Applied Coalescent Theory and Statistical Phylogenetics

- Likelihood Calculations and Mutation Models
- Genealogies
  - Call for Multiple Mergers Genealogies/Models
- COMET *(GLBIO 2015 Poster/Abstract Accepted)*
  - COalescence with Multiple mergers Employing Thermodynamic Integration
  - Thermodynamic Integration (TI) for marginal probabilities, Bayes Factor (BF), and model selection
- Preliminary Analysis Results from Comet
- Milestones, Plans, Next Steps…?

Models of Nucleic Acid Substitution

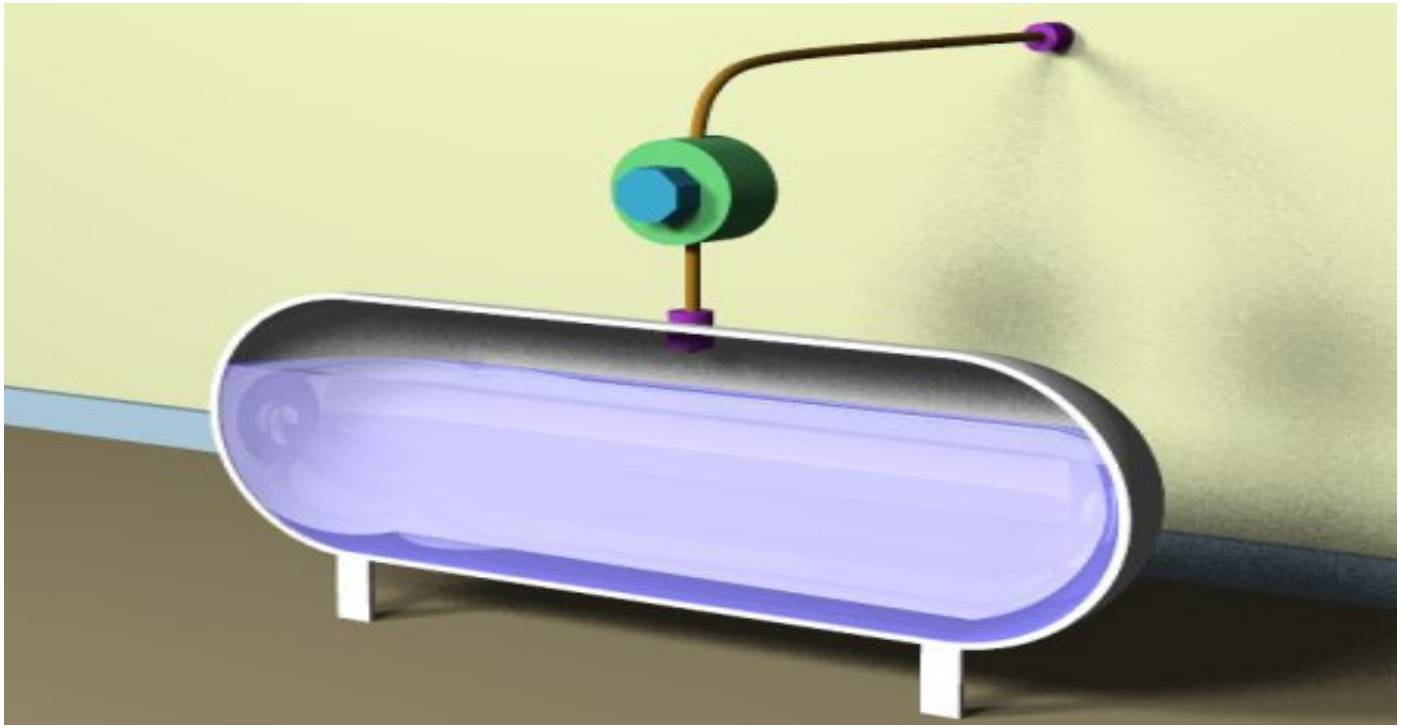


Image from http://www.arachnoid.com/calculus/volume2.html

Nucleic acid substitution models are like having four tanks (one per A,C,G,T, base) of “liquid probability” and a calculus (differential equations) initial-value problem. One unit of “liquid probability” is in the tank of an observed base. As time passes the liquid passes among the tanks, indicating probability of a mutation. Normalization is done to account for fast mutations and long times or slow mutations and short times (rate matrix rows each sum to zero ; these define diagonals) [1].

