A Distance Measure for Genome Phylogenetic Analysis

Minh Duc Cao[1] Lloyd Allison[1,2] Trevor I. Dix[1]

[1]Clayton School of Information Technology,
Monash University.

[2]National ICT Australia
Victorian Research Laboratory, University of Melbourne.

The 22nd Australasian Joint Conference
on Artificial Intelligence

Outline

- Phylogenetic analysis.
- The proposed distance measure.
- Experimental results.
- Conclusions.

Phylogenetic Analysis

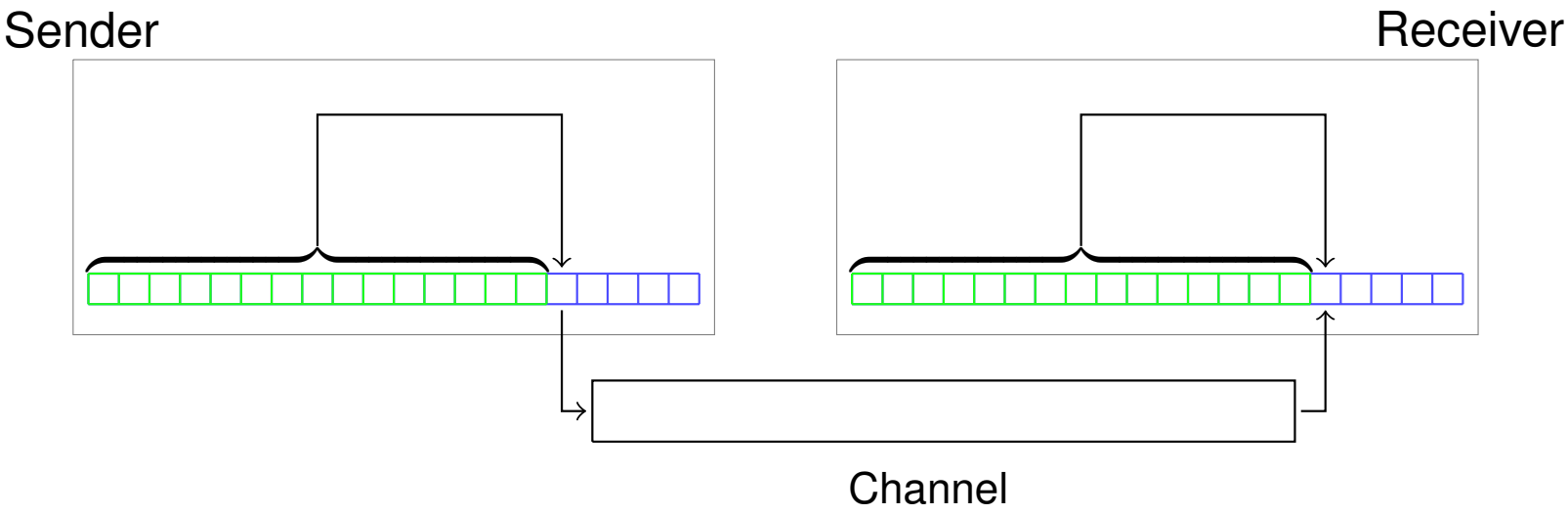
- Assemble an evolutionary relationship of a set of species.
- Classical phylogenetic analysis:
  - Select a homolog that presents in all the species genomes.
  - Generate an alignment of the sequences.
  - Apply a tree construction method to build the tree.
  - Perform a statistical test on the tree generated.
- Problem: trees from different genes may be inconsistent.
  - Variation rates of evolution among genes.
  - Some genes arisen through forms other than inheritance.

Phylogenetic Analysis from Genomes

- Reliable alignment of whole genomes is almost impossible.
- Character-based tree building methods: too computationally expensive.
- Distance-based tree building methods: Need a distance measure.
- Kolmogorov complexity, approximated by compression was proposed (Li at. al., 2001, Otu and Saywood, 2003), but...
- General compression algorithms fail to compress biological sequences.
- Existing special purpose compression algorithms cannot handle long sequences.

