**Copy of proposal**

**An improvement - Relative Bayes Factors**

In an attempt to improve statistical stability, we define and utilize in our calculations **a reference model - Mref.** A reference model is a demographic model which is a generalization (to be defined) of the models to be compared. Using Mref we compute the likelihood of every Mi relative to Mref. Then for every pair of models Mi and Mj we can compute their likelihood ratio.

[III]

…

…

…

For now we completely omitted likelihood of parameter priors - , to be handled later in our research. We calculated instead an altered version of formula [III] –

[V]

…

…

…

Formula [IV] is the genealogy likelihood calculation of Mroot under the Kingman Coalescent Model -

[IV]

We extended G-PhoCS to calculate sufficient statistics for Mroot, by having it emit in each iteration the aggregations of and across all loci. This allowed us to later calculate for any .

…

…

…

**TODO**

* Add RBFhat somewhere to indicate that the sampler is not exact
* formal definition of M –
  + Tree with random variable properties (aka priors)
    - Edges have priors for width and migration rate+target, vertices have priors for time
      * Definition that says – when an edge has migrations rates & destinations, when exactly does it start and stop
* Formal definition of Gh & Θh
  + Θh – set of RRs….
  + Gh – definition of a local genealogy. Gh is a group of LGs, one for each locus.
* Formal definition of

**Parameter mappings**

Consider a phylogenetic population model M consisting of a model structure T and its corresponding parameters Θ, M = <T, Θ>.

T is a binary tree describing a hierarchy on populations. Edges in T are populations and nodes are divergence events. Θ is the set of random variable parameter priors of the model. An edge *v* in T has a parameter גv representing population size. A node *k* has parameter *t*k representing divergence time. Some edges *u & v* also have parameters *muv & mvu* representing migration rates between corresponding populations. .

**Definition**: **a merge operation on model M** is the act of contracting two adjacent terminal edges *u*& *v* in T and recalibrating in some manner the relevant parameters - גv, גu, גw, mvu, muv & tw.

**Definition**: a phylogenetic model Mmerge is a **merge-model of M** if Tmerge can be created using only successive merge operations on M.

**Definition**: Assuming Mi is a merge-model of M, we define a **merge-preserving mapping from M to Mi** as a function s.t. -

* is well defined and tractable

**Definition**: a **mock prior of model M** is a random variable , which has no effect on the distribution of genealogies or data in model M.

For the purpose of estimating RBF we construct the following merge-preserving mapping to create M’ref:

* M’ref has the same structure as Mh. set
* M’ref has the same amount of parameters as Θh , created in the following manner-
  + For any edge ,
  + For any two edges , set &
  + For every parameter corresponding to an edge or node which was dropped during a merge, and , created an equally distributing **mock parameter,** and

Let’s rewrite formula [III] with the new reference…

**Solution to “Handling Priors”**

A “model” comes with its own set of parameters, defining prior distributions on population divergence times, population sizes and migration rates. Two models with the same topological structure but with different parameters are distinct. Denote the parameters of model Mh - Θh, and denote its genealogies Gh.

The G-PhoCS MCMC sampler samples genealogies and parameters from the distribution [GΘh|XMh]. We use this distribution to estimate both and . A problem with this approach is that the values we sample from Mh are not applicable to Mref, i.e. P(Θh|Mref) is not well-defined. We solve this by substituting Mref with an altered version of it, M’ref, such that and is well defined and calculable.

Predefinition: a “mock prior” of model Mi is a random variable , which has no effect on the distribution of genealogies or data in model Mh.

The following is a construction of M’ref which assumes :

* Create a model M’ref with a topology equal to the topology of Mref
* Map the parameters of M’ref from parameters Θh in the following manner –

For Every parameter h,

* + if than goes unchanged to ,
  + Else if than create a mock prior with the same distribution as and place it in ,

Our construction preserves the structure of Mref and the parameters which determine data likelihood. Since these are preserved and only mock parameters are added to the model, we get . Since our construction defines a mapping of parameters from Mh to M’ref, we can now calculate .

