Speed Dating using Least-Squares

Thu Hien TO, Matthieu JUNG, Samantha LYCETT, Olivier GASCUEL

Bioinformatique Evolutive, C3BI USR3756 Institut Pasteur & CNRS, Paris

Institut de Biologie Computationnelle, Montpellier – France

Institute of Evolutionary Biology, Edinburgh – United Kingdom

Institut Pasteur

Centre de Bioinformatique, Biostatistique et Biologie Intégrative

Institut de biologie computationnelle

Speed Dating using Least-Squares

- **A deluge of data**
- **Fast algorithms are needed**
- **We must rely on simple models**

Speed Dating using Least-Squares

- **A deluge of data**
  - Dozens of thousands of virus sequences (eg 40,000 in the UK HIV database)
  - Origin of epidemics, phylodynamics, resistance mutations, surveillance
  - Dating is essential in all of these tasks
- **Fast algorithms are needed**
  - Linear in time and space (i.e. proportional to the number of taxa)
- **We must rely on simple models**
  - Gaussian, (truncated) normal distribution of the noise
  - Strict molecular clock (SMC), but robust

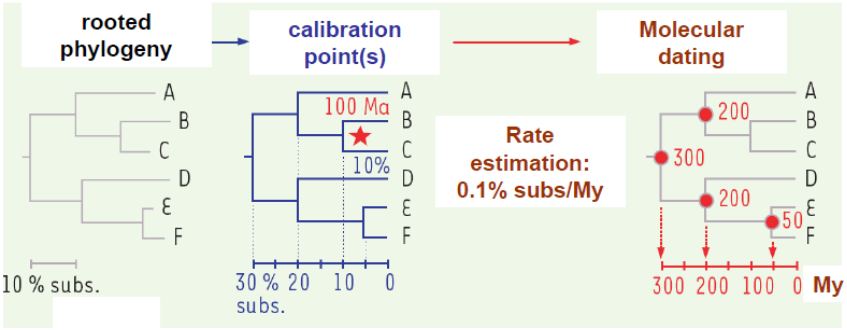
Speed Dating using Least-Squares

- **A deluge of data**
  - Dozens of thousands of virus sequences (eg 40,000 in the UK HIV database)
  - Origin of epidemics, phylodynamics, resistance mutations, surveillance
  - Dating is essential in all of these tasks
- **Fast algorithms are needed**
  - Linear in time and space (i.e. proportional to the number of taxa)
- **We must rely on simple models**
  - Gaussian, (truncated) normal distribution of the noise
  - Strict molecular clock (SMC), but robust
- **Suprizingly accurate!**

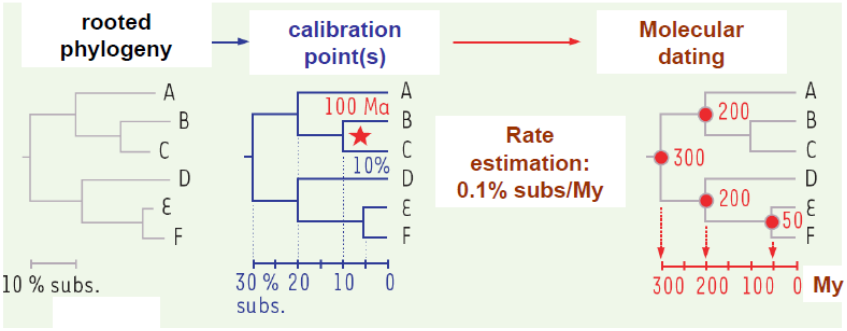
Speed Dating using Least-Squares

- Quick survey of dating models and methods
- The distance-based approach, root-to-tip regression and LF model
- A simple (but robust) Gaussian model
- Dating using linear algebra (LD, unconstrained)
- Quadratic programming dating (QPD, temporal constraints)
- Tree rooting
- Simulation results
- Application to a large H1N1 influenza data set
- Discussion