1. Z. Yang, 2006, Computational Molecular Evolution, Oxford, UK

Likelihood Calculation for Trees

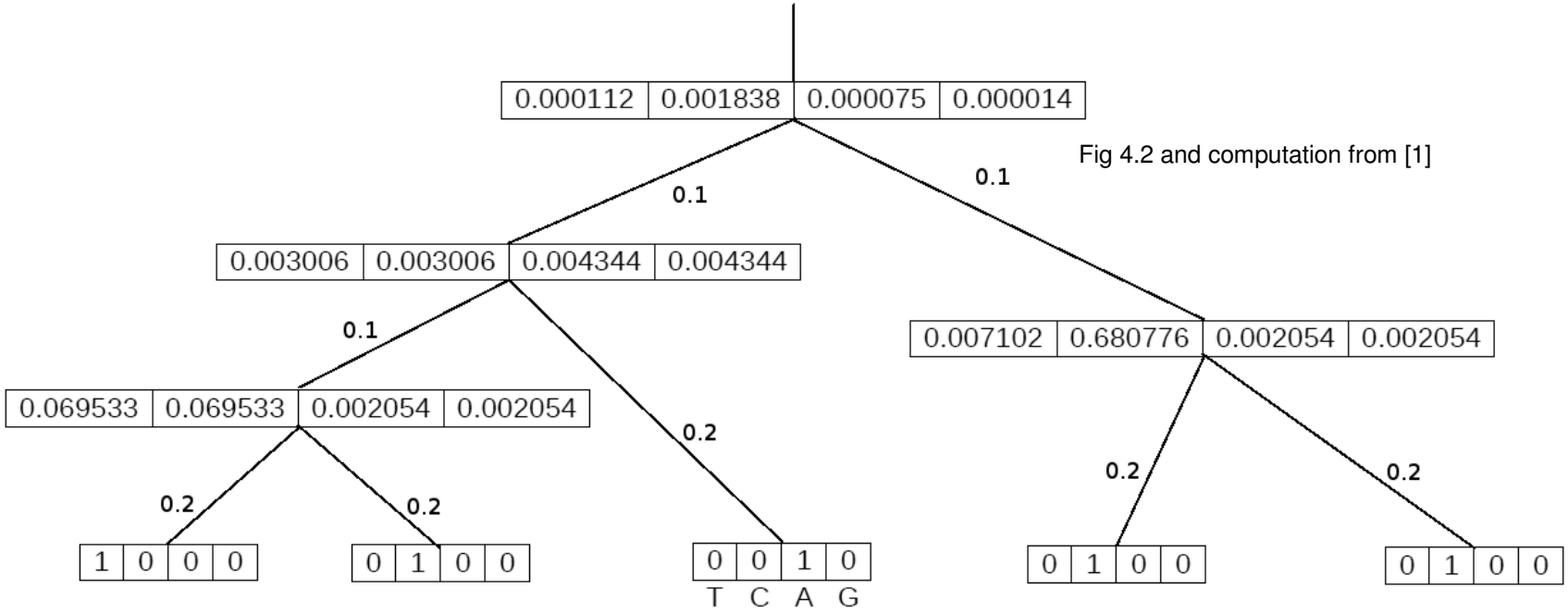


Fig 4.2 and computation from [1]

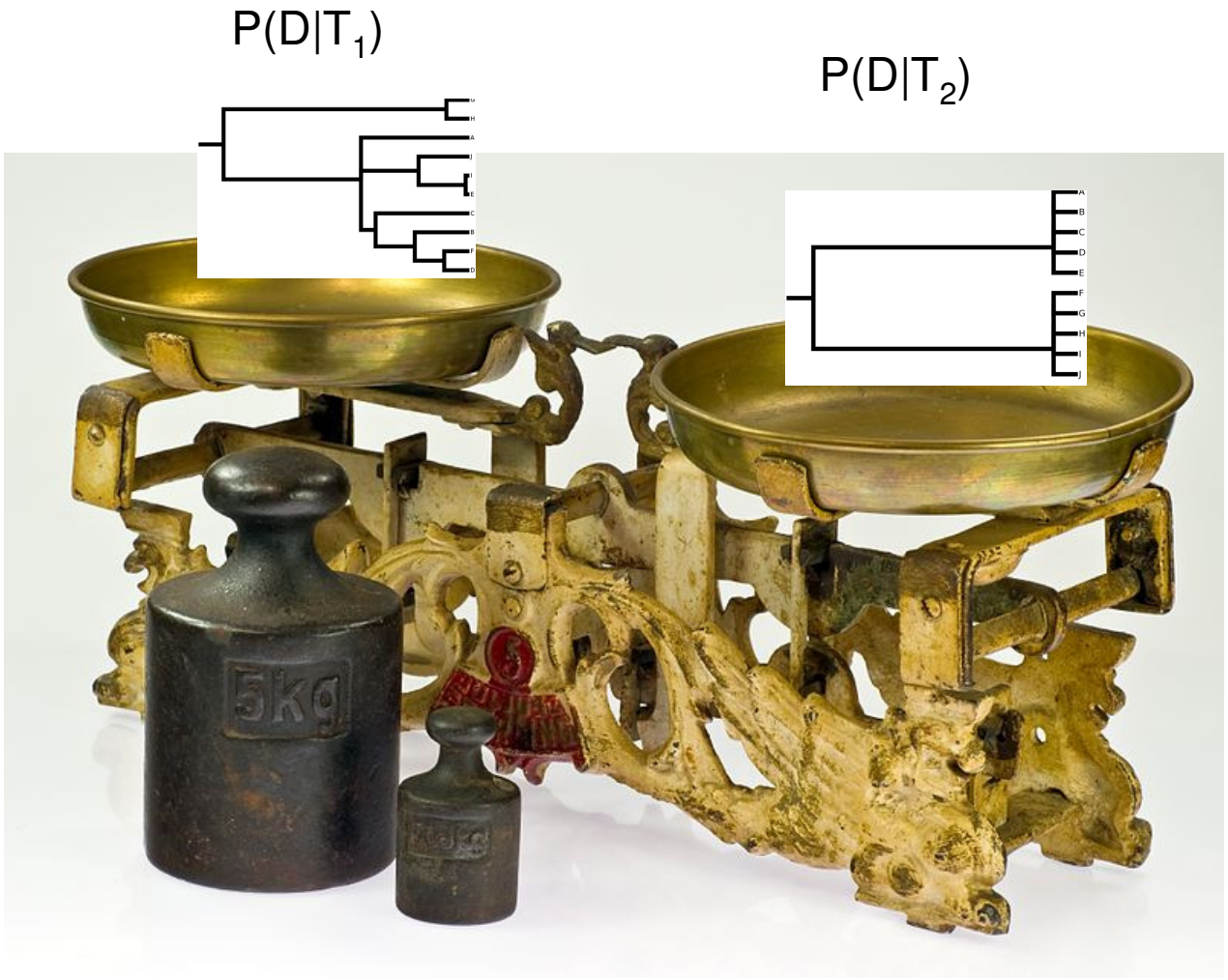
L=0.000509843
ln(L)=(-7.581408)

The "pruning" algorithm [1] uses rules of probability to calculate the likelihood *P(Data|Tree)*

