Advanced Bayesian Phylogenetics: Phyloalignment

Philippe Lemey and Marc A. Suchard

Rega Institute

Department of Microbiology and Immunology

K.U. Leuven, Belgium, and

Departments of Biomathematics and Human Genetics

David Geffen School of Medicine at UCLA

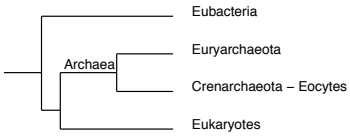
Department of Biostatistics

UCLA School of Public Health

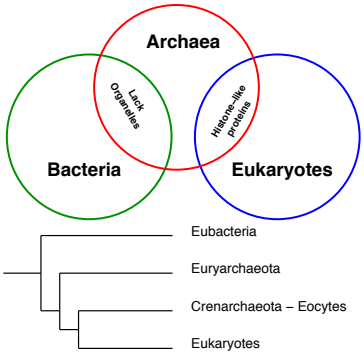
Resolving Early Branches in the Tree of Life

3? Domains of Life (Woese et al. 1990)

Contentious issue in genomics:
Do the Archaea form a single clade
(Rivera and Lake, 1992)?



Archaeal Tree



Eocyte Tree

Early evidence based on phylogenetic reconstruction techniques:

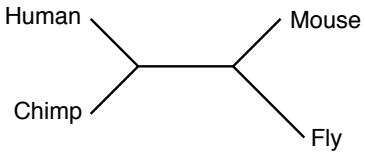
- Model how biologic sequences mutated over time
- Infer branching patterns based on “shared" substitutions

Traditional Phylogenetic Reconstruction

Reconstruction Example

Human	–	T	T	C	C	T	G	G	A	A	T
Chimp	–	T	A	C	C	T	G	G	A	A	T
Mouse	–	A	A	C	C	T	–	–	T	A	T
Fly	–	A	A	G	A	T	C	G	T	A	T
Site:		1	2	3	4	5	6	7	8	9	10

Along Molecular Sequence



- **Substitution**: single residue replaces another
- **Insertion/deletion**: residues are inserted or deleted

Statistical Model

Assume: Homologous sites are iid and site patterns (e.g. dotted box)

$$XY \ldots Z \sim \text{Multinomial}(p_{XY \ldots Z})$$

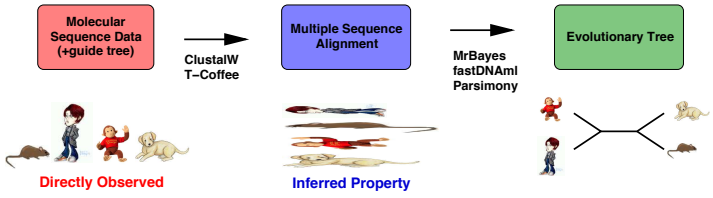
where $p_{XY \ldots Z}$ is determined by an unknown tree τ and a continuous-time Markov chain model (for residue substitution) given by infinitesimal rate matrix $\mathbf{Q}$

$$\text{P}(X \to Y \text{ in time } t) = \left\{ e^{t\mathbf{Q}} \right\}_{XY}$$

Calculating $p_{XY \ldots Z}$ integrates out unobserved states (internal nodes, gaps).

Fundamental Difficulty: Sequential Estimation

Current phylogenetic reconstruction methods:



Issues: Poor alignment biases phylogeny (Lake data: EF-1α/Tu)

- Use guide tree and naive evolutionary models (**Trouble!**)

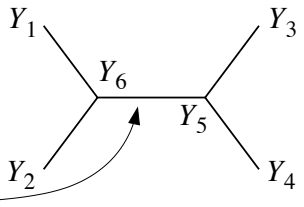
Solution: Infer alignment and phylogeny **simultaneously**

Previous approaches: Limited

- Optimization alignment, parsimony-based
- TFK91/92, forbidden positional homologies, inefficient