Quick survey – Basic principle



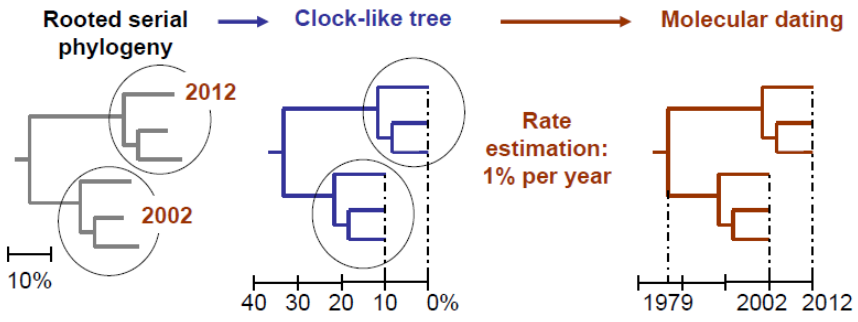
Quick survey – Basic principle



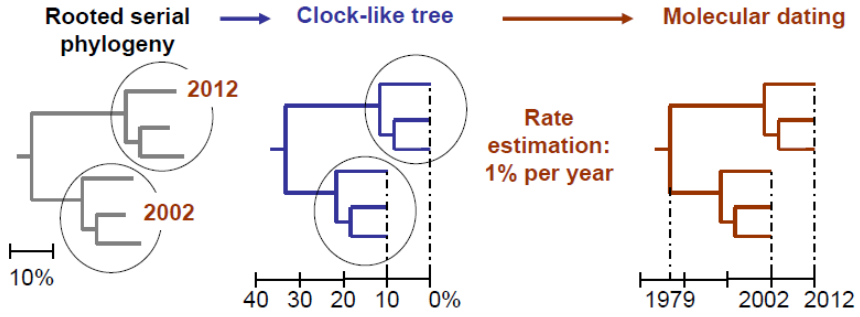
Much more difficult than this with real data:

- Phylogenetic uncertainty
- Non molecular clock (unrooted) trees
- Several (incompatible) calibration points
- High uncertainty depending on the calibration point position

Quick survey – Basic principle



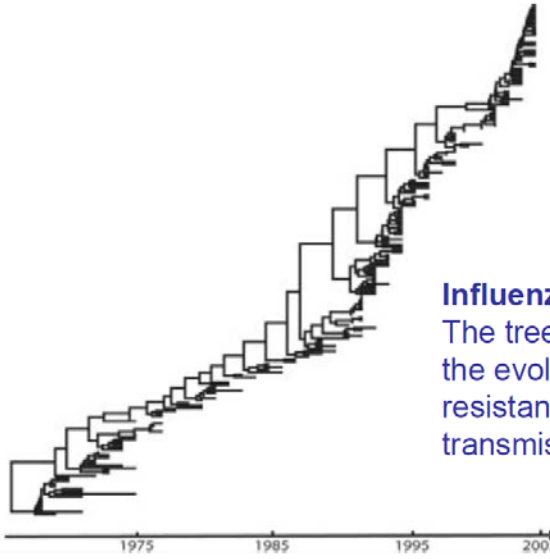
Quick survey – Basic principle



Much more difficult than this with real data:

- Phylogenetic uncertainty
- Non molecular clock (unrooted) tree
- Several (incompatible) sampling times
- High uncertainty depending on sampling times, tree shape …

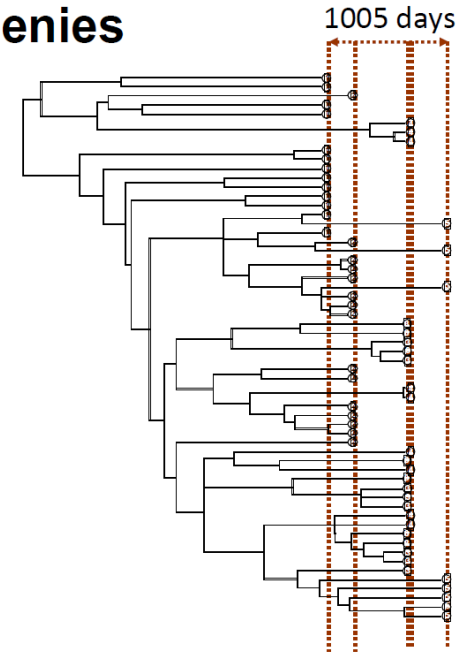
Serial virus phylogenies



already time scaled!

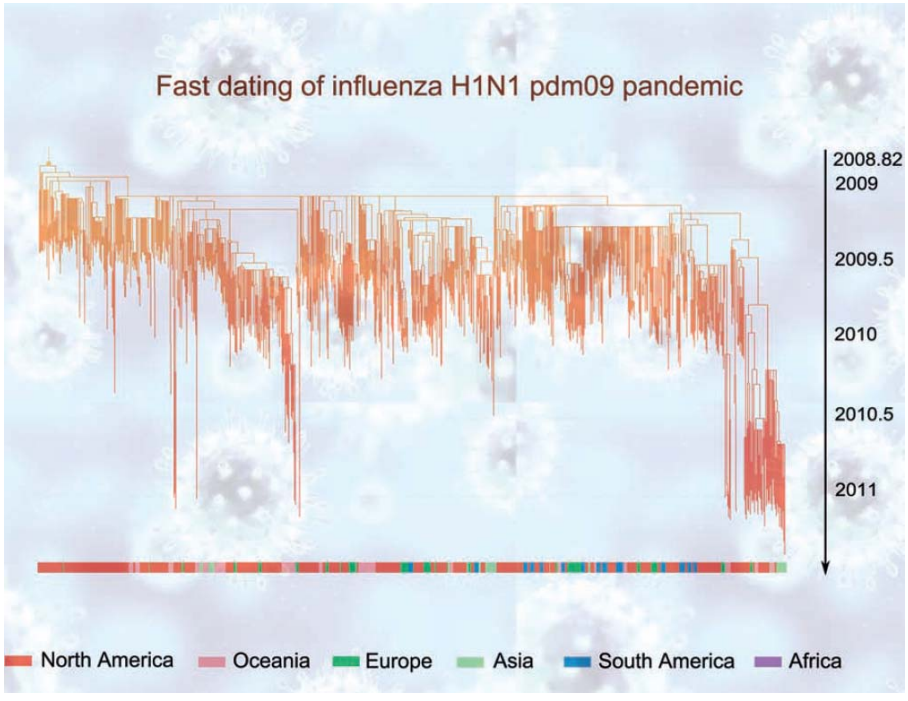
Influenza A H3N2
The tree shape is explained by the evolutionary pressure (human resistance) and the mode of transmission (short life time)

