Comparative genomics FuMiSh ORF Predictor

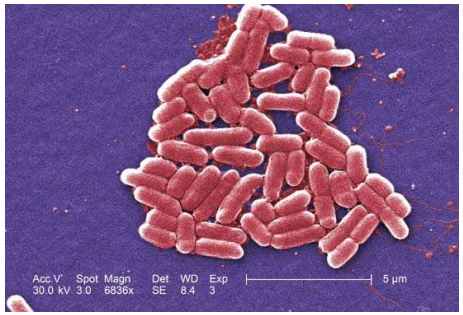
Fuqi Xu, Milda Valiukonytė, Shuhan Xu

Contents

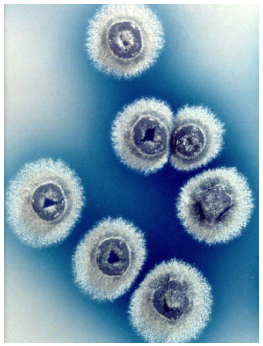
- Genomes
- Genetic composition analysis
- ORF prediction
- Evolutionary relationship

Genomes

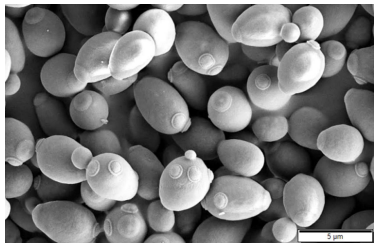
Species	Family	Chromosome	Genome length
Escherichia coli	*Enterobacteriaceae*	All	5443340
Streptomyces coelicolor	*Streptomycetaceae*	All	9054847
Saccharomyces cerevisiae	*Saccharomycetaceae*	IV	1531933
Rubrobacter xylanophilus	*Rubrobacteraceae*	All	3225748
Spiribacter curvatus	*Ectothiorhodospiraceae*	All	1926631



Escherichia coli

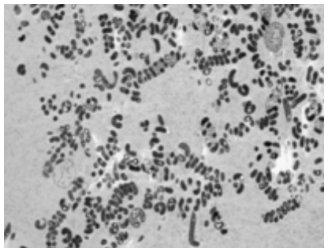


Streptomyces coelicolor



Saccharomyces cerevisiae

Rubrobacter xylanophilus



Spiribacter curvatus

Genetic composition analysis

GC content and dinucleotide frequency calculation

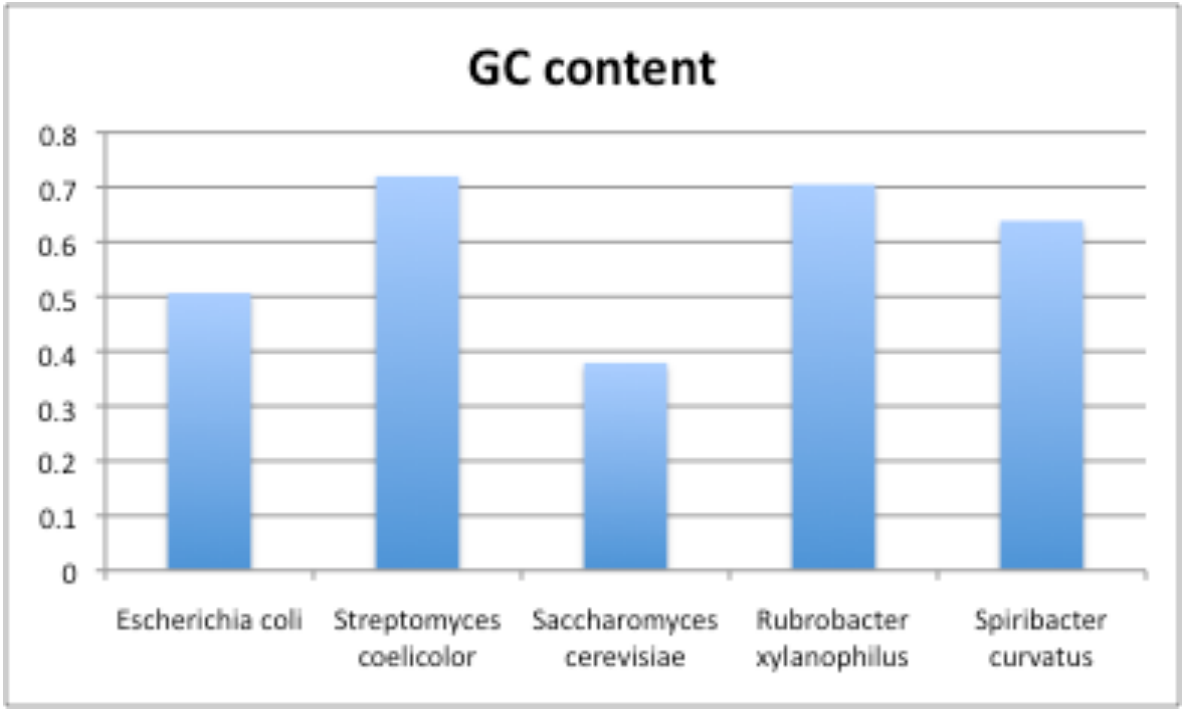
$$GC = \frac{count(G) + count(C)}{length(genome)}$$

$$frequency(dinucleotide) = \frac{count(dinucleotide)}{length(genome) - 1}$$

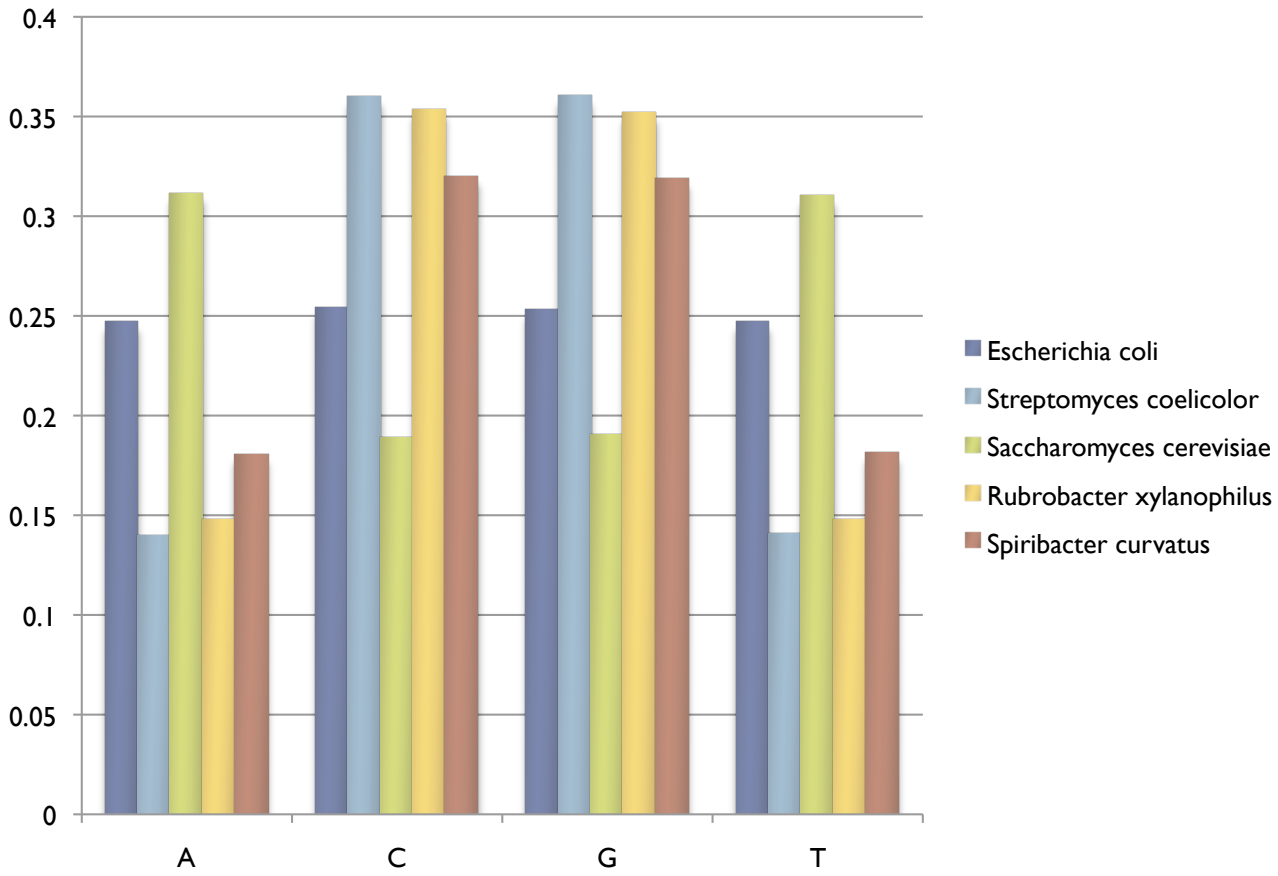
GC content and dinucleotide frequency calculation

```
import sys

filename = sys.argv[1]

with open (filename, "r") as genome:
    genome=genome.readlines()
    genome=genome[1]

# Nucleotide frequencies #

G=genome.count("G")
C=genome.count("C")
A=genome.count("A")
T=genome.count("T")

length=len(genome)

A_freq=A/length
C_freq=C/length
G_freq=G/length
T_freq=T/length

# GC content #

f=open("GC_freq_%s" % filename,"w")
G_and_C=G+C
GC_content=G_and_C/length
print (GC_content, file=f)

# Dinucleotide frequencies #

dList = [genome[x:x+2] for x in range(length)]

dNumber = length-1

dinucleotides= ["AA", "AC", "AG", "AT", "CA", "CC", "CG", "CT", "GA", "GC", "GG", "GT", "TA", "TC", "TG", "TT"]

for x in dinucleotides:
    freq = dList.count(x)/dNumber
    print (freq, file=f)

f.close()
```

