

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

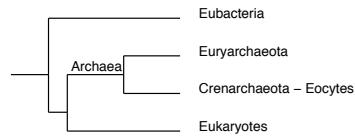
UCLA School of Public Health

SISMID – p.1

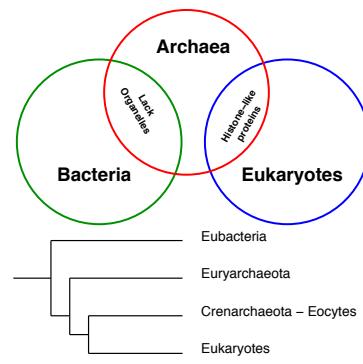
Resolving Early Branches in the Tree of Life

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

SISMID – p.2

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										

```

graph LR
    Human --- Node(( ))
    Node --- Chimp
    Node --- Mouse
    Node --- Fly
    style Node fill:none,stroke:none
    
```

- **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)

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

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

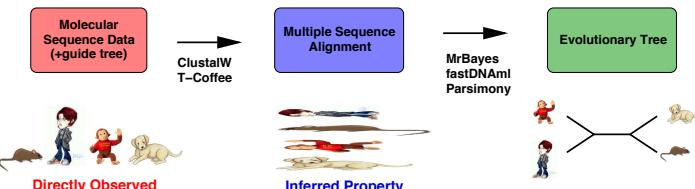
$$\mathbb{P}(X \rightarrow Y \text{ in time } t) = \left\{ e^{t\mathbf{Q}} \right\}_{XY}$$

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

SISMID – p.3

Fundamental Difficulty: Sequential Estimation

Current phylogenetic reconstruction methods:



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

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

Solution: Infer alignment and phylogeny **simultaneously**

Previous approaches: Limited

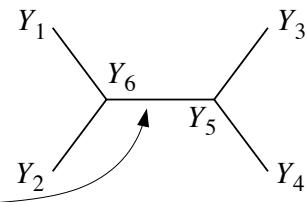
- Optimization alignment, parsimony-based
- TFK91/92, forbidden positional homologies, inefficient

SISMID – p.4

Alignment as a Random Variable

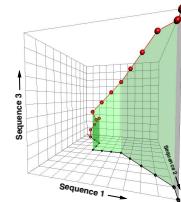
	Y
Observed Data	$Y_1 = (A, T, T, C)$
	$Y_2 = (A, T, T, G)$
	$Y_3 = (T, C, T, G)$
	$Y_4 = (T, C, T)$
Missing Data	$Y_5 = (\star, \star, \star)$
	$Y_6 = (\star, \star, \star, \star)$

M(A)	f
1 2 - 3 4	A T - T C
1 2 - 3 4	A T - T G
- 1 2 3 4	- T C T G
- 1 2 3 -	- T C T -
- 1 2 3 -	- * * * -
1 2 - 3 4	* * - * *



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

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



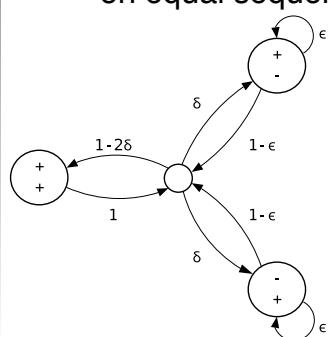
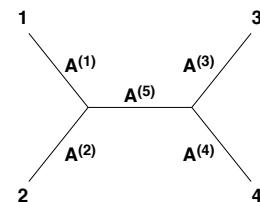
Note substitution process depends only on $\mathbf{Y}_{\text{obs}} \Rightarrow$ separates substitution and indel processes into (substitution likelihood \times gap prior)

SISMID – p.5

Gap Model along a Branch

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

- \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 $\Lambda = (\delta, \epsilon)$