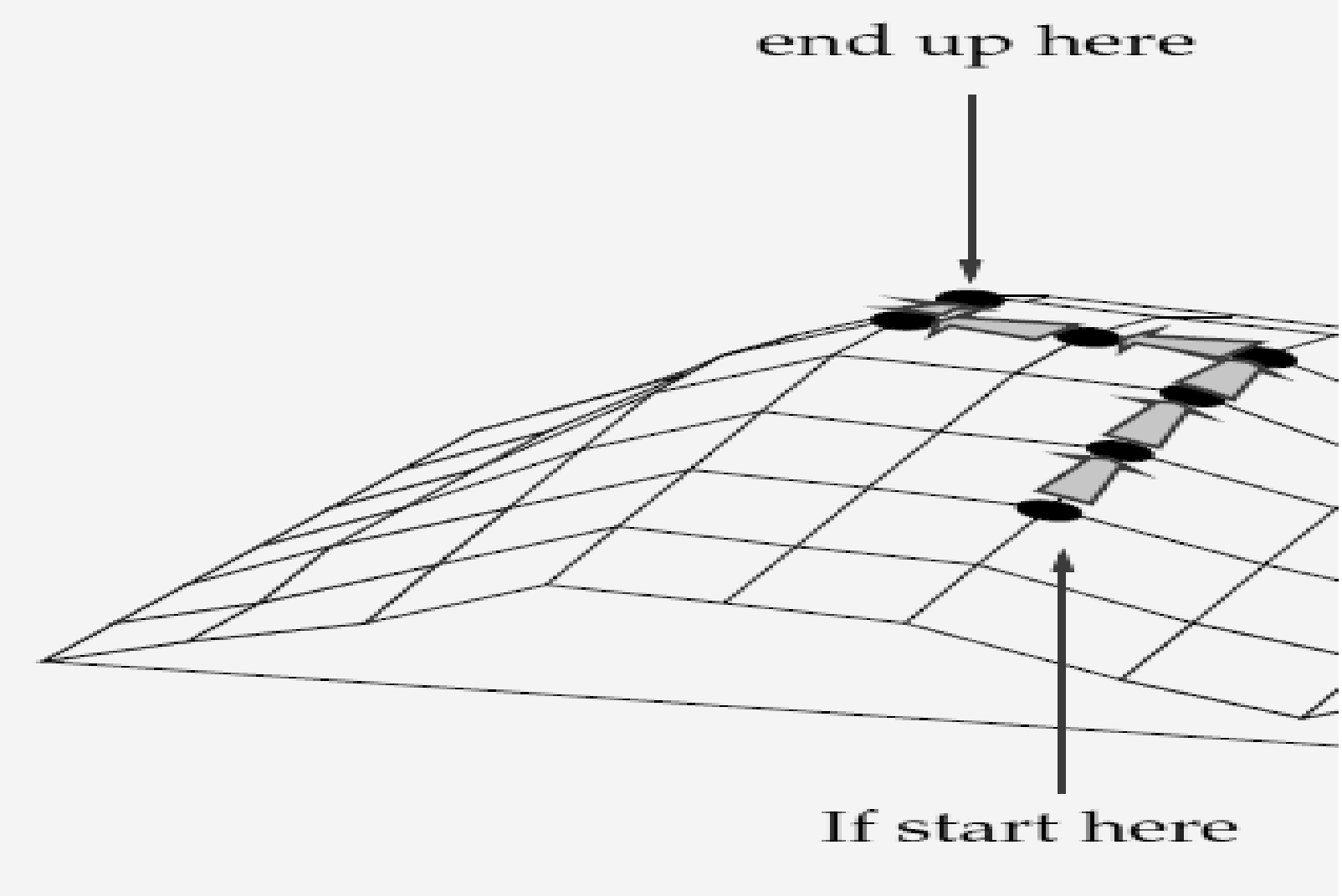
[1] Z. Yang, 2006, Computational Molecular Evolution, Oxford, UK

Likelihood Comparison for Tree Selection

PHYML : Guindon S, Gascuel O., A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst Biol. 2003 Oct;52(5):696-704