The Expert Model: A Predictive Compression Model

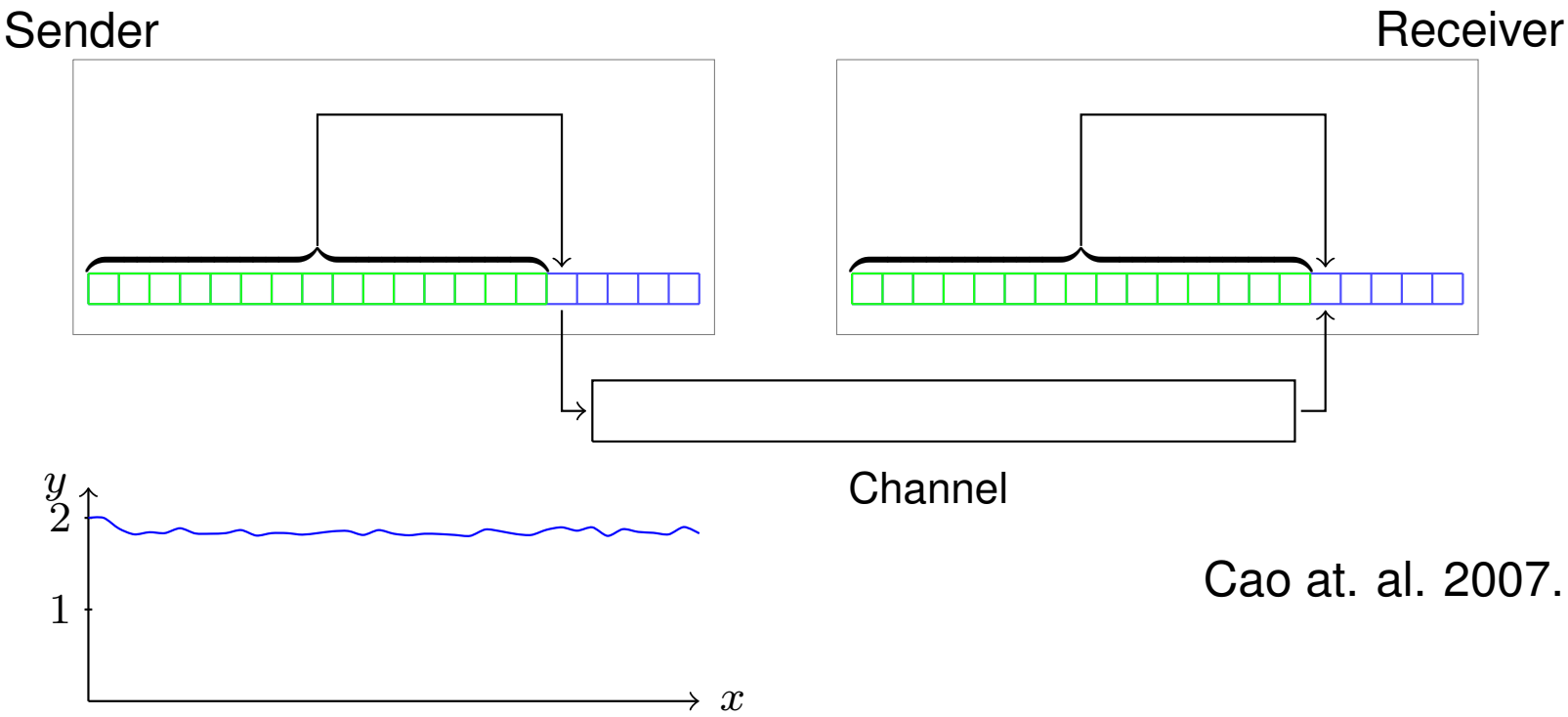
- Compress a sequence by employing predictive experts.
- Combine expert predictions by Bayesian averaging.



Cao at. al. 2007.