Serial virus phylogenies



HIV within patient
env gene, sampled
over 5 time points

Lader shape still visible,
but dating is more difficult



Quick survey – Input data

- Sequences/pairwise distances/topology/phylogeny
- Outgroup/ingroup only
- Rooted/unrooted phylogeny
- Internal calibration points/tips sampled through time

Quick survey – Main attempts

- Estimating the global rate of evolution
- Estimating several rates (before/after treatment)

  Constraints needed!

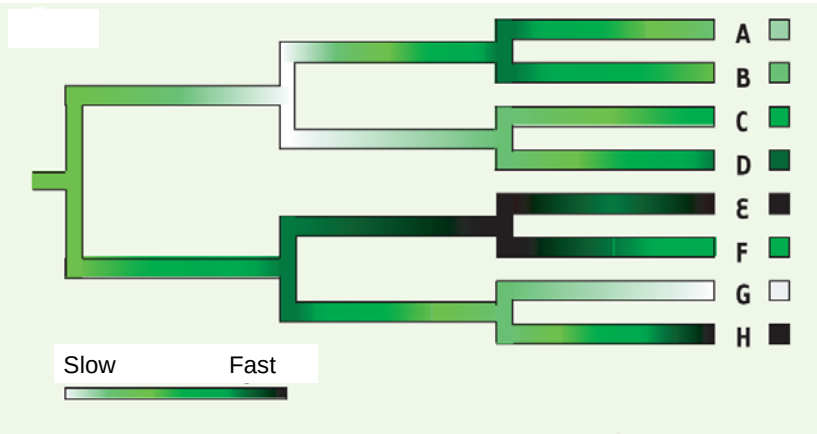
- Estimating the root position and its date
- Estimating the dates of all nodes in the tree

- Estimating a complete, time-scaled tree (e.g. BEAST)

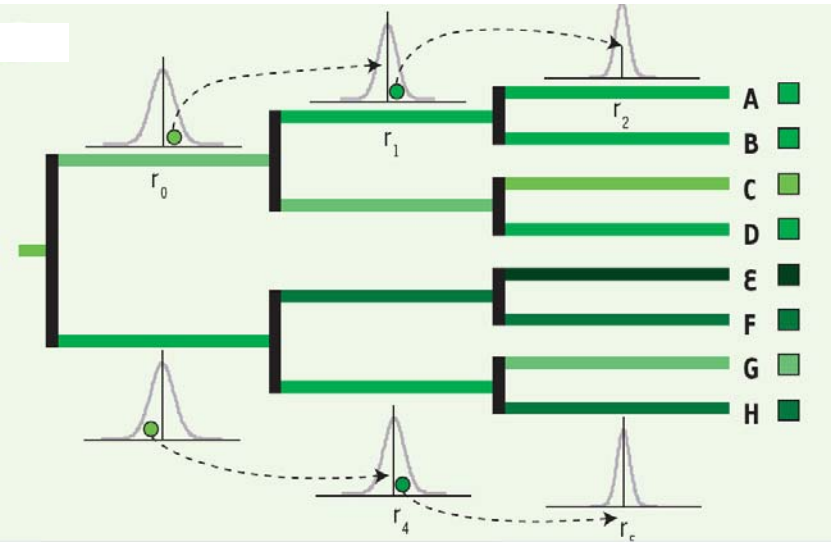
Quick survey – Clock models

- Strict molecular clock: the time is proportional to the number of substitutions per site (plus noise)
- Uncorrelated rates, with known distribution (e.g. lognormal, with mean and variance to be estimated)
- Correlated under some model (e.g. the mean of daughter branch is drawn from a distribution with mean equal to mother's rate)

Relaxed, correlated clock models



Relaxed, correlated clock models



Quick survey – Clock models

- Strict molecular clock: the time is proportional to the number of substitutions per site (plus noise)
- Uncorrelated rates, with known distribution (e.g. lognormal, with mean and variance to be estimated)
- Correlated under some model (e.g. the mean of daughter branch is drawn from a distribution with mean equal to mother's rate)
- Models of increasing complexity, typically requiring MCMC or ABC algorithms, usually slow and limited to a few hundred taxa-sequences

