Al buio non si trova:

Principled phylodynamics for pandemic preparation

Luiz Max Carvalho

FGV EMAp

Acknowledgments

Andrew Rambaut
UoE

Guy Baele
KU Leuven

Marc Suchard
UCLA

Rodrigo B. Alves
FGV EMAp

Remco Bouckaert
Auckland

Cristiana Couto
ICMC USP

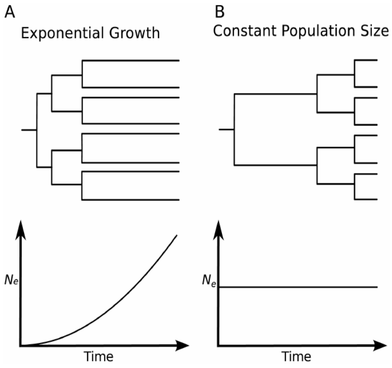
Motivation

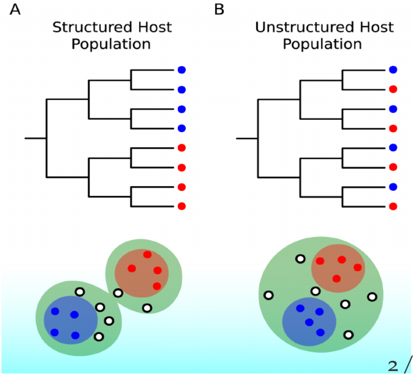
Phylodynamics of fast-evolving viruses

Inferring spatial and temporal dynamics from genomic data:

Phylogenies*!

* plus complicated models





Plan for today

Statistical Problem(s)

Central object, inference, algorithms

Principled priors

Being Bayesian is great, but it ain't free

MCMC in tree space

A journey through a strange land

How to tell if phylogenetic MCMC

A) Is correct;

B) Works better than the state-of-the-art.

Central object: time-calibrated trees

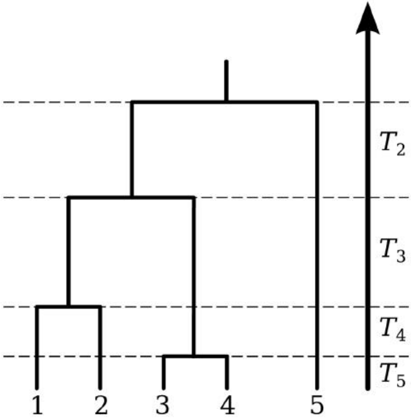


Figure: Figure 4 from Volz et al. (2013).

Let T_n denote the time for n lineages to *coalesce*, i.e., merge into one ancestral lineage, in a population of size N_e. Then:

$$\Pr(T_n = t) = \lambda_n e^{-\lambda_n t}$$

$$\lambda_n = \binom{n}{2} \frac{1}{N_e} = \binom{n}{2} \frac{1}{\theta\tau}$$

where N_e is the effective population size and τ is the generation time. Let T_{mrca} denote the age of the most recent common ancestor:

$$\begin{aligned} E[T_{\text{mrca}}] &= E[T_n] + E[T_{n-1}] + \ldots + E[T_2] \\ &= 1/\lambda_n + 1/\lambda_{n-1} + \ldots + 1/\lambda_2 \\ &= 2N_e \left(1 - \frac{1}{n}\right) \end{aligned}$$

“Just chuck any prior”

Consider:

$$t_k \mid N_e \sim \text{Exponential}\left(\binom{n}{2}\frac{1}{N_e}\right).$$

If you pick $\pi_N(N_e) \propto 1/N_e$, i.e. the Jeffreys’s-type prior, you get that the marginal prior for t_k is $\pi_T(t_k) \propto 1/t_k$.

$$\boldsymbol{P}(t_k) = \exp(t_k\boldsymbol{Q}) = \sum_{i=0} \frac{(t_k\boldsymbol{Q})^i}{i!}.$$

Lemma

If $\boldsymbol{Q}$ is diagonalisable, the posterior for t_k is improper[1] under a Jeffreys’s prior for N_e.

[1] A measure-theoretic proof of a very similar result is given in the Appendix of Drummond et al. (2004).

