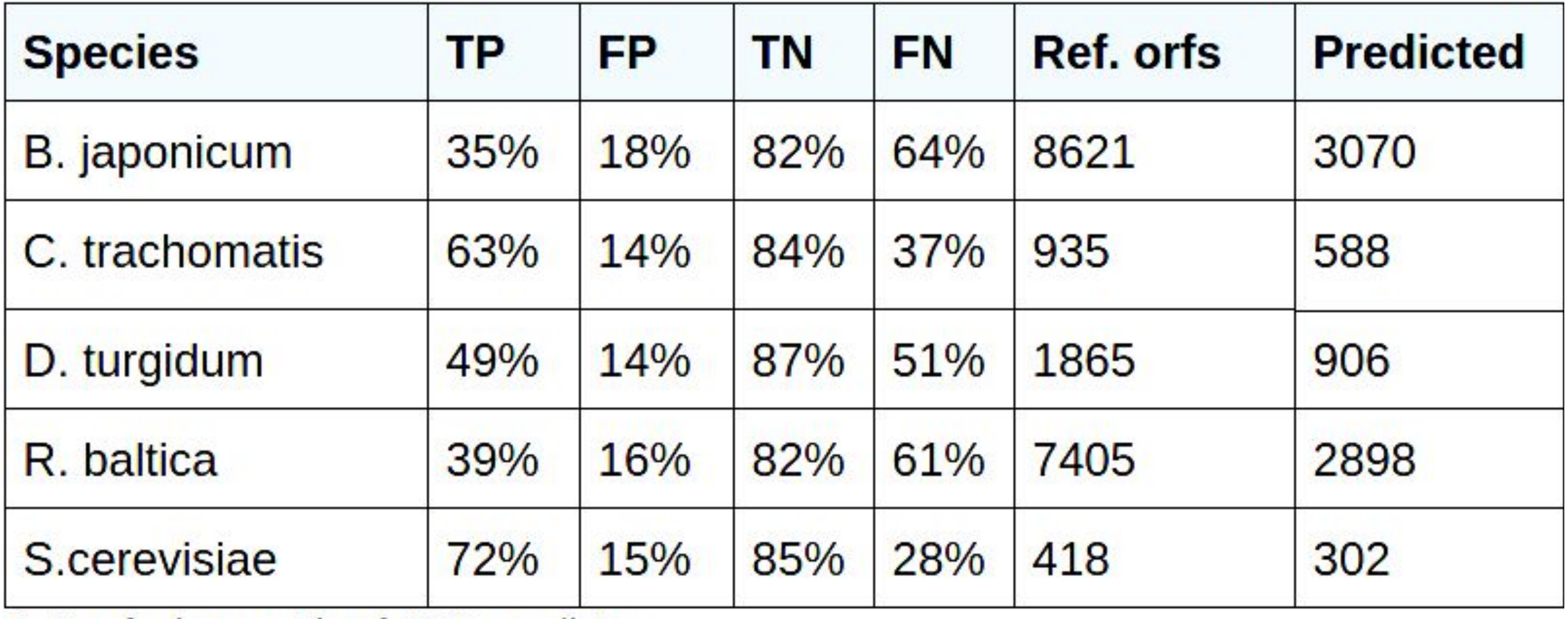
- ATG - start codon
- TAA, TAG, TGA - stop codon

Refinement

- 225+ bases
- Shannon entropy

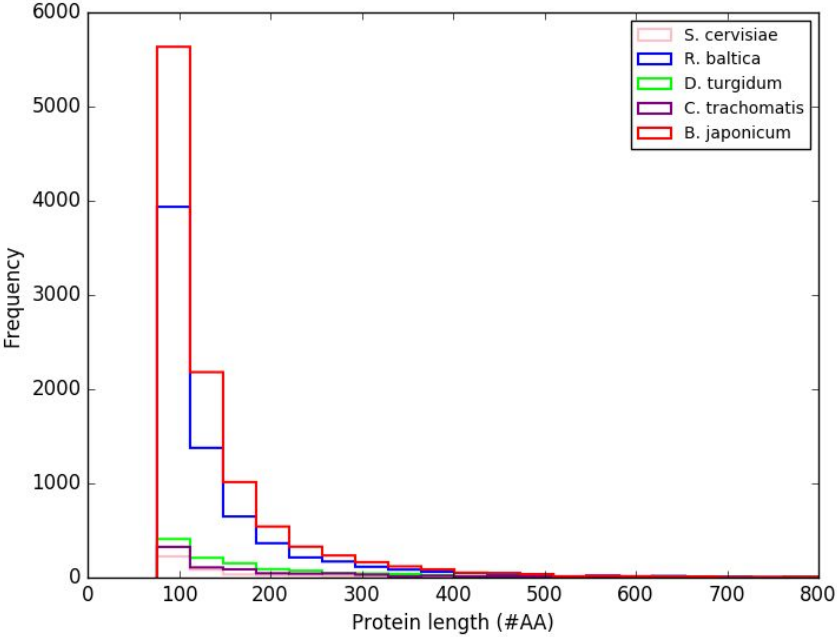
$$H = \sum P(\text{Dinucleotide in sequence}) * \text{log2}\, P(\text{Dinucleotide in genome})$$

