Learning to Count: Tests for Evolutionary Innovation and Robust Sequence Distance Estimation

Philippe Lemey and Marc A. Suchard

Rega Institute

Department of Microbiology and Immunology

K.U. Leuven, Belgium, and

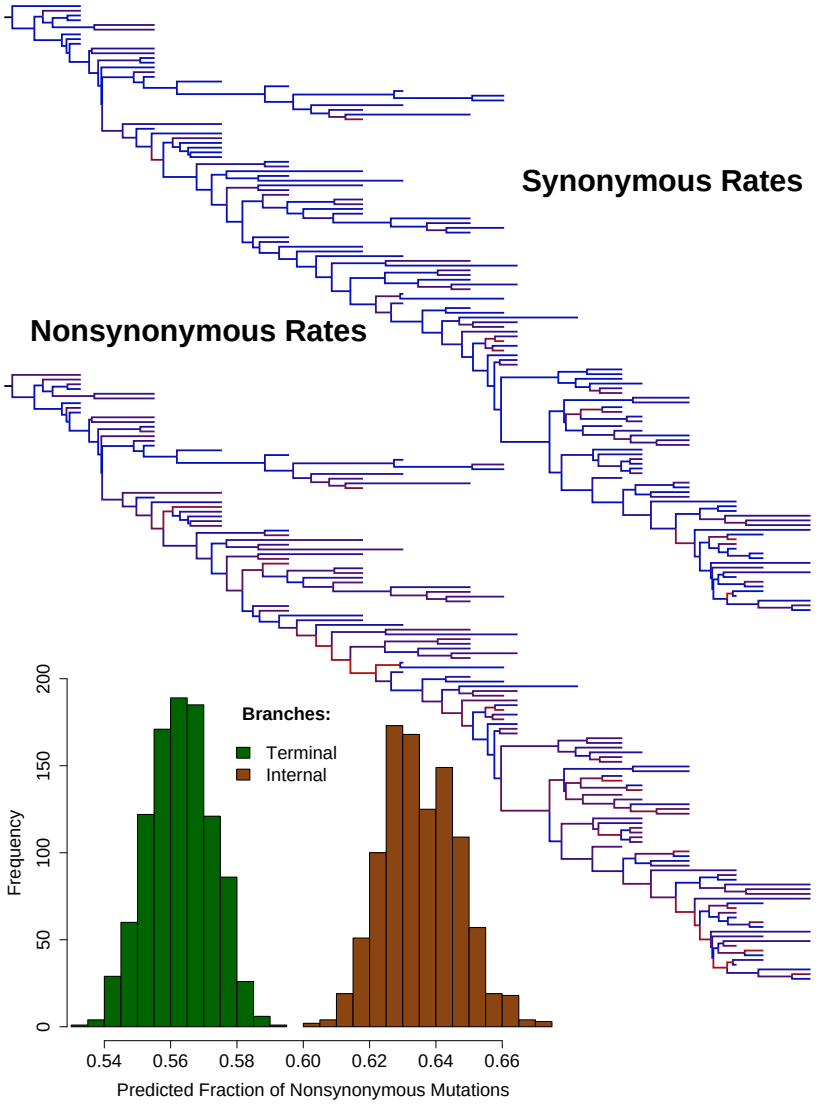
Departments of Biomathematics and Human Genetics

David Geffen School of Medicine at UCLA

Department of Biostatistics

UCLA School of Public Health

Classic Problem: Detecting Adaptation in Intrahost HIV Evolution



- **Data**: 129 HIV variants from one patient
- **Question**: Does adaptation occur along the backbone of evolution? (Suggests violations of neutrality)
- **Difficulty**: Branch/time-specific synonymous/non-synonymous rate models are too unwieldy
- **Solution?**: Count the expected # of labeled transitions

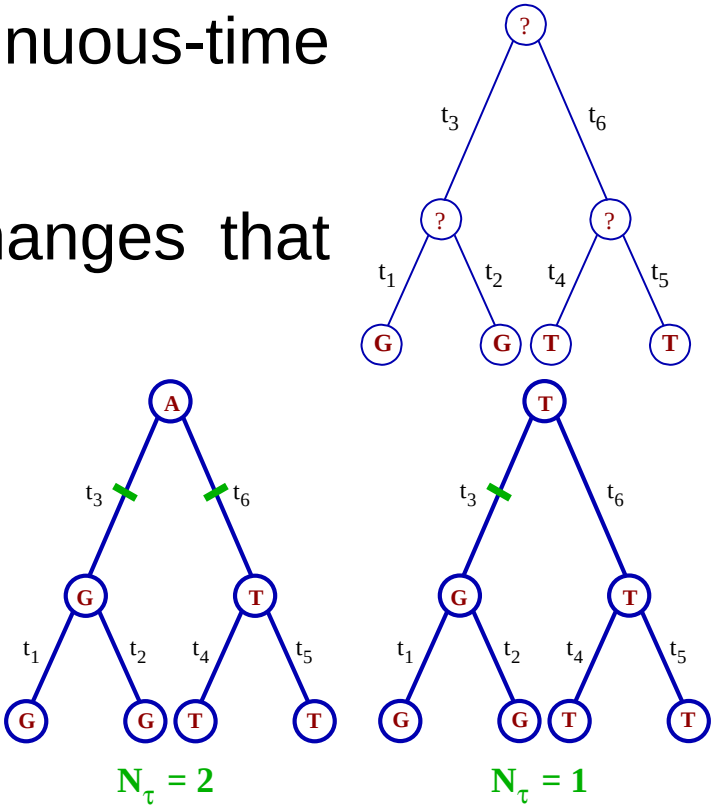
Evolutionary Counting Processes: Current Approaches

Model trait evolution as a continuous-time Markov chain $\mathbf{\Lambda}$, and

Infer the number N_τ of state-changes that occur along the tree τ via

Stochastic Mapping (Nielsen, 2002):

- Simulation-based
- Uses **rejection sampling**



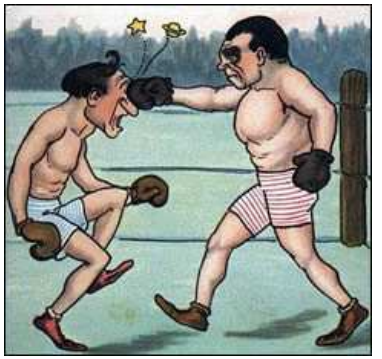
Can do one better? Analytic solutions for $\Pr(N_\tau = n|\mathbf{Y})$ and $\mathrm{E}(N_\tau^k|\mathbf{Y})$ enable (computationally) efficient statistical tests.

