In our model-selection problem we are given two (or more) phylogenetic population models, and aligned DNA data, and are required to select which model better describes the data.

In our work we introduce a “reference model”. This is a less restrictive, intermediate population model, which we utilize in the comparison of the two target models –

A population model describes generation of DNA using a layer of hidden random variables. These are the model parameters (population sizes, divergence times & migration rates) and the specific loci genealogies.

The G-PhoCS framework uses an MCMC algorithm to simulate integration over the parameters of Mh. In our model-selection algorithm we attempt to use this integration on ph (parameters of Mh) to also, during a single execution of the G-PhoCS MCMC sampler, integrate over pr. The following chapter lays down the justifications for integrating over a ‘small’ vector space (pr) by employing integration on a larger vector space (ph), this using an extension of the small space and a mapping between the two spaces.

Consider two models, Mh and Mr. Mh has j parameters y1..j and Mr has i params, z1..i. Assume i<j. denote the models’ vector space of parameters Ph & Pr.

Given are i functions { *fi* }, each mapping a parameter from Mh to a parameter of Mr - *fi*:<h1..j>->ri.

The aggregate function F:<h1, h2, …,hj> ---> < *f*1(<h1..j, …,hj >),..., *f*i(<h1..j, …,hj >)> is a surjective (but not bijective) function from Ph on Pr. This F is a mapping from the parameter space of Mh to that of Mr

We now artificially extend Pr to a j-dimensional space by adding j-i random variables to Mr. denote this new space P’r.

We extend F to a bijection by creating j-i more j-dimensional functions {*fi+1, ..*fj}, such that F’:<p1, p2, …,pj> ---> < *f*1(<p1..j, …,pj >),..., *f*i(<p1..j, …,pj >, *f*i+1(<p1..j, …,pj >),..., *f*j(<p1..j, …,pj >)> is a bijective function from Ph on Pr’.

Let’s now assume, cause I need it, that is an injective differentiable function with continuous partial derivatives, the Jacobian of which is nonzero for every  in . Denote , the Jacobian of .

Given also is a conditional probability function which holds

1. ~~If for some , then there exist s.t. for in .~~
2. ~~For every assignment to r~~~~1..i~~~~, the integral over (integrating over all assignments to ) is 1.~~

NOW FOR THE MATH:::

\*- extend Pr to P’r

\*\*- change integration variables from hi to *f*(hi)

\*\*\*-

In the above formula, *K* states the phylogenetic model selection problem. X is the aligned Dna Data. Mh is the hypothesis model and Mr is a reference model, used as a connector to later select between Mh and some other hypothesis model Mh’.

The models are parameterized by their genealogies Gh and Gr and by the population tree priors . For convenience, denote the parameters of models Mh & Mr as vectors . Denote the sizes of the vectors respectively. Since Mr is a reference model, assume .

To calculate we integrate over the models’ parameters -

We now artificially increase the dimension of the integral to by integrating over more mock-parameters, . These are “parameters” but have no effect on data likelihood. We also introduce a probability function, , and integrate over these factors -

Let’s now assume we have a mapping function from , the parameters of , onto , the parameters and mock-parameters of . This mapping is an injective differentiable function with continuous partial derivatives, the Jacobian of which is nonzero for every . Denote , the Jacobian of -

Using the Mapping Function, we change the integration variables from to -

Let’s also expand the denominator of using bayes theorem:

Finally we complete our model-selection equation: