Cool things to do wtih BEAST 2

Remco Bouckaert

Seili, September 2017

Computational
Evolution Group

MAX PLANCK INSTITUTE FOR
THE SCIENCE OF HUMAN HISTORY

Some slides based on material provided by Alexei Drummond, Chieh-Hsi Wu, Denise Kühnert, Tim Vaughn

Vision

To provide a framework for **computational evolution** that is

- *easy to use*, that is, well documented, have intuitive user interfaces with small learning curve.
- *open access*, that is, open source, open xml format, facilitating reproducibility of results, runs on many platforms.
- *easy to extend*, by having extensibility in design

Scope

Efficient inference and model-based hypothesis testing for sequence data analysis involving phylogenetic tree models.

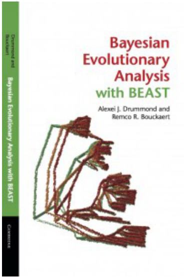
What BEAST 2 does

- The kind of Bayesian analysis as per citations on the BEAST 1 wiki
- BEAUti 2: GUI to specify analysis
- Sequence generator for simulation studies
- Some post processing tools: log analyser, log combiner, DensiTree
- Documentation for all the above – from user to developer, XML tweaker, etc.

What BEAST 2 does that BEAST 1 doesn't...

...hence why you want to use BEAST 2

- Can resume runs when a chain is not mixing well
- BEAUti 2: reload existing specifications – reduced need for XML hacking
- Logs model with trace – allows looking up where the trace comes from
- Provide a platform to develop packages - powerful interface, easy extensible XML, templates for BEAUti.
- Book available



BEAST packages

Consider BEAST 2 as a library for MCMC and phylogenetics

A BEAST 2 package is a library based on BEAST 2

Why package:

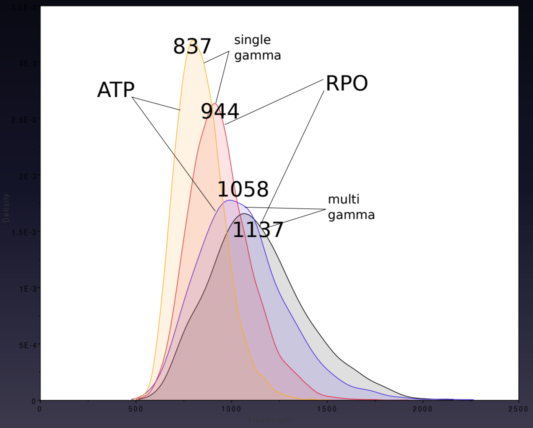
- Making work easier citable
- Making the core easier to learn – it's a lot smaller / cleaner
- Separating out stable / experimental code / dead code
- ...

Tutorial on writing packages: see BEAST wiki
`http://beast2.cs.auckland.ac.nz`

What to do with heterotachy?

- Single Gamma site model (Yang)
- Multi Gamma Site Model: one α per branche
- Relaxed Gamma Site model: sample α from (log normal) distribution
- Much better fit for many models
- Significantly different root ages

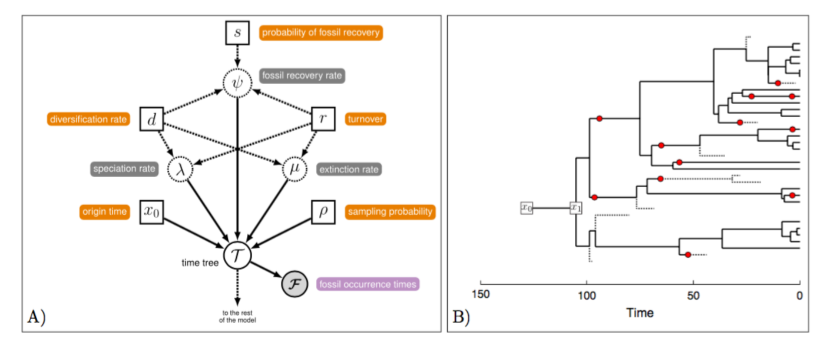
Algae/green plants age



Bouckaert & Lockhart, bioRxiv, 2015, package: MGSM

How to represent fossil calibrations?