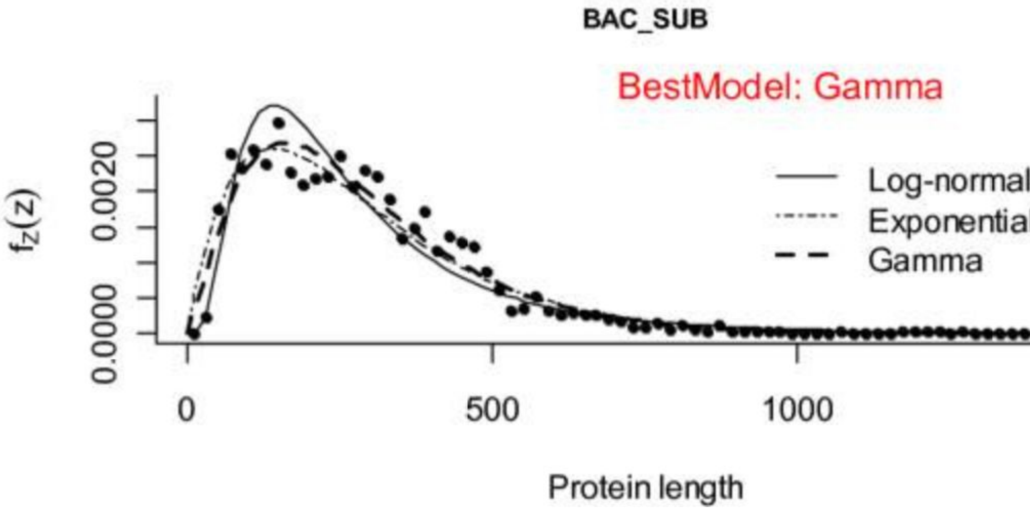
Performance



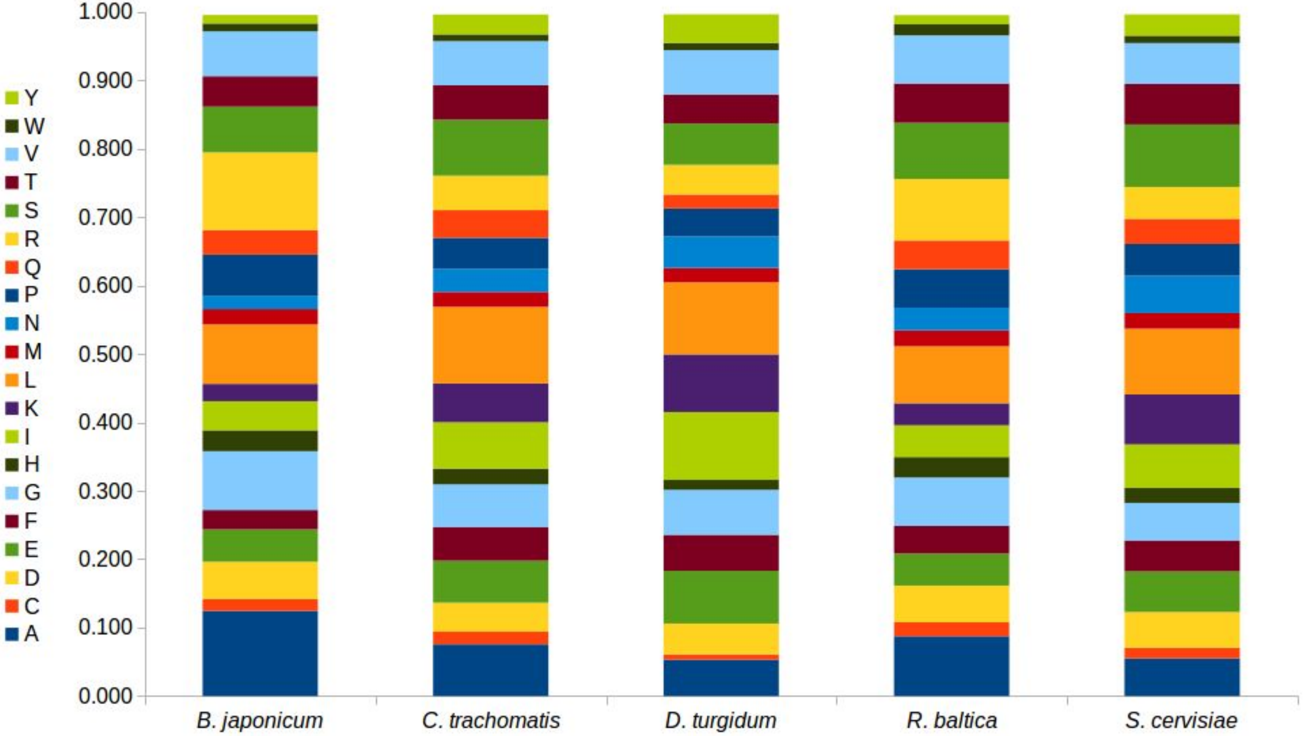
Species	TP	FP	TN	FN	Ref. orfs	Predicted
B. japonicum	35%	18%	82%	64%	8621	3070
C. trachomatis	63%	14%	84%	37%	935	588
D. turgidum	49%	14%	87%	51%	1865	906
R. baltica	39%	16%	82%	61%	7405	2898
S.cerevisiae	72%	15%	85%	28%	418	302

Predicted protein length distribution



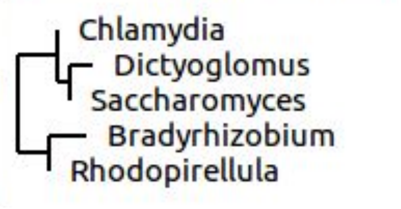