The Expert Model: A Predictive Compression Model

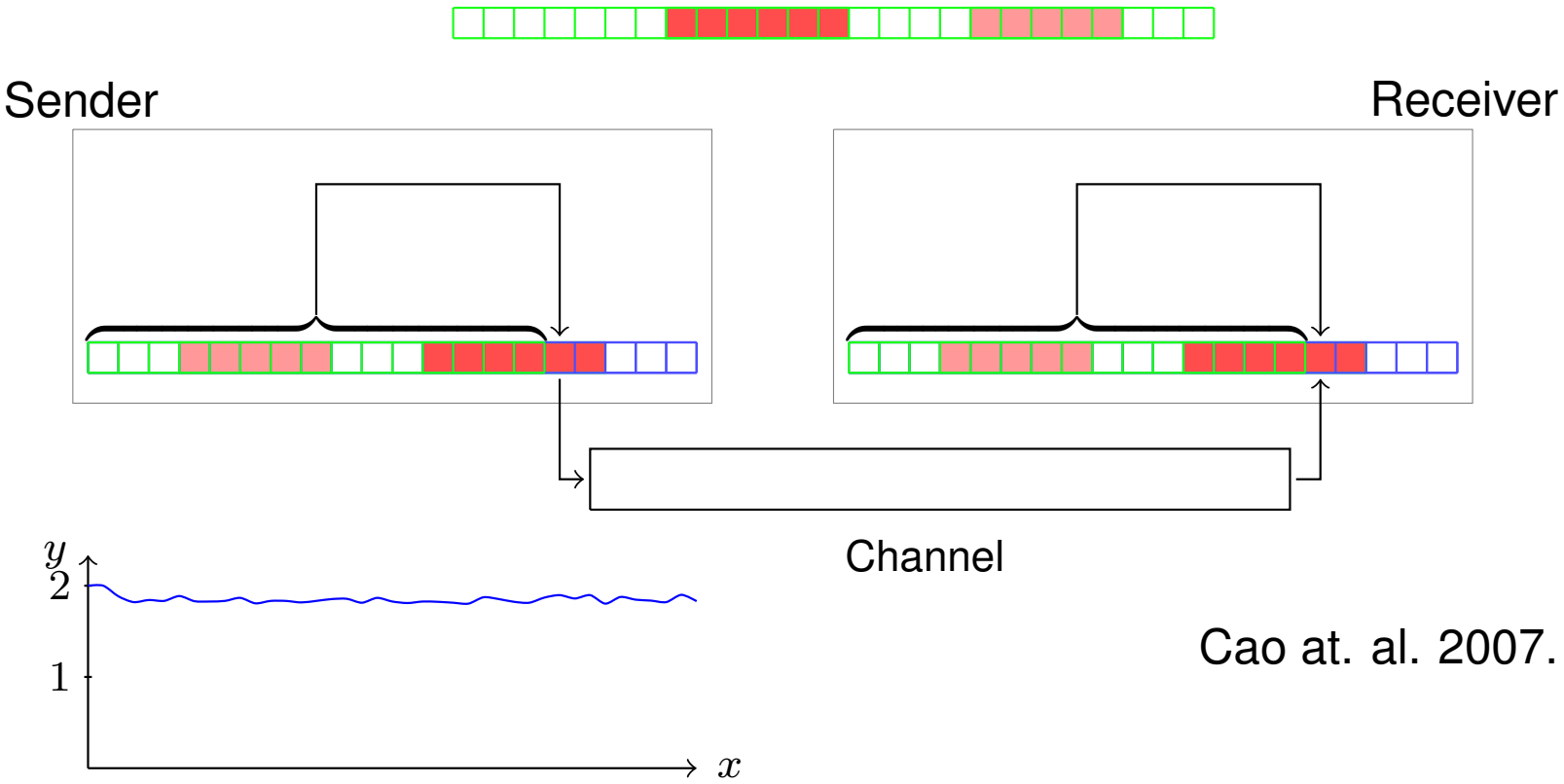
- Compress a sequence by employing predictive experts.
- Combine expert predictions by Bayesian averaging.



