



# Ancestral Sequence Reconstruction for Co-evolutionary models

Edwin Rodríguez-Horta, Alejandro Lage-Castellanos, Roberto Mulet

## ► To cite this version:

Edwin Rodríguez-Horta, Alejandro Lage-Castellanos, Roberto Mulet. Ancestral Sequence Reconstruction for Co-evolutionary models. *Journal of Statistical Mechanics: Theory and Experiment*, 2021, 2022 (1), pp.013502. 10.1088/1742-5468/ac3d93 . hal-03556943

**HAL Id: hal-03556943**

**<https://hal.sorbonne-universite.fr/hal-03556943>**

Submitted on 4 Feb 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Ancestral Sequence Reconstruction for Co-evolutionary models

Edwin Rodríguez Horta,<sup>1,2,\*</sup> Alejandro Lage-Castellanos,<sup>2</sup> and Roberto Mulet<sup>2,†</sup>

<sup>1</sup>*Sorbonne Université, CNRS, Institut de Biologie Paris-Seine,  
Laboratoire de Biologie Computationnelle et Quantitative – LCQB, Paris, France*

<sup>2</sup>*Group of Complex Systems and Statistical Physics,  
Department of Theoretical Physics,  
University of Havana, CP10400. La Habana, Cuba*

(Dated:)

## Abstract

The ancestral sequence reconstruction problem is the inference, back in time, of the properties of common sequence ancestors from measured properties of contemporary populations. Standard algorithms for this problem assume independent (factorized) evolution of the characters of the sequences, which is generally wrong (e.g. proteins and genome sequences). In this work, we have studied this problem for sequences described by global co-evolutionary models, which reproduce the global pattern of cooperative interactions between the elements that compose it. For this, we first modeled the temporal evolution of correlated real valued characters by a multivariate Ornstein-Uhlenbeck process on a finite tree. This represents sequences as Gaussian vectors evolving in a quadratic potential, who describe the selection forces acting on the evolving entities. Under a Bayesian framework, we developed a reconstruction algorithm for these sequences and obtained an analytical expression to quantify the quality of our estimation. We extend this formalism to discrete valued sequences by applying our method to a Potts model. We showed that for both continuous and discrete configurations, there is a wide range of parameters where, to properly reconstruct the ancestral sequences, intra-species correlations must be taken into account. We also demonstrated that, for sequences with discrete elements, our reconstruction algorithm outperforms traditional schemes based on independent site approximations.

---

\* Correspondence to: Edwin Rodríguez Horta, [edwin@fisica.uh.cu](mailto:edwin@fisica.uh.cu)

† Correspondence to: Roberto Mulet, [mulet@fisica.uh.cu](mailto:mulet@fisica.uh.cu)

## I. INTRODUCTION

The ancestral reconstruction (AR) problem is the inference, back in time, of the properties of common ancestors using as data set the measured properties of contemporary populations [1]. Ancestral reconstruction rests in a phylogeny, a tree that orders the populations. The leaves of the tree form the contemporary populations (the observed elements). They are connected to common ancestors at branching points, or nodes. The goal of AR is to estimate the internal characteristics of these nodes. Notice that although the construction of this tree (the phylogeny) is a problem itself [2], for many applications of AR this tree is assumed as known, and we follow this approach here.

Ancestral reconstruction relies also on a model of evolution. But, since the actual evolutionary process is rarely known, the proper selection of the model is fundamental for the outcome. In general, the use of simple models deteriorates the inference faster with increasing evolutionary time, but more realistic models are more difficult to calculate. It is in the researcher's ability to properly fine tune the complexity of the model for a specific application.

Ancestral sequence reconstruction (ASR) ([3, 4]) is a sub-problem of AR in which the characteristics of the system under study are described by a sequence of elements. These elements are encoded by character states (residues). In biological applications, these characters are usually defined within a finite set, such as nucleotides for genome sequences or amino-acids for proteins. In this case, the main goal is to infer each of these characters for each of the sequences that conform to the ancient populations.

Modern methods for ASR are based on the Maximum Likelihood (ML) framework [5, 6]. In this context the elements of the sequences are chosen, given the model of evolution and a phylogenetic tree, maximizing the probability of occurrence of the data (sequences) at the bottom of the tree. However, even in systems of moderate sizes, the huge number of possible configurations makes the problem intractable in general and one must introduce extra assumptions to approach real problems. A widely used hypothesis considers the evolution of the sequences as single site independent processes [3]. This simplification reduces the computational cost of the inference, and is at the basis of every advanced algorithm currently in use ([6, 7]). These algorithms may vary by type of biological information, i.e. the evolutionary model and the tree used, and on the strategy to reach the globally optimal

solution of the problem. For example, marginal and joint reconstruction or empirical and hierarchical Bayesian method [8]. The assumption of single site evolution, however, has been unavoidable, and its effect on the ASR remains unclear.

