1 Simple model for phylogeny

Assume a very simple system, with one single spin $\sigma \in \{-1, +1\}$, and hamiltonian $\mathcal{H}^0 = h\sigma$. The equilibrium distribution is obviously $P^{eq} = e^{h\sigma}/Z^0$ with $Z^0 = 2\cosh(h)$. $\mathcal{H}^0$ is the model we would like to infer. With a sample of *i.i.d.* configurations from P^{eq}, this is a trivial problem.
Now, consider the case where the available sample comes from the process sketched in Fig. 1. A number of initial configurations σ^α, $a \in \{1 \dots K\}$, are chosen – either at random or from P^{eq}. For each initial configuration, a number $N(a)$ of samples are drawn in the following manner: with probability μ, σ is drawn from P^{eq}, and with probability $(1-\mu)$ it stays equal to σ^a. In other words, the distribution of σ inside one "cluster" a is

$$P(\sigma|a) = (1-\mu)\delta_{\sigma,\sigma^a} + \mu P^{eq}(\sigma) \tag{1}$$

The available sample is the union of all configurations of all clusters, resulting in the following distribution

$$P^{phylo}(\sigma) = \sum_{a=1}^{K} p(a) P(\sigma|a) \tag{2}$$

where $p(a)$ is the fraction of configurations in cluster a. In the following, it is assumed that the repartition of configurations in clusters is known. In other words, one knows the cluster-wise distributions $P(\sigma|a)$.

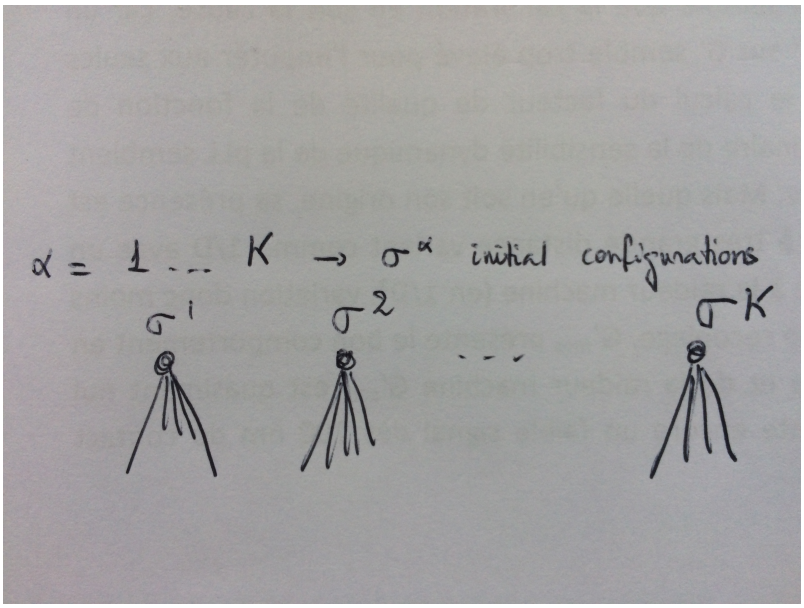

Figure 1

2 Naive hidden variable model

Clearly, if one wants to recover the underlying hamiltonian $\mathcal{H}^0$, the biais due to phylogeny needs to be accounted for. The idea of the hidden variable of is to model P^{phylo} with a hamiltonian including a supplementary variable which corresponds to the different phylogenetic clusters. Something like $\tilde{\mathcal{H}} = \mathcal{H}^0(\sigma) + \mathcal{H}^{HN}(\sigma, \alpha)$ – in the case of a general N spins Ising model, $\mathcal{H}^{HN}$ would not include any interactions between spins, but only with the hidden variable α. In practice, one tries to fit P^{phylo} with a distribution of the form

$$\tilde{P}(\sigma, \alpha) \propto \exp\left(\xi\sigma + J(\alpha)\sigma + \kappa(\alpha)\right). \tag{3}$$