PhyML for Maximum Likelihood Tree-space search



The Kingman Coalescent

Based on a certain assumptions about population size

Induces a distribution of genealogy trees given a set of individuals

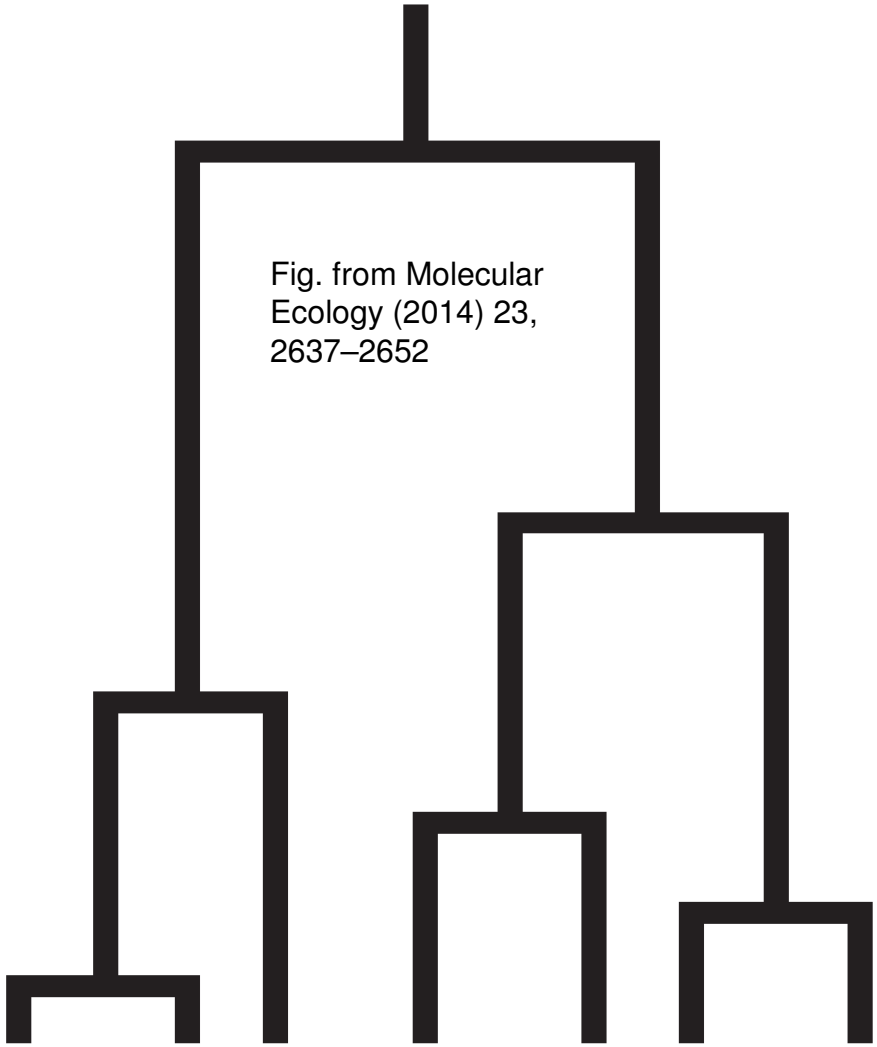
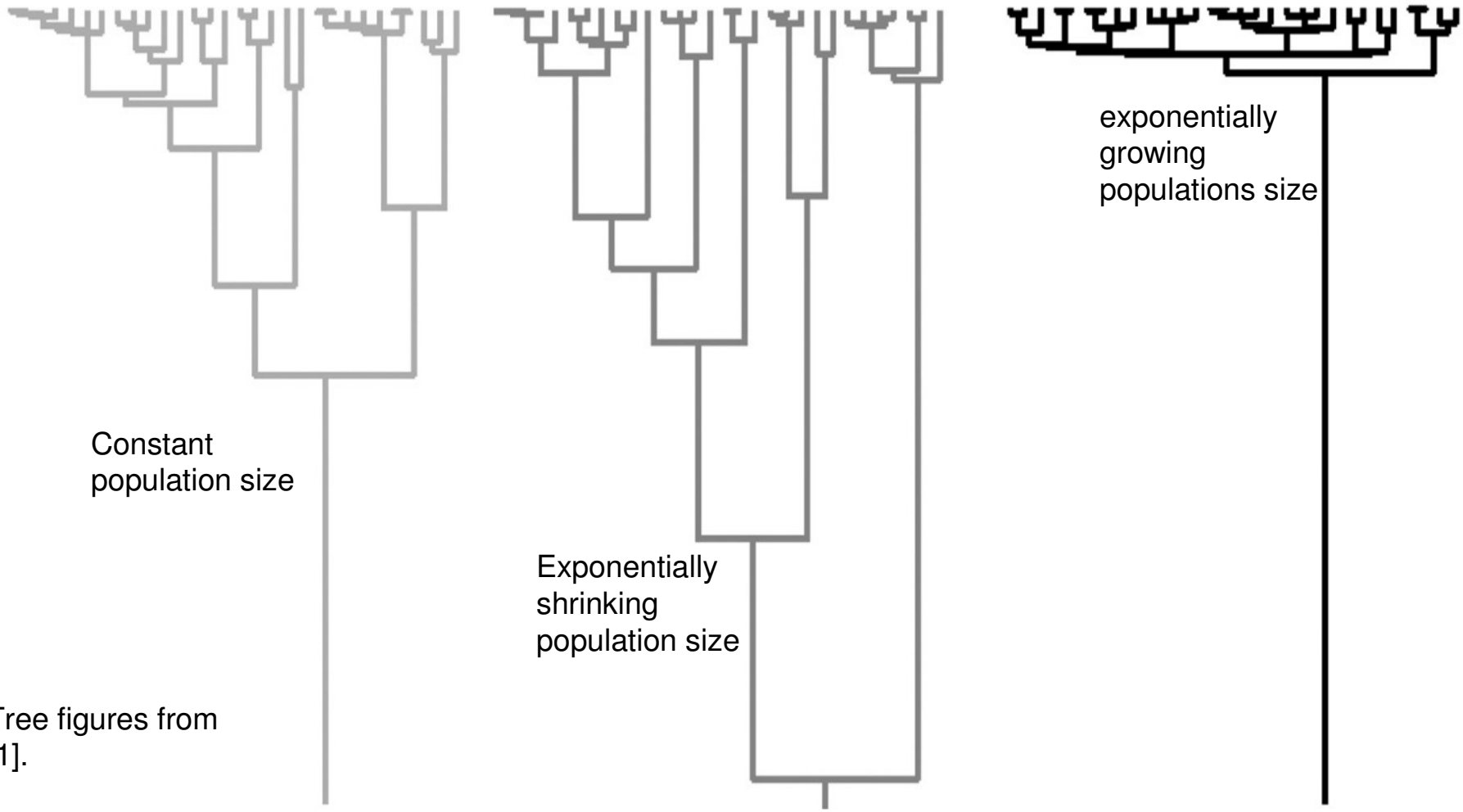


Fig. from Molecular Ecology (2014) 23, 2637–2652

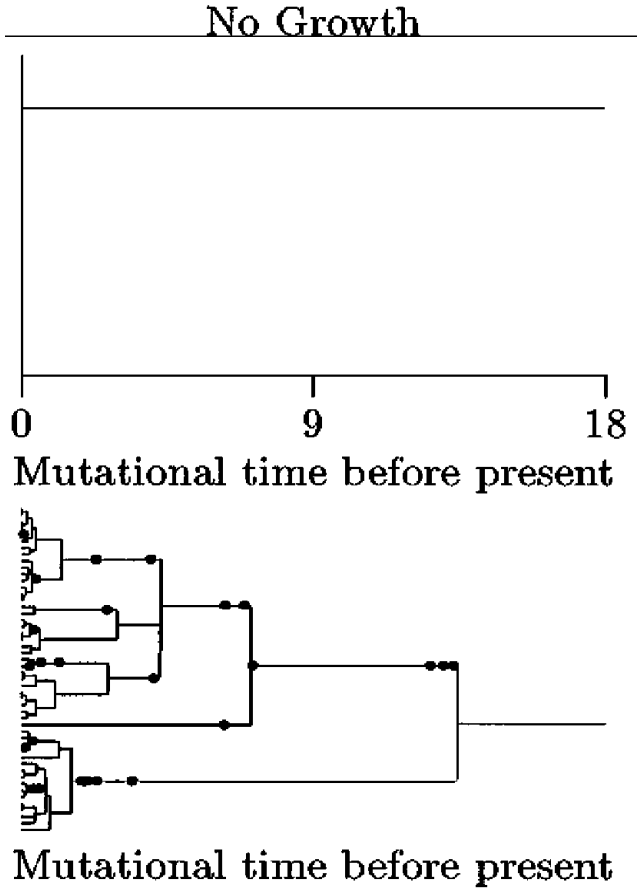
Significance of Branch Lengths in Genealogies