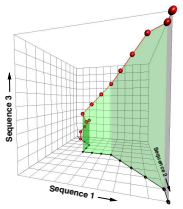
Alignment as a Random Variable

		Y	M(A)	f
Observed Data	Y_1	(A,T,T,C)	1 2 – 3 4	$A\ T\ –\ T\ C$
	Y_2	(A,T,T,G)	1 2 – 3 4	$A\ T\ –\ T\ G$
	Y_3	(T,C,T,G)	– 1 2 3 4	$–\ T\ C\ T\ G$
	Y_4	(T,C,T)	– 1 2 3 –	$–\ T\ C\ T\ –$
Missing Data	Y_5	$(\star,\star,\star)$	– 1 2 3 –	$–\ \star\ \star\ \star\ –$
	Y_6	$(\star,\star,\star,\star)$	1 2 – 3 4	$\star\ \star\ –\ \star\ \star$



Just over **1 billion** possible alignments for $\mathbf{Y}_{\text{obs}}$

Explore space via **Forward-Backward algorithm (DP)** (Scott, 2002) to consider all possible alignments (and phylogenies) in polynomial time, weighted by posterior probability

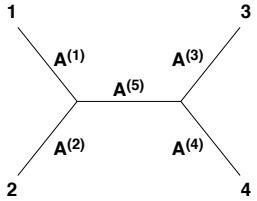


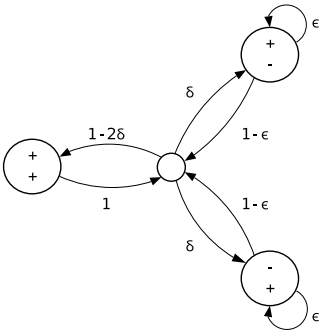
Note substitution process depends only on $\mathbf{Y}_{\text{obs}} \Rightarrow$ separates substitution and indel processes into (substitution likelihood $\times$ gap prior)

Gap Model along a Branch

Let the **multiple alignment** $\mathbf{A} = (A^{(1)}, \ldots, A^{(B)})$

- $\mathbf{A}$ is composed of pairwise alignments along each branch
- Pairwise alignment distribution follows a pair hidden Markov model (pair-HMM) **conditional** on equal sequence lengths at internal nodes





Pair-HMM parameterized by $\mathbf{\Lambda} = (\delta, \epsilon)$

- δ: Probability of indel
- ϵ: Probability of extending an indel

Affine gap penalty $\approx [\log \delta] + (\ell - 1)[\log \epsilon]$

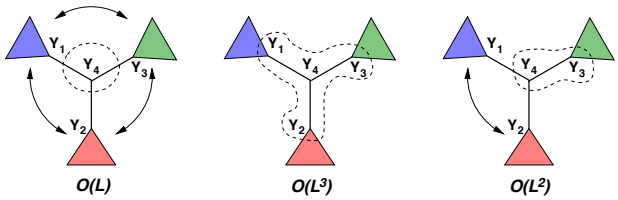
Choosing the Blocks: Efficient Sampling

Gibbs cycle over smaller blocks in alignment $\mathbf{A}$:

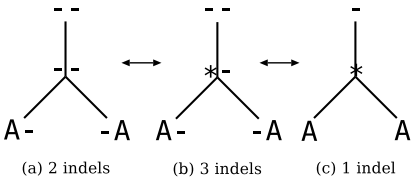
$$A^{(1)}, A^{(2)}, (A^{(3)}, A^{(4)}), (A^{(5)}, A^{(6)}, A^{(7)}), \ldots$$

Let $L =$ length of alignment:

- $O(L)$, too cold (Holmes and Bruno 2001)
- $O(L^3)$, too hot (Jensen and Hein 2005)
- $O(L^2)$, just right?