Punch-Line: Simulation Methods . . . Could Be Better

Computational efficiency / **accuracy** comparison:

Simulants	Slow Evolving Site $N_\tau \approx 4$		Fast Evolving Site $N_\tau \approx 15$	
	Rejections/Simulant	Error	Rejections/Simulant	Error
100	100	0.0598	38845	0.4624
500	105	0.0255	39247	0.3319
1000	102	0.0259	42075	0.2905
10000	106	0.0205	40805	0.2809

- 61-state Markov chain (codon model) on 129-tip tree (HIV evolution)
- Counting labeled subsets of changes (synonymous/non-synonymous) on "internal" vs. "external" branches

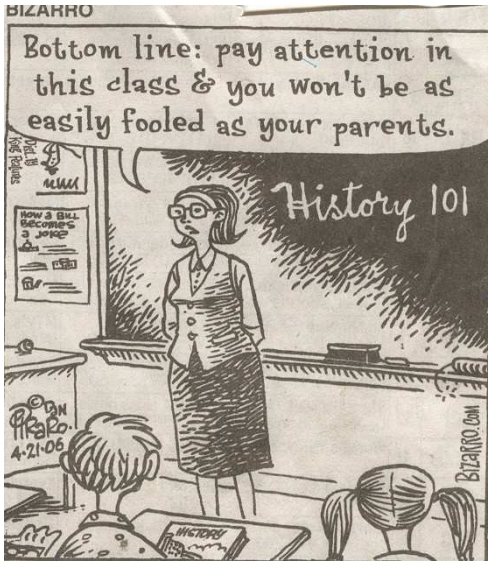


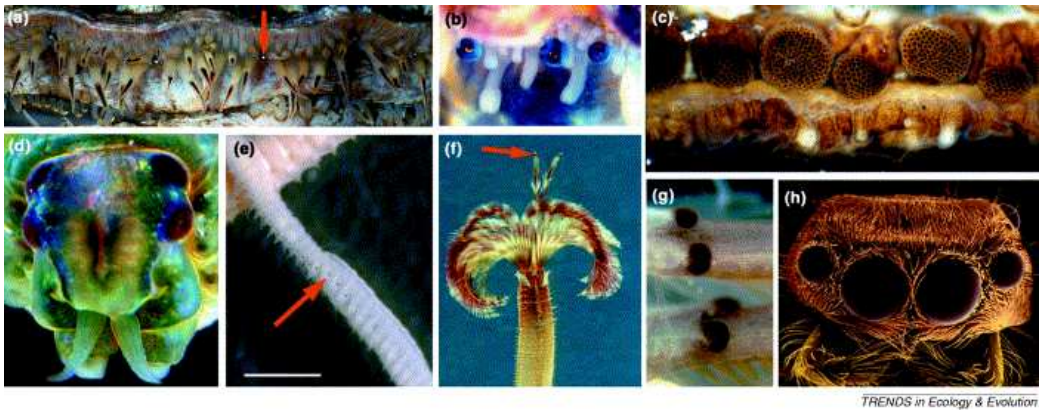
1 min vs. 10 hrs

Those Who Forget Their History . . .

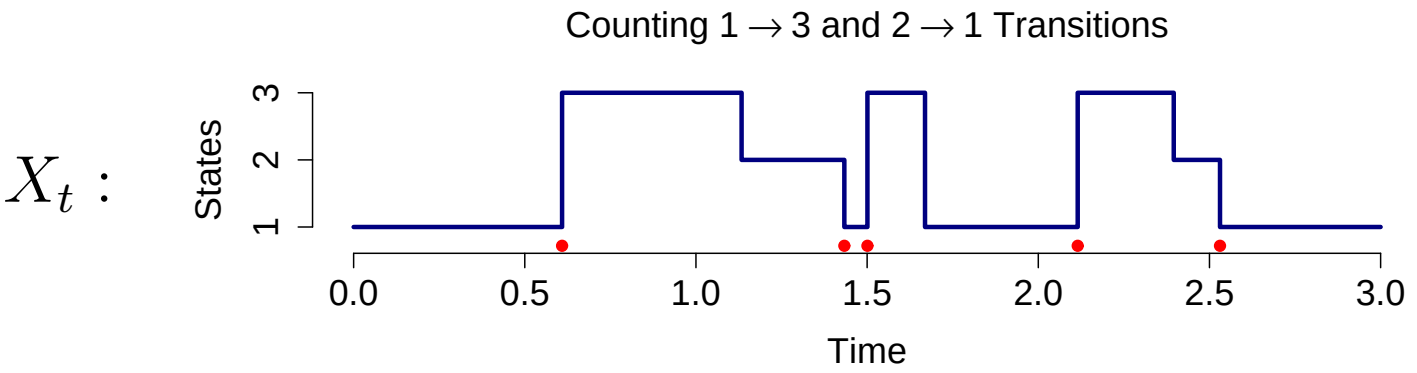
Three Major Approaches:

- Examine process at stationarity/no conditioning on start/end points: Ball, Neuts (ion channel physics)
- Label only one specific transition: Guttorp, Bruno, Hobolth
- Via uniformization: Siepel





