Joint Bayesian Estimation of Alignment and Phylogeny

Benjamin D. Redelings and Marc A. Suchard

Outline

- Introduction
- Existing Methods
- Bayesian Coestimation and Inference
- Results

Introduction

- There are problems with existing alignment and phylogeny methods that can produce inaccurate results

- Introduce a new method based on Bayesian inference to simultaneously estimate the alignment and phylogeny

- Results show that trees remain consistent when more taxa are added

Existing Methods

- Two stages:
  - Sequence alignment
  - Phylogeny reconstruction

Existing Methods

- Works well when alignments are well resolved
- Spurious or no results when alignments are uncertain
- Exaggerated support for phylogenies
- Biased towards the alignment guide tree

Existing Methods

- Solutions:
  - Remove ambiguous regions from alignments
  - Code ambiguity information into the alignment
  - Optimization alignment: coestimation of alignments and phylogenies in a parsimony framework

Bayesian Coestimation

- Joint estimation of alignments and phylogenies
- Weighted by posterior probabilities
- No external guide tree—eliminates bias
- More accurate substitution and indel models

Bayesian Coestimation

- Markov chain Monte Carlo (MCMC) sampling
- Simplifying assumption: indels occur independently of branch length
  - Can be modeled using a symmetrical pair-hidden Markov model (pair-HMM)

$$\mathrm{P}(\mathbf{H} \mid \mathbf{E}) = \frac{\mathrm{P}(\mathbf{E} \mid \mathbf{H}) \cdot \mathrm{P}(\mathbf{H})}{\mathrm{P}(\mathbf{E})}$$

Bayes' Theorem

$\mathbf{Y}$ = set of sequences

$\mathbf{A}$ = multiple alignment

τ = unrooted tree topology

$\mathbf{T}$ = branch lengths

$\boldsymbol{\Theta}$ = substitution parameters

$\boldsymbol{\Lambda}$ = indel parameters

Variables

$$\mathrm{P}(\mathbf{Y}, \tau, \mathbf{T}, \mathbf{\Theta} \mid \mathbf{A}) = \mathrm{P}(\mathbf{Y} \mid \tau, \mathbf{T}, \mathbf{\Theta}, \mathbf{A}) \times \mathrm{P}(\tau, \mathbf{T}) \times \mathrm{P}(\mathbf{\Theta})$$

Traditional Probabilistic Model

$$\begin{aligned}\mathrm{P}(\mathbf{Y}, \mathbf{A}, \tau, \mathbf{T}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}) = \mathrm{P}(\mathbf{Y} \mid \mathbf{A}, \tau, \mathbf{T}, \boldsymbol{\Theta}) \times \mathrm{P}(\mathbf{A} \mid \tau, \boldsymbol{\Lambda}) \\ \times \mathrm{P}(\tau, \mathbf{T}) \times \mathrm{P}(\boldsymbol{\Theta}) \times \mathrm{P}(\boldsymbol{\Lambda})\end{aligned}$$

Coestimation Probabilistic Model

Posterior Sampling

- Sample from the model posterior P(A, τ, T, Θ, Λ | Y) using Markov Chain Monte Carlo (MCMC)

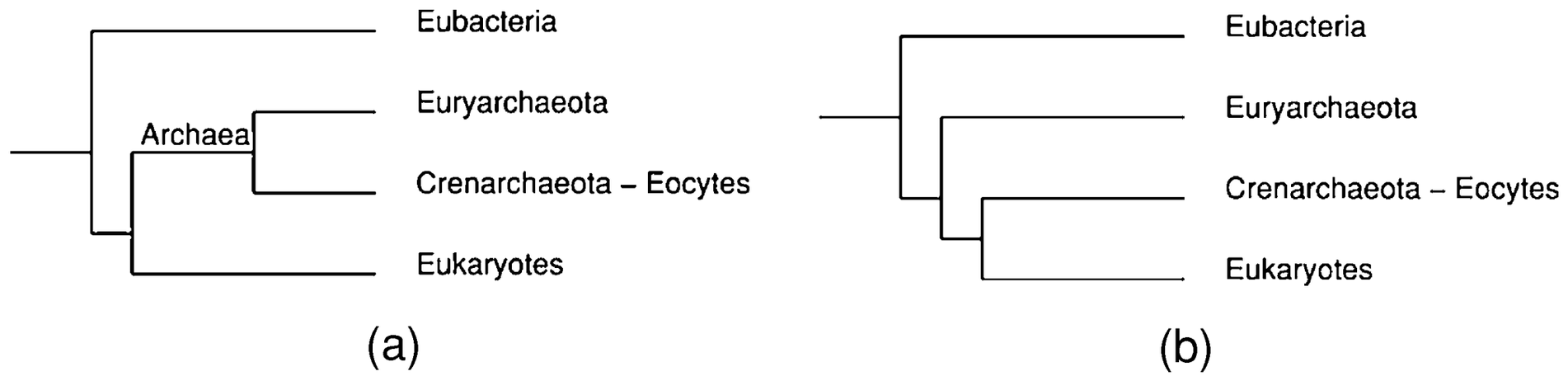
- Update the topology and each branch length, pairwise branch alignment, and internal node sequence length and alignment at least once each iteration