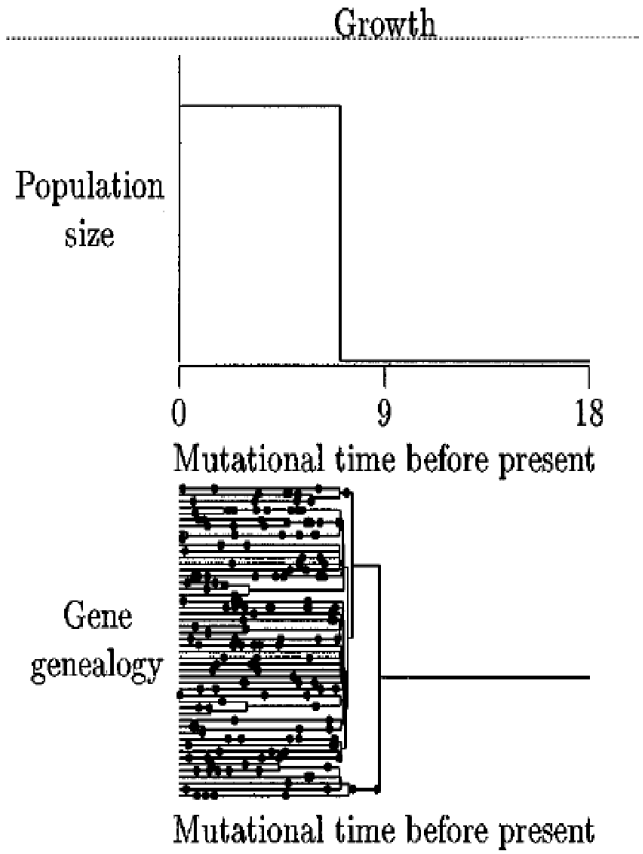


Tree figures from [1].

1. Coalescent genealogy samplers: windows into population history, Kuhner, Mary K. Trends in Ecology & Evolution , Volume 24 , Issue 2 , 86 - 93

Significance of Branch Lengths in Genealogies





H. Harpending and A. Rogers, Genetic Perspectives on Human Origins and Differentiation, Annu. Rev. Genomics Hum. Genet. 2000. 1:361–85

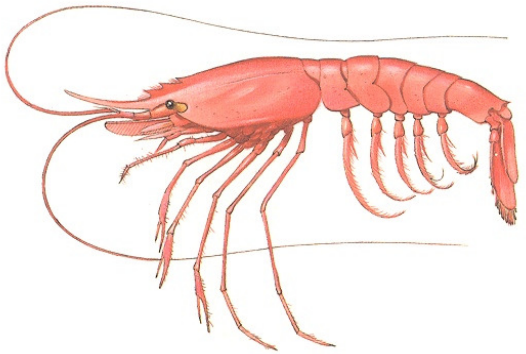
A Call for Multiple Mergers Models

"The large variance in the reproductive success determines that gene genealogies are better described by the so called multiple mergers models, where multiple coalescent events per generation can occur, unlike in the Kingman's n-coalescent……

We also highlight the importance in the next future to extend and apply model based on multiple mergers coalescent to check if they can provide a better interpretation of the data observed. Comparing models based on Kingman's coalescent and the new extension of coalescent theory would be of crucial importance in the next future."[1] (p 12)

1. Marra A, Mona S, Sà RM, D'Onghia G, Maiorano P (2015) Population Genetic History of Aristeus antennatus (Crustacea: Decapoda) in the Western and Central Mediterranean Sea. PLoS ONE 10(3)

2. Tellier A., Lemaire C., Coalescence 2.0: a multiple branching of recent theoretical developments and their applications , Molecular Ecology (2014) Volume 23, Issue 11, pp 2637–2652

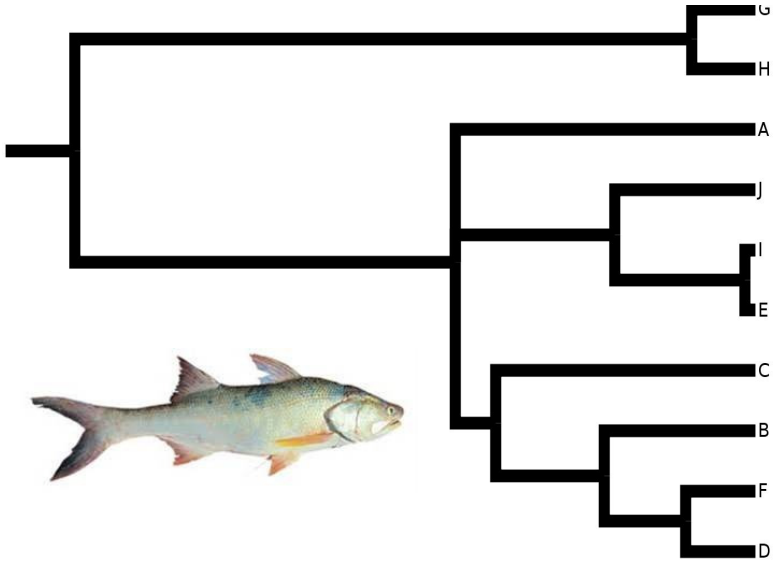


A. antennatus (http://w3.ualg.pt/~madias/geocrust/CamVer-b.jpg)

Algorithms for Multiple Mergers Genealogies

1. 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

Photo of N. Unicornis from John B. Horne, PhD



Tree generated by Algorithm #1[1], Y=2, Φ=0

Tree generated by Algorithm #2[1], Y=1, Φ=.71, c=4

Introducing COMET

MCMC-based sampling of multiple-mergers genealogies

COalescence with
Multiple mergers
Employing
Thermodynamic
Integration

Incorporates Kingman Model and two multiple mergers models (Algorithms #1 and #2 from[1])