General Framework: Labeled Changes



- R is a set of ordered index pairs that **label transitions**
- $\mathbf{\Lambda}$ - generator, $\mathbf{\Lambda}_R = \{\lambda_{ij} \times 1_{\{(i,j)\in R\}}\}$, and $\mathbf{\Lambda}_{\overline{R}} = \mathbf{\Lambda} - \mathbf{\Lambda}_R$

- Matrix $\mathbf{Q}(n,t)$ of probabilities $(N_t = n, X_t = j \mid X_0 = i)$
- Matrix $\mathbf{M}^{[k]}(t)$ of restricted, factorial moments $(N_t^{[k]} 1_{\{X_t=j\}} \mid X_0 = i)$

Derivation Sketch: A Moment's Reflection on One Branch

Start with **Kolmogorov's Forward** equation:

$$\frac{d}{dt}\mathbf{Q}(n,t) = \mathbf{Q}(n,t)\mathbf{\Lambda}_{\overline{R}} + \mathbf{Q}(n-1,t)\mathbf{\Lambda}_R$$

and the **matrix probability generating function**:

$$\mathbf{G}(r,t) = \sum_{n=0}^{\infty} r^n \mathbf{Q}(n,t)$$

Then $\frac{\partial}{\partial t}\mathbf{G}(r,t) = \mathbf{G}(r,t)\left(\mathbf{\Lambda}_{\overline{R}} + r\mathbf{\Lambda}_R\right) \Rightarrow \mathbf{G(r,t) = e^{(\Lambda_{\overline{R}} + r\Lambda_R)t}}$ and $\mathbf{M}^{[k]}(t) = \frac{\partial^k}{\partial r^k}\mathbf{G}(r,t)|_{r=1}$ - hard unless $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_R$ commute!

Use **integration** instead $\mathbf{M}^{[k]}(t) = k\int_0^t \mathbf{M}^{[k-1]}(t)\mathbf{\Lambda}_R e^{\mathbf{\Lambda}(t-\theta)}d\theta$

$$\mathbf{M}^{[1]}(t) = \sum_{i,j}\mathbf{B}_i\mathbf{\Lambda}_R\mathbf{B}_j I_{ij}(t), \quad I_{ij}(t) = \begin{cases} te^{d_i t} & \text{if } d_i = d_j, \\ \dfrac{e^{d_i t} - e^{d_j t}}{d_i - d_j} & \text{if } d_i \neq d_j. \end{cases}$$

Reaping the Rewards of Counting over Trees

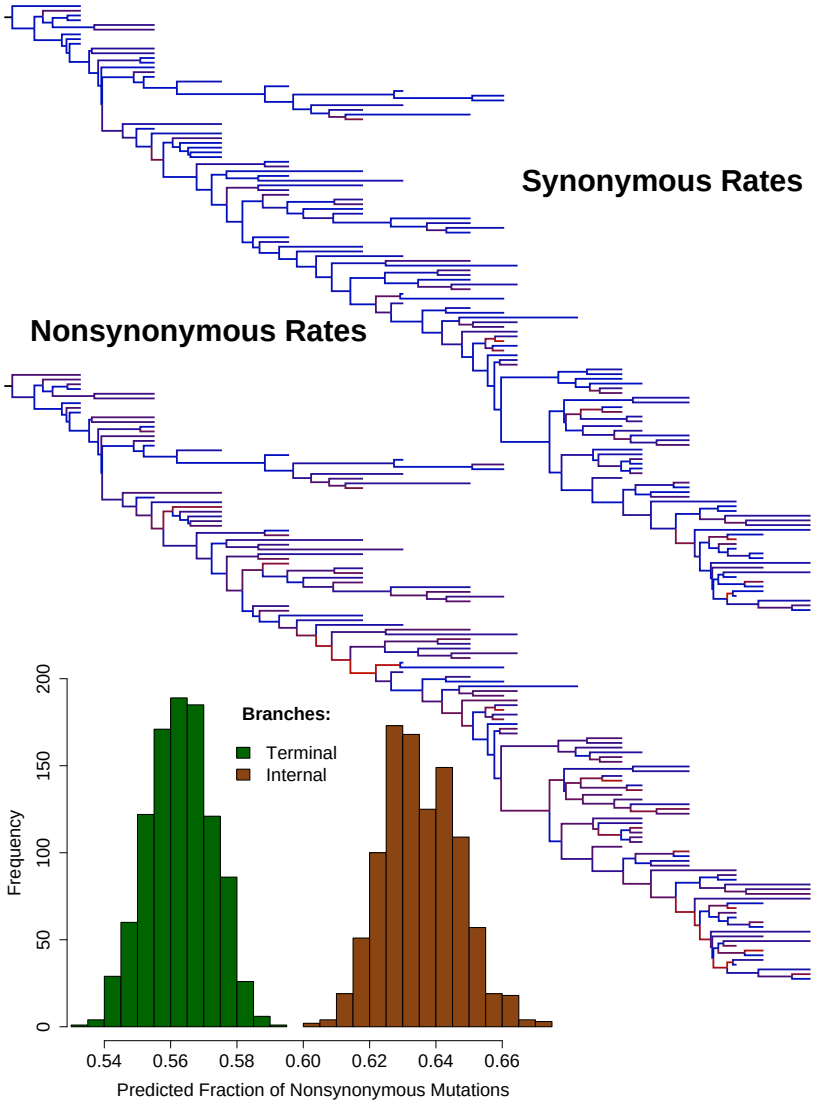
Let $\mathbf{H} = \sum_b h(X_t^{(b)})$ be an additive **summary**, where $h(\cdot)$ **counts/rewards** (on possibly select branches), e.g.,

- Transitions
- Final states
- Dwell times
- And others . . .

