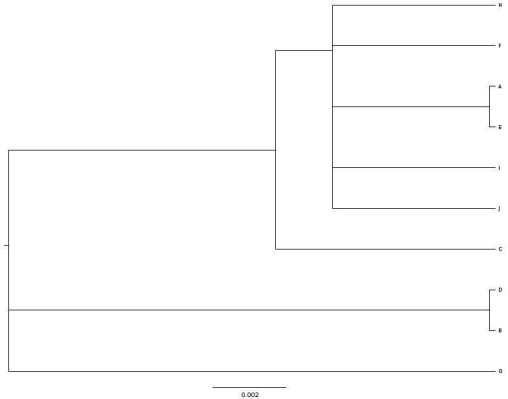
Recent Developments in

Proposal schemes for MCMC sampling of multiple mergers genealogies



Edward A. Salinas
John B. Horne
Rita Castilho

Monday, Nov. 9th, 2015

The importance of MCMC sampling in applied statistical coalescent snalysis

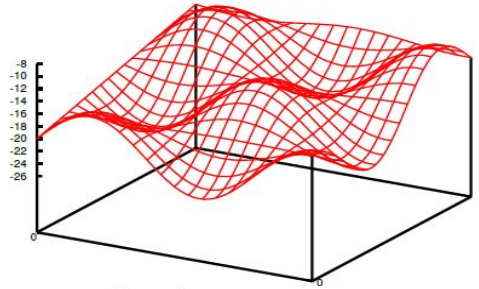
“Use of a prior enables us to interpret the result as the distribution of the quantity given the data.”[1]

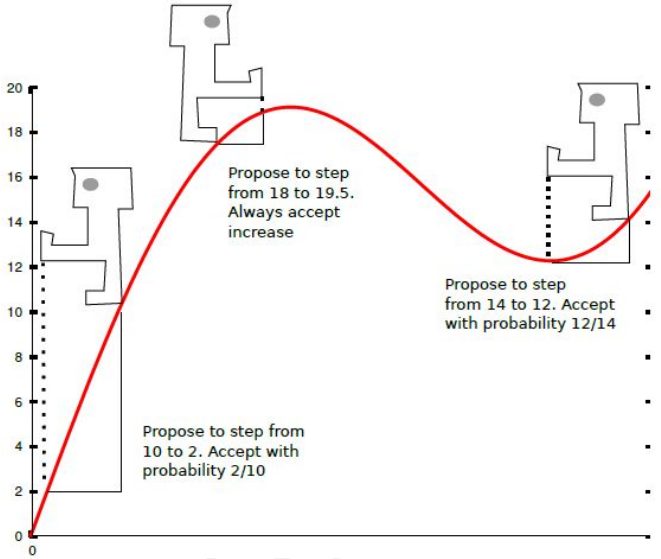
The analysis referred to is a Bayesian one such as MCMC and the distribution refers to the distribution of trees. The MCMC allows for many trees to be considered and ranked according to likelihood.

1. J. Felsenstein - Inferring Phylogenies, 1st Ed., Ch. 18, pg 288

The implementation and efficiency of proposal schemes is important

"It is a fine art to design and compose proposal distributions and the operators that implement them. The efficiency of MCMC algorithms crucially depends on a good mix of operators forming the proposal distribution." [2]





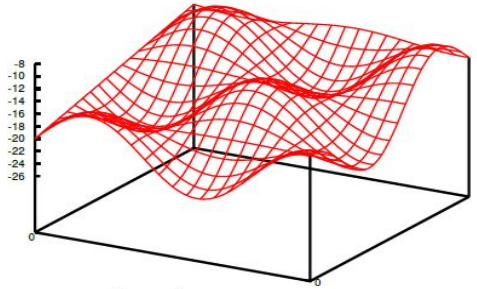
2. Bayesian evolutionary analysis with BEAST Alexei J. Drummond and Remco R. Bouckaert. Cambridge University Press, 2015