Of course, the repartition of sequences in the different clusters a needs to be known in advance.
However, this fitting procedure will not lead to $\xi = h$ in general. It is sufficient to examine the basic case of $K = 2$ cluster with initial configurations σ^1 and σ^{-1}, where one tries to find parameters J and κ so that

$$\frac{\exp\left(\xi\sigma + J\sigma\alpha + \kappa\alpha\right)}{Z(\xi, J, \kappa)} = P^{phylo}(\sigma|a = 1),$$

where hidden variable takes values in $\{-1,+1\}$. The value of ξ for this fitting procedure can be derived analytically :

$$\xi = \frac{1}{4}\log\frac{\left((1-\mu)\delta_{1,\sigma^1}+\mu e^h/Z^0\right)\left((1-\mu)\delta_{1,\sigma^{-1}}+\mu e^h/Z^0\right)}{\left((1-\mu)\delta_{-1,\sigma^1}+\mu e^{-h}/Z^0\right)\left((1-\mu)\delta_{-1,\sigma^{-1}}+\mu e^{-h}/Z^0\right)} \neq h \tag{4}$$

This means that the result of this specific parametrization of the hidden variable will not yield the hamiltonian that corresponds to P^{eq}.
Number of parameters (corresponding to the hidden node, *i.e.* supplementary parameters with respect to the normal DCA inference) for this method, in the case of q states spins :

- $J(\alpha,\sigma)$: $(K-1)*(q-1)$ (because of gauge invariance) – $\times N$ for a full Ising model,
- $\kappa(\alpha)$: $(K-1)$.

3 Other parametrization for a hidden variable

One can design a parametrization that better corresponds to equation 2:

$$\tilde{P}(\sigma,\alpha) = \sum_{a=1}^{K} p(\alpha)\left(\mu_\alpha\delta_{\sigma,\sigma^\alpha}+(1-\mu_\alpha)\frac{e^{\xi\sigma}}{2\cosh(\xi)}\right) \tag{5}$$

The parameters corresponding to the hidden variable are σ^α, which can have q values for a given α, and the scalars μ_α. If fitted perfectly to P^{phylo}, this parametrization will obviously give $\xi = h$, recovering the $\mathcal{H}^0$ model. The number of parameters corresponding to the hidden variable in this case is (assuming $p(\alpha)$ is known since sequences are divided into clusters):

- σ^α: $K*q$ (not sure if there is some gauge invariance for this kind of model) – $\times N$ for a full Ising model,
- μ_α: K.

This is (almost) the same number as in the previous case. However, those parameters now have an interpretation in terms of phylogeny – *i.e.* ancestral sequences –, and is more adapted to this very crude phylogeny model.

4 Calculations

In the case $K=2$, the naive hidden variable model is

$$\tilde{P}(\sigma,\alpha) = Z^{-1}\exp(\xi\sigma + J\alpha\sigma + \kappa\alpha)$$

with hidden variable taking values in $\{-1,+1\}$.
If one knows the repartition of configurations in cluster, then the distribution of data is

$$P^{phylo}(\sigma,a) = p(a)\left((1-\mu)\delta_{\sigma,\sigma^a}+\mu\frac{e^{h\sigma}}{Z^0}\right).$$

One tries to find values $\{\xi,J,\kappa\}$ such that $\tilde{P}(\sigma,\alpha) = P^{phylo}(\sigma,\alpha)$. Writing this equality for all combinations of σ and α, one gets

$$Z^{-1}e^{\xi+J+\kappa} = P^{phylo}(1,1) \tag{6}$$
$$Z^{-1}e^{\xi-J-\kappa} = P^{phylo}(1,-1) \tag{7}$$
$$Z^{-1}e^{-\xi-J+\kappa} = P^{phylo}(-1,1) \tag{8}$$
$$Z^{-1}e^{-\xi+J-\kappa} = P^{phylo}(-1,-1). \tag{9}$$
$$\tag{10}$$

This immediately gives

$$\xi = \frac{1}{4}\log\frac{P^{phylo}(1,1)P^{phylo}(1,-1)}{P^{phylo}(-1,1)P^{phylo}(-1,-1)}.$$

In the particular case where $p(a)=1/2$, this is the same as equation 4.