Cao at. al. 2007.

The Expert Model: A Predictive Compression Model

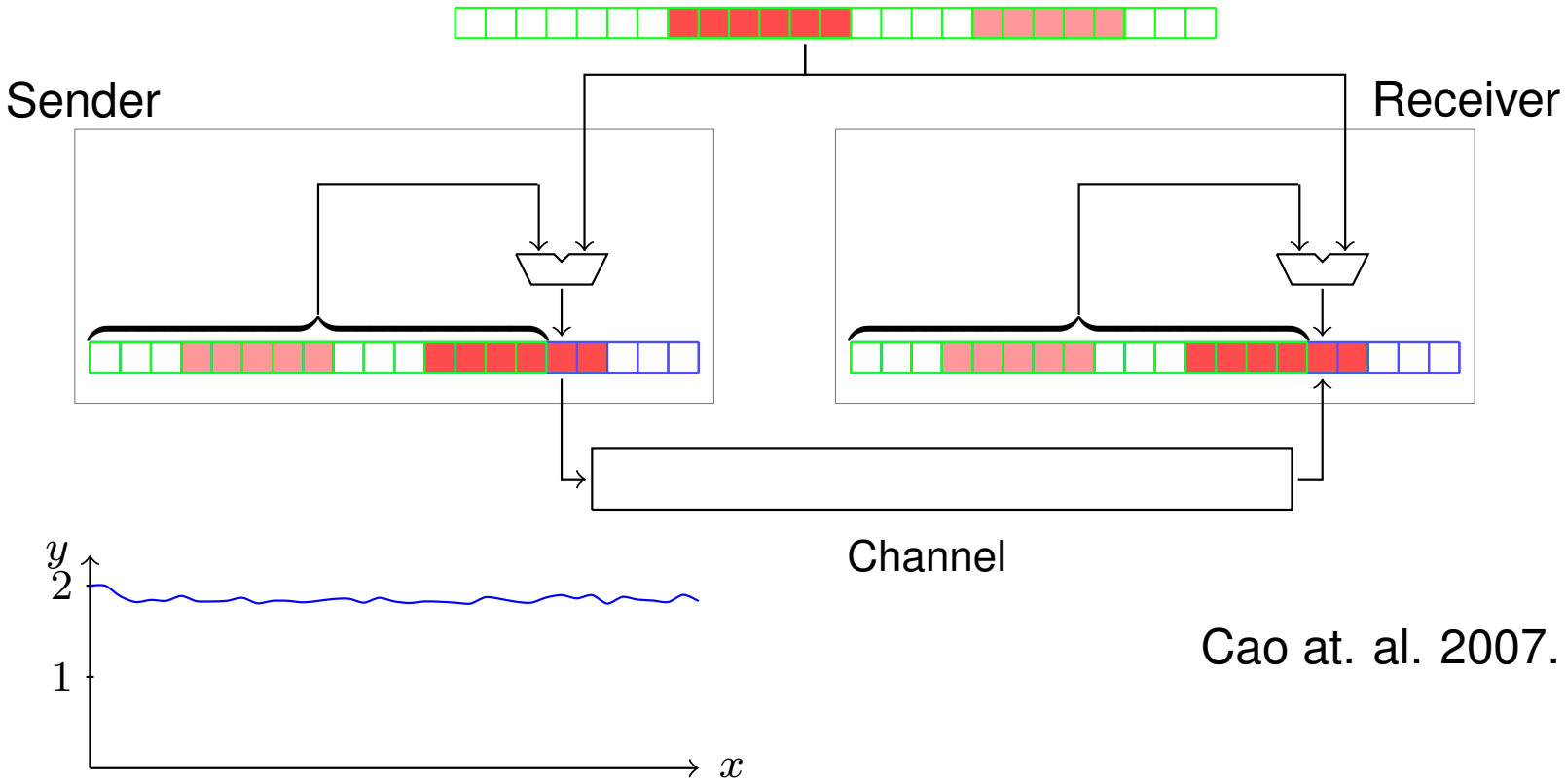
- Compress a sequence by employing predictive experts.
- Combine expert predictions by Bayesian averaging.