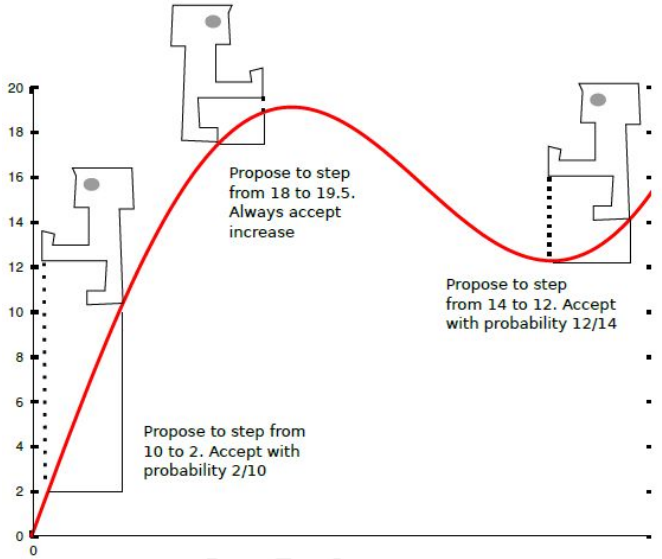
The implementation and efficiency of proposal schemes is important

Better schemes for faster and more effective mixing to achieve runtime efficiency!

All else equal, faster is better!

Help the chain to avoid being stuck in local maxima/modes





Challenges in implementing MCMC proposal schemes of multiple mergers models genealogies

Proposals must be constrained by the multiple mergers model

they can and should have elements of randomness, but not so random as to violate the model and leave it contraints

Proposals must update and maintain parameters of the model and tree

new and old “SREs” (successful reproduction events) must be updated, noted, and recorded properly; proposals must update the parameters of the SRE’s (Y, phi values)

The Proposal acceptance rate should fall in the range ~20%-70% for sufficient “mixing” and acceptable performance

Proposed proposals for multiple mergers models

Schemes presented in the next four slides have been implemented in v0.1.94 of COMET and have shown significant promise to deliver efficiency

- branch-length adjustment
- parent swapping/selection
- overlap extension
- overlap retraction