Then $\mathsf{E}(\mathbf{H})$ also has an analytic solution $\Rightarrow$ no simulation of internal-node states or conditioning on ancestral reconstruction.

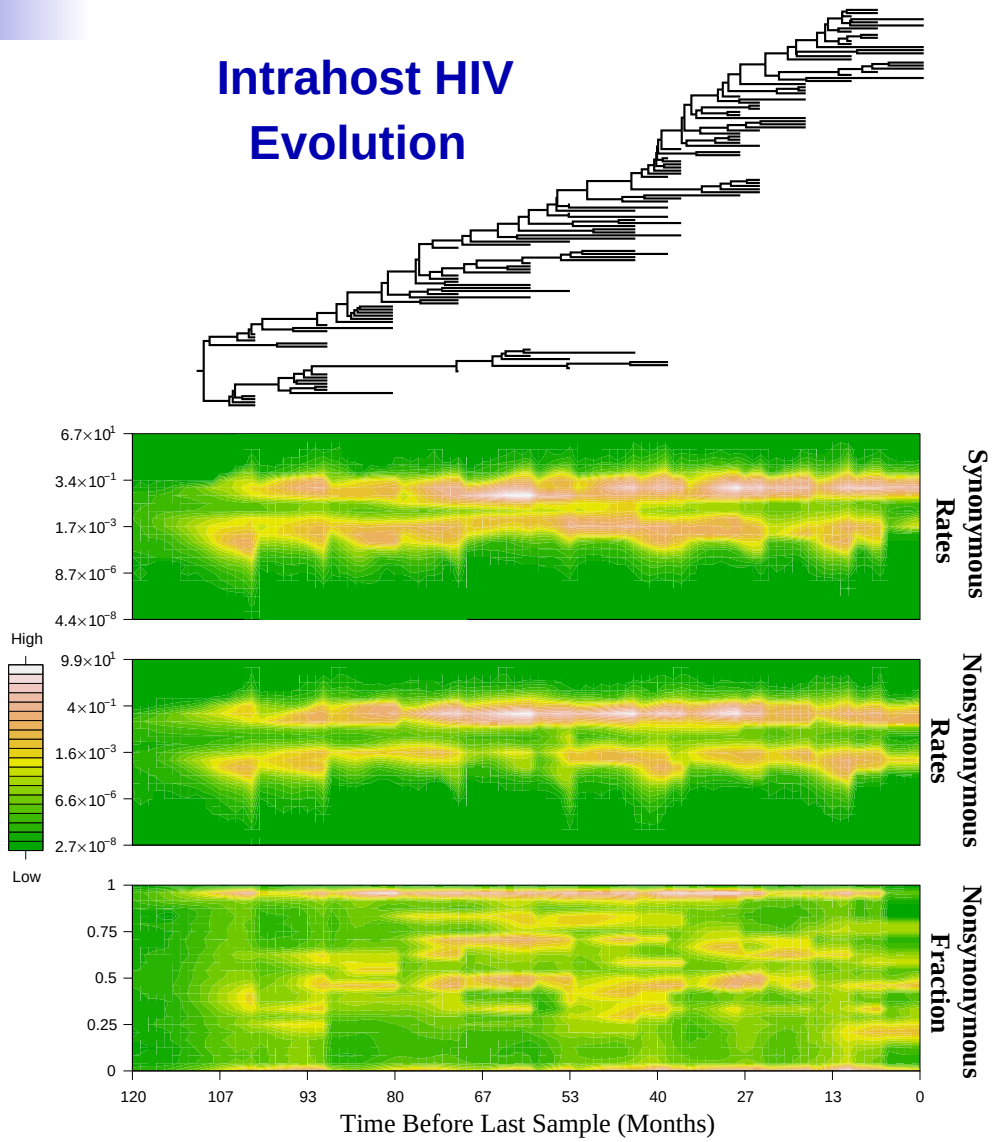
Noteworthy: Integrating $\mathbf{H}$ **generalizes** Felsenstein's Pruning Algorithm, the work-horse of modern phylogenetics.

Detecting Adaptation in Intrahost HIV Evolution



- **Data**: 129 HIV variants from one patient
- **Question**: Does adaptation occur along the backbone of evolution? (Suggests violations of neutrality)
- **Difficulty**: Branch-specific synonymous/non-synonymous rate change models are too unwieldy
- **Key:** Requires posterior simulation from only a **simple**, homogeneous rate model

Temporal Rate Variation in Intrahost HIV Evolution

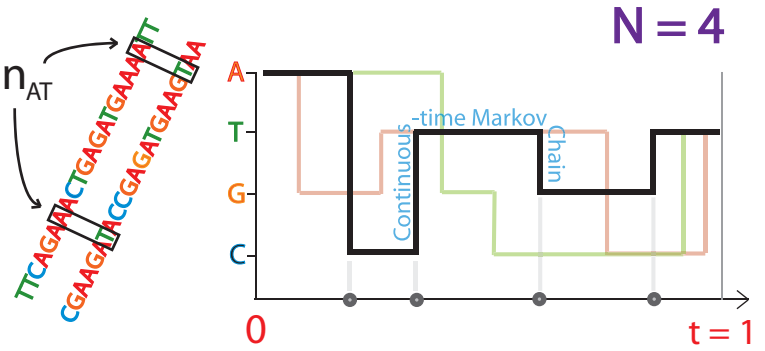


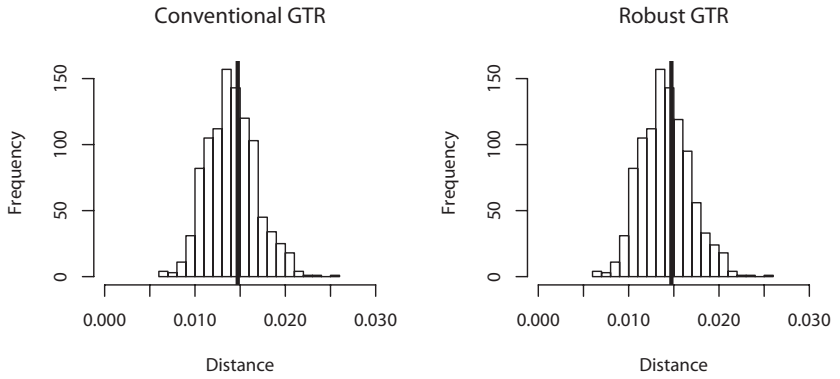
- Branch-specific counts enable rate projection onto **real time**
- **Bimodality** of both synonymous and non-synonymous distributions
- Early adaptation, followed by weakening selection