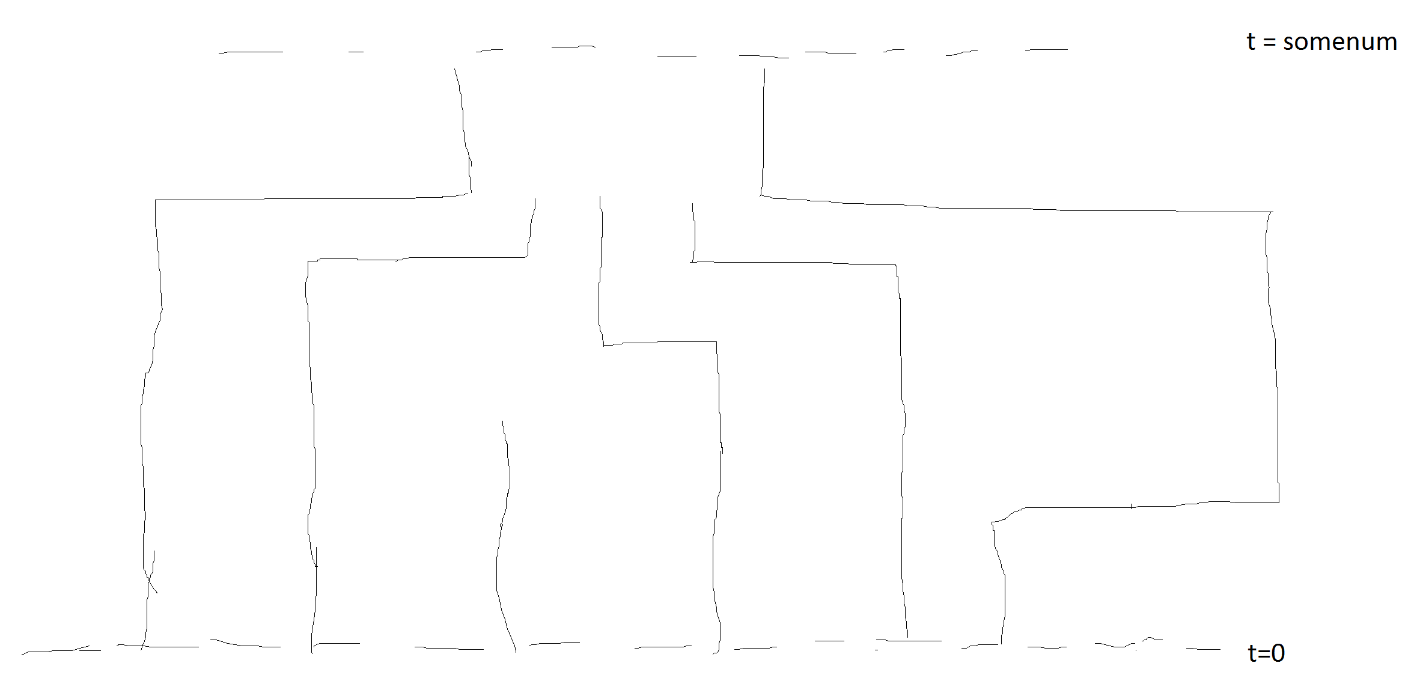
Let’s rewrite formula [III] with the new reference model –

Thoughts about Sufficient Statistics & G-PhoCS

Phylogenetic Topologies

To figure out how to expand the sufficient statistics for support of general reference models, we need to decide what these ‘general reference models’ look like.

We know reference models are generalizations of models. So what can models do?



1. Populations split into two new populations
2. Populations change size
3. Two populations “migrate” (exchange individuals)

Okay, that’s too many possibilities. For now we’ll allow only divergences- Instead of supporting ‘general reference models’ we’ll support only reference models which can split (see operation [a]) into the hypothesis model.

When collecting sufficient statistics in gphocs we will –

1. For every population in Mh, calculate sufficient statistics using formula [IV]
2. Merge (the opposite off split [TODO – explain merge & split]) sibling populations pi and pj into a single population pij and calculate sufficient statistics for it.
3. Repeat step 2 until all populations are merged and Mroot is reached.
4. Emit all statistics for post-processing by MC3

Below is an illustration of supported reference models for some model Mh -

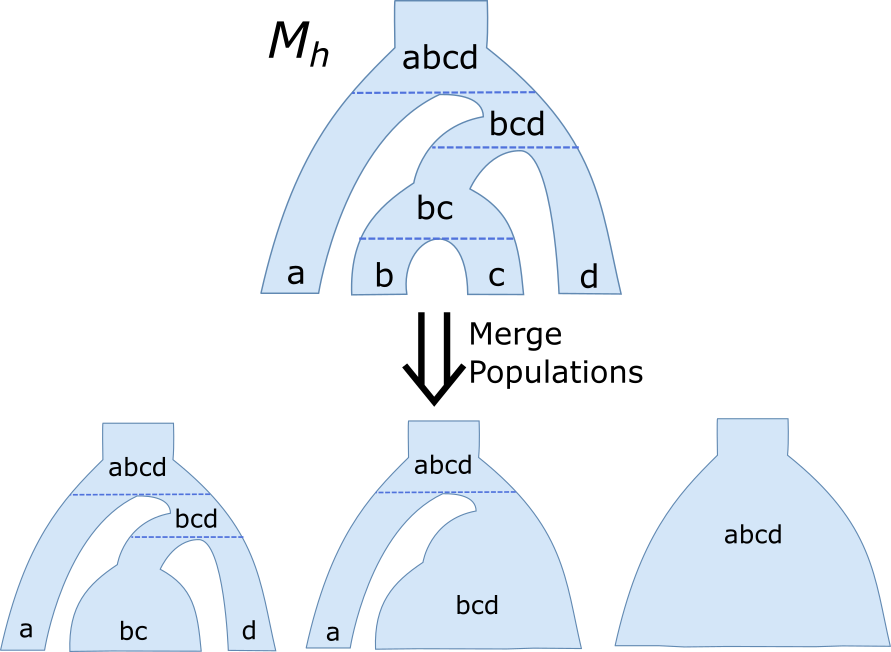


Figure – Generating the Set of Candidate Reference Models of Mh