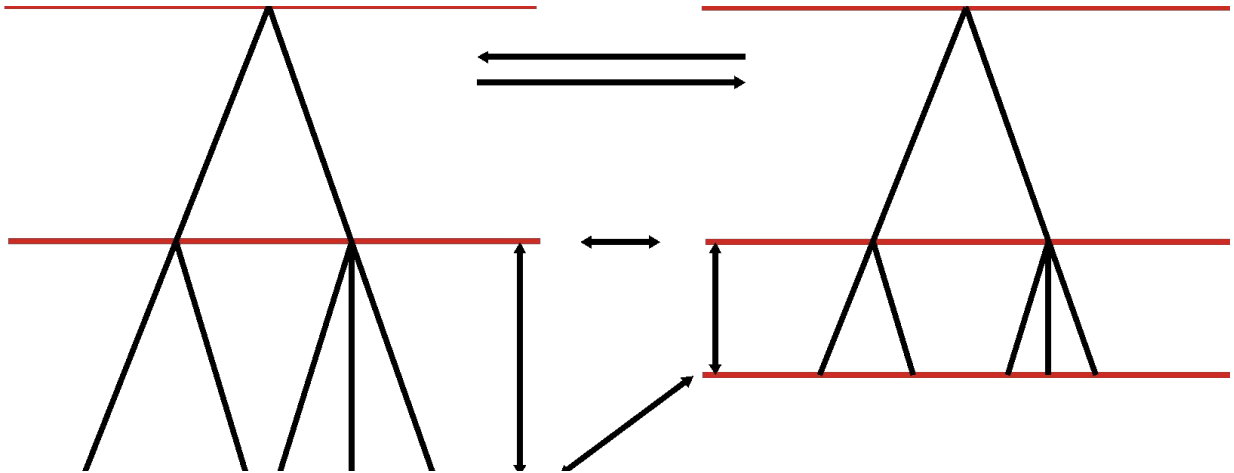
Branch Length Adjustment Proposal

<u>An Intra-SRE Length Change</u>

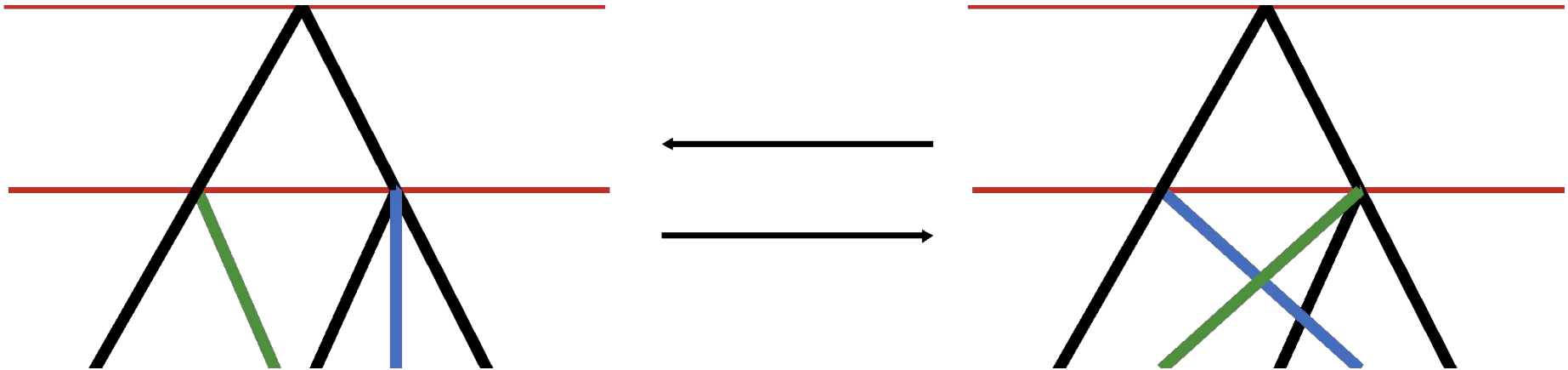
SRE Length/Wait time is adjusted

NO topology changes ; just branch lengths changed

The exponential distribution, with truncation for improved acceptance ratios, has been used



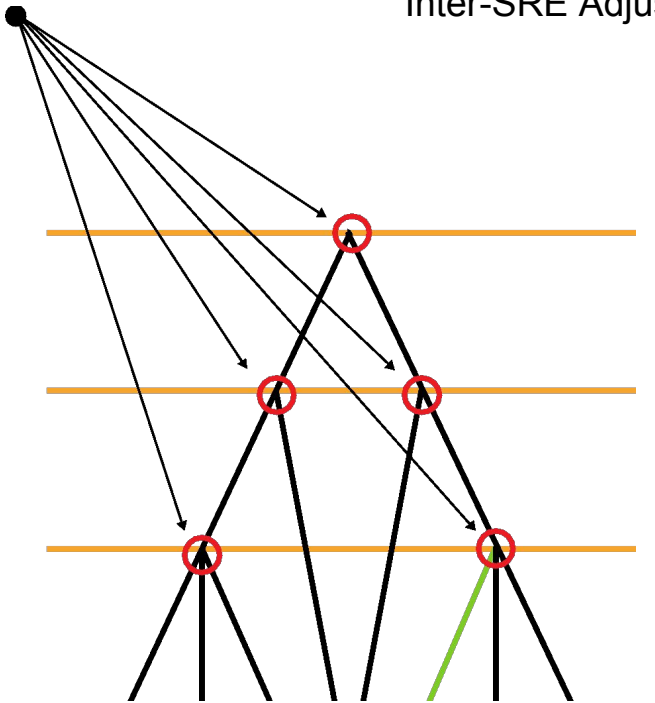
Intra-SRE Lineage/Parent-swapping and reselection