♪Smooth operator ♪

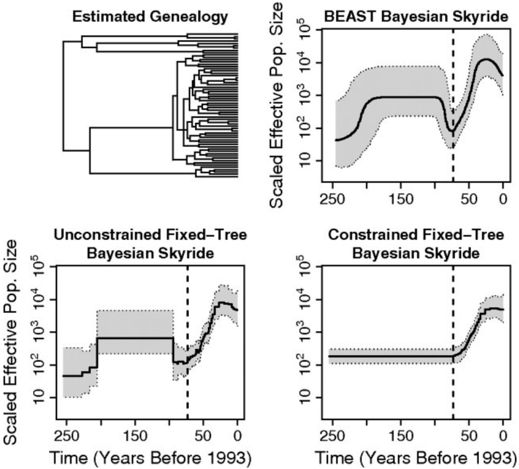


Figure: HCV in Egypt [2].

[2] Minin et al. (2008). See also Karcher et al. (2020)

Gaussian Markov random fields to the rescue

Denote the population sizes by $\boldsymbol{\theta} = (\theta_2, \ldots, \theta_n)$, the likelihood becomes

$$\Pr(\boldsymbol{s}|\boldsymbol{\theta}) = \prod_{k=2}^{n} \frac{n_{k0}(n_{k0}-1)}{2\theta_k} \exp\left(-\sum_{j=0}^{j_k} \frac{n_{kj}(n_{kj}-1)s_{kj}}{2\theta_k}\right),$$

$$\Pr(\boldsymbol{\gamma}|\tau) \propto \tau^{(n-2)/2} \exp\left(-\frac{\tau}{2}\sum_{k=2}^{n-1} \frac{(\gamma_{k+1}-\gamma_k)^2}{\delta_k}\right),$$

where $\gamma_k = \log(\theta_k), k = 2, \ldots, n$, δ_k is the (1d) distance between intervals and τ is the precision parameter associated with the smoothing.

Penalising complexity

Simpson et al. (2017) propose proper priors that penalise deviations from a simple base model ("complexity"). For the GMRF precision, this prior is a Gumbel type II family:

$$\pi_2(\tau \mid a, b) = ab \cdot \tau^{-a-1} \exp\left(-b\tau^{-a}\right), \; \tau > 0. \tag{1}$$

We set $a = 1/2$ and b such that $\Pr(1/\sqrt{\tau} > S) = p$, where the value S and the probability p are to be chosen on substantive grounds – e.g. $S = 1$ and $p = 0.1$. We can then find $b = -\ln(p)/S$.

Some reconstructions are sensitive to the prior

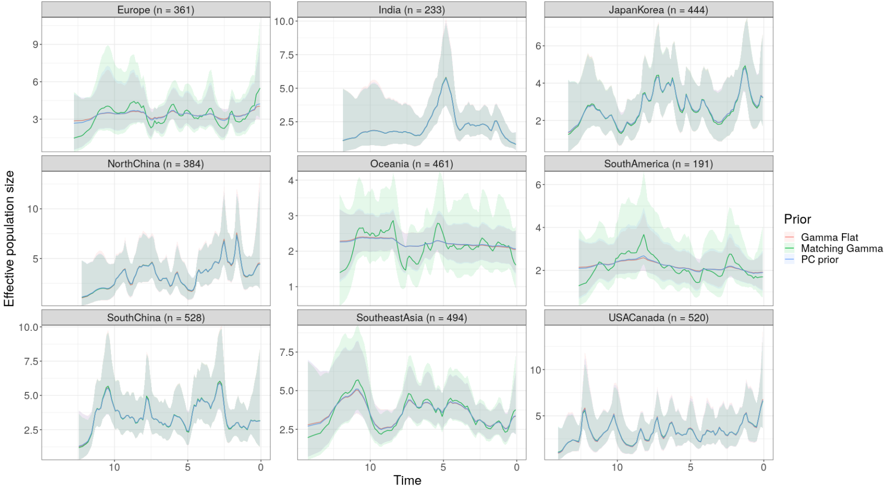


Figure: Regional Influenza

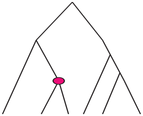
The phylogenetic target

$$p(t, \boldsymbol{b}, \boldsymbol{\omega} | D) = \frac{f(D | t, \boldsymbol{b}, \boldsymbol{\omega}) \pi(t, \boldsymbol{b}, \boldsymbol{\omega})}{\sum_{t_i \in \boldsymbol{T}_n} \int_{\boldsymbol{B}} \int_{\boldsymbol{\Omega}} f(D | t_i, \boldsymbol{b}_i, \boldsymbol{\omega}) \pi(t_i, \boldsymbol{b}_i, \boldsymbol{\omega}) d\boldsymbol{\omega} d\boldsymbol{b}_i}. \quad (2)$$

