Research!

June 11, 2016

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 this work we introduce the notion of a "Reference Model". This, in short, is a less restrictive, intermediate population model, which we utilize in the comparison of the target models –

$$K := \frac{P(X|M_r)}{P(X|M_h)}$$

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

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

Consider two models, M_h and M_r. Model M_h has j parameters, $z_{1..j}$ and model M_r has i parameters, $y_{1..i}$. Assume $i < j$. Given are i functions $\{f_i\}_{1..i}$, each mapping the parameters of M_h to a parameter of M_r — $\mathbf{f_i} : \langle \mathbf{y_{1..j}} \rangle \mapsto \mathbf{z_i}$.

Denote F the application of all $\{f_i\}$ functions —

$$F : y_{1..j} \mapsto \langle f_1(y_{1..j}), .., f_1(y_{1..j}) \rangle$$

This F is a surjective ~~, differentiable, with continuous partial derivatives~~ mapping from M_h on to M_r. We intend to use F during a change of variables from $y_{1..i}$ to $z_{1..j}$, so F must be turned into a bijection. We do this by creating $j - i$ new mock-parameters (parameters who have no effect on data likelihood) of M_r, $\{z'_{i+1..j}\}$, and $j - i$ new j-dimensional functions $f'_{i+1}, ..., f'_j$ onto the new parameters, such that F' is a bijection —

$$F' : y_{1..j} \mapsto \langle f_1(y_{1..j}), ..., f_i(y_{1..j}), f'_{i+1}(y_{1..j}), ..., f'_j(y_{1..j}) \rangle$$

During the creation of the new random variables we are also given a conditional probability function $P(z'_{i+1..j}|z_{1..i})$, to be used when adding the mock-parameters to the integration.

We now apply these structures to our Model-Selection problem —

$$\mathbf{K(M_h, M_r|X)} := \frac{\mathbf{P(X|M_r)}}{\mathbf{P(X|M_h)}} =$$

$$= \int \frac{P(z_{1..i}, X|M_r)}{P(X|M_h)} dz_{1..i} = \int \frac{P(z_{1..i}, X|M_r)}{P(X|M_h)} \int P(z'_{i+1..j}|z_{1..i}) dz'_{i+1..j} dz_{1..i} =$$

$$= \int \frac{P(F(y_{1..j}), X|M_r)}{P(X|M_h)} P(f'_{i+1..j}(y_{1..j})|f_{1..i}(y_{1..j})) J_{F'} dy_{1..j} =^* \int \frac{P(F(y_{1..j}), X|M_r)}{P(X|M_h)} PF'(y_{1..j}) J_{F'} dy_{1..j} =$$

$$= \int \frac{P(F(y_{1..j}), X|M_r)}{P(y_{1..j}, X|M_h)} PF'(y_{1..j}) J_{F'} P(y_{1..j}|X, M_h) dy_{1..j} = E_{y_{1..j}|X,M_h} \left[\frac{P(F(y_{1..j}), X|M_r)}{P(y_{1..j}, X|M_h)} PF'(y_{1..j}) J_{F'} \right]$$

$$^* PF'(y_{1..j}) := P(f'_{i+1..j}(y_{1..j})|f_{1..i}(y_{1..j}))$$