Thermodynamic Integration (TI) for marginal probabilities models, Bayes Factor (BF) for models, and then model selection

Collaborations among and with
Edward A. Salinas, Independent Software Engineer/Researcher
John B. Horne, Dalhousie Univ., and
Rita Castilho, CCMAR, Univ. of Algarve, Portugal

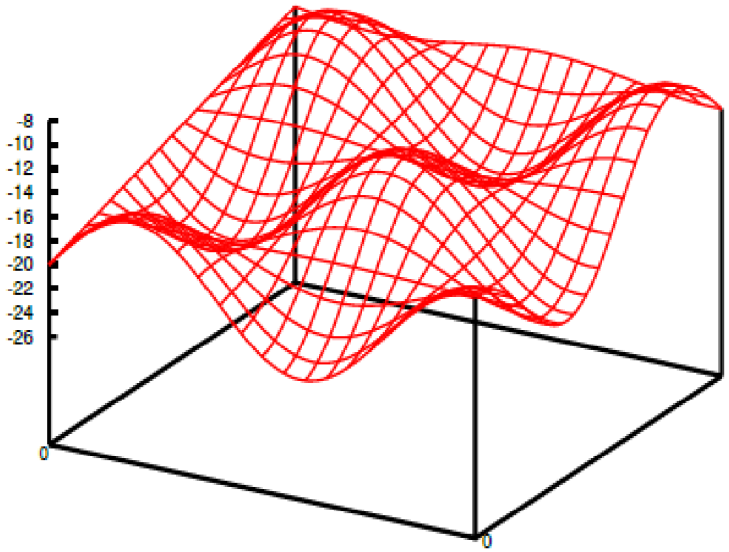
1. Ori Sargsyan, John Wakeley, A coalescent process with simultaneous multiple mergers for Approximating the gene genealogies of many marine organisms, Theoretical Population Biology 74 (2008) 104–11

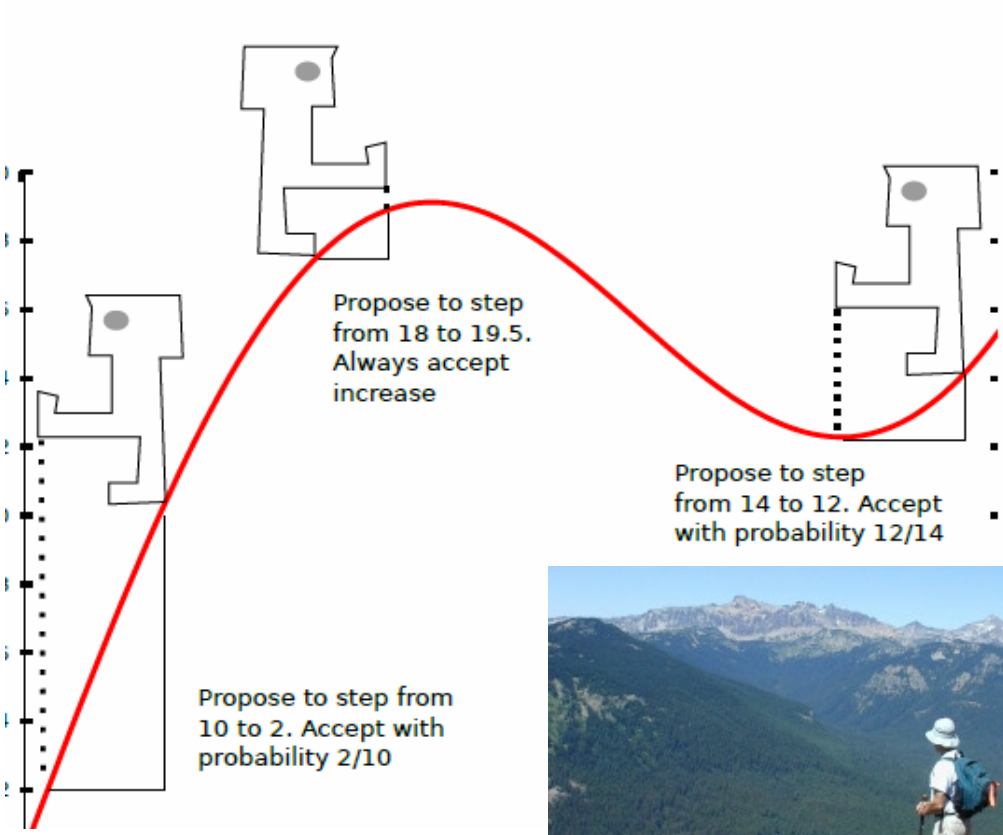
For more information see http://www.eddiesalinas.com/COMET

MCMC Sampling of Trees

A “robot” walks on “a tree likelihood surface” generating logs of trees explored.

Fig. 1.6 from [1]