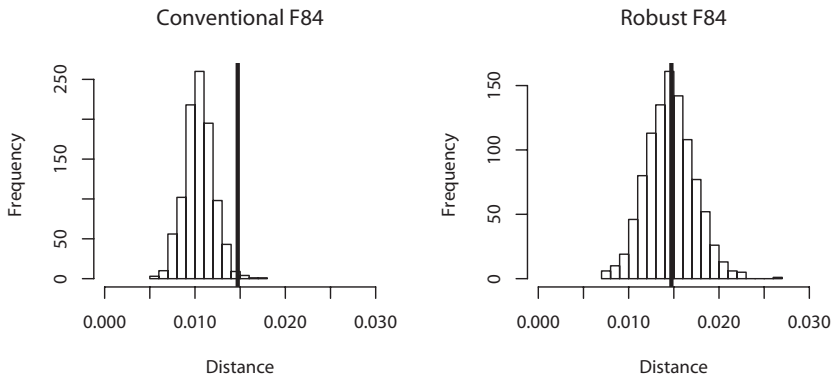
Pair-wise Robust Distance Estimation

Pairwise distance = E(N) w.r.t.

- **Stationary**$(\pi_i \times p_{ij})$ distribution, vs.
- **Empirical**(f_{ij}) distribution (**robust**) further straight-forward to **label**







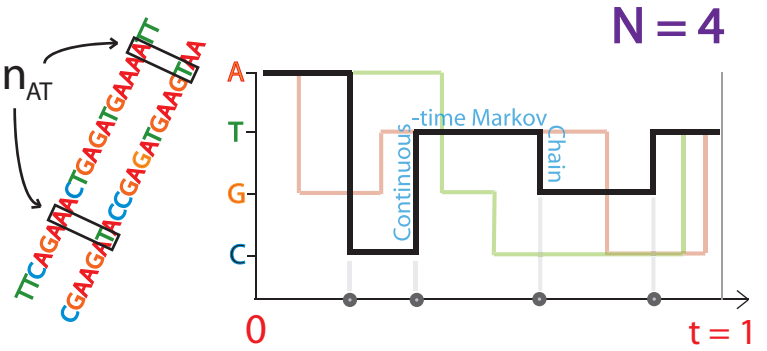
Nucleotide simulation:

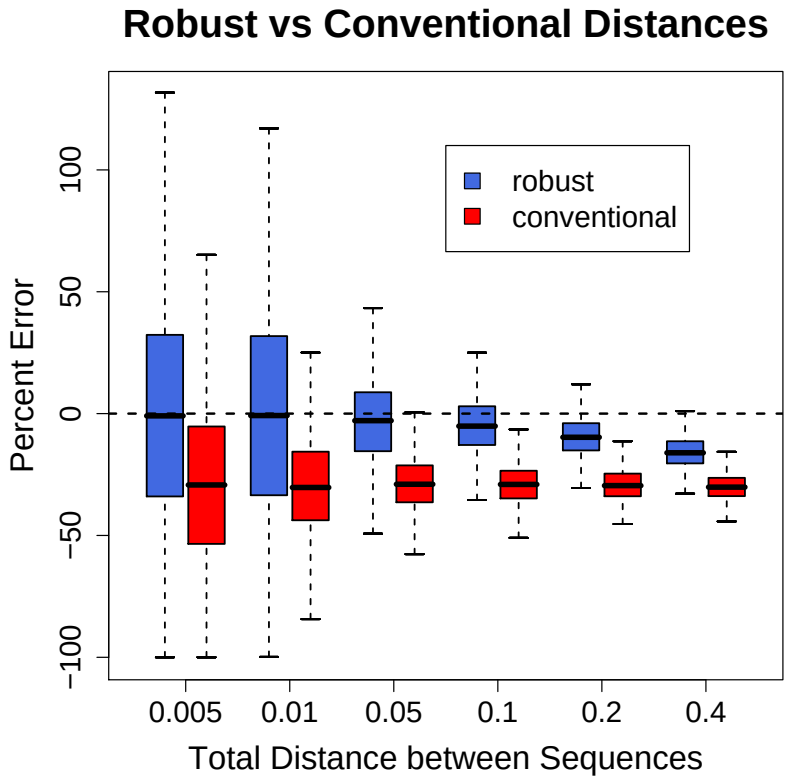
- True model = GTR
- Robust estimator using F84 (analytic calculations) performs as well as estimators under GTR

Pair-wise Robust Distance Estimation - II

Pairwise distance = E(N) w.r.t.