- D: observed sequence (DNA) data;
- $\boldsymbol{T}_n$: set of all binary ranked trees ($\mathbb{G}^{(2n-3)!!}$);
- $\boldsymbol{b}_k$: set of branch lengths of $t_k \in \boldsymbol{T}_n$ ($\mathbb{R}_+^{2n-2}$, kind of) ;
- $\boldsymbol{\omega}$: set of parameters of interest such as substitution model parameters, migration rates, heritability coefficients, etc.

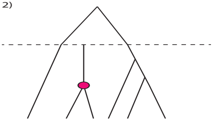
Traversing treespace: SubTreeLeap (STL)

1)



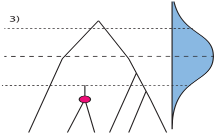
Pick a node

2)



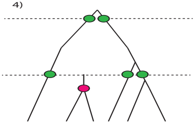
Disconnect its parent

3)



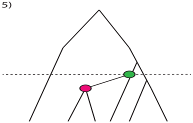
Draw a new height from a normal centred on old height of parent. Also consider the symmetrical height above or below the old height.

4)



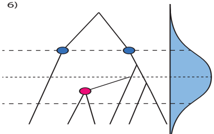
Pick uniformally from branches subtending that height and the symmetrical height above or below (in this case 5).

5)



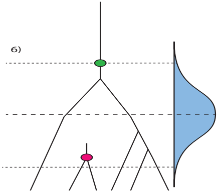
Attach parent to the chosen location.

6)



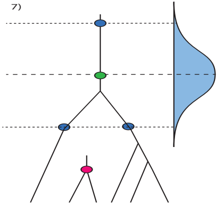
Hastings ratio: ratio of reverse probability (1 / number of reverse locations, i.e., 1/2) to forwards probability (i.e., 1/5). Hastings ratio = 5 / 2

6)



There is always at least 1 target location (above the root).

7)



In this case the HR would be 1/3

STL ergodicity

Carvalho (2019), Chapter 2.

Lemma

Assume strictly positive branch lengths. Then SubTreeLeap induces an irreducible Markov chain on $\mathbb{G}$.

Sketch: Starting at $x \in \mathbb{G}$, notice there exists $\delta_y^\star > 0$ such that $P\left(x \to y \mid \delta_y^\star\right) > 0$ for any tree $y \in \mathbb{G}$ in the SPR neighbourhood of x.

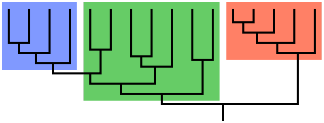
Theorem

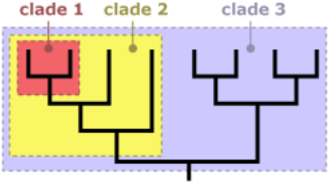
Assume the target satisfies $p(A) > 0$ for all $A \subset \mathbf{\Psi}$. Then, SubTreeLeap induces an ergodic Markov chain on $\mathbf{\Psi}$.

Sketch: Employ the remark to get to the case where $d_{\text{SPR}}(x, y) = 0$ and then establish Harris recurrence.

A lower-dimensional projection

A clade is a partition of the set of leaves and two clades $A = A_1|A_2$ and $B = B_1|B_2$ are said to be compatible if at least one of $A_i \cap B_j$, $i, j = 1, 2$ is empty. Here's a picture[3]:





[3] Pictures taken from Wikipedia and from https://evolution.berkeley.edu/evolibrary/news/080301_elephantshrew

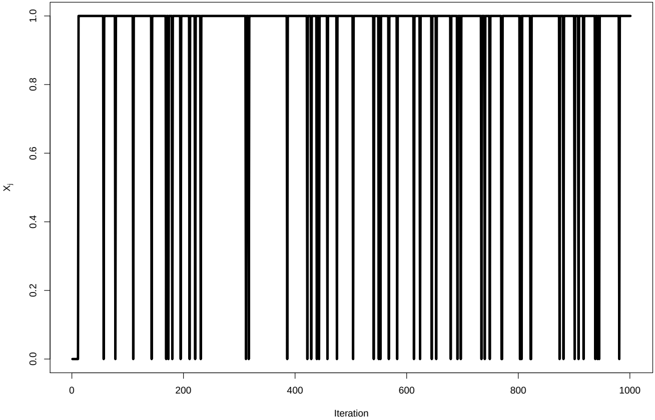
Why clades?