Cao at. al. 2007.

The Expert Model: A Predictive Compression Model

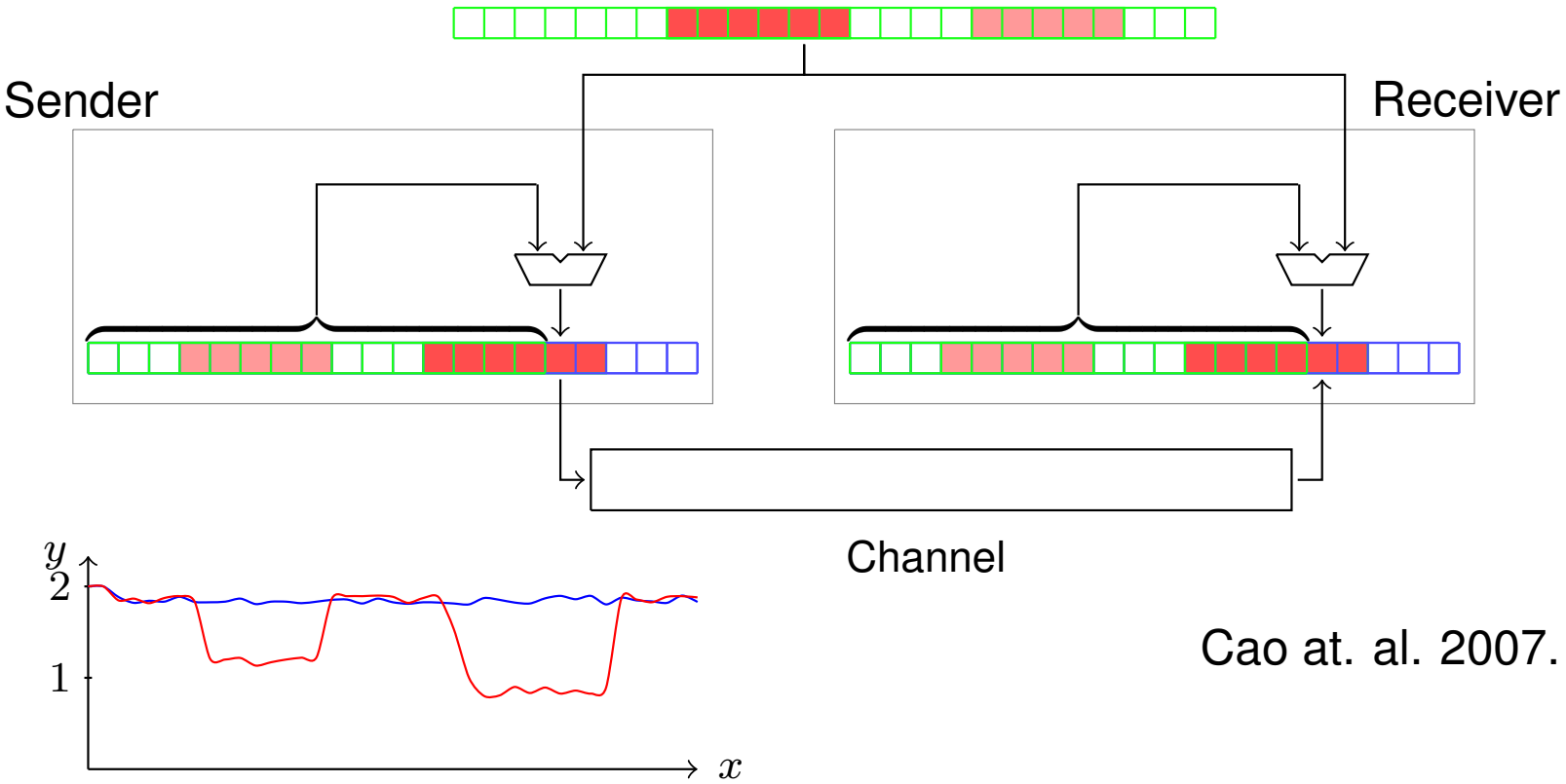
- Compress a sequence by employing predictive experts.
- Combine expert predictions by Bayesian averaging.