Two lineages “swap parents” or select new parents

Overlap Extension

3 candidate SREs and
5 branch point candidates
here

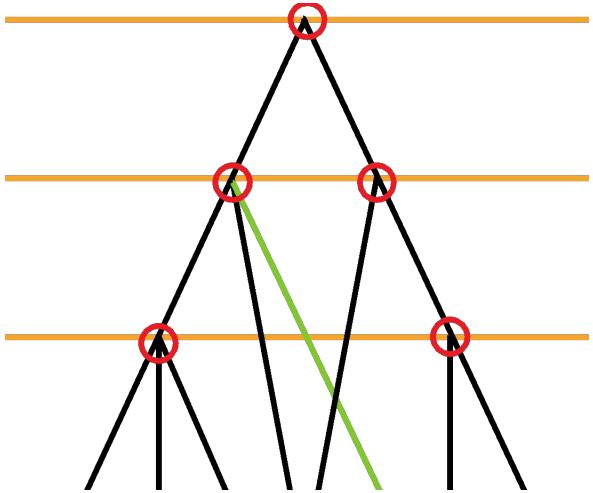


Inter-SRE Adjustment for “Overlap Extension”

SRE #3
Y=1;Φ=1

SRE #2
Y=2;Φ=1

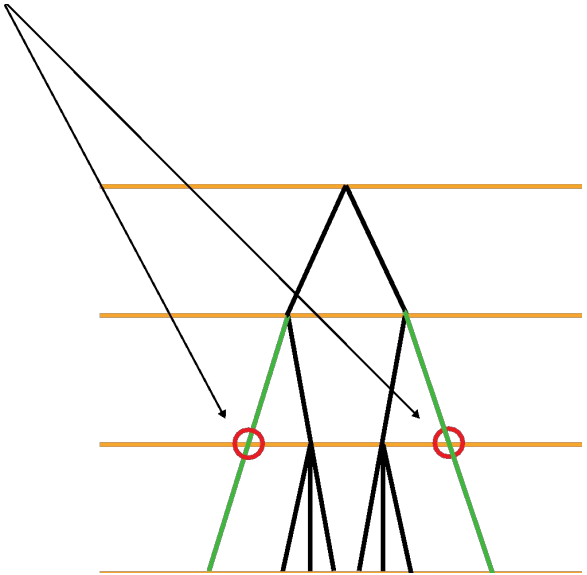
SRE #1
Y=2(pre)
Y=2(post)
Φ=¾(pre)
Φ=⅝(post)
(Φ **decreased**)



Overlap Contraction and Lineage Adoption

Inter-SRE Adjustment Overlap “Retraction” or “Contraction”

Two candidates here

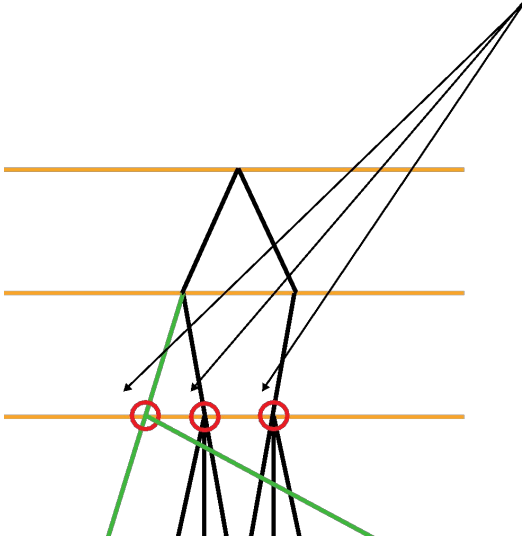


SRE #3
Y=1;Φ=1

SRE #2
Y=1;Φ=1

SRE #1
Y=2(pre) Y=3 (post)
Φ=¾(pre)
Φ=⅞(post)
(Φ **increased**)