- **Dimension**! $|\mathbb{T}_n| = (2n-3)!!$ *vs* $|\mathbb{C}_n| = 2^{n-1} - 1$
- Interpretability;
- Under simplifying assumptions, clades are independent (Larget, 2013[4]);
- Clade distribution is known under popular prior distributions.

[4] but see Whidden & Matsen, 2015 and Zang & Matsen, 2018.

Clade indicators during MCMC

Let $X_j^{(i)} \in \{0,1\}$ be the indicator of whether clade j in the tree sampled at the i-th iteration and $\hat{p}_j = M^{-1} \sum_{i=1}^{M} X_j^{(i)}$ be a simple MCMC estimator of its marginal success probability.



Playing pretend

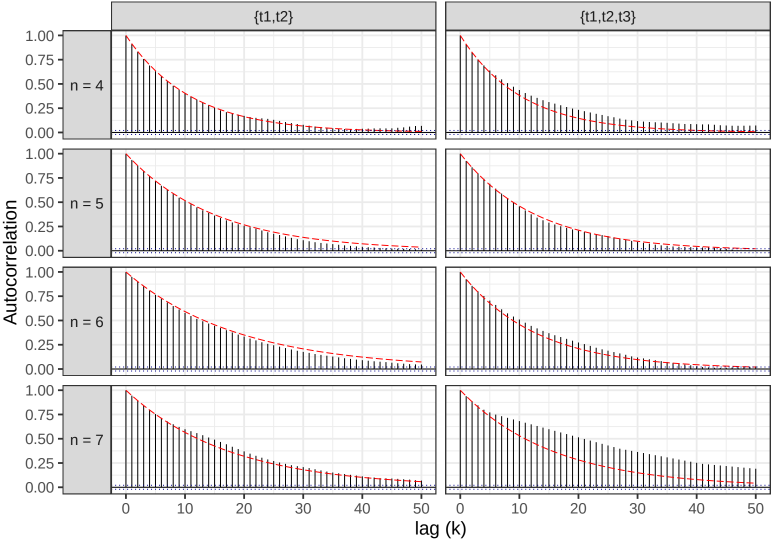
Pretend for a second $\left(X_j^{(i)}\right)_{i\geq 0}$ is Markov on $\mathcal{X} = \{0,1\}$ and reparametrise the usual two-state model as

$$\tilde{\boldsymbol{P}}_x := \begin{bmatrix} 1-\alpha & \alpha \\ \alpha\frac{1-p}{p} & \frac{p-\alpha(1-p)}{p} \end{bmatrix}, \tag{3}$$

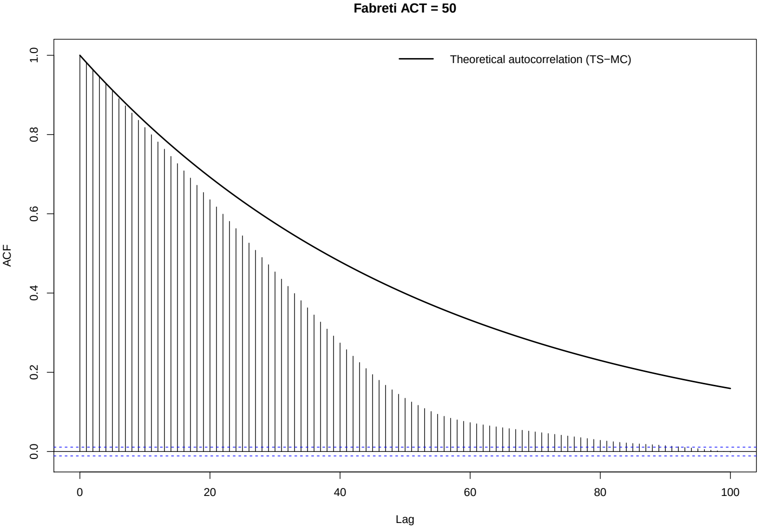
where p is the marginal success probability and a α controls the "flipping rate" of the chain. Then

$$\begin{aligned} \text{ESS} &= \frac{M}{1+2\sum_{t=1}^{\infty}\rho_t}, \\ &= \frac{M}{1+2\frac{p-\alpha}{\alpha}}, \\ &= \frac{\alpha}{2p-\alpha}M. \end{aligned}$$