Cao at. al. 2007.

The Expert Model: A Predictive Compression Model

- Compress a sequence by employing predictive experts.
- Combine expert predictions by Bayesian averaging.



Cao at. al. 2007.

Distance Measure

- Similarity measure:

$$S_{X,Y} = \frac{(\mathcal{I}_X - \mathcal{I}_{X|Y}) + (\mathcal{I}_Y - \mathcal{I}_{X|Y})}{\mathcal{I}_X + \mathcal{I}_Y} = 1 - \frac{\mathcal{I}_{X|Y} + \mathcal{I}_{X|Y}}{\mathcal{I}_X + \mathcal{I}_Y}$$

- Or distance measure:

$$D_{X,Y} = 1 - S_{X,Y} = \frac{\mathcal{I}_{X|Y} + \mathcal{I}_{X|Y}}{\mathcal{I}_X + \mathcal{I}_Y}$$

- From the distance matrix, apply a distance-based construction tree method such as *Neighbour Joining*.

Experiment 1: *Plasmodium* phylogeny

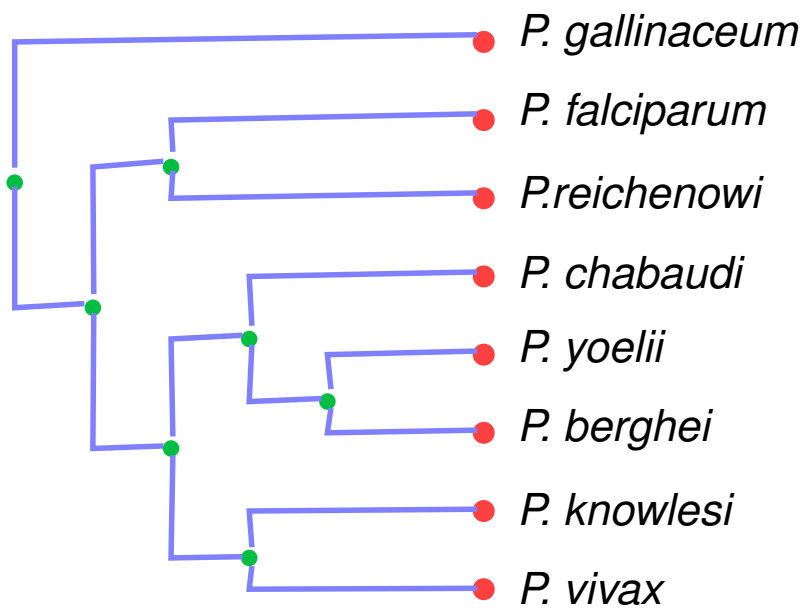
- The genomes of eight *Plasmodium* species causing malaria in various hosts.
- Variation of evolutionary rates among genes: Phylogenetic analyses from different genes resulting in different trees.
- The composition distributions are greatly different:

Species	Host	Size	A+T Content
P. falciparum	Human	23.3 Mb	80.64%
P. vivax	Human	27.0 Mb	57.72%
P. knowlesi	Monkey	22.7 Mb	61.17%
P. reichenowi	Chimpanzee	7.4 Mb*	77.81%
P. berghei	Rodent	18.0 Mb	76.27%
P. chabaudi	Rodent	16.9 Mb	75.66%
P. yoelii	Rodent	20.2 Mb	77.38%
P. gallinaceum	Bird	16.9 Mb	79.37%

* Incomplete.

Experiment 1: *Plasmodium* phylogeny

- The tree is consistent with most phylogenetic studies.
- Support the hypotheses that *P. falciparum* is the sister species of *P. reichenowi* and that *P. vivax* resulted from host switch from *P. knowlesi*.