- **Stationary**$(\pi_i \times p_{ij})$ distribution, vs.
- **Empirical**(f_{ij}) distribution (**robust**) further straight-forward to **label**





Nucleotide simulation:

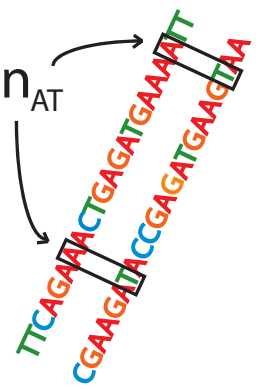
- Vary “true” sequence distance from 0.005 to 0.4
- Robust inference **decreases bias** caused by model misspecification

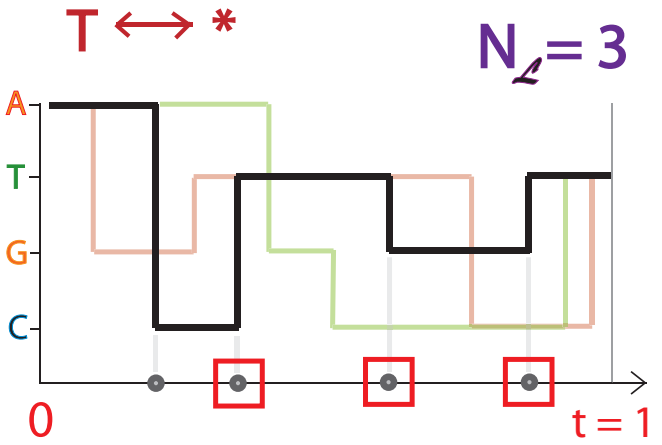
Robust Labeled Distance Estimation

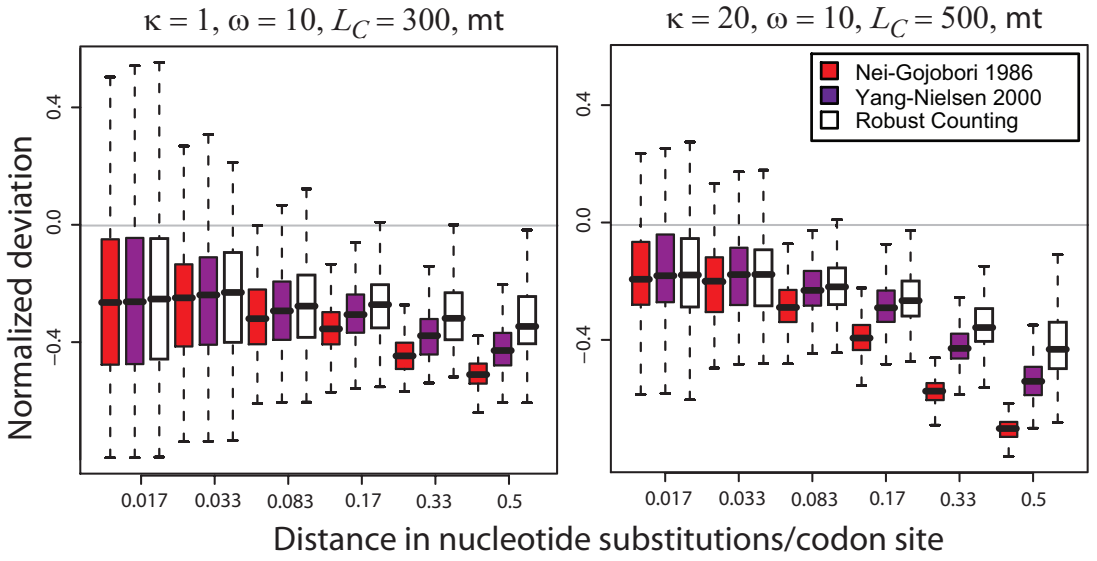
Labeled distance = $\mathrm{E}(N_{\mathcal{L}})$

Go **robust** with a silly codon model:

- Composition $3\times$F84s
- **No numerical optimization**







Synonymous distance estimation outperforms:

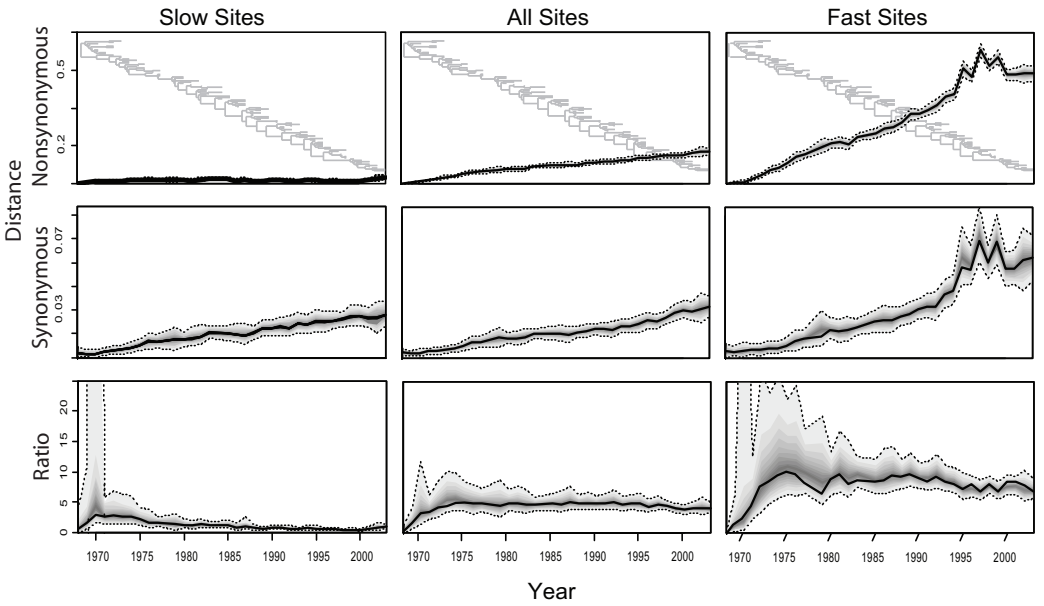
- Nei and Gojobori (1986)
- Yang and Nielsen (2000) - specially tailored estimator

"Neutral" reconstruction possible?