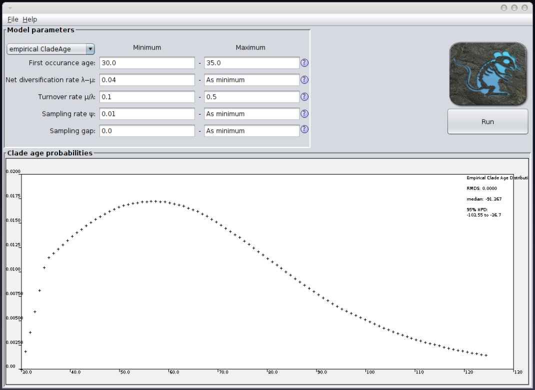
Fossilised Birth Death



Gavryushkina et al, PloS Comp Bio, 2014, package: SA

How to represent fossil calibrations?

CladeAge



empirical CladeAge

	Minimum		Maximum
First occurance age:	30.0	-	35.0
Net diversification rate λ−μ:	0.04	-	As minimum
Turnover rate μ/λ:	0.1	-	0.5
Sampling rate ψ:	0.01	-	As minimum
Sampling gap:	0.0	-	As minimum

Empirical Clade Age Distributi

RMDS: 0.0000

median: -91.367

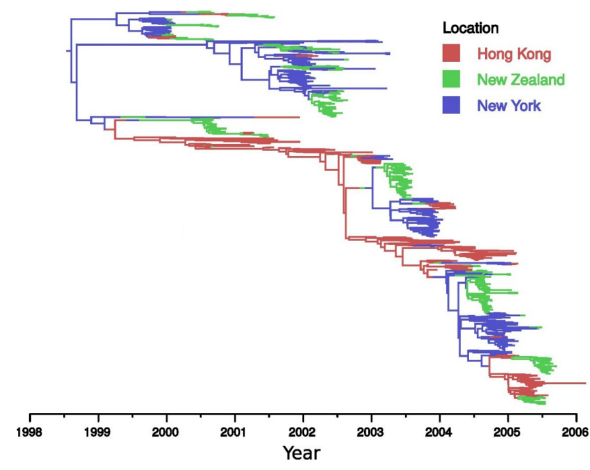
95% HPD:
-103.55 to -36.7

Matschinger et al, Sys Bio, In press, 2016, package: CA

How many states can I have in a structured coalescent ?

- MultiTypeTree, up to 4 demes
- Bayesian Structured Coalescent Approximation: BASTA
- up to 11 demes?
- ongoing work to increase nr of demes

MASCOT, SCOTTI

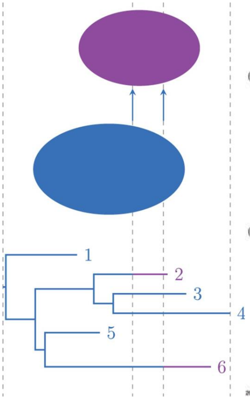


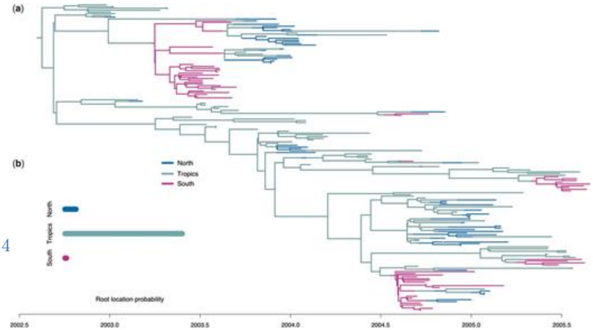
Vaughan et al, Bioinformatics, 2014, package: MultiTypeTree