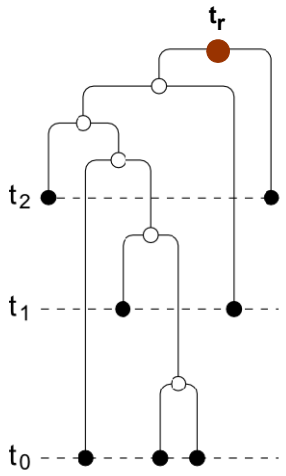
Quick survey – Clock models

- Strict molecular clock: the time is proportional to the number of substitutions per site (plus noise)
- Uncorrelated rates, with known distribution (e.g. lognormal, with mean and variance to be estimated)
- Correlated under some model (e.g. the mean of daughter branch is drawn from a distribution with mean equal to mother's rate)
- No evidence that correlated models are useful for viruses (Drummond et al. 2006)

Quick survey – Clock models

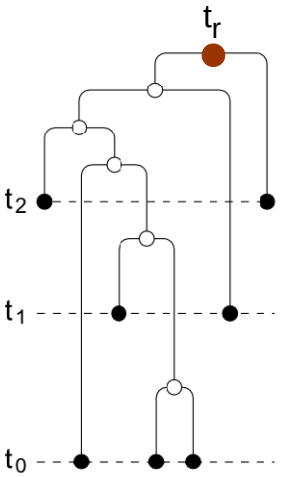
- Strict molecular clock: the time is proportional to the number of substitutions per site (plus noise)
- Uncorrelated rates, with known distribution (e.g. lognormal, with mean and variance to be estimated)
- Correlated under some model (e.g. the mean of daughter branch is drawn from a distribution with mean equal to mother's rate)
- No model, just smoothing (e.g. PathD8)

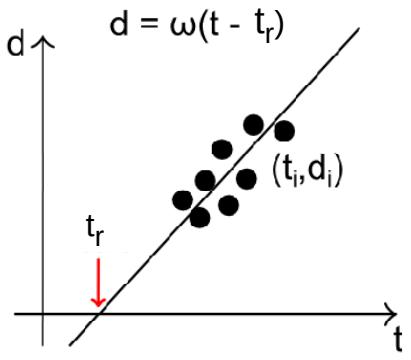
Distance-based approach: root-to-tip regression



- Input: rooted tree, dated tips
- Strict molecular clock
- Model: root-to-tip distances are affected by i.i.d. normal noise
- Output: rate (ω) and root date
- Simple and fast (O(n))
- Highly sensitive to root position
- Evolutionary correlation not accounted for

Distance-based approach: root-to-tip regression





- Standard regression (GLS does not work)
- Able to select the root position in $O(n^2)$

Distance-based, Langley-Fitch (LF) model - r8s

- **Input: a rooted tree, with branch lengths and dated tips**
- **Output: substitution rate (ω) and all nodes dates**
- **Strict molecular clock**
- **Substitutions on each tree branch (*i*, *a*(*i*)) follow a Poisson distribution with mean $s\omega(t_i - t_{a(i)})$**
- **Multi-dimensional optimisation of the likelihood function, using the Powell algorithm (r8s, Sanderson 2003)**
- **Relatively fast** *(but not fast enough for tree rooting)*

A simple Gaussian approximation of LF model

- **The length b_i of branch (*i*, *a*(*i*)) is normally distributed**

$$b_i = \omega(t_i - t_{a(i)}) + N\left(0, \sigma_i^2\right)$$

$$\sigma_i^2 = \frac{\omega\left(t_i - t_{a(i)}\right)}{s} \propto E\left(b_i\right)$$

$$\widehat{\sigma_i^2} \propto \hat{b}_i + C/s$$

Pseudo-count

Robust to some violation of SMC

- **Uncorrelated, normal, relaxed clock model**

$$\omega_i = \omega + N\left(0, \xi^2\right)$$

$$b_i = \omega_i\left(t_i - t_{a(i)}\right) + N\left(0, \frac{\omega\left(t_i - t_{a(i)}\right)}{s}\right)$$

$$b_i = \omega\left(t_i - t_{a(i)}\right) + N\left(0, \xi^2\left(t_i - t_{a(i)}\right)^2 + \frac{\omega\left(t_i - t_{a(i)}\right)}{s}\right)$$

***b_i* is still normally distributed**
its variance is again an increasing function of *b_i*

Least-squares criterion – Temporal constraint

- **Log-Likelihood (Weighted Least Squares) criterion:**

$$LL(\omega, t_1, \ldots, t_{n-1}) \propto \sum_i \frac{1}{\sigma_i^2}\left(b_i - \omega(t_i - t_{a(i)})\right)^2$$

$$\propto \sum_i \left(\frac{1}{b_i + C/s}\right)\left(b_i - \omega(t_i - t_{a(i)})\right)^2$$