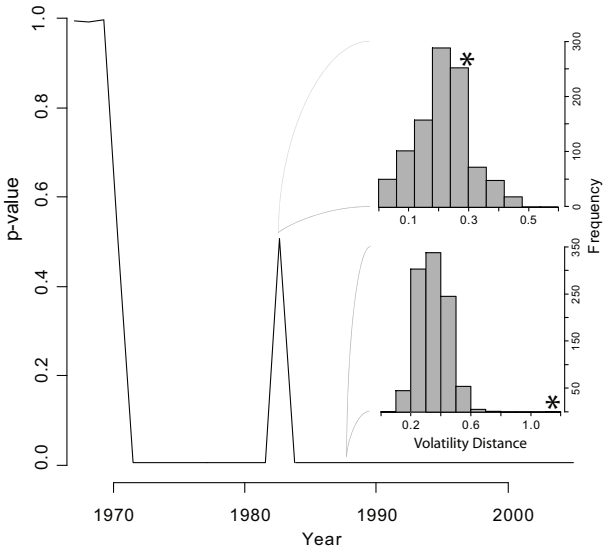
Arbitrarily Sophisticated Distances: Codon Volatility Change in Influenza A H3N2

Hypothesis: **Codon volatility** correlates with selective pressures (Plotkin and Dushoff, 2003)

S/N distances for 96 HA sequences:



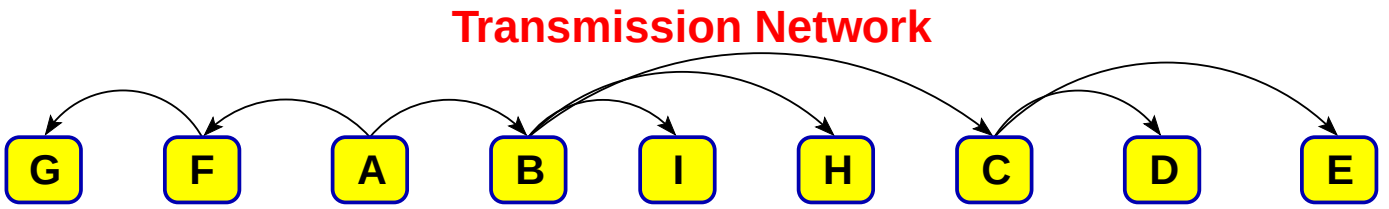
Distribution (epitope vs. elsewhere):



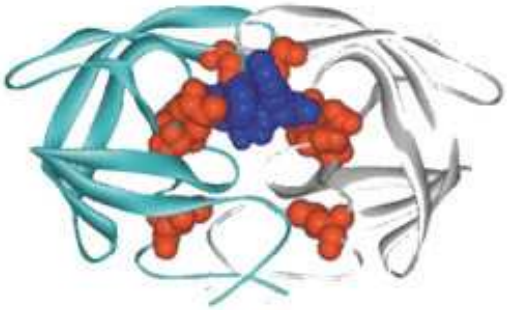
Antigenic shift in 1982

Question: Do volatility changes differ in the antibody interaction sites (consistent with the volatility hypothesis)?