- Topology is updated with nearest-neighbor interchange

- Parameters sampled a Poisson number of times, with mean 1 + B/3, where B is the number of branches

Results

- Early Branching in the Tree of Life: Resolving the Archaea
- Archaeal hypothesis: Archaea form a monophyletic group
- Eocyte hypothesis: Archaea are paraphyletic and that the eocyte Archaea are more closely related to Eukaryotes than to other Archaea



Archaeal vs Eocyte Hypothesis

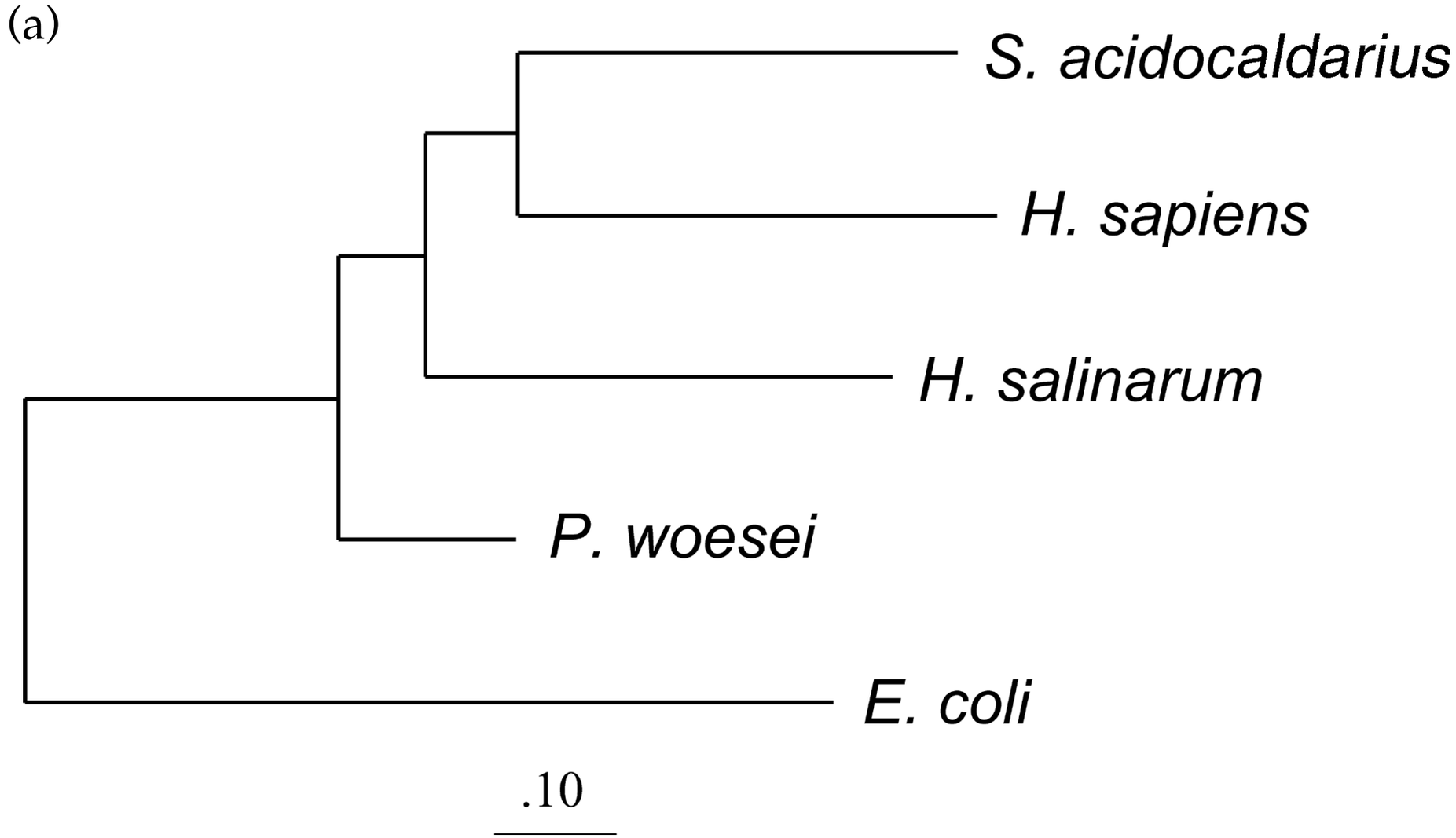
(a): Archaeal
(b): Eocyte

Species Examined

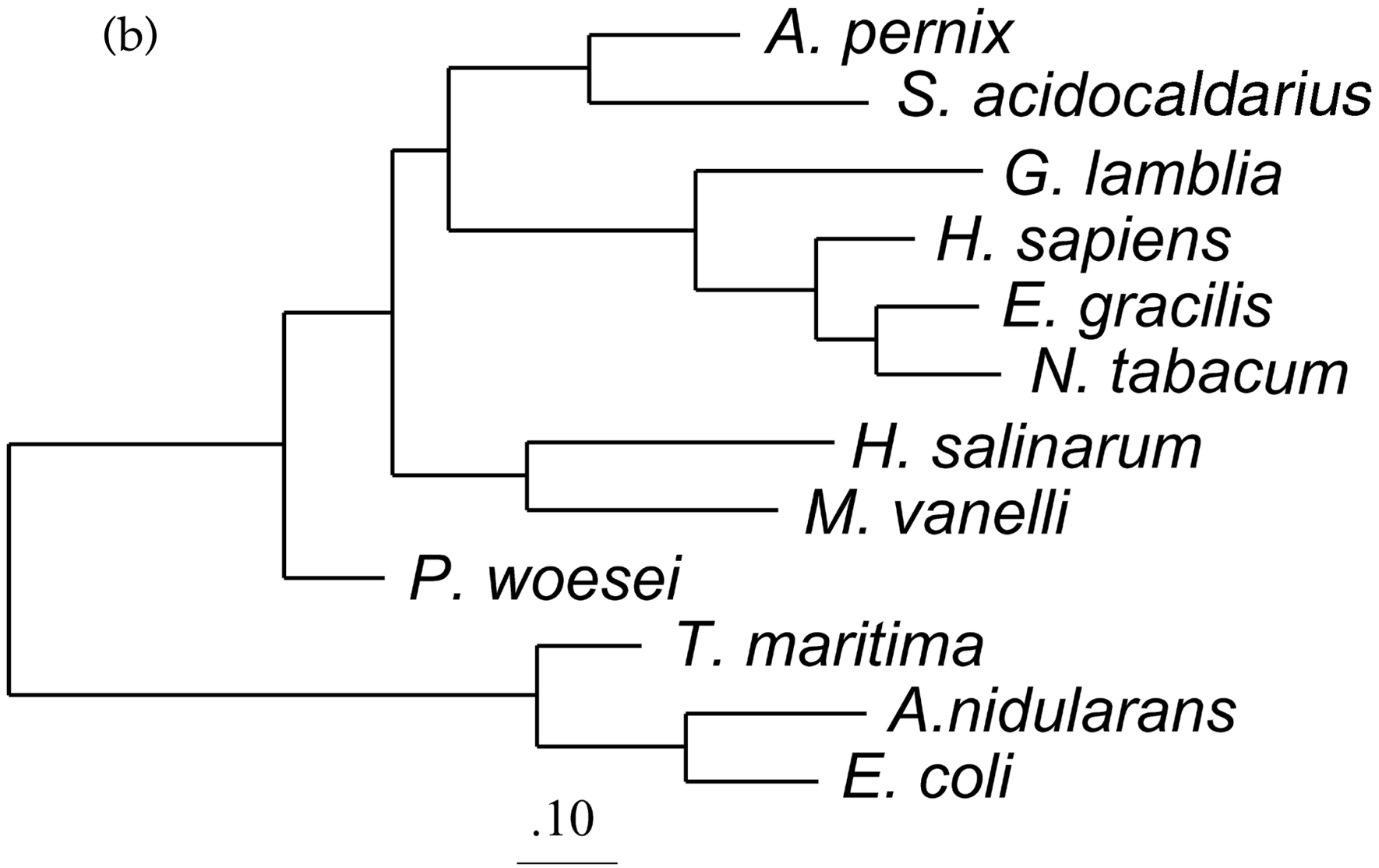
Taxa	Domain	Order	Note
Homo sapiens	**Eukaryotes**	**Metazoa**	**Human beings**
Nicotiana tabacum	Eukaryotes	Plantae	Tobacco plant
Euglena gracilis	Eukaryotes	Protista	Photosynthetic single cell
Giardia lamblia	Eukaryotes	Diplomonadida	Intestinal protist
Sulfolobus acidocaldarius	**Archaea**	**Crenarchaeota**	**High pH thermophile**
Aeropyrum pernix	Archaea	Crenarchaeota	Anaerobic thermophile
Pyrococcus woesei	**Archaea**	**Euryarchaeota**	**Thermophile**
Halobacterium salinarum	**Archaea**	**Euryarchaeota**	**Methanogen**
Methanococcus vannelli	Bacteria	Aquificae	Thermophile
Thermotoga maritima	Bacteria	Thermotogales	Thermophile
Anacystis nidularans	Bacteria	Cyanobacteria	Photosynthetic, aquatic
Escherichia coli	**Bacteria**	**Proteobacteria**	**Intestinal symbiont**

Results

- 5– and 12–taxon datasets
- EF-1α/Tu protein: at least 26% conserved
- Near threshold of 20~25% where homology becomes difficult to detect



5-Taxon Tree



12-Taxon Tree