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

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

SISMID – p.6

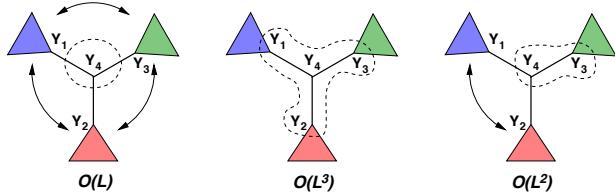
Choosing the Blocks: Efficient Sampling

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

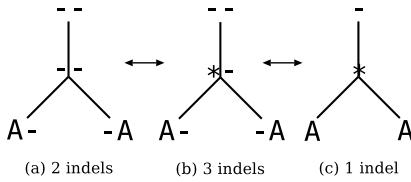
$$A^{(1)}, A^{(2)}, (A^{(3)}, A^{(4)}), (A^{(5)}, A^{(6)}, A^{(7)}), \dots$$

Let $L = \text{length of alignment}$:

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



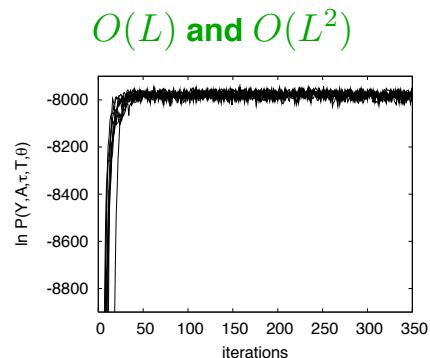
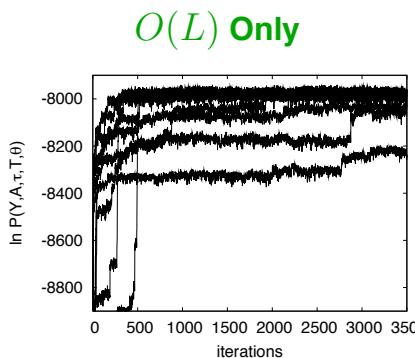
Possible **poor mixing** with $O(L)$ algorithm:



Must move through unfavorable intermediate to align/unalign sequence fragments

SISMID – p.7

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



Enabling MCMC transition kernel decreases:

- Burn-in time
- Autocorrelation

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

SISMID – p.8

Sequential vs. Simultaneous Illustration

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

Sequential (ClustalW) alignment of hypervariable region:

	***	*****	****		**	*****	***	****	*****	***	****	*****
ref	A	A	T	C	A	C	A	A	T	A	C	A
S1	A	A	T	C	A	C	A	A	C	A	C	A
S10	A	A	C	C	A	T	C	A	A	C	A	A
S11	A	A	T	C	A	C	A	A	C	A	C	A
S15	A	A	T	C	A	C	A	A	C	A	C	A
S16	A	A	T	C	A	C	A	A	C	A	C	A
S20	A	A	T	C	A	C	A	A	C	A	C	A
S5	A	A	T	C	A	C	A	A	C	A	C	A
S9	A	A	T	C	A	C	A	A	C	A	C	A

Simultaneous MAP alignment:

	***	*****	****	*	**	*****	***	****	*****	***	****	*****
ref	A	A	T	C	A	C	A	A	T	A	C	A
S1	A	A	T	C	A	C	A	A	C	A	C	A
S10	A	A	C	C	A	T	C	A	A	C	A	A
S11	A	A	T	C	A	C	A	A	C	A	C	A
S15	A	A	T	C	A	C	A	A	C	A	C	A
S16	A	A	T	C	A	C	A	A	C	A	C	A
S20	A	A	T	C	A	C	A	A	C	A	C	A
S5	A	A	T	C	A	C	A	A	C	A	C	A
S9	A	A	T	C	A	C	A	A	C	A	C	A

SISMID – p.9

Sequential vs. Simultaneous Illustration

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

Sequential (ClustalW) alignment of hypervariable region:

	***	*****	****		**	*****	***	****	*****	***	****	*****
ref	A	A	T	C	A	C	A	A	T	A	C	A
S1	A	A	T	C	A	C	A	A	C	A	C	A
S10	A	A	C	C	A	T	C	A	A	C	A	A
S11	A	A	T	C	A	C	A	A	C	A	C	A
S15	A	A	T	C	A	C	A	A	C	A	C	A
S16	A	A	T	C	A	C	A	A	C	A	C	A
S20	A	A	T	C	A	C	A	A	C	A	C	A
S5	A	A	T	C	A	C	A	A	C	A	C	A
S9	A	A	T	C	A	C	A	A	C	A	C	A

Sampled alignment (1):