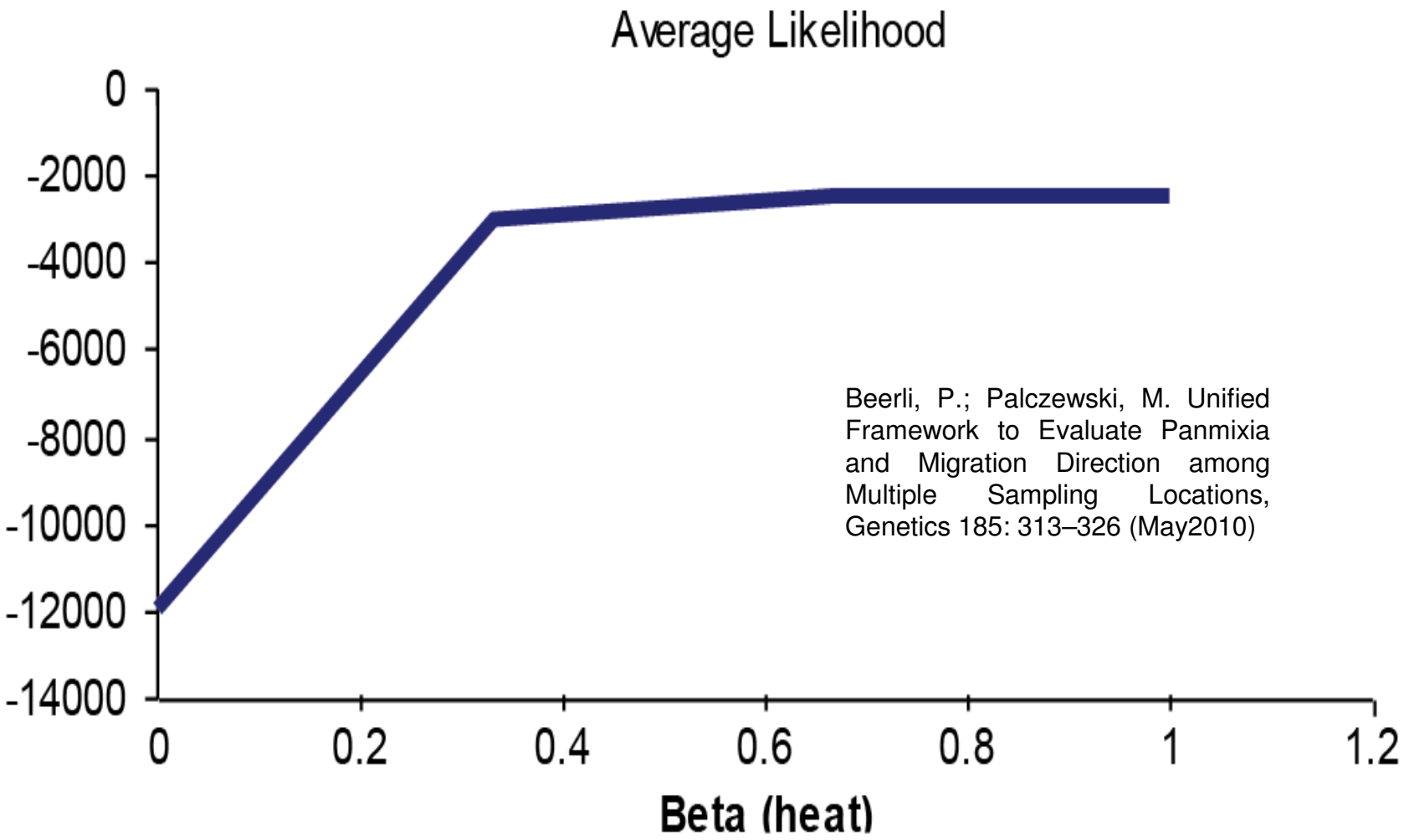




Gifford Pinchot National Forest www.fs.usda.gov

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

Thermodynamic Integration for Marginal Likelihoods : *P(D|M)*



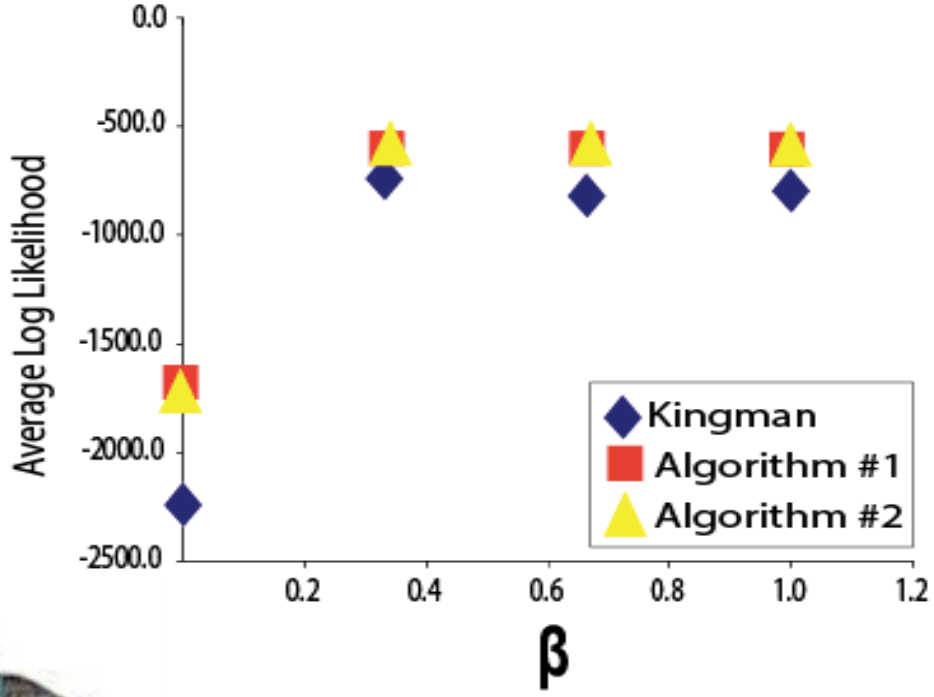
Beerli, P.; Palczewski, M. Unified Framework to Evaluate Panmixia and Migration Direction among Multiple Sampling Locations, Genetics 185: 313–326 (May2010)

Bayes Factors for Model Selection

Not a single trees but *classes* of trees

An Analysis suggesting Multiple Mergers are a better fit

	Marginal Likelihoods	
Model	TI P(D|M)	HM P(D|M)
Kingman	-805.13	underflow/0
Algorithm #1	-618.30	-585.29
Algorithm #2	-625.62	-587.84



Horne, J.B.; Momigliano, P.; Welch, D.J.; Newman, S.J.; van Herwerden, L. Searching for common threads in threadfins: phylogeography of Australian polynemids in space and time, Mar Ecol Prog Ser 449: 263–276, 2012. Genbank Accessions: NT0199K,NT0200K,NT0201K,NT0202K,NT0203K,NT0211K,NT0213K,NT0216K,NT0219K,NT0220K

BEAGLE for Likelihood Calculations

The
Broad-platform
Evolutionary
Analysis
General-
Likelihood
Evaluator [1]

- API for likelihood calculations
- Threads (*OpenMP*), GPUs (CUDA)
- for speed!
- Currently in COMET, but unused/uncalled; needs improved integration

Image Source: Paul Roberts Flickr

[1] D.L.Ayres, et al., BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics, Syst Biol. 2012 Jan; 61(1): 170–173