De Maio et al, PLoS Genet. 2015, package: BASTA

Multitype Birth Death

Phylodynamics with migration

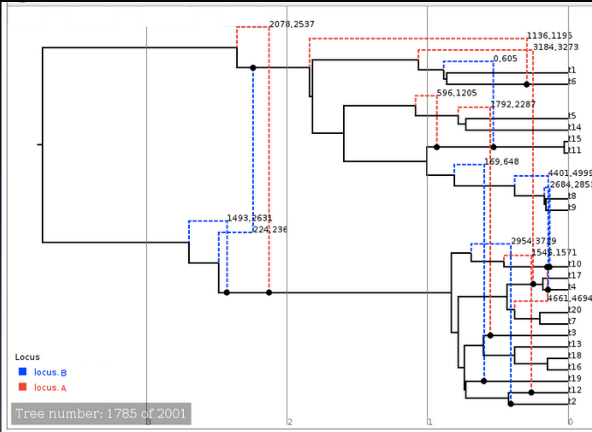




Do I have recombination in my bacterial data?

BACTER

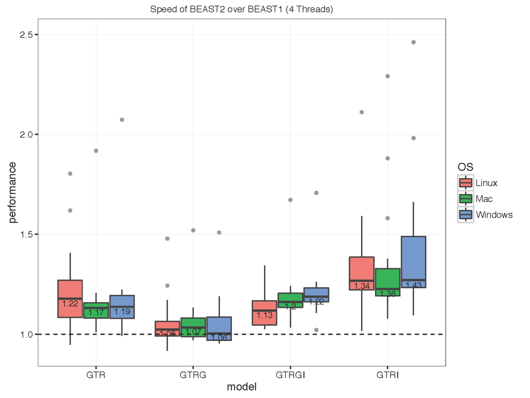
- ancestral recombination graphs
- estimate recombination rate
- estimate expected tract length associated with ARG



Vaughan, bioRxiv?, 2016. Didelot et al, Genetics , 2010. package: BACTER

How to speed things up?

- almost no overhead when using proportion invariant category
- use threads
- use BEAGLE library with -beagle_SSE option



http://beast2.org/2016/04/05/beast-1-vs-2-performance-benchmarking/

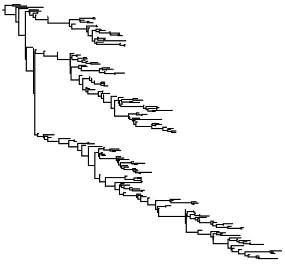
BEASTLabs package

Utilities including

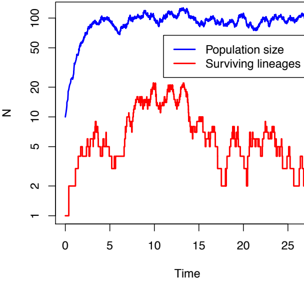
- multi chain MCMC
- MCMCMC
- particle swarm MCMC
- multi monophyletic constraints + appropriate operators
- a spread-sheet like GUI interface to interrogate and edit BEAST 2 models
- multi epoch models

MASTER: Easy stochastic simulation

- package offering easy specification of population genetics models directly inside BEAST 2 XML.



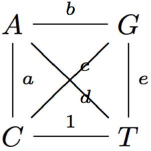
```
<model spec="Model" id="model">
  <population spec="Population" populationName="X" id="X"/>
  <reaction spec="InheritanceReactionString"
          reactionName="Birth" rate="1.0">
    X{1} =: X{1} + X{1}
  </reaction>
  <reaction spec="InheritanceReactionString"
          reactionName="Death" rate="0.01">
    X{1} =: 0
   </reaction>
</model>
```



- Versitile application: generate simulated population size histories, moment estimates, simulated genealogies all from the same model specification.
- Extensible implementation: additional simulation algorithms easy to incorporate.