- **Precedence constraint for every node/leaf *i* (except the root):**

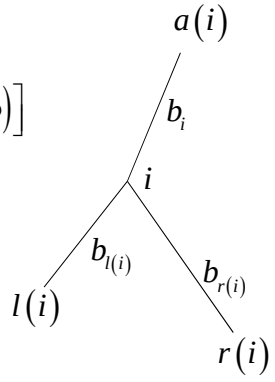
$$t_i \geq t_{a(i)}$$

LD (unconstrained)

The unique, optimal (OLS) solution satisfies

$$t_i = \frac{1}{3}\left[\left(t_{l(i)} - b_{l(i)}/\omega\right) + \left(t_{r(i)} - b_{r(i)}/\omega\right) + \left(t_{a(i)} + b_{a(i)}/\omega\right)\right]$$

$$t_{root} = \frac{1}{2}\left[\left(t_{l(root)} - b_{l(root)}/\omega\right) + \left(t_{r(root)} - b_{r(root)}/\omega\right)\right]$$



A linear system that is solved in linear time (using bottom-up and top-down tree traversals – just as with parsimony), thus providing the value of t_i given ω:

$$t_i = c_i + k_i/\omega$$

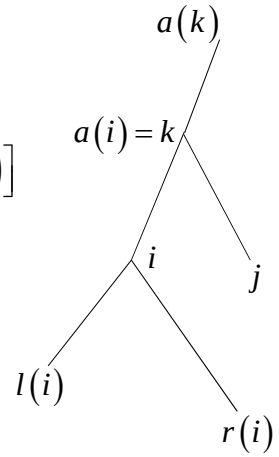
We use these equalities in WLS criterion to obtain in linear time ω , and then all dates t_i

LD (unconstrained)

The unique, optimal (OLS) solution satisfies

$$t_i = \frac{1}{3}\left[\left(t_{l(i)} - b_{l(i)}/\omega\right) + \left(t_{r(i)} - b_{r(i)}/\omega\right) + \left(t_{a(i)} + b_{a(i)}/\omega\right)\right]$$

$$t_{root} = \frac{1}{2}\left[\left(t_{l(root)} - b_{l(root)}/\omega\right) + \left(t_{r(root)} - b_{r(root)}/\omega\right)\right]$$



A linear system that is solved in linear time (using bottom-up and top-down tree traversals – just as with parsimony), thus providing the value of t_i given ω:

$$t_i = c_i + k_i/\omega \quad \downarrow$$

$$\uparrow \quad t_i = w_i t_{a(i)} + v_i/\omega + u_i$$

We use these equalities in WLS criterion to obtain in linear time ω , and then all dates t_i

QPD (with temporal constraints)

Quadratic function of the (changed) variables:

$$LL = \sum_i (b_i - \omega(t_i - t_{a(i)}))^2$$

$$= \sum_{i \in \text{leaves}} (b_i - \omega t_i + \beta_{a(i)})^2 + \sum_{i \in \text{internal}} (b_i - \beta_i + \beta_{a(i)})^2$$

$$\beta_i = \omega t_i \text{ for the internal nodes}$$

Subject to: internal nodes: $\beta_i \geq \beta_{a(i)}$

tree leaves: $\omega t_i \geq \beta_{a(i)}$

Unique solution, obtained using an active set method

QPD (with temporal constraints)

Active set method (summary)

1. **Run LD**
2. **All violated constraints are put in the active set** $\left(t_i = t_{a(i)}\right)$
3. **Compute the optimal solution x* and the Lagrange multipliers corresponding to the active constraints**
   Use a variant of LD on the collapsed tree $\left(b_i = 0\right)$
4. **If x* is feasible and all constraints are useful, then output x*, else remove the most useless constraint** $\left(\lambda_i < 0\right)$ **and go to 3**
5. **If x* is not feasible, add to the active set the most violated constraint and go to 3**

Time complexity O(n x k)
k = # iterations << n (~70 with ~900 influenza strains)

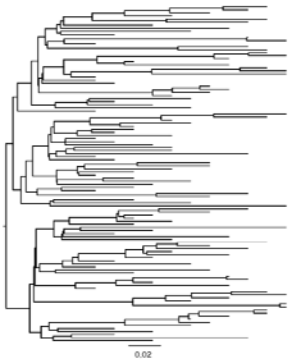
Tree rooting

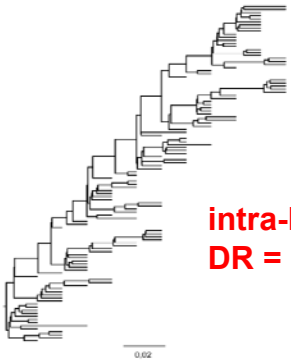
- For any given edge, we use a slightly modified versions of LD and QPD to find the best rooting position on that edge (i.e. minimizing WLS).
- Run LD or QPD on every edge of the tree, and find the best root position in $O(n^2)$
- Still quite fast with LD
- With QPD, we first run LD to find an initial solution, and then run QPD in a hill-climbing fashion to improve that solution (most of the time LD solution is best, or nearly best)