Possible **poor mixing** with $O(L)$ algorithm:

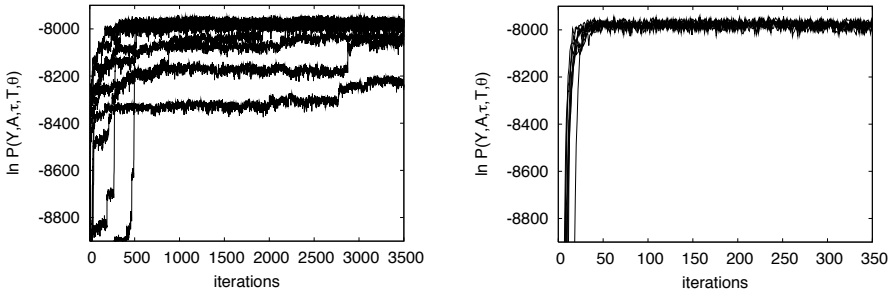


(a) 2 indels (b) 3 indels (c) 1 indel

Must move through unfavorable intermediate to align/unalign sequence fragments

Improved Alignment Mixing with $O(L^2)$ Sampling

$O(L)$ **Only** $O(L)$ **and** $O(L^2)$



Enabling MCMC transition kernel decreases:

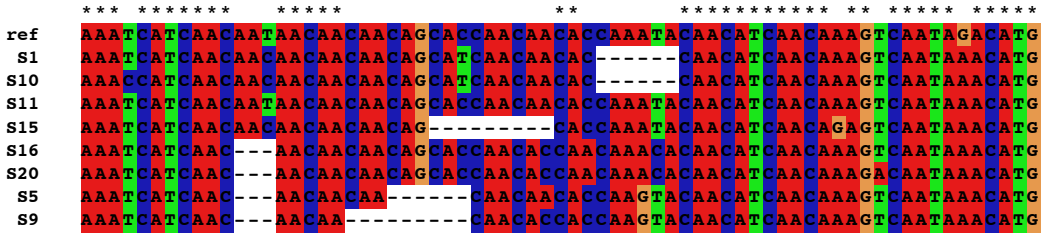
- Burn-in time
- Autocorrelation

Result: $> 70\times$ improvement shown here (12-taxon EF-1α/Tu dataset).

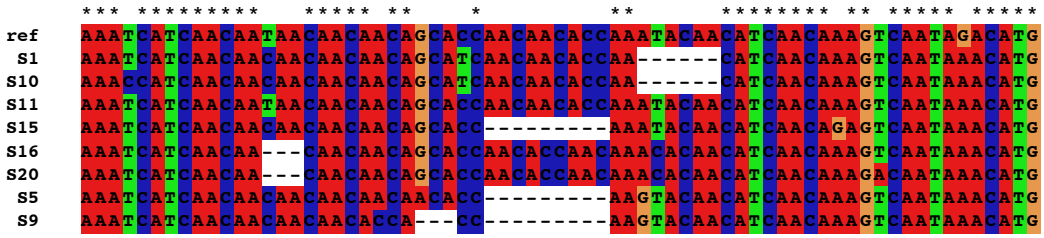
Sequential vs. Simultaneous Illustration

SIVmac251 partial *env* sequences from Cheynier et al (2001)

Sequential (ClustalW) alignment of hypervariable region:



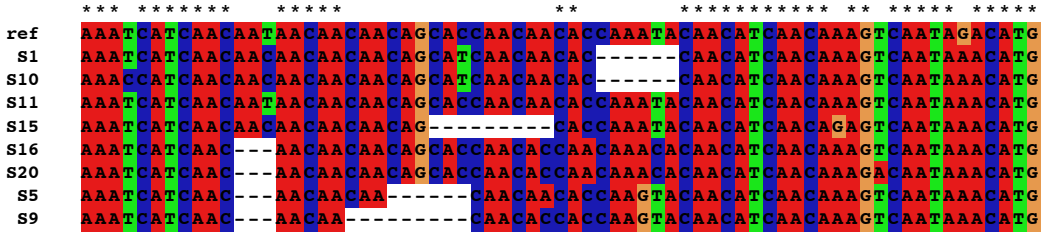
Simultaneous MAP alignment:



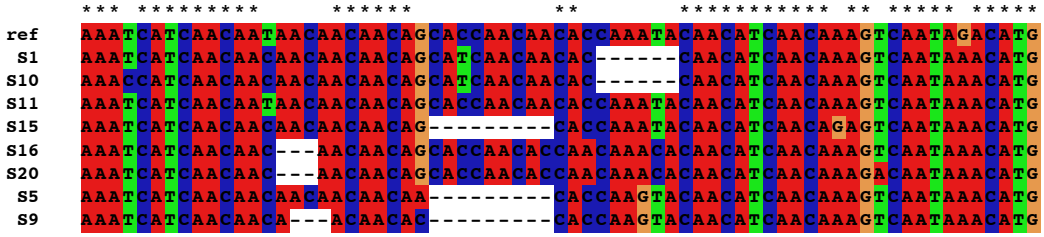
Sequential vs. Simultaneous Illustration

SIVmac251 partial *env* sequences from Cheynier et al (2001)

Sequential (ClustalW) alignment of hypervariable region:



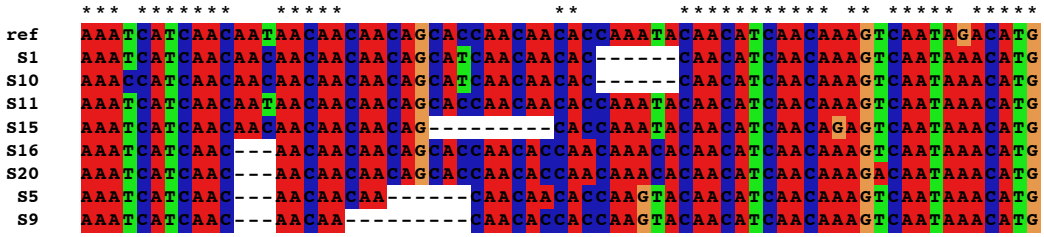
Sampled alignment (1):



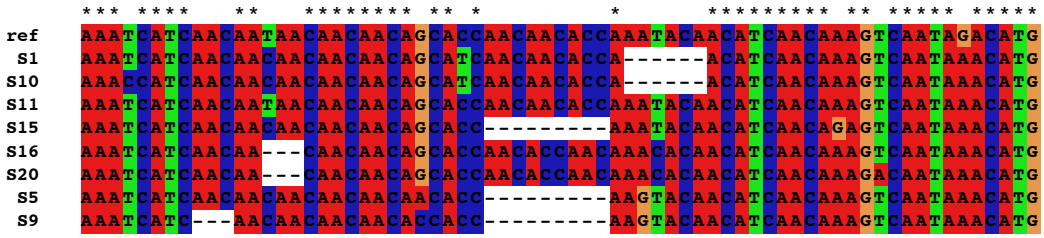
Sequential vs. Simultaneous Illustration

SIVmac251 partial *env* sequences from Cheynier et al (2001)

Sequential (ClustalW) alignment of hypervariable region:



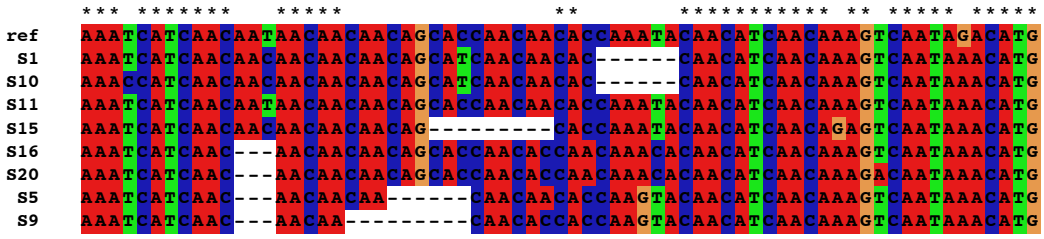
Sampled alignment (2):