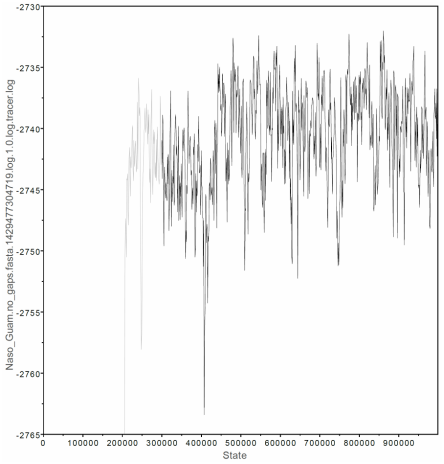
Effective Data Mining by improving COMET's MCMC Mixing/Convergence with New and Improved Proposal Strategies

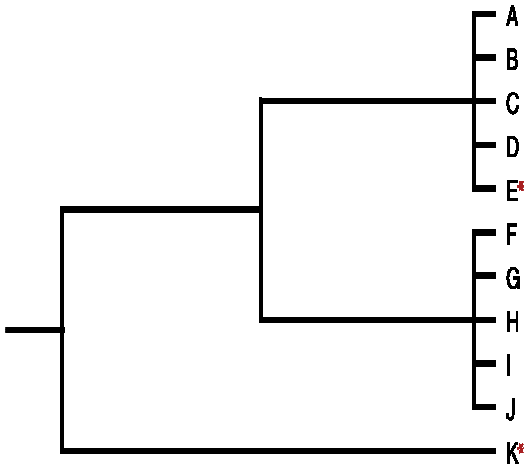
Mixing/convergence

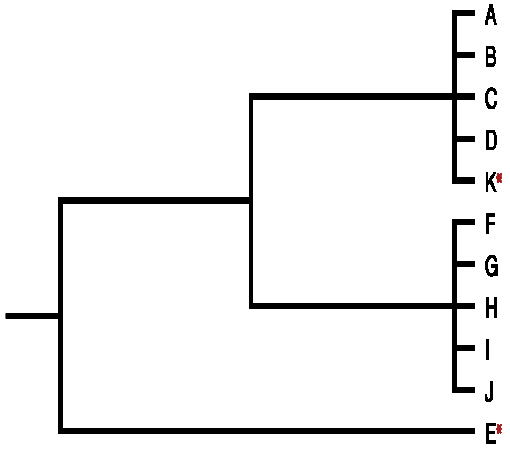
Enough sampling?

Logs viewed with TRACER [2] for assessment

"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." (p. 16) [3]







Figures generated by COMET,FigTree [1], and Tracer [2]

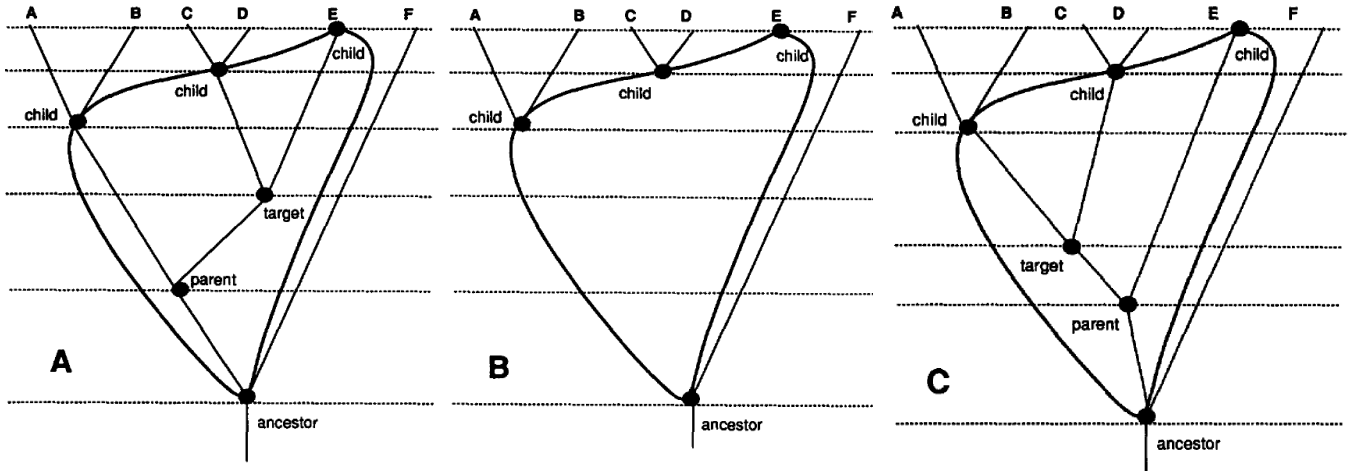
[1] FigTree, available from http://tree.bio.ed.ac.uk/software/figtree/

[2] Rambaut A, Suchard MA, Xie D & Drummond AJ (2014) Tracer v1.6, Available from http://beast.bio.ed.ac.uk/Tracer

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

Effective and Faster Data Mining by improving COMET's MCMC and Mixing/Convergence with improved MCMC Strategies

- PLL
  - The Phylogenetic Likelihood Library [1]
- *MCMCMC=(MC)³*
  - "Metropolis-Coupled Markov Chain Monte Carlo" [2]
  - May help in exploring multi-modal posterior distributions
- Local Neighborhood Rearrangement [3]



[1] T. Flouri, et al., The Phylogenetic Likelihood Library, Syst Biol. 2015 Mar; 64(2): 356–362

[2] Ziheng Y., Molecular Evolution, A Statistical Approach, Oxford Univ. Press, 2014

[3] Kuhner, M. K., Yamato, J. and Felsenstein, J. (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics, 140, 1421–1430

COMET Milestones/Plans

1. GLBIO 2015 Abstract/Poster ✓

May 18-20, Purdue Univ., IN

2. COMET public deployment (in progress)

Improve documentation, sample data, mixing/convergence improvement(s)

3. Winter/Spring COMET Paper **?**

Work planned for Summer/Fall 2015

4. Further enhancements?