	***	*****	****		**	*****	***	****	*****	***	****	*****
ref	A	A	T	C	A	C	A	A	T	A	C	A
S1	A	A	T	C	A	C	A	A	C	A	C	A
S10	A	A	C	C	A	T	C	A	A	C	A	A
S11	A	A	T	C	A	C	A	A	C	A	C	A
S15	A	A	T	C	A	C	A	A	C	A	C	A
S16	A	A	T	C	A	C	A	A	C	A	C	A
S20	A	A	T	C	A	C	A	A	C	A	C	A
S5	A	A	T	C	A	C	A	A	C	A	C	A
S9	A	A	T	C	A	C	A	A	C	A	C	A

SISMID – p.9

Sequential vs. Simultaneous Illustration

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

Sequential (ClustalW) alignment of hypervariable region:

	***	*****	****		**	*****	*****	*	*****	*****	*****
ref	A	A	T	C	A	C	A	A	T	C	A
S1	A	A	T	C	A	C	A	A	T	C	A
S10	A	A	C	C	A	T	C	A	A	T	C
S11	A	A	T	C	A	C	A	A	T	C	A
S15	A	A	T	C	A	C	A	A	T	C	A
S16	A	A	T	C	A	C	A	A	T	C	A
S20	A	A	T	C	A	C	A	A	T	C	A
S5	A	A	T	C	A	C	A	A	T	C	A
S9	A	A	T	C	A	C	A	A	T	C	A

Sampled alignment (2):

	***	****	**	*****	****	*	*	*****	*****	*	*****	*****
ref	A	A	T	C	A	C	A	A	T	C	A	A
S1	A	A	T	C	A	C	A	A	T	C	A	A
S10	A	A	C	C	A	T	C	A	A	T	C	A
S11	A	A	T	C	A	C	A	A	T	C	A	A
S15	A	A	T	C	A	C	A	A	T	C	A	A
S16	A	A	T	C	A	C	A	A	T	C	A	A
S20	A	A	T	C	A	C	A	A	T	C	A	A
S5	A	A	T	C	A	C	A	A	T	C	A	A
S9	A	A	T	C	A	C	A	A	T	C	A	A

SISMID – p.9

Sequential vs. Simultaneous Illustration

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

Sequential (ClustalW) alignment of hypervariable region:

	***	*****	*****		**	*****	*****	*	*****	*****	*****
ref	A	A	T	C	A	C	A	A	T	C	A
S1	A	A	T	C	A	C	A	A	T	C	A
S10	A	A	C	C	A	T	C	A	A	T	C
S11	A	A	T	C	A	C	A	A	T	C	A
S15	A	A	T	C	A	C	A	A	T	C	A
S16	A	A	T	C	A	C	A	A	T	C	A
S20	A	A	T	C	A	C	A	A	T	C	A
S5	A	A	T	C	A	C	A	A	T	C	A
S9	A	A	T	C	A	C	A	A	T	C	A

Sampled alignment (3)

	***	***	*	*****	*****	*	*****	*****	*	*****	*****
ref	A	A	T	C	A	C	A	A	T	C	A
S1	A	A	T	C	A	C	A	A	T	C	A
S10	A	A	C	C	A	T	C	A	A	T	C
S11	A	A	T	C	A	C	A	A	T	C	A
S15	A	A	T	C	A	C	A	A	T	C	A
S16	A	A	T	C	A	C	A	A	T	C	A
S20	A	A	T	C	A	C	A	A	T	C	A
S5	A	A	T	C	A	C	A	A	T	C	A
S9	A	A	T	C	A	C	A	A	T	C	A

SISMID – p.9

Sequential vs. Simultaneous Illustration

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

Sequential (ClustalW) alignment of hypervariable region:

	***	*****	****		**	*****	*****	*	*****	*****
ref	A	A	T	C	A	C	A	A	G	C
S1	A	A	T	C	A	C	A	A	C	G
S10	A	A	C	C	A	A	C	A	C	A
S11	A	A	T	C	A	C	A	A	C	G
S15	A	A	T	C	A	C	A	A	C	G
S16	A	A	T	C	A	C	A	A	C	G
S20	A	A	T	C	A	C	A	A	C	G
S5	A	A	T	C	A	C	A	A	C	G
S9	A	A	T	C	A	C	A	A	C	G