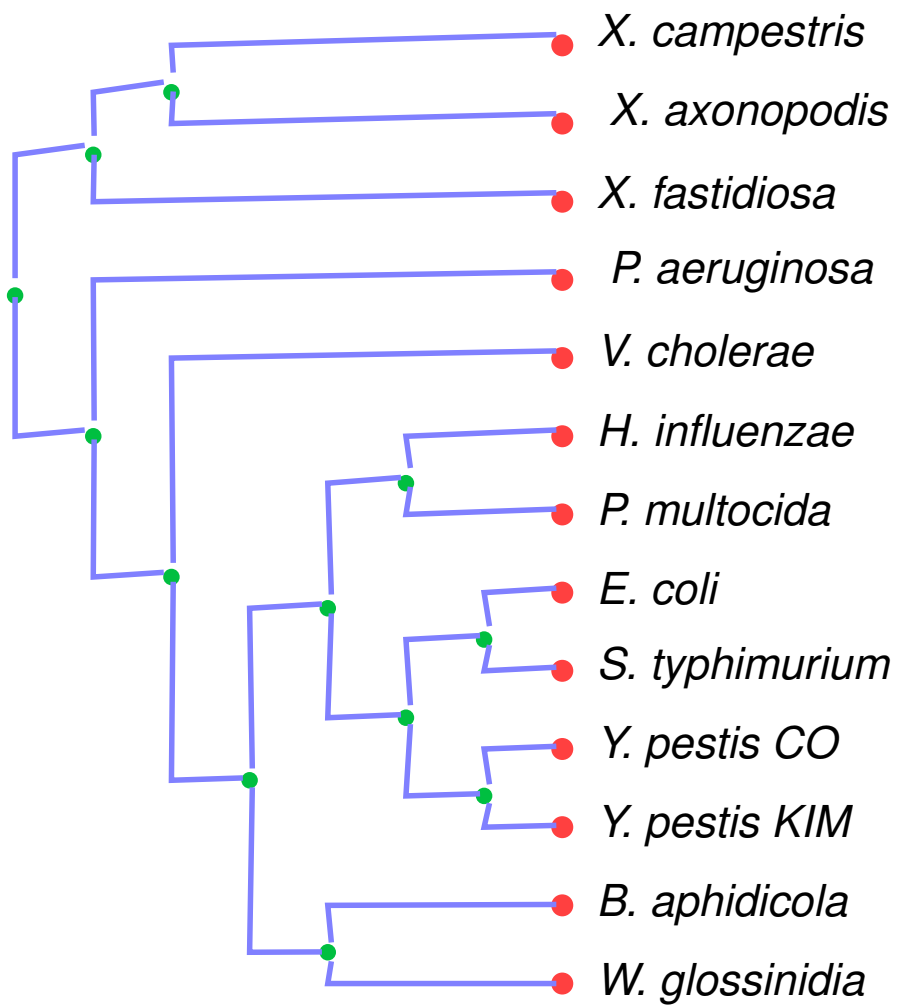


Experiment 2: Bacteria Phylogeny

- A dataset of 13 bacteria genomes in the γ-*Proteobacteria* group.
- Genome sizes range 1.3Mb - 7Mb, total size 45Mb.
- Known for abundant of horizontal genes transfer.
- Out of 14,158 gene families in 13 genomes, only 275 families present in all 13 genomes, and 205 families contain one gene in each genome.
- An earlier study from 205 gene families resulted in 13 tree topologies.
- Four most likely trees are consistent with 180 gene alignments.

Experiment 2: Bacteria Phylogeny

- Close to the four most probable trees, only different at the positions of *B. aphidicola, W. glossinidia* and *V. cholerae*.



Conclusions

- An information theoretic method to measure distances between genomes.
- A predictive model for compression of genomes.
- Scalable with large datasets.
- Work well with obscure data and produce highly confident phylogenetic trees.