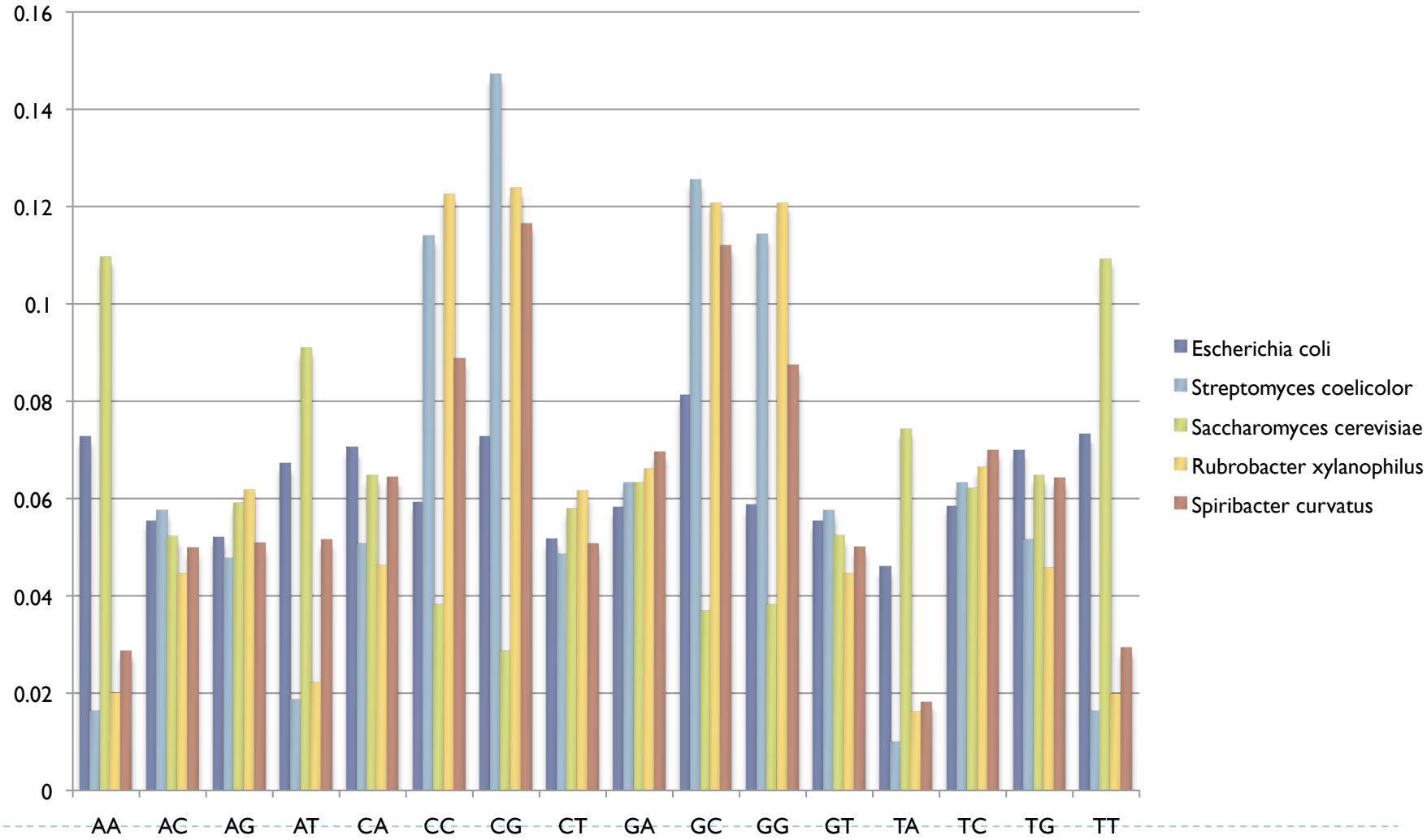
GC content


GC content
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Escherichia coli
Streptomyces coelicolor
Saccharomyces cerevisiae
Rubrobacter xylanophilus
Spiribacter curvatus