Vaughan & Drummond, 2014, MBE

Reversible-jump Based (RB) substitution model for nucleotides



$$R = \left| \begin{array}{cccc} - & \alpha & \beta & \gamma \\ \alpha & - & \delta & \epsilon \\ \beta & \delta & - & \omega \\ \gamma & \epsilon & \omega & - \end{array} \right|$$

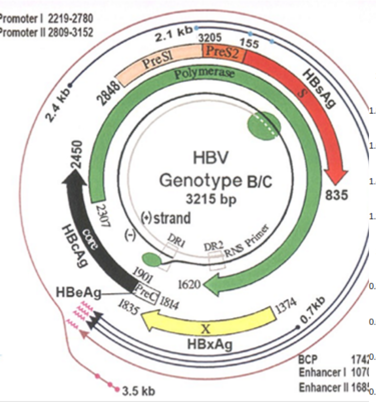
frequencies π
$Q = \pi R$

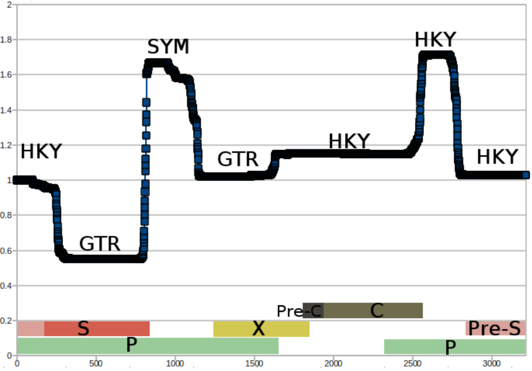
transition prob.
$P(t) = e^{Qt}$

	α	β	γ	δ	ϵ	ω	# dimensions
F81 (JC69)	1	1	1	1	1	1	0
HKY85 (K80)	a	1	a	a	1	a	1
TN93	a	b	a	a	1	a	2
TIM	a	b	c	c	1	a	3
new	a	b	c	d	1	a	4
GTR (SYM)	a	b	c	d	1	e	5

Increase dimension by drawing a new parameter value from $\Gamma(0.2, 5)$

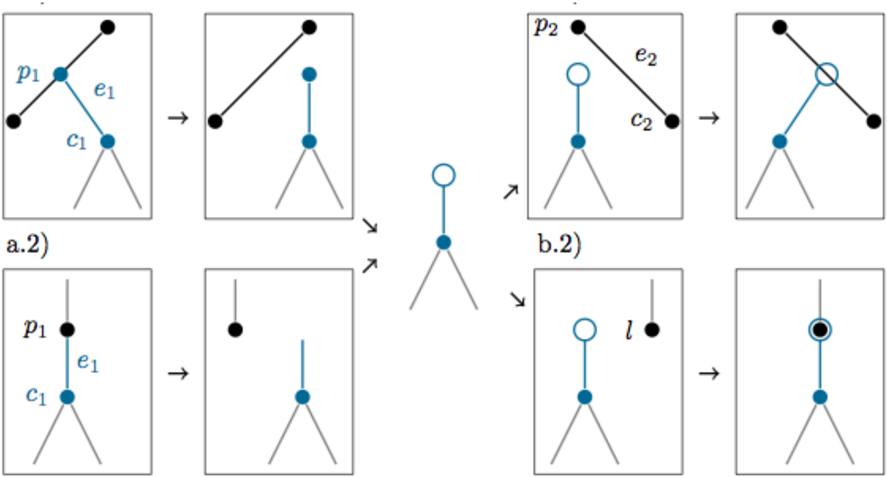
Autopartition + RB substitution model





Test for ancestrality

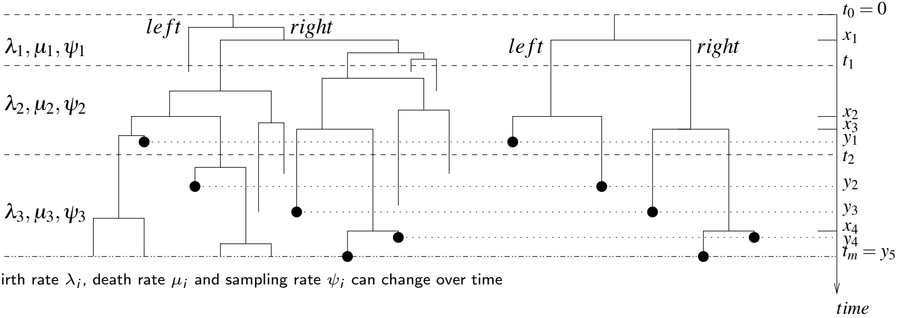
Sampled ancestors model (Gavryushkina et al. 2014) uses reversible jump