Lumpability in clade space



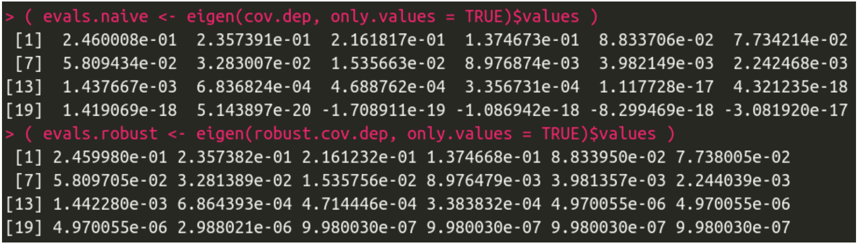
Doesn’t always work



Measuring efficiency

Thus, we can employ the idea from Vats, Flegal & Jones (2019): Magee et al, 2021 point out that trees are fundamentally multivariate objects.

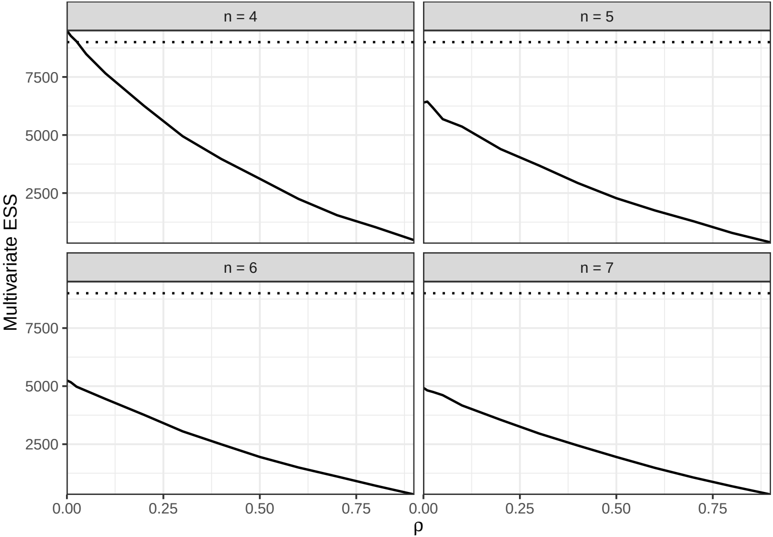
$$\text{mESS} = M \left(\frac{\det(\mathbf{\Lambda})}{\det(\mathbf{\Sigma})} \right)^{1/p}.$$



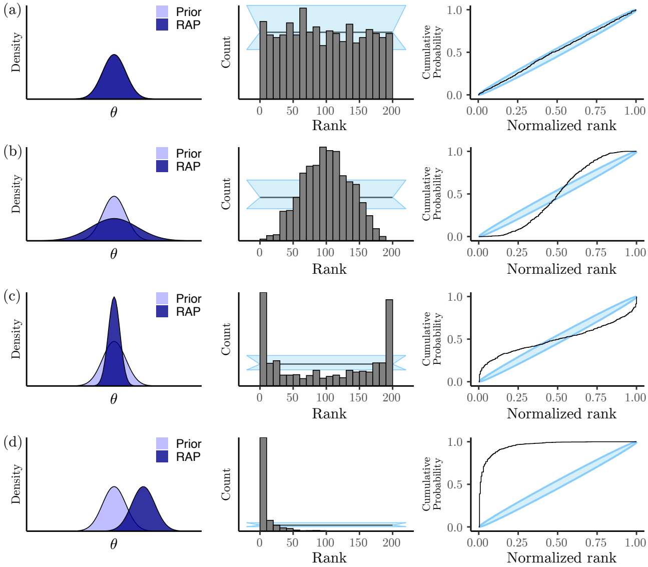
```
> ( evals.naive <- eigen(cov.dep, only.values = TRUE)$values )
 [1]  2.460008e-01  2.357391e-01  2.161817e-01  1.374673e-01  8.833706e-02  7.734214e-02
 [7]  5.809434e-02  3.283007e-02  1.535663e-02  8.976874e-03  3.982149e-03  2.242468e-03
[13]  1.437667e-03  6.836824e-04  4.688762e-04  3.356731e-04  1.117728e-17  4.321235e-18
[19]  1.419069e-18  5.143897e-20 -1.708911e-19 -1.086942e-18 -8.299469e-18 -3.081920e-17
> ( evals.robust <- eigen(robust.cov.dep, only.values = TRUE)$values )
 [1] 2.459980e-01 2.357382e-01 2.161232e-01 1.374668e-01 8.833950e-02 7.738005e-02
 [7] 5.809705e-02 3.281389e-02 1.535756e-02 8.976479e-03 3.981357e-03 2.244039e-03
[13] 1.442280e-03 6.864393e-04 4.714446e-04 3.383832e-04 4.970055e-06 4.970055e-06
[19] 4.970055e-06 2.988021e-06 9.980030e-07 9.980030e-07 9.980030e-07 9.980030e-07
```