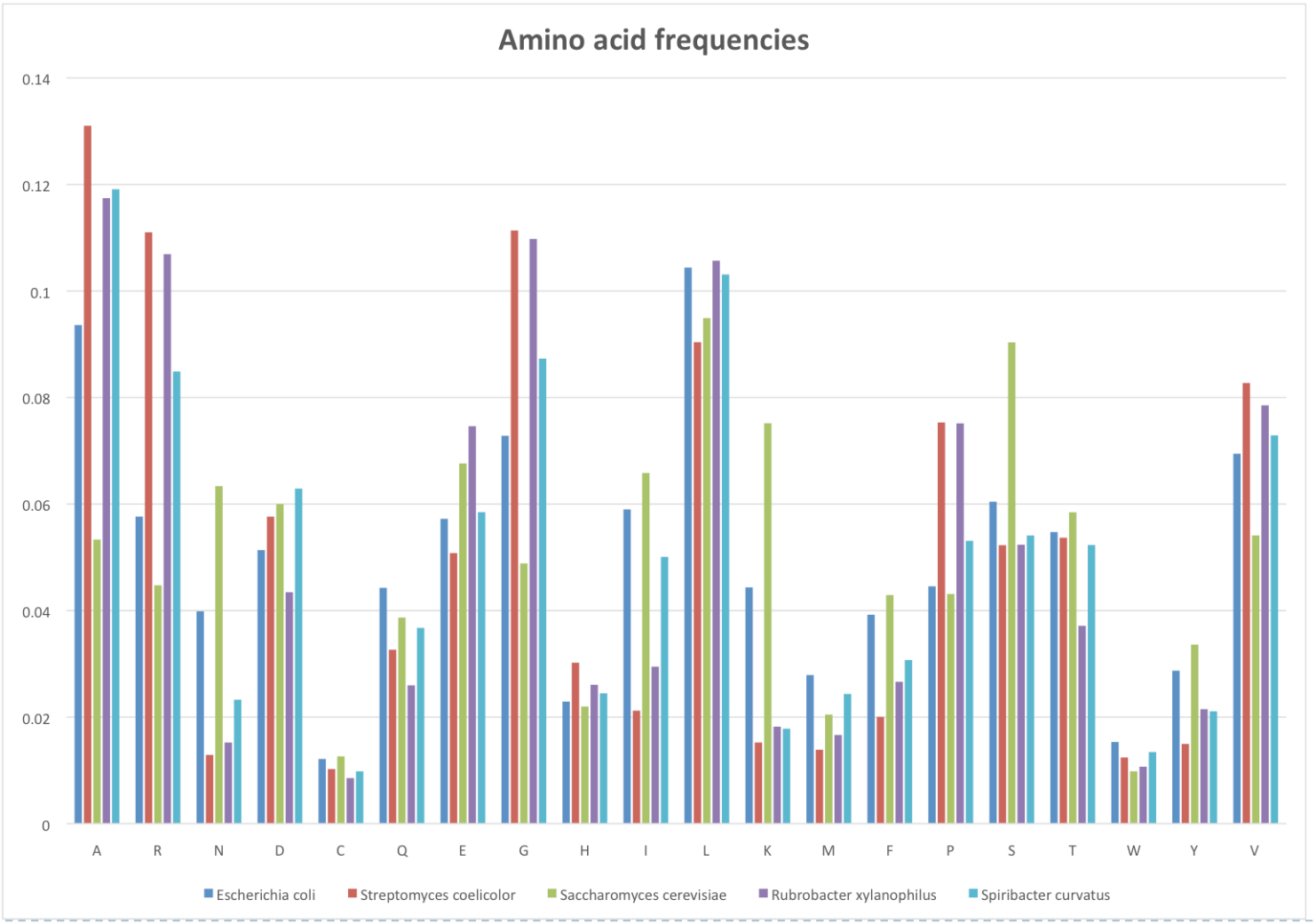
Nucleotide frequencies



Dinucleotide frequencies



Amino acid frequencies

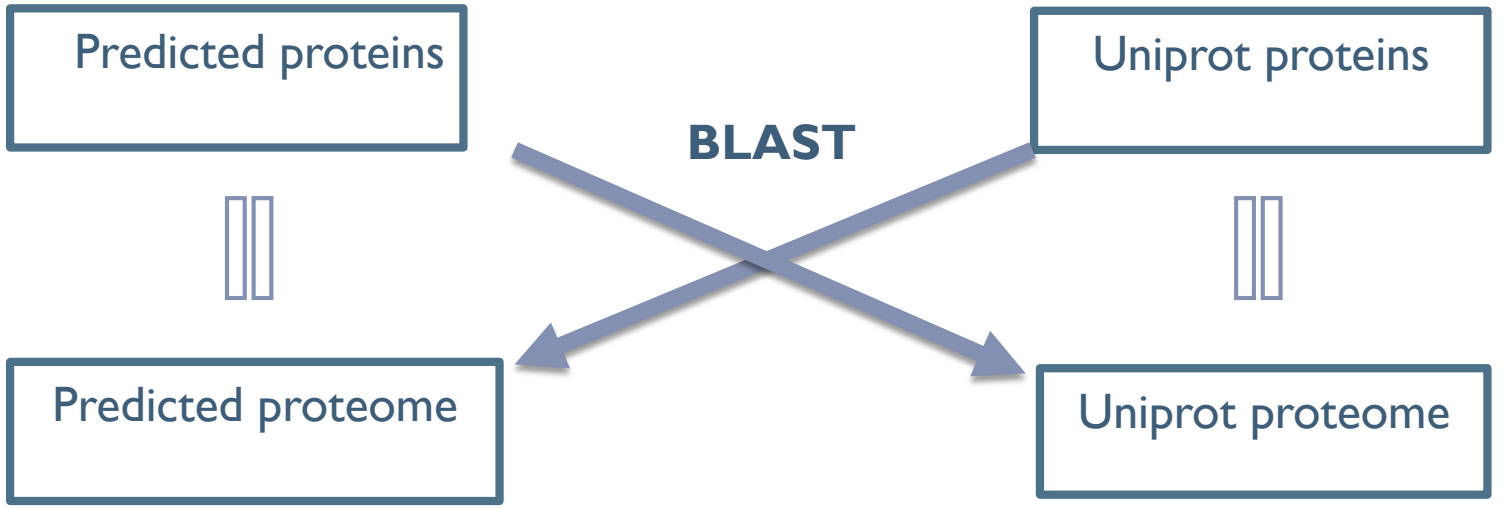


ORF prediction

Assumptions

- ORF begins with a start codon and ends with a stop codon
- If more than one ORF share the same stop codon, the longest one is kept
- Minimum ORF length of 200 bp for prokaryote and 300 bp for eukaryote
- Maximum overlap of 60 bp
- Longer ORF is kept if overlap occurs between two ORFs

Comparison with Uniprot proteome



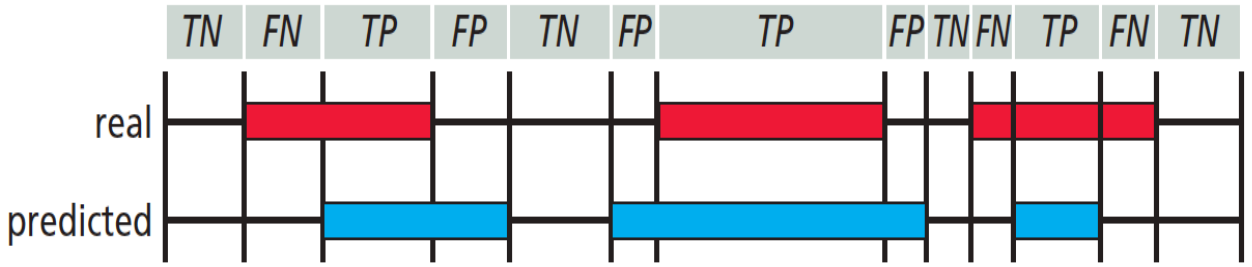
Cross BLAST identifies true positive predictions.
TP: Proteins with E-value <0.001 in both BLAST tests.

Comparison with Uniprot proteome

Genome ID	Species	Uniprot protein number	Predicted protein number	Reciprocal best hit number (TP)	Specificity	Sensitivity
9	Escherichia coli	4313	5079	3691	0.856	0.727
17	Streptomyces coelicolor	7731	7172	4009	0.519	0.559
24	Saccharomyces cerevisiae ch4	724	724	711	0.925	0.982
49	Rubrobacter xylanophilus	2778	2778	1945	0.622	0.700
51	Spiribacter curvatus	1752	1758	1460	0.785	0.830

Comparison with Glimmer ("ground truth")

- Use minimum gene length of 110 bp
- Use maximum overlap of 50 bp
- Analyzed nucleotide level prediction accuracy
- For each reading frame, each nucleotide is classified as True Positive (TP), True Negative (TN), False Positive (FP) or False Negative (FN)



Source: Zvelebil and Baum, Understanding Bioinformatics, Figure 9.12

Comparison with Glimmer ("ground truth")

- Sum up all TP, TN, FP and FN for all six reading frames
- Sensitivity = TP / (TP + FN)
- Specificity = TP/ (TP + FP)
- Calculate approximate correlation coefficient (AC)

$$ACP = \frac{1}{4}\left[\frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FP} + \frac{TN}{TN+FN}\right]$$