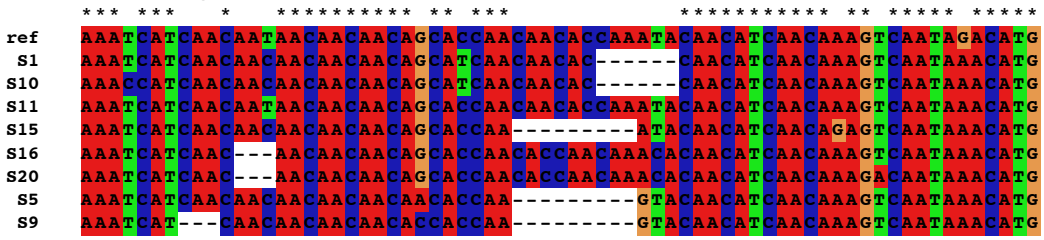
Sequential vs. Simultaneous Illustration

SIVmac251 partial *env* sequences from Cheynier et al (2001)

Sequential (ClustalW) alignment of hypervariable region:



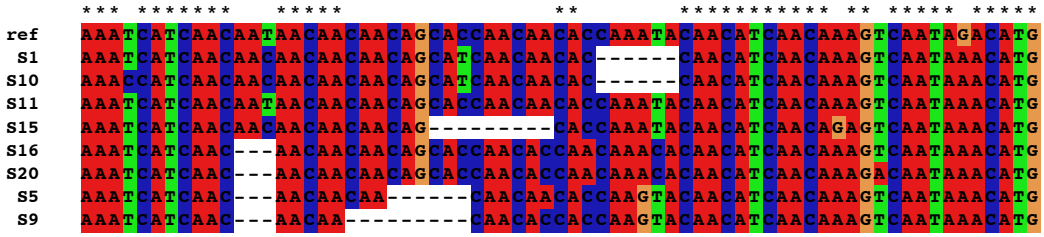
Sampled alignment (3)



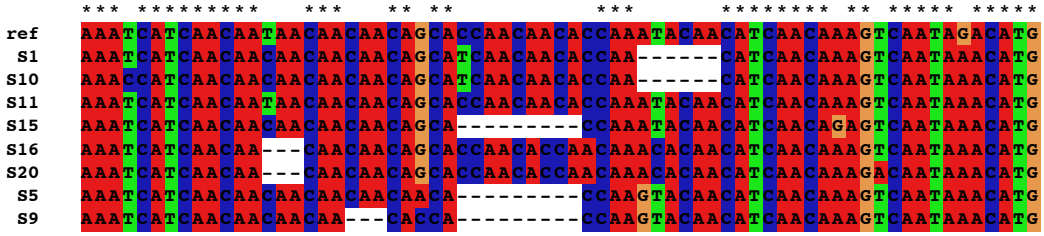
Sequential vs. Simultaneous Illustration

SIVmac251 partial *env* sequences from Cheynier et al (2001)

Sequential (ClustalW) alignment of hypervariable region:



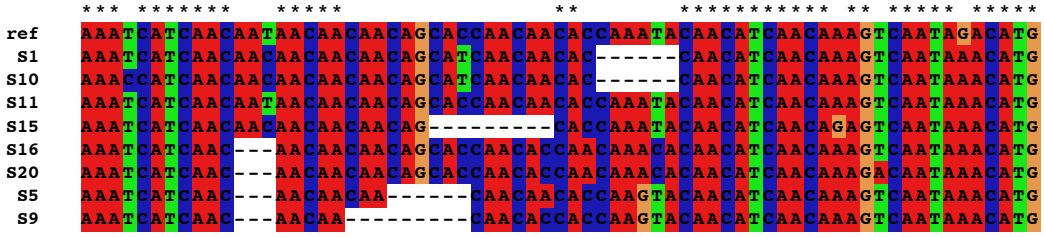
Sampled alignment (4)