Three prospective adoptive parents



Algorithm #1 MCMC code improvement

The aforementioned proposal schemes have been incorporated into COMET v0.1.94 and into the code for the multiple merger model for Algorithm #1 [3]

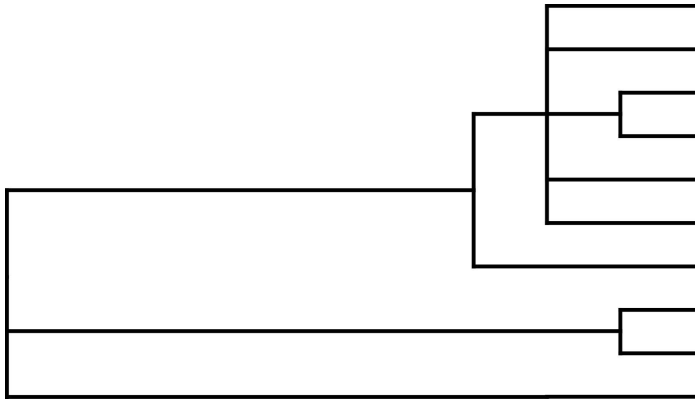
Happily, with a trial-and-error approach, acceptance ratios within the 20%-70% window have been achieved

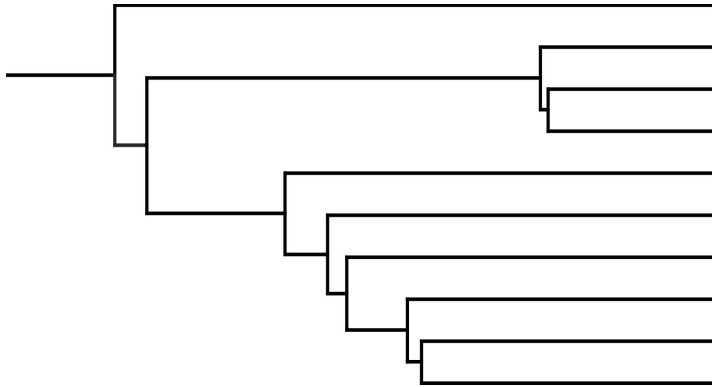
from which we garner some confidence in our approach

Useable in COMET for computing Thermodynamically Integrated Marginal probabilities of P(D|M) where M is the model indicating the space of multiple mergers genealogies induced by algorithm #1

3. Ori Sargsyan, John Wakeley, A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms, Theoretical Population Biology 74 (2008) 104–11

Multiple mergers models may be a better fit for King Threadfin





Algorithm #1 : Multiple Mergers	Kingman	
-564.494888	-607.459840	Most Likely Tree ML Value
-643.387065497	-651.179637941	TI Marginal Likelihood

Introduction of MCMC proposals into a hybrid Kingman-multiple-merger model: Algorithm #2

Algorithm #2 is a hybrid model along a spectrum of multiple-merger and kingman models/genealogies [4]. The *c* parameter indicates the mix.

The Kingman coalescent trees have specific proposal schemes[5, 6]

Just as Algorithm #2 is an Algorithm#1/Kingman hybrid algorithm, so too therefore do we imagine that proposal schemes for it should also be hybrid between Algorithm #1 and Kingman proposal schemes

Such schemes are the next steps and directions for taking COMET!

4. Ori Sargsyan, John Wakeley, A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms, Theoretical Population Biology 74 (2008) 104–11
5. Wakely, John, 2009, Coalescent Theory: An Introduction. Roberts and Company Publishing, 2009
6. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. M K Kuhner, J Yamato, and J Felsenstein Genetics August 1995 140:1421-30

Algorithm #2: the next proposal steps...