Simulation results

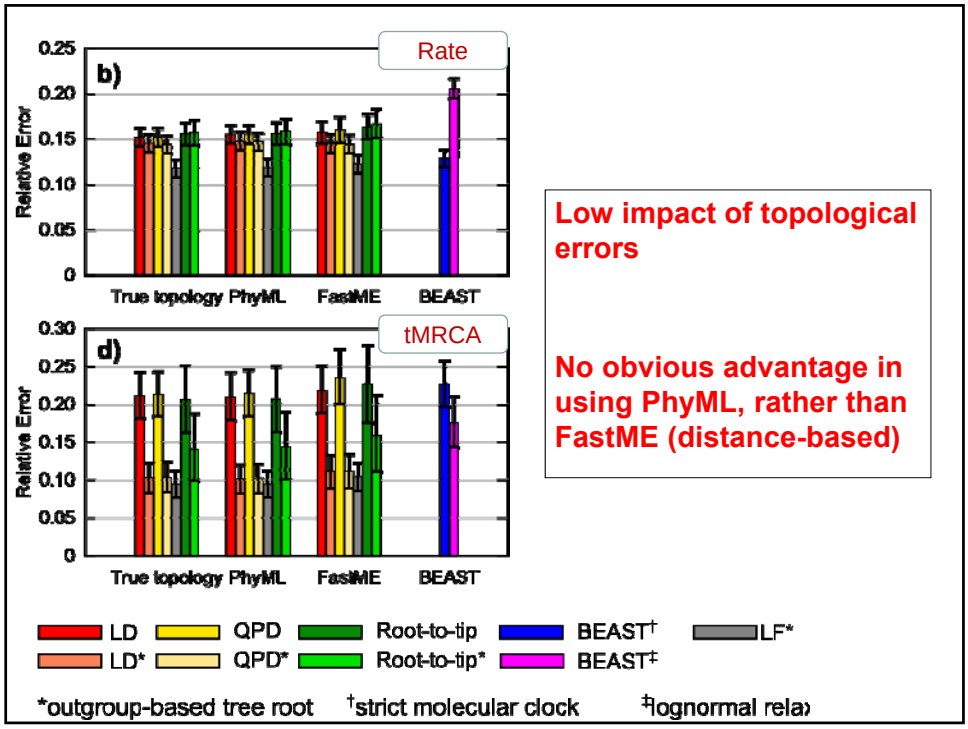
- Birth-death trees with various death rates (DR), 70 to 110 taxa
- Uncorrelated, log-normal relaxed clock model
- F84+Γ substitution model, 500 sites
- "HIV" parameters (in between *Pol* and *env*)

inter-host
DR = 0.75



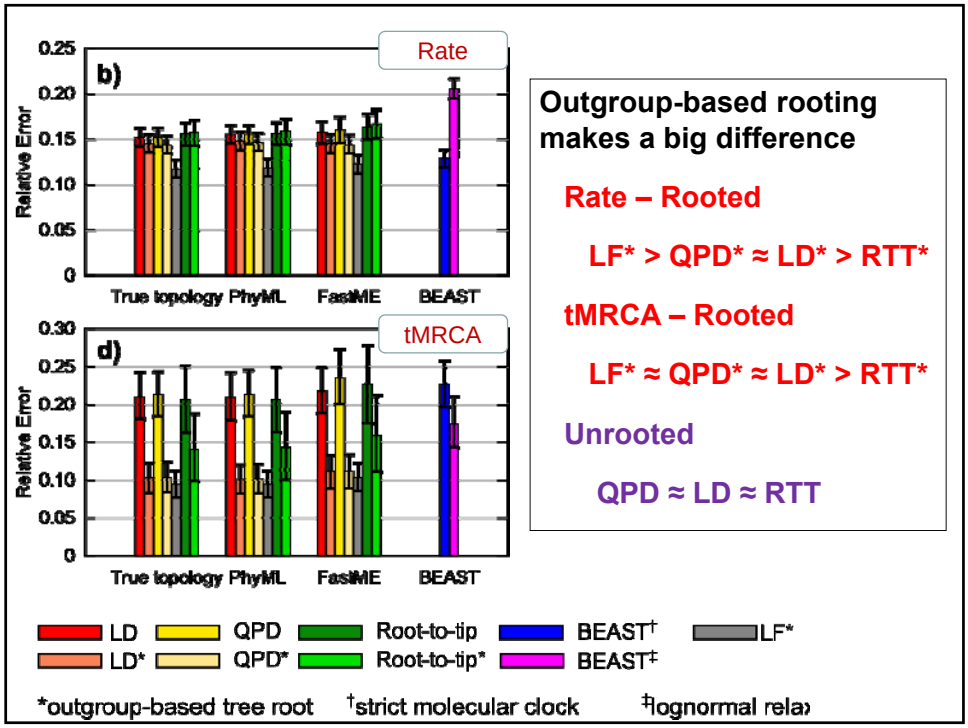


intra-host
DR = 0.995



Low impact of topological errors

No obvious advantage in using PhyML, rather than FastME (distance-based)