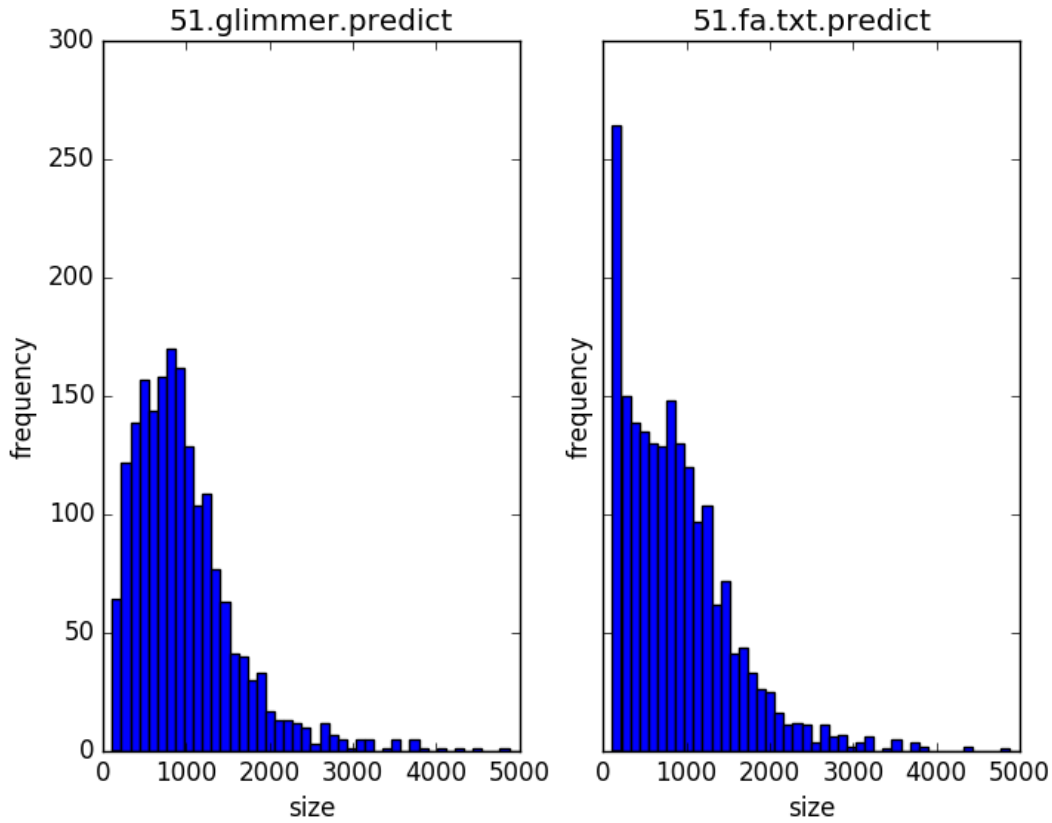
$$AC = 2 \times ACP - 1$$

Source: Zvelebil and Baum, Understanding Bioinformatics, Box 10.1

Comparison with Glimmer ("ground truth")

Genome ID	Species	true ORF	predicted ORF	average true length	average predicted length
9	Escherichia coli	5323	6329	888	773
17	Streptomyces coelicolor	8548	8175	939	954
24	Saccharomyces cerevisiae ch4	921	1604	1235	782
49	Rubrobacter xylanophilus	3375	3167	879	869
51	Spiribacter curvatus	1862	1955	973	897

Comparison with Glimmer ("ground truth")



Average gene length of *Spiribacter curvatus*

973bp in Glimmer 897bp in FuMiSh

Comparison with Glimmer ("ground truth")

Genome ID	Species	Sensitivity	Specificity	AC
9	Escherichia coli	0.95	0.917	0.922
17	Streptomyces coelicolor	0.59	0.61	0.529
24	Saccharomyces cerevisiae ch4	0.97	0.88	0.91
49	Rubrobacter xylanophilus	0.623	0.671	0.586
51	Spiribacter curvatus	0.841	0.869	0.829

Comparison with Glimmer ("ground truth")

Genome ID	Species	Sensitivity	Specificity	AC
9	Escherichia coli	0.95	0.917	0.922
17	Streptomyces coelicolor	0.59	0.61	0.529
24	Saccharomyces cerevisiae ch4	0.97	0.88	0.91
49	Rubrobacter xylanophilus	0.623	0.671	0.586
51	Spiribacter curvatus	0.841	0.869	0.829

Analysis of comparison with Glimmer

```
>./49.fa.txt_orf00005
LSAHIRAIRLVNFRNYAGATALLSPGLNVLVGENAQGKTNLLEALAFVVSGSSPRTPNDS
EVVRWGEGFVRVEARVVDGGHERRLAVGYAPGSRKRLTVDGAPVESLARYAAGVAGVRAV
TFFPDDLRVVKGSPSDRRSFLDALLSSLRPAYARAAAEYARAVQQRNQLLRRIRDGLSSE
RTLATWDRKVVELGLVLLEGRAAAAAPLDEHFRASMRALYGPQKAAVGYSYSATPERYAQ
ALREAHSADIERGITSVGPHRDDLRILLEGVDLTTYGSQGQQRLATLALKFAARDYIRDA
TGQDPVLLFDDVMSELDERRRDYLAGCFLESTQAVISTTNLRYFEPGALRRARVLGISGG
SISQTTKSGVDNGG
```

Analysis of comparison with Glimmer

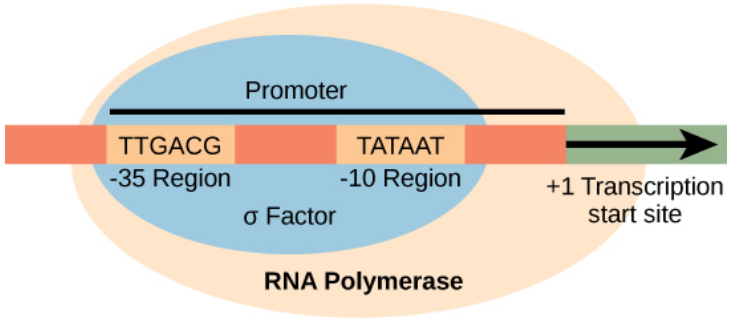
Genome ID	Species	Fraction of GTG and TTG start codons	Fraction of GTG start codons
9	Escherichia coli	0.198760098	0.13995867
17	Streptomyces coelicolor	0.45414132	0.407814693
24	Saccharomyces cerevisiae ch4	0.194353963	0.10640608
49	Rubrobacter xylanophilus	0.56237037	0.437925926
51	Spiribacter curvatus	0.249194415	0.186358754

Improvements

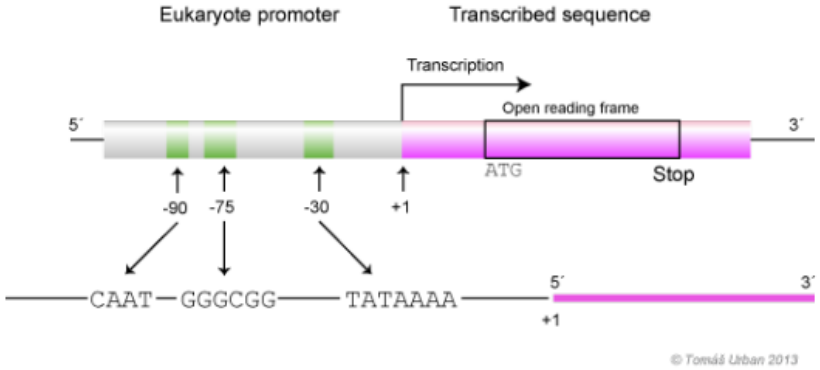
- Our improvements should mainly focus on:
  - filtering small ORFS
  - Selecting overlapping ORFs
  - Improve start codons