Sampled alignment (4)

	***	*****	****		**	***	*****	*	*****	*****
ref	A	A	T	C	A	C	A	A	G	C
S1	A	A	T	C	A	C	A	A	C	G
S10	A	A	C	C	A	A	C	A	C	A
S11	A	A	T	C	A	C	A	A	C	G
S15	A	A	T	C	A	C	A	A	C	G
S16	A	A	T	C	A	C	A	A	C	G
S20	A	A	T	C	A	C	A	A	C	G
S5	A	A	T	C	A	C	A	A	C	G
S9	A	A	T	C	A	C	A	A	C	G

SISMID – p.9

Sequential vs. Simultaneous Illustration

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

Sequential (ClustalW) alignment of hypervariable region:

	***	*****	****		**	*****	*****	*	*****	*****
ref	A	A	T	C	A	C	A	A	G	C
S1	A	A	T	C	A	C	A	A	C	G
S10	A	A	C	C	A	A	C	A	C	A
S11	A	A	T	C	A	C	A	A	C	G
S15	A	A	T	C	A	C	A	A	C	G
S16	A	A	T	C	A	C	A	A	C	G
S20	A	A	T	C	A	C	A	A	C	G
S5	A	A	T	C	A	C	A	A	C	G
S9	A	A	T	C	A	C	A	A	C	G

Sampled alignment (5)

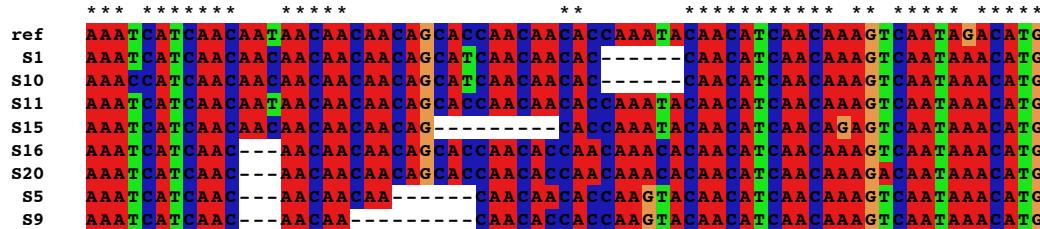
	***	*****	*	****		**	*****	*	*****	*****
ref	A	A	T	C	A	C	A	A	G	C
S1	A	A	T	C	A	C	A	A	C	G
S10	A	A	C	C	A	A	C	A	C	A
S11	A	A	T	C	A	C	A	A	C	G
S15	A	A	T	C	A	C	A	A	C	G
S16	A	A	T	C	A	C	A	A	C	G
S20	A	A	T	C	A	C	A	A	C	G
S5	A	A	T	C	A	C	A	A	C	G
S9	A	A	T	C	A	C	A	A	C	G

SISMID – p.9

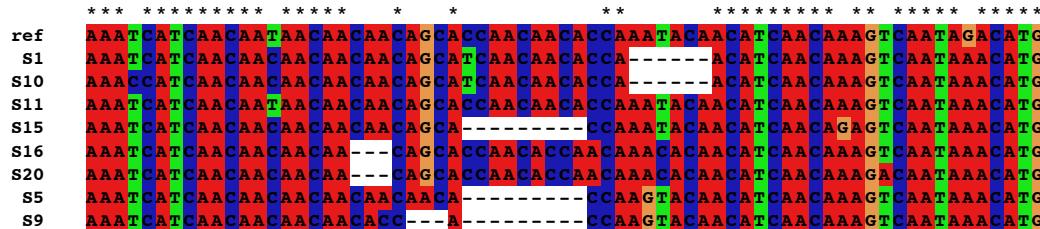
Sequential vs. Simultaneous Illustration

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

Sequential (ClustalW) alignment of hypervariable region:



Sampled alignment (6)