Outgroup-based rooting makes a big difference

Rate – Rooted

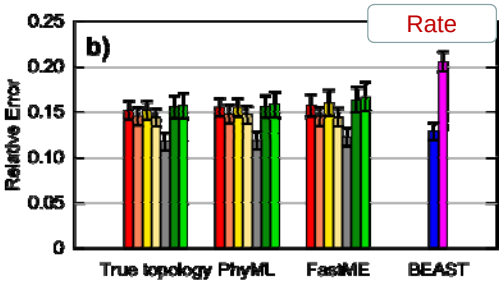
LF* > QPD* ≈ LD* > RTT*

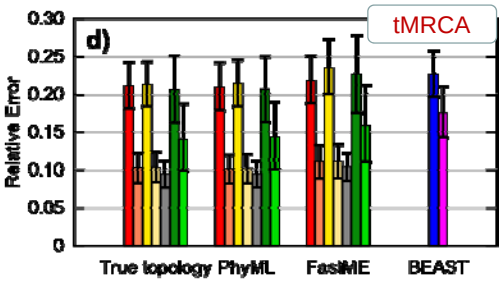
tMRCA – Rooted

LF* ≈ QPD* ≈ LD* > RTT*

Unrooted

QPD ≈ LD ≈ RTT

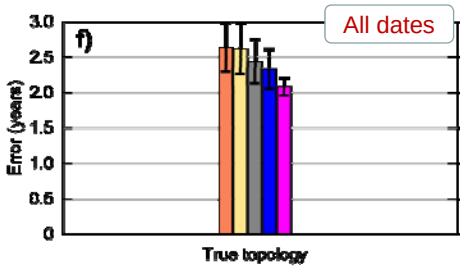




Desappointing results with BEAST (no outgroup, complete time-scaled tree)

LD QPD Root-to-tip BEAST† LF*
LD* QPD* Root-to-tip* BEAST‡

*outgroup-based tree root †strict molecular clock ‡lognormal rela



True rooted tree topology

BEAST-RMC is consistently best to estimate all node dates

(but tiny differences, and still some trouble with rate estim.)

LD QPD Root-to-tip BEAST† LF*
LD* QPD* Root-to-tip* BEAST‡

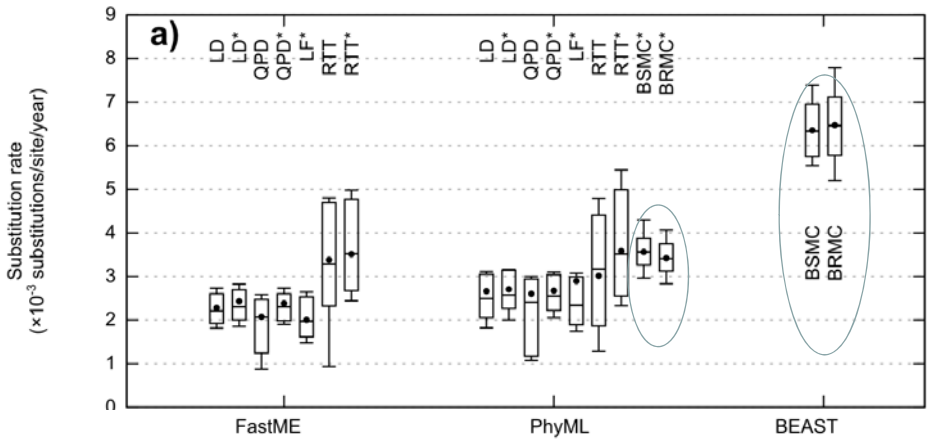
*outgroup-based tree root †strict molecular clock ‡lognormal rela

Computing times (in seconds - 110 taxa)

	750/11×10
Phylogeny inference	
DNAdist+FastME	5
PhyML	8mn
Dates and rate estimation	
LD	0.1
LD*	<0.1
QPD	0.2
QPD*	<0.1
Root-to-tip	<0.1
Root-to-tip*	<0.1
LF*	3.5
BEAST with a strict molecular clock	4h
BEAST with a relaxed molecular clock	17h