Useful for ancient DNA

Open question: how to specify hyper priors

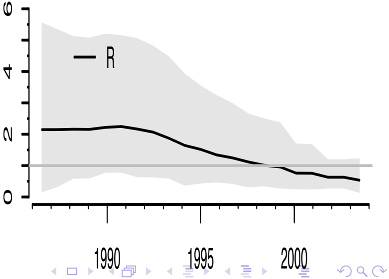
The birth-death skyline model (for serially sampled data)



Birth rate λ_i, death rate μ_i and sampling rate ψ_i can change over time

Reparametrization of birth-death process:

- Reproduction number $R_i = \frac{\lambda_i}{\psi_i + \mu_i}$
- Become-uninfectious rate $\delta_i = \mu_i + \psi_i$
- Sampling proportion $s_i = \frac{\psi_i}{\psi_i + \mu_i}$
- Fixed number of intervals m



OBAMA

- OBAMA for Bayesian Amino acid Model Averaging
- Like bModelTest but for amino acids
- Seems to work well for selecting substitution model, but perhaps less so for rate heterogeneity/invariable sites

BEASTvntr

- Microsattelite data
- Sainudiin
- Sainudiin vanilla
- Sainudiin Computed Frequencies

Break-away model

- Phylogegraphy model
- Assumes one population stays/one population goes
- Implies root is at one of the sample locations
- Does not necessarily be close to the 'centre' of samples

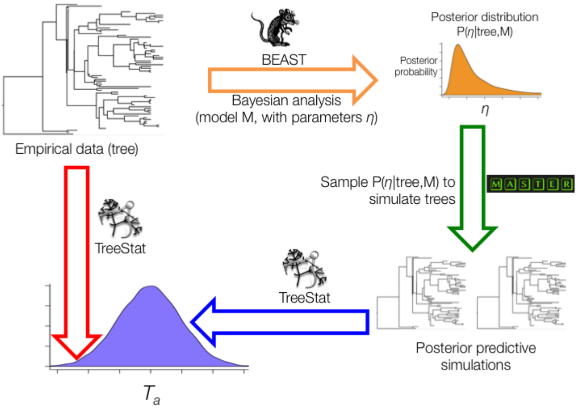
Language Seq Gen

- simulate language data – cognates CTMC, Covarion, SDollo
- allow for borrowing – distinguish between local and global borrowing
- can generate missing data/meaning classes
- useful in simulation studies

DENIM

- Divergence Estimation Notwithstanding ILS and Migration
- = multi species coalescent with migration
- For details, see documentation in pacakge

Tree model adequacy – TMA



- tests whether the tree prior is adequate for the data
- posterior predictive simulation
- differs from model selection:

Duchene et al, under review

Tree Stat

- calculate tree statistics
- B1Statistic BetaTreeDiversityStatistic CherryStatistic CladeMRCAAttributeStatistic CladeMeanAttributeStatistic CollessIndex DeltaStatistic ExternalBranchRates ExternalInternalRatio FuLiD GammaStatistic InternalBranchLengths InternalBranchRates InternalNodeAttribute IntervalKStatistic LineageCountStatistic LongestBranchLength LttSlopeRatio MRCAOlderThanStatistic Nbar NodeHeights NumberOfTips RankBranchLength RelativeTrunkLength RootToTipLengths SAStatistic SamplingTimesInterval SingleChildCountStatistic SingleChildTransitionCounts SummaryStatisticDescription TMRCASummaryStatistic TimeMaximumLineages TopologyStringStatistic TreeHeight TreeLength TreenessStatistic

Others...

- Protein evolution model (US)
- Lie-Markov models/Non-reversible substitution models (Tasmania)
- Protracted speciation (Netherlands)
- Correlated evolution models (UK)
- Density dependent speciation (Zurich)
- Substitution model adequacy (Sydney)
- Tree Set analysis (Auckland)
- ....

BEAST clinic is open