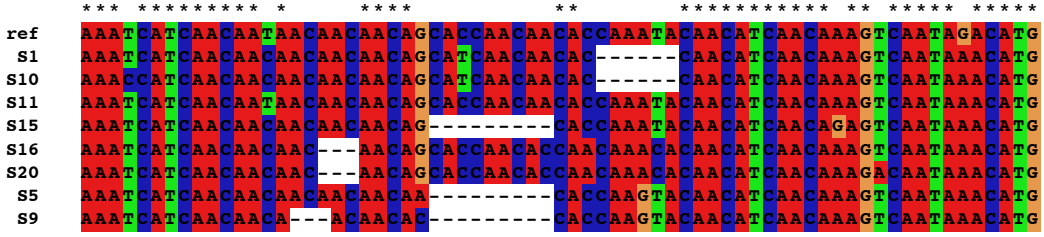
Sequential vs. Simultaneous Illustration

SIVmac251 partial *env* sequences from Cheynier et al (2001)

Sequential (ClustalW) alignment of hypervariable region:



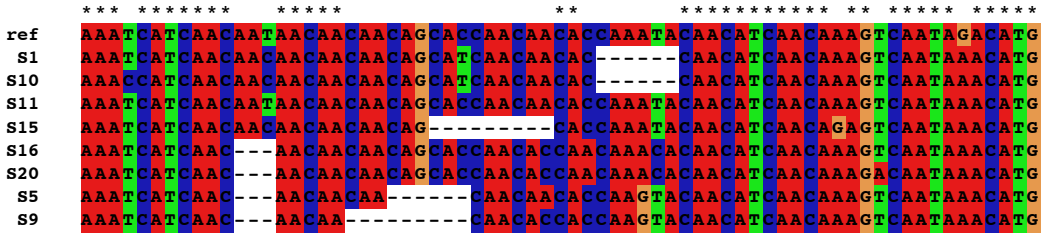
Sampled alignment (5)



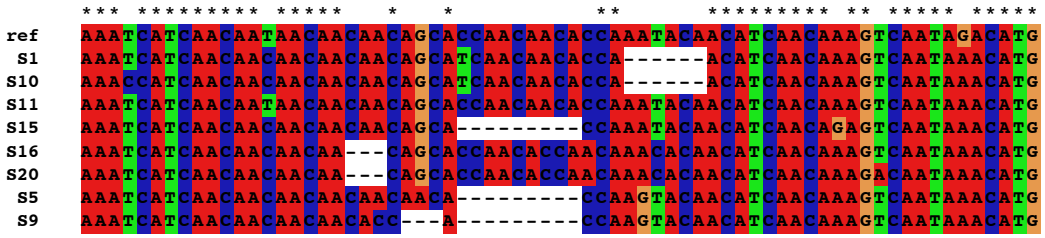
Sequential vs. Simultaneous Illustration

SIVmac251 partial *env* sequences from Cheynier et al (2001)

Sequential (ClustalW) alignment of hypervariable region:



Sampled alignment (6)

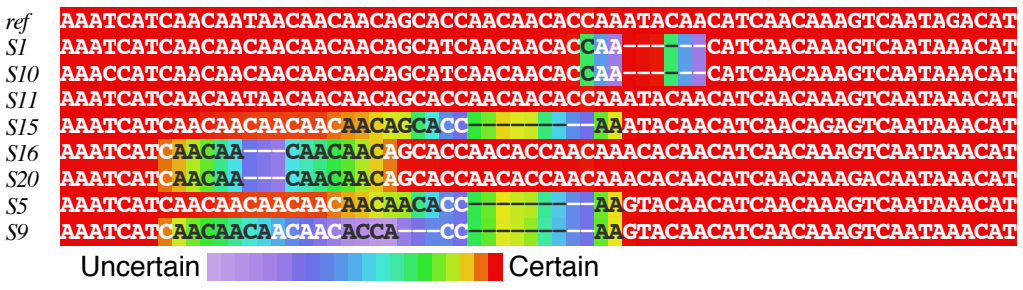


Sequential vs. Simultaneous Illustration

SIVmac251 partial *env* sequences from Cheynier et al (2001)