Improvements

- Adding promoter information



Pribnow box in prokaryotes



GC box and TATA box in eukaryotes

Distance(TSS,start codon) = 20-40 bp

Distance(TATA box,TSS) = 25-30 bp

So we assume:
Distance(TATA box,start codon) = 45-70bp

```
if args.type == 'euk':
    tataBox = ['TATAAAA','TATATAT','TATATAA','TATAAAT']
    tata_locs = []
    # Locating TATABoxs
    for i in tataBox:
        tata_loc = [m.start() for m in re.finditer(i, geno)]
        tata_locs.extend(tata_loc)
    # Distance(TATA box,start codon) = 45-70bp
    for startcodon in genome :
        if startcodon in [(tata_loc-45),(tata_loc-70):
            trueStartcodon.append(startcodon)
```


1. Mendoza-Vargas, Alfredo, et al. "Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli
2. Zvelebil, Marketa J., and Jeremy O. Baum. Understanding bioinformatics. Garland Science, 2007.

Improvement

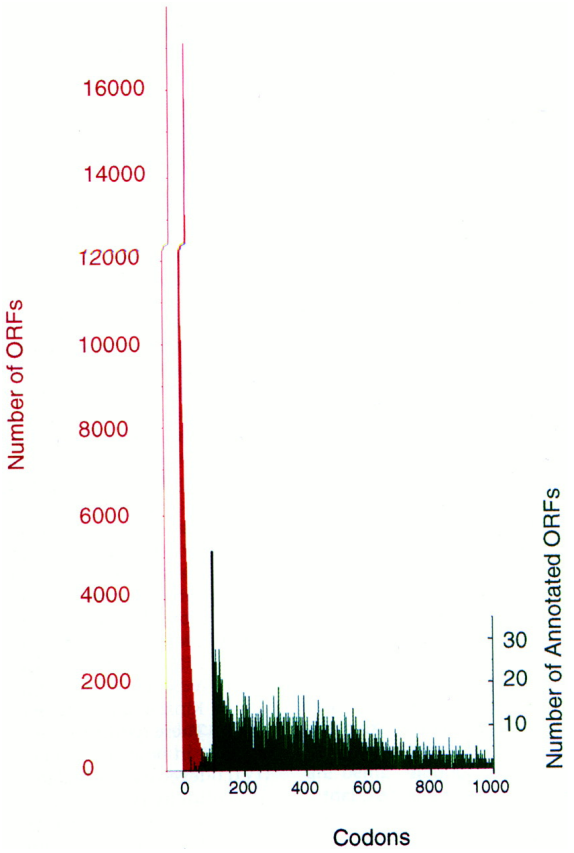
- It is difficult to assign small ORFs and set the minimum gene length.

In Ecoli:

The average ORF size = 317 amino acids

There are 381 ORFs that are smaller than 100 amino acids.

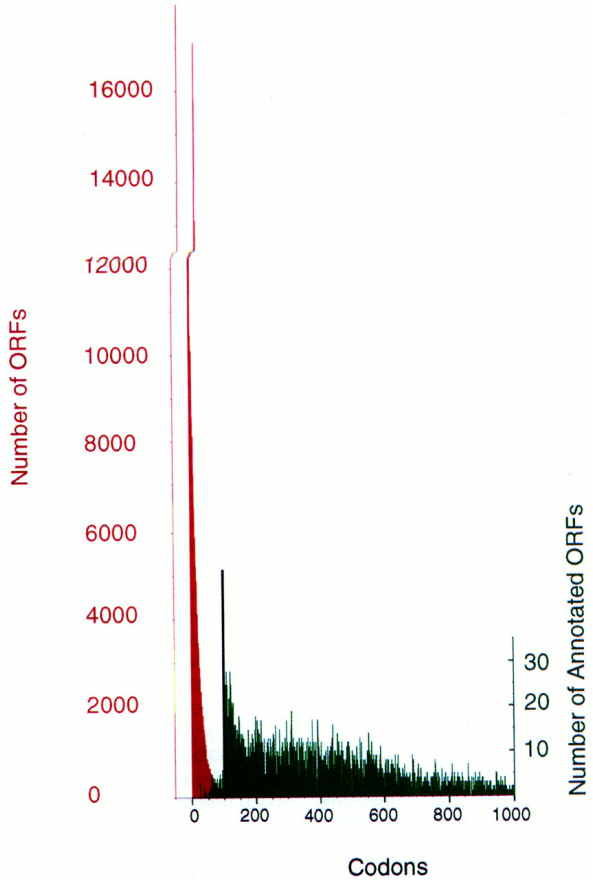
In yeast:

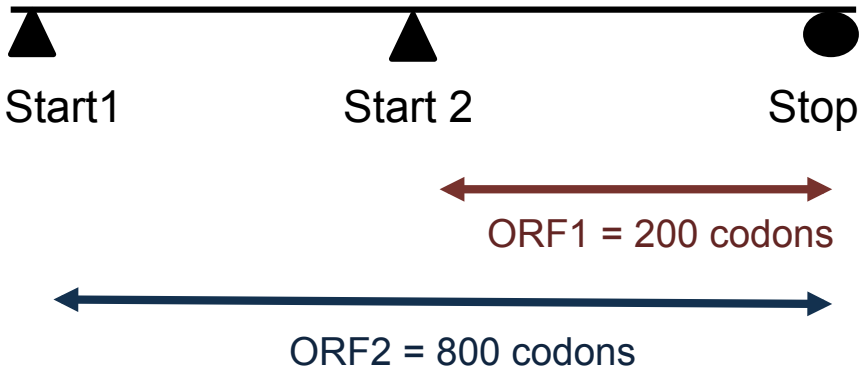


Blattner, Frederick R., et al. "The complete genome sequence of Escherichia coli K-12." *science* 277.5331 (1997): 1453-1462.