Handling Convergent Evolution in HIV

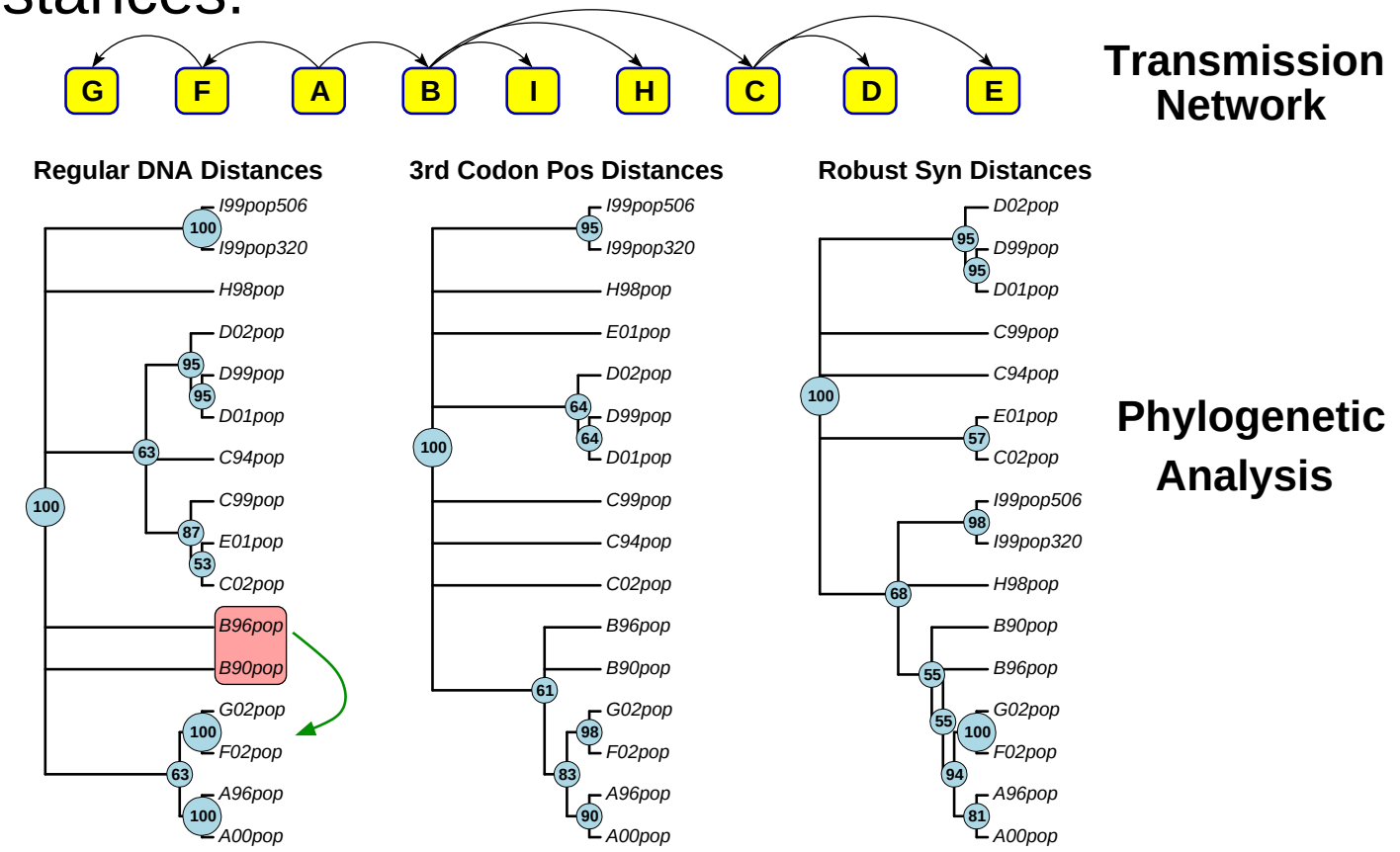


- Lemey et al. (2005) examine the genetic signature of a known HIV transmission network in the face of **convergent evolution**
- HIV *pol* and *env* sequences from 9 subjects
- Distance-based reconstructions using NG86 measures
- **Trouble:** *pol* phylogenies **conflict** with network



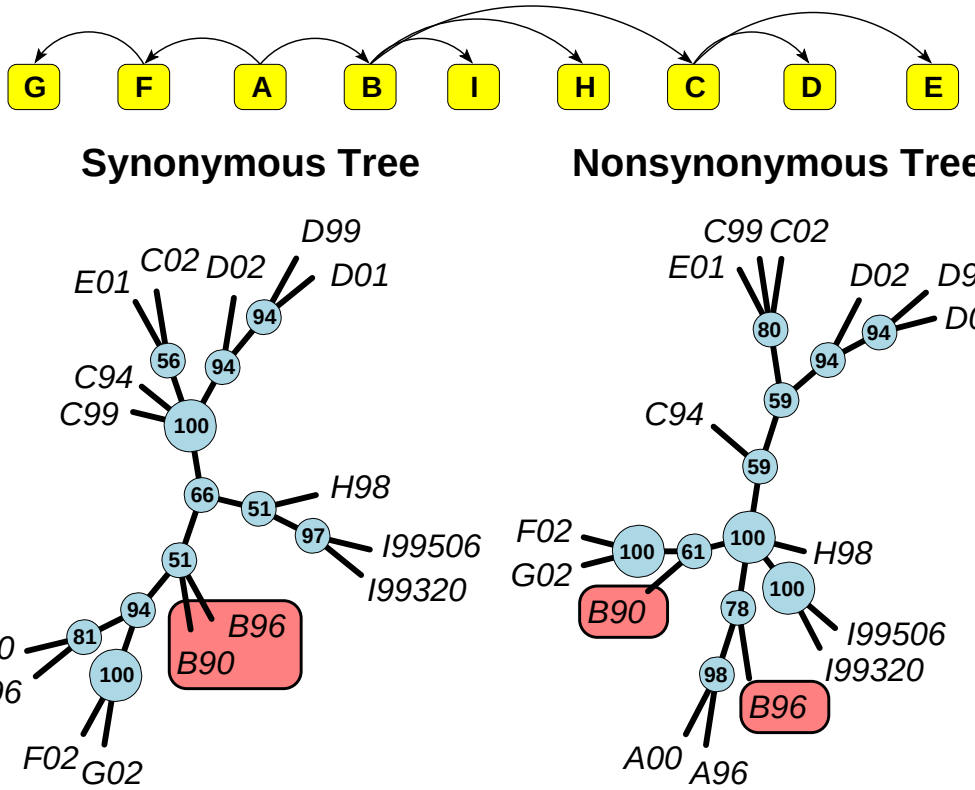
HIV Convergent Evolution-II

- Compare **"synonymous tree"** estimates using 3rd codon positions (**throw away data**) vs our robust synonymous distances:



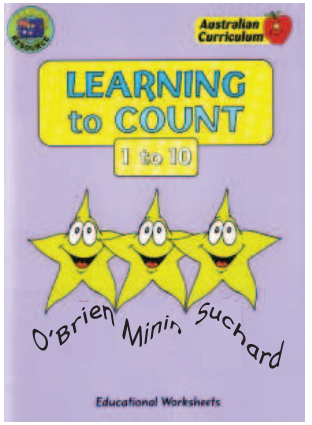
HIV Convergent Evolution-III

- Compare **"synonymous"** and **"nonsynonymous"** trees:



A Few Summary Comments

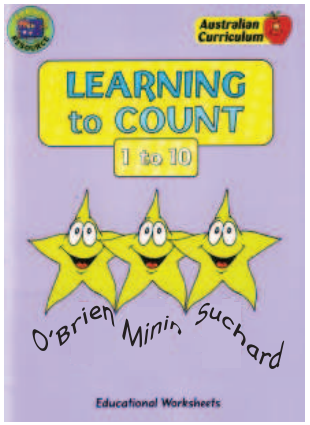
- Analytic expressions for evolutionary counting processes are derivable and flexible to use:
  - For moderate $|\mathbf{\Lambda}|$: **substantial increase** in computational efficiency over simulation-based methods
  - For large $|\mathbf{\Lambda}|$: now **tractable**
- Complex posterior p-value tests require simulation only under the null (simple) model for which **standard software exists**



- Minin and Suchard, *Journal of Mathematical Biology*, 2008
- Minin and Suchard, *Proceedings of the Royal Society B*, 2008

A Few (More) Summary Comments

- We introduce a general framework for computing abitrary **labeled distances**; appears **robust to model misspecification**
- How robust??? $\Leftrightarrow$ an open question
- O'Brien, Minin and Suchard, *Molecular Biology and Evolution*, 2009



- R package **markovjumps**
- Recently integrated into **BEAST** and ready for release; anyone want to try it?