*outgroup-based rooted tree

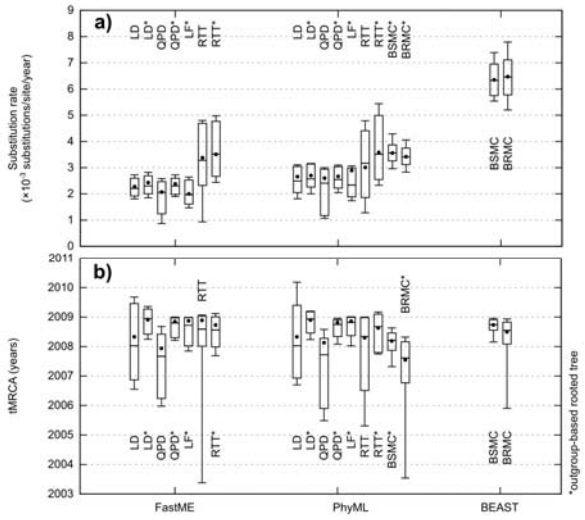
1,195 H1N1 influenza strains + outgroup



Same methods and options as with simulated data

We also ran BEAST with fixed rooted PhyML topology

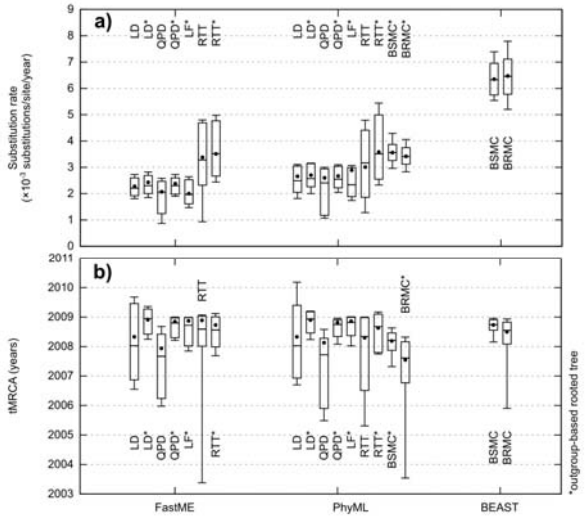
100 bootstrap replicates to obtain confidence intervals



Results are mostly consistent with simulations

Large intervals with unrooted input tree (LD, QPD, RTT)

QPD* and LF* are very close, and compatible with BEAST*



LD+PhyML : ~7% of violated temporal constraints (> 1month)

QPD has a clear advantage!

BEAST + TreeAnnotator : ~2% - BEAST* + TreeAnnotator : 0%

Computing times (with 100 boostrap rep.)

BEAST :	5 (*) to 20 days (Beagle, GPU …)
PhyML :	4 days (desktop, not parallelized)
FastME :	1 hour
RTT, LD, QPD, LF* :	1 hour
QPD* :	2 mn
RTT*, LD* :	10 sec.

Summary

Ability to deal with rooted and unrooted trees

Provide estimates for the rate and all node dates

Similar accuracy as LF (despite normal approximation) and BEAST (still unexplained)

Fast and already used with very large datasets

- Mourad et al. (AIDS 2015), transmission of resistance mutations in HIV, 24,000 strains, rooted tree, ~30 minutes (LF > 2 weeks)
- PANGEA_HIV consortium to estimate phylodynamics parameter from rooted/unrooted trees (→ 20,000 strains)

To be done - To be finished-published

Fast confidence intervals (e.g. based on the second derivative of the likelihood function, parametric bootstrap …)

Extension to time calibration points (see also Xia 2011)

Analyse the LS residues (e.g. to check for MC)

Extend to correlated rate models (Sanderson 2002)

Fast Dating Using Least-Squares Criteria and Algorithms

THU-HIEN TO[1], MATTHIEU JUNG[1,2], SAMANTHA LYCETT[3], AND OLIVIER GASCUEL[1,*]

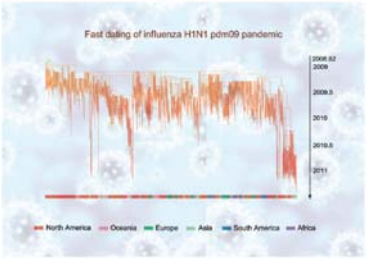
[1]*Institut de Biologie Computationnelle, LIRMM, UMR 5506 CNRS – Université de Montpellier, France;* [2]*IGBMC (Institut de Génétique et de Biologie Moléculaire et Cellulaire), INSERM, U596, CNRS, UMR7104, Université de Strasbourg, Illkirch, France;* [3]*Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Edinburgh, UK*

**Correspondence to be sent to: Institut de Biologie Computationnelle, LIRMM, UMR 5506 CNRS – Université de Montpellier, 161 rue Ada, 34392 Montpellier, France; Email: gascuel@lirmm.fr*

Systematic Biology

A JOURNAL OF THE Society of Systematic Biologists

http://www.atgc-montpellier.fr/LSD/



VOLUME 65
NUMBER 1
JANUARY 2016
ONLINE ISSN 1076-836X
PRINT ISSN 1063-5157