Improvement

- The distribution of gene size helps with the selection of overlapping ORFs.

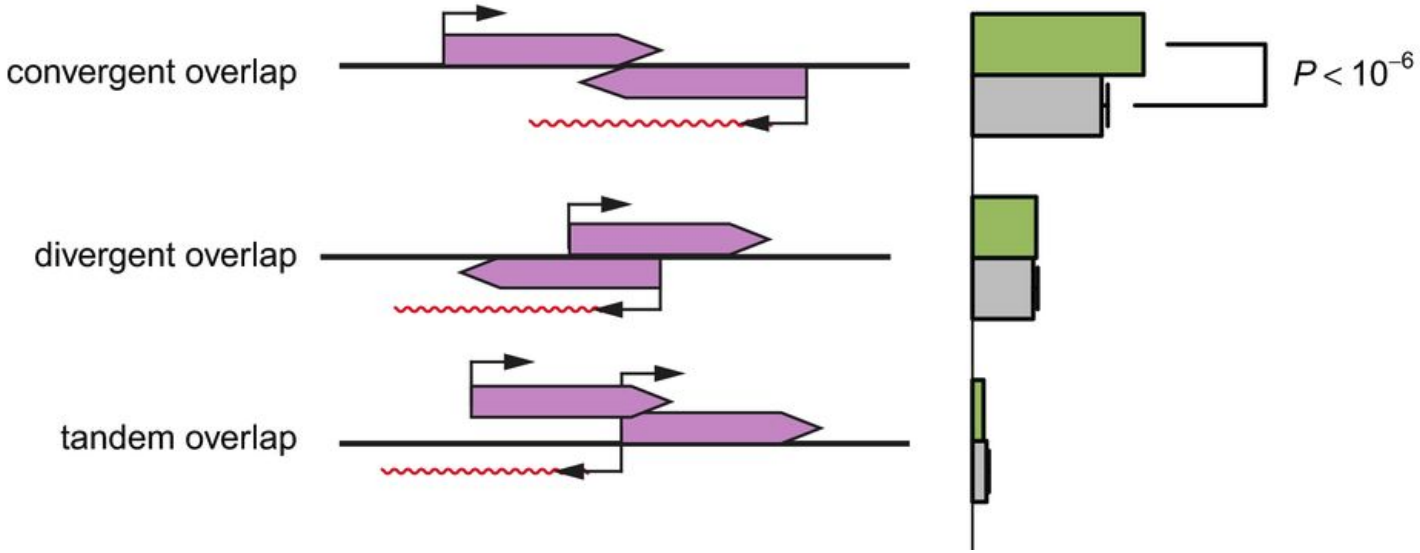




Basrai, Munira A., Philip Hieter, and Jef D. Boeke. "Small open reading frames: beautiful needles in the haystack." Genome research 7.8 (1997)

Improvement

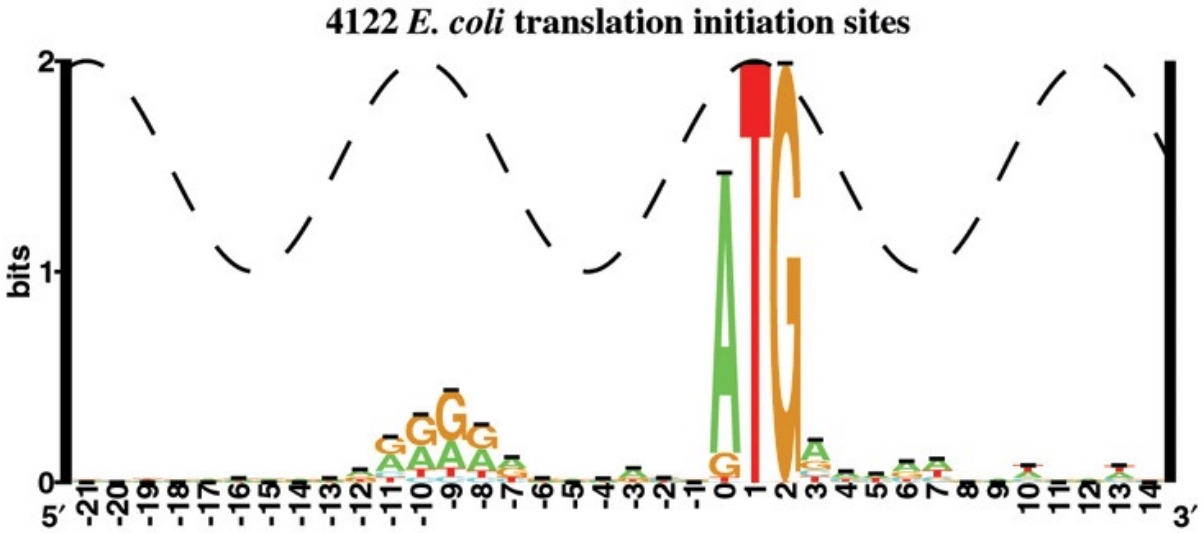
- Both orientations and lengths should be considered in selecting overlapping ORFs in different reading frames.
- A maximal overlap of 60 bp is allowed between two genes on the same strand in FuMiSh.



Clément-Ziza, Mathieu, et al. "Natural genetic variation impacts expression levels of coding, non-coding, and antisense transcripts in fission yeast." Molecular systems biology 10.11 (2014): 764.
Hyatt, Doug, et al. "Prodigal: prokaryotic gene recognition and translation initiation site identification." BMC bioinformatics 11.1 (2010): 119.