Figure: Eigenvalues can be numerically unstable.

True mESS



Simulation-based calibration

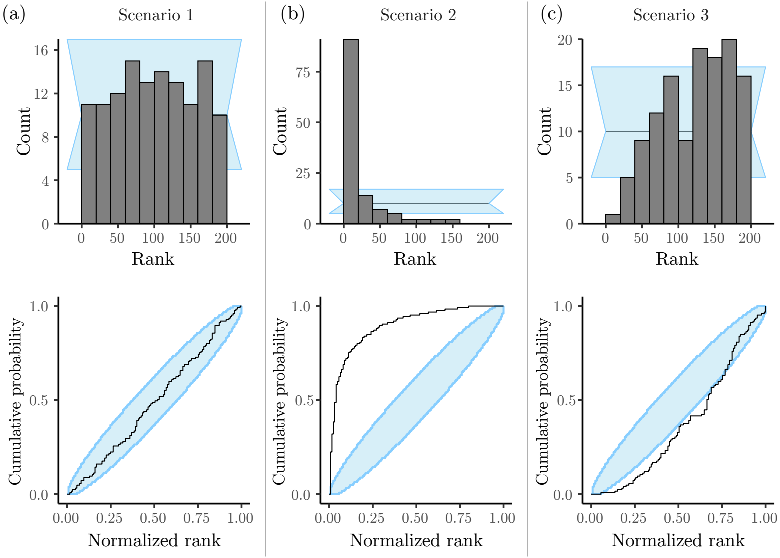


SBC for trees

See Mendes et al. (2024) for more details.

0. Generate a reference tree from the prior $\bar{\tau}_0 \sim \pi_T(\tau|\boldsymbol{\gamma})$;
   for each iteration in 1:N, **do**:
1. Generate $\bar{\tau} \sim \pi_T(\tau|\boldsymbol{\gamma})$;
2. Compute the distance $\bar{\delta} = d_\sigma(\bar{\tau}, \bar{\tau}_0)$ according to the metric of choice;
3. Generate some (alignment) data $\tilde{y} \sim p(y|\bar{\tau}, \boldsymbol{\alpha})$;
4. Draw (approximately) $\boldsymbol{\tau}_s = \{\tau_s^{(1)}, \tau_s^{(2)}, \ldots, \tau_s^{(L)}\}$ from the posterior $\pi(\tau|\tilde{y})$;
5. Compute distances $\boldsymbol{\delta}_s = \{\delta_1, \delta_2, \ldots, \delta_L\}$ with $\delta_i = d_\sigma(\tau_s^{(i)}, \bar{\tau}_0)$;
6. Compute the rank $r(\boldsymbol{\delta}_s, \bar{\delta}) = \sum_{i=1}^{L} \mathbb{I}(\delta_i < \bar{\delta})$.

Simulation-based calibration: results



Take home

Principled priors

Prior calibration, proper priors for generative modelling.

Principled simulation methods

Ascertaining correctness and efficiency

Major methodological challenges (as I see them)

A) Thinking carefully about priors, especially as regularisers;

B) Efficient (preferrably on-line) methods for phylogeny reconstruction;

C) Incorporate mathematical models to link to other data (model-driven data integration).

THE END