Gaussian Phylogenetic reconstruction of correlations

Alejandro Lage-Castellanos

February 8, 2018

1 Direct problem: phylogenetic tree under gaussian evolution

We consider a genome characterized by a vector $\vec{x}$ with an equilibrium probability distribution given by

$$P(\vec{x}) = \frac{1}{Z} \exp H(\vec{x}) \qquad \text{with } \mathcal{H}(\vec{x}) = \vec{x}^T J \vec{x} + \vec{h} \cdot \vec{x} \tag{1}$$

where the degrees of freedom $x_i \in \mathbb{R}, i \in [1 \dots N]$ are continuous variables, and the interactions among them J_{ij} are fixed, but randomly selected from a given ensemble of symmetric negative definite matrices. Notice the absense of the usual minus sign in front of the Hamiltonian. So, usually the vector $\vec{x}$ will be near the maximum of $\mathcal{H}(\vec{x})$

We will assume that the matrix J have some eigenvalues close to zero, so the genetic sequences can drift with relatively low cost in the subsapce generated by such eigenvectors.

1.1 Evolution process

We consider a phylogenetic tree construction starting from a vector $\vec{x}_0$ extracted randomly from the equilibrium distribution (1). Then at every point of our evolution algorithm we do the following

The step named Monte-Carlo-evolve($\vec{x}, d$), carries out a Monte Carlo evolution starting at configuration $\vec{x}$ and until the configuration found $\vec{x}^t$ is at distance

$$d = \frac{1}{N} ||\vec{x} - \vec{x}^t||_2$$

This means that all new generation sequences are at the same distance of their parents.

Algorithm 1 Generates tree of evolved vectors $T = \{\vec{x}_0, \vec{x}_{1,1}, \vec{x}_{1,2}, \ldots\}$

Require: Root $\vec{x}_0$, number of generations G, distance to mutants d.
Ensure: Returns T
  $T = \{\vec{x}_0\}$
  $L = \{\vec{x}_0\}$ {Leaves of the tree (the last added nodes)}
  for $(g = 1; g < G; g+ = 1;)$ **do**
    $L_{new} = \{\}$
    for ($\vec{x}$ in L) **do**
      $\vec{x}_{\text{child1}} = \text{Monte-Carlo-evolve}(\vec{x}, d)$
      $\vec{x}_{\text{child2}} = \text{Monte-Carlo-evolve}(\vec{x}, d)$
      Append $T \leftarrow= \{\vec{x}_{\text{child1}}, \vec{x}_{\text{child2}}\}$
      Append $L_{new} \leftarrow= \{\vec{x}_{\text{child1}}, \vec{x}_{\text{child2}}\}$
    end for
    $L = L_{new}$
  end for
  return $T = \{\vec{x}_0, \vec{x}_{1,1}, \vec{x}_{1,2}, \ldots\}$

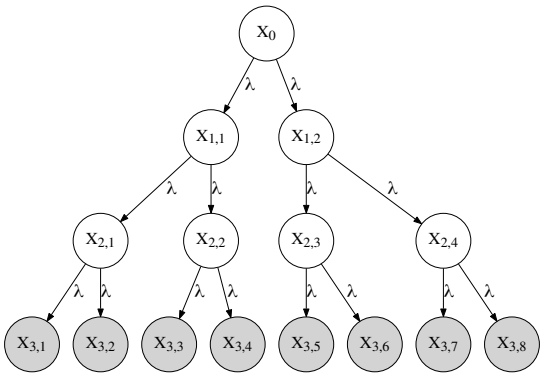

Figure 1: Phylogenetic model for genetic drifting

1.2 Likelihood of observed sequences

In figure 1 we represent the phylogenetic tree generated by an evolution process as described in algorithm 1. The observed sequences are those of the last generation $\vec{x}_{G,\cdot}$ and are portrayed in gray. We assume as known also the model $J, \vec{h}$ and the distances d at which the tree was created, but we ignore the set of sequences $\vec{x}_{g,\cdot}$ in previous generations $g < G$.

Let us define the part of the interactions concerning a node and its children as follows

$$K(\vec{x}_0, \vec{x}_{1,1}, \vec{x}_{1,2}) = \exp\left(\mathcal{H}(\vec{x}_0) + \lambda_0 \vec{x}_0 \cdot \vec{x}_{1,1} + \lambda_0 \vec{x}_0 \cdot \vec{x}_{1,1}\right) \tag{2}$$

where the coupling constants λ are there to ensure that the distance between the father and each children is consistently fixed to have the given expected value.

The full probability distribution of the process in figure 1 is

$$\begin{aligned} P(T) &= \frac{1}{Z} K(\vec{x}_0, \vec{x}_{1,1}, \vec{x}_{1,2}) K(\vec{x}_{1,1}, \vec{x}_{2,1}, \vec{x}_{2,2}) K(\vec{x}_{1,2}, \vec{x}_{2,3}, \vec{x}_{2,4}) \\ & K(\vec{x}_{2,1}, \vec{x}_{3,1}, \vec{x}_{3,2}) K(\vec{x}_{2,2}, \vec{x}_{3,3}, \vec{x}_{3,4}) \\ & K(\vec{x}_{2,3}, \vec{x}_{3,5}, \vec{x}_{3,6}) K(\vec{x}_{2,4}, \vec{x}_{3,7}, \vec{x}_{3,8}) \end{aligned} \tag{3}$$

If this is correct, the marginal over any of these variables have to be the same distribution (1). BUT IS NOT

1.3 Reconstruction of hidden sequences