Displaying the posterior distribution of alignments

Alignment uncertainty (Au) plot:

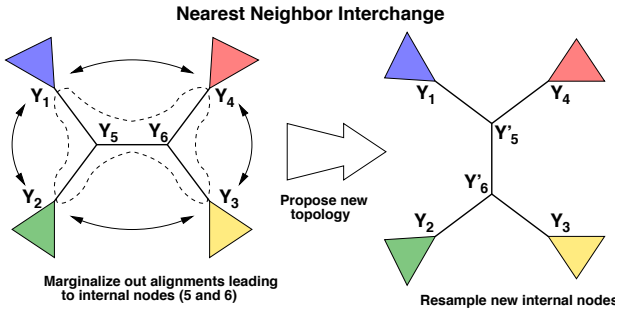


- Take “MAP” alignment as template
- Guild each residue with color reflecting probability of aligning to “root"

Trees and Alignments: Collapsed Gibbs Sampling

Problem: Tree and alignment are **highly** correlated

Further important aspect: Alignment-aware tree τ sampling



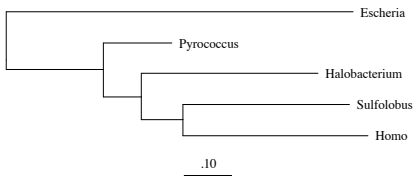
Generate $(\tau, \mathbf{A}) \mid \mathbf{Y}, \boldsymbol{\theta}$ by $1.\ \tau \mid \mathbf{Y}, \boldsymbol{\theta}$ (collapsed)

$2.\ \mathbf{A} \mid \tau, \mathbf{Y}, \boldsymbol{\theta}$

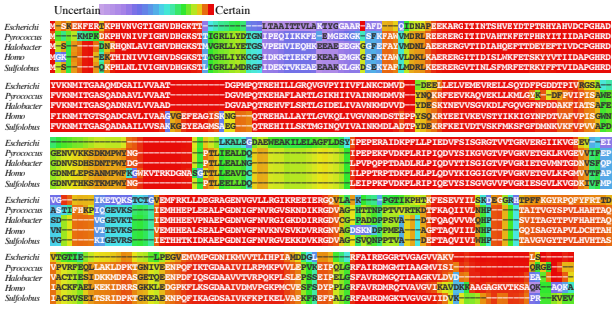
Similar procedure available for **global changes** (SPR moves)

EF-1α/Tu Strongly Supports Eocyte Hypothesis

(*Homo*,*Sulfolobus*) clade supported at $\geq 99.9\%$ (sampling resolution):



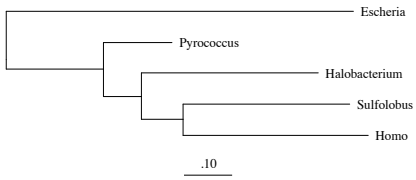
Alignment Uncertainty (Au, "gold") plot:



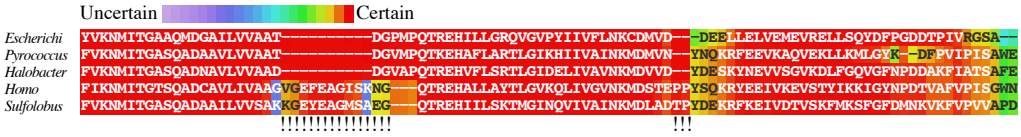
- Regions of marked homology (conservation)
- Uncertain regions
- Shared indels