Improvement

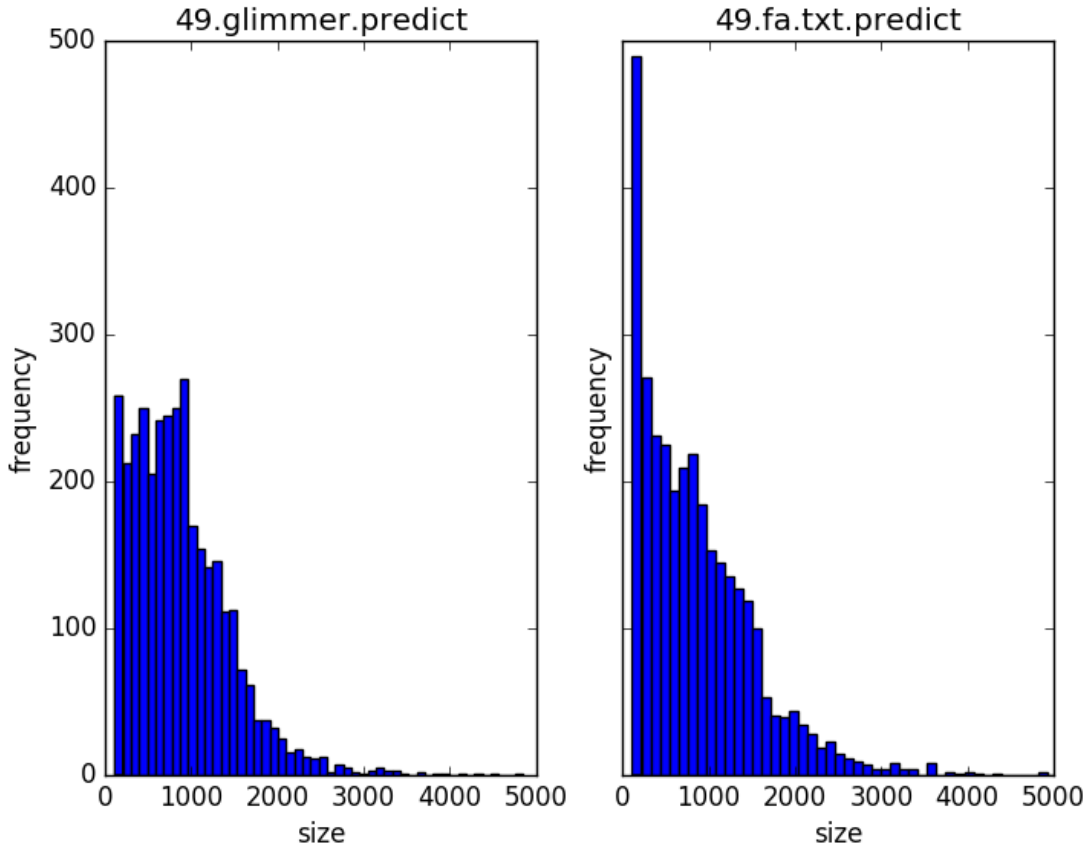
Different start codons.



With the increase in GC content,
the ORFs are more likely to start with GTG.

Validation

We can also use the gene size distribution to test the performance of different predictors statistically, using Chi square test.



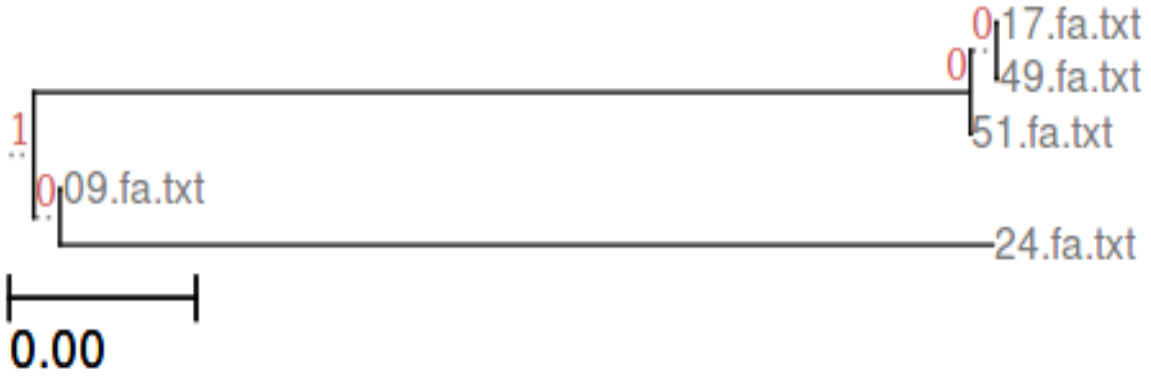
Improvement

- Algorithms

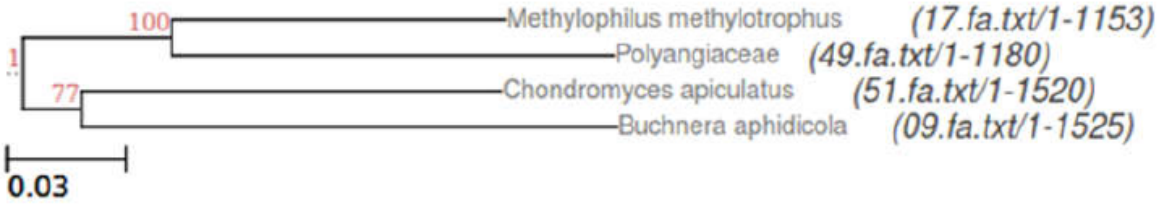
Classification, Decision tree, HMM and Neuro Network

Evolutionary relationships

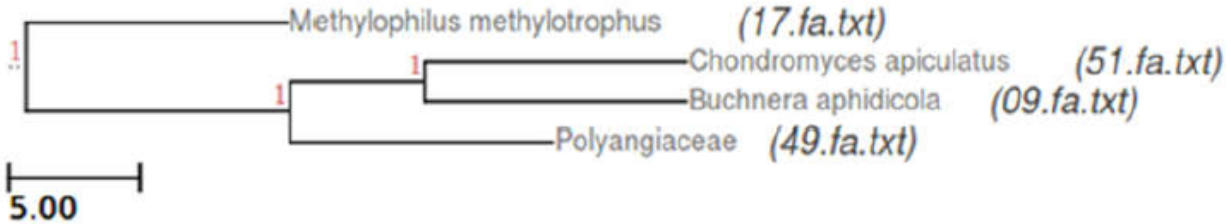
Evolutionary relationship



Tree based on dinucleotide frequencies



Tree based on 16S sequences



Consensus tree based on ten ortholog clusters

Thank you!