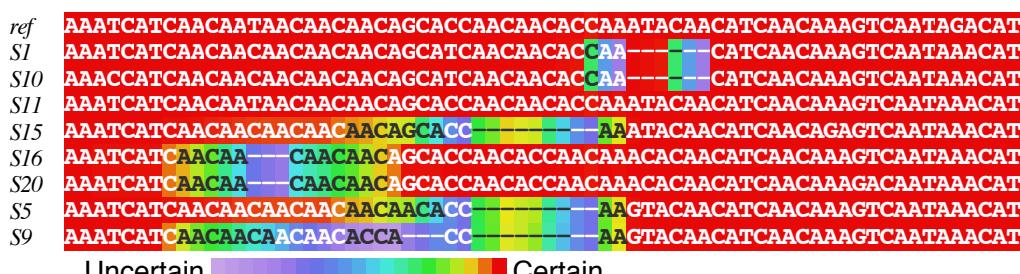
SISMID – p.9

Sequential vs. Simultaneous Illustration

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

Displaying the posterior distribution of alignments

Alignment uncertainty (Au) plot:



Uncertain Certain

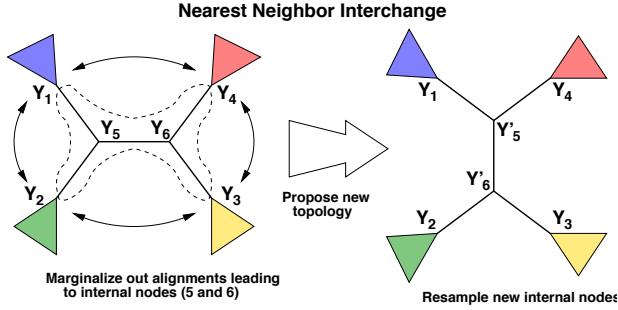
- Take “MAP” alignment as template
- Guild each residue with color reflecting probability of aligning to “root”

SISMID – p.9

Trees and Alignments: Collapsed Gibbs Sampling

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

Further important aspect: Alignment-aware tree τ sampling



Generate $(\tau, \mathbf{A}) | \mathbf{Y}, \theta$ by

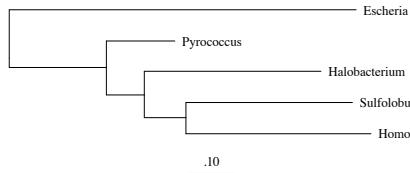
- $\tau | \mathbf{Y}, \theta$ (collapsed)
- $\mathbf{A} | \tau, \mathbf{Y}, \theta$

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

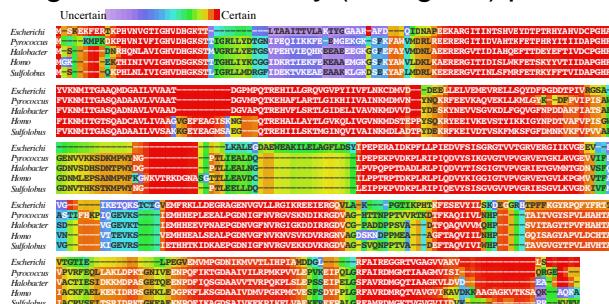
SISMID – p.10

EF-1 α /Tu Strongly Supports Eocyte Hypothesis

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



Alignment Uncertainty (Au, “gold”) plot:

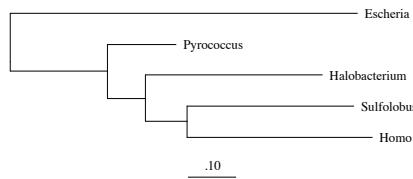


- Regions of marked homology (conservation)
- Uncertain regions
- Shared indels

SISMID – p.11

EF-1 α /Tu Strongly Supports Eocyte Hypothesis

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



MIND THE GAP

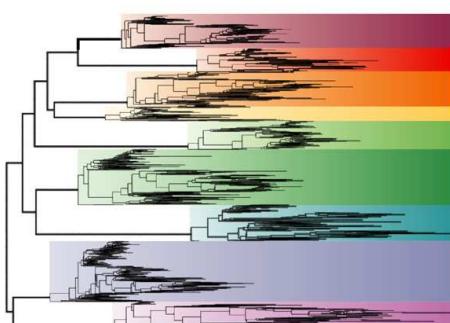
Alignment Uncertainty (Au, “gold”) plot:



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

SISMID – p.11

Future Directions: Intra-Host Viral Evolution



Nature Reviews | Genetics

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

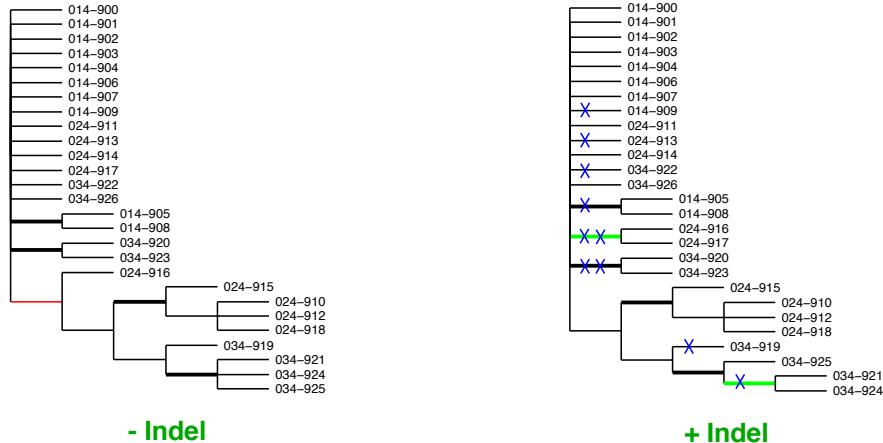
Improve resolution using joint model:

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

SISMID – p.12

Improved Resolution – I Can See!

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



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

SISMID – p.13

Codon Models: Not Always A Good Thing

<pre> 14-00 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-02 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-03 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-04 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-07 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 24-13 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 34-22 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 34-26 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-09 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 34-20 AGTACTTGGG----- CTTTGAATTAATGTTACTGAA 34-23 AGTACTTGGG----- CTTTGAATTAATGTTACTGAA 24-11 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-05 AGTACTTGGGATAATAGTACCTCTGAATTAATGTTACTGAA 14-08 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-01 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-09 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 34-20 AGTACTTGGG----- CTTTGAATTAATGTTACTGAA 34-23 AGTACTTGGG----- CTTTGAATTAATGTTACTGAA 24-11 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-05 AGTACTTGGGATAATAGTACCTCTGAATTAATGTTACTGAA 14-08 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-06 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 24-14 AGTACTTGGGATAATAGTACCTTAAATTAATGTTACTGAA 24-16 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 24-17 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 24-15 AGTATTGGGATAATAGTACCTTGAATTAATGTTACTGAA 24-10 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 24-12 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 24-18 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 34-19 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 34-25 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 34-21 AGTACTTGGGATA----- CTTTGAATTAATGTTACTGAA 34-24 GGTACTTGGGATA----- CTTTGAATTAATGTTACTGAA </pre>	<pre> 14-00 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-02 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-03 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-04 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-07 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 24-13 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 34-22 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 34-26 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-09 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 34-20 AGTACTTGGG----- CTTTGAATTAATGTTACTGAA 34-23 AGTACTTGGG----- CTTTGAATTAATGTTACTGAA 24-11 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-05 AGTACTTGGGATAATAGTACCTCTGAATTAATGTTACTGAA 14-08 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-01 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-09 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 34-20 AGTACTTGGG----- CTTTGAATTAATGTTACTGAA 34-23 AGTACTTGGG----- CTTTGAATTAATGTTACTGAA 24-11 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-05 AGTACTTGGGATAATAGTACCTCTGAATTAATGTTACTGAA 14-08 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 14-06 AGTACTTGGGATAATAGTACCTTAAATTAATGTTACTGAA 24-14 AGTACTTGGGATAATAGTACCTTAAATTAATGTTACTGAA 24-16 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 24-17 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 24-15 AGTATTGGGATAATAGTACCTTGAATTAATGTTACTGAA 24-10 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 24-12 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 24-18 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 34-19 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 34-25 AGTACTTGGGATAATAGTACCTTGAATTAATGTTACTGAA 34-21 AGTACTTGGGATA----- ACTTGAATTAATGTTACTGAA 34-24 GGTACTTGGGATA----- ACTTGAATTAATGTTACTGAA </pre>
--	--

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

SISMID – p.14