EF-1α/Tu Strongly Supports Eocyte Hypothesis

(*Homo*,*Sulfolobus*) clade supported at $\geq 99.9\%$ (sampling resolution):

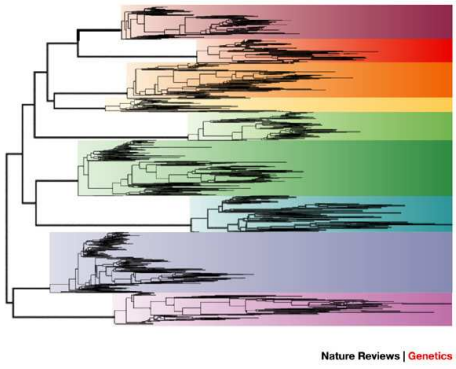


Alignment Uncertainty (Au, “gold”) plot:



- Automatic detection of indels shared by descent vs. by state
- Two indels shared by *Homo* and *Sulfolobus* contribute support for Eocyte Hypothesis

Future Directions: Intra-Host Viral Evolution



1195 *env* sequences from 9 HIV+ patients [taken from Rambaut et al. (2004)]

Retroviruses (and HBV) exist as a **quasi-species** within infected patients:

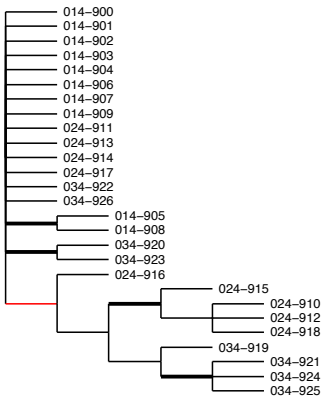
- Shared substitutions may be insufficient to resolve intra-host phylogenies

Improve resolution using joint model:

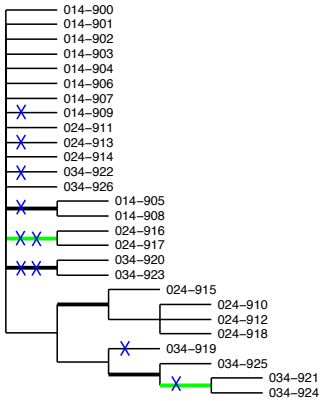
- Indel rates $\geq$ substitution rates
- Opportunity to detect intra-host recombination

Improved Resolution – I Can See!

Shankarrapa (1999) Pt #1: 3 time-points, **90% consensus** trees



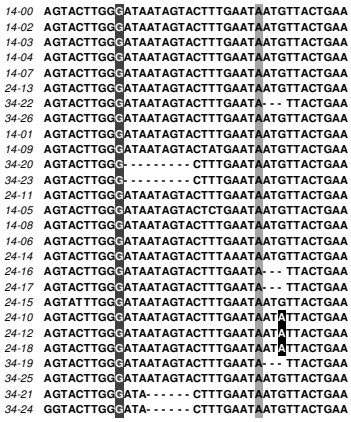
- Indel



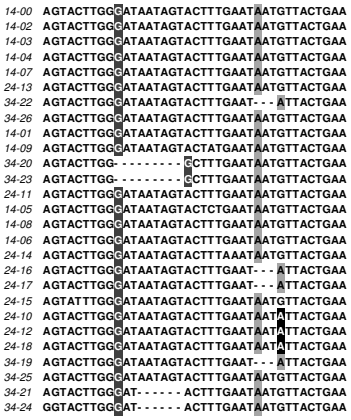
+ Indel

- Indel events ($\times$) $\Rightarrow$ **2** additional bi-partitions supported.

Codon Models: Not Always A Good Thing



HKY$\times$1 (-1556)



HKY$\times$3 (-1580)

- Codon model M0: $\omega = 1.0\,(0.9, 1, 2) \approx$ HKY$\times$3
- **Singlet** model more likely. **Triplet** model **shifts** indels and **misaligns** residues