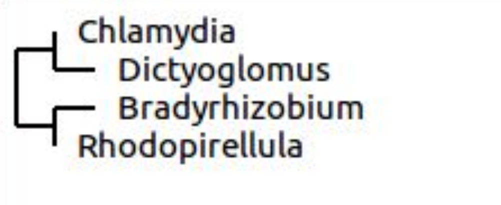
Predicted ORF Amino Acid frequency



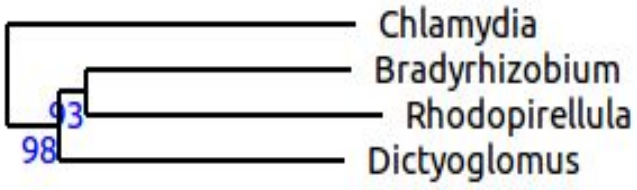
Distance matrix and tree

$$D = \sqrt{(GC\ genome\ 1 - GC\ genome\ 2)^2}$$

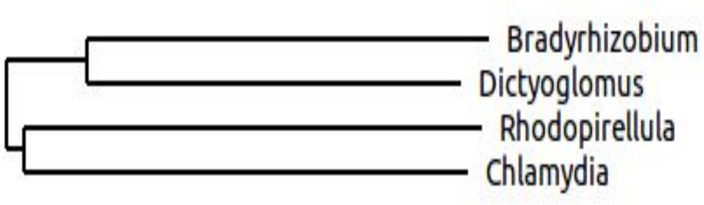




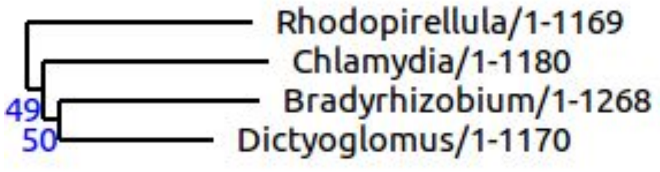
Alignment of 10 orthologs



Synteny



rRNA alignment



Discussion?

Writing ORF prediction software in 2018, colorized

Source: https://netrunnerdb.com

Thanks!!