We know that in many biological scenarios this assumption is not correct. For amino acids sequences of proteins there is an abundant evidence of epistasis [9, 10], due to structural constraints imposed by the three-dimensional fold of the protein [11, 12]. In the case of genome sequences, epistasis is reflected in the distribution over genotypes in a population evolving with sufficient amount of exchange of genetic material (recombination, or any form of sex), a phenomenon called Quasi-Linkage Equilibrium and discovered by M. Kimura [13, 14]. Then, it is already well accepted that global co-evolutionary models are necessary to correctly represent relevant statistical features of biological sequences ([15–17]). They have been fundamental in the prediction of non-trivial structural contacts in the protein fold [18], designing novel functional protein sequences [19] and the inference of gene interactions [20].

Of course the use of the single-site approximation is not exclusive of the ASR methods. It has been identified as a limitation in other important inference problems in biology. This simplification has been shown to decrease the accuracy of the inferred phylogenetic trees in the presence of non-independent sites [21, 22]. However, data sets with strong functional or structural constraints are often analyzed within phylogenetic frameworks that assume independence among sites. Moreover, one of the most important computational problems in biology: the construction of sequence alignments is typically addressed through profile models, which capture position specificities like conservation in sequences but assume an independent evolution of different positions. A recent attempt to overcome the limitations of profile models and to include co-evolution among positions was carried out by Muntoni et al in [23]. The search for fast statistical tools, able to improve over the single site approximation is a research line that is still in its infancy.

In this contribution, we assess the impact of intra-species traits correlations on the performance of ancestral reconstruction. We first study the ASR problem for sequences whose elements are continuous characters co-evolving through an Ornstein-Uhlenbeck (OU) process on a phylogenetic tree. Although, at first, this model may seem distant from realistic biological applications, it has been used in the field of phylogenetic comparative methods (PCM) [24, 25] and, in a recent contribution [26], to extract intrinsic signals from hierarchically correlated data. At the same time it opens the way for algorithmic solutions

as well as analytical calculations clarifying the relevance of the parameters of the model and the affordability of the technique. With this understanding, we extend the approach to the more biologically relevant case of sequences defined on discrete values. We will show that the main picture devised for the OU process continues to be valid, and also that our reconstruction process outperforms standard methods that assume single site evolutionary processes. Notice that, as far as we know, a study about the robustness of current ASR methods to the violation of the hypothesis of independent substitutions has not been carried out. However a similar study exists for the phylogenetic inference problem [22] where the authors found that, for all the methods studied, even small amounts of dependencies can lead to significant errors in estimating the actual topologies. Then, it is very plausible that ASR methods present similar robustness problem. As we will see below our results point into this direction.

The rest of the document is organized as follows. In section II we define mathematically the problem. Then, in III we show how this formalism translates into a Bayesian framework assuming a co-evolutionary model with an Ornstein-Uhlenbeck dynamics. This section includes an analytical expression for the error in the inference that successfully compares with results from numerical simulations. Next we show how to exploit the mapping from discrete to continuous variables proposed in [27] to use the results from the previous sections to recover sequences on a finite alphabet subject to a stochastic dynamics. Finally we present the conclusions of our work.

## II. STATEMENT OF THE PROBLEM

Let us consider a set of observed sequences  $\{\underline{X}^m\}_{m=1,\dots,M}$  phylogenetically related by an evolutionary process on a tree  $\mathcal{T}$  that we will always assume known. Each sequence  $\underline{X} = (X_1, \dots, X_L)$  has length  $L$  and character states  $X_i$  take values on a discrete alphabet  $\mathbf{X}$  with size  $q = |\mathbf{X}|$ . For genome or protein sequences, the alphabet would be that of the 4 bases of DNA or the 20 amino acids respectively. We assume the sequences to be aligned, conforming a matrix of dimensions  $M \times L$  called multiple sequences alignment (MSA).

These contemporary sequences correspond to the terminal nodes of the tree (see Fig. 1) and are assumed to be the result of an evolutionary process initialized from a common ancestor at root  $\underline{X}^0$ . The evolution is mathematically defined by a propagator  $P(\underline{X}^j | \underline{X}^i, \Delta t_{ij})$

representing the probability of observing sequence  $\underline{X}^j$  as the result of the evolution from  $\underline{X}^i$  in a time  $\Delta t_{ij}$ . The precise form of the evolutionary propagator, as well as the inference of the phylogenetic tree are problems themselves, but to the purpose of this research, they are considered known.

Our goal is to infer the set of ancestral sequences at the internal nodes of the tree  $\{\underline{A}^m\}$  from which the observed contemporary sequences  $\{\underline{X}^m\}$  evolved. Bayesian methods use to compute the maximum-a-posteriori (MAP) estimate, by maximizing the joint probability of ancestral configurations given the contemporary sequences and the details of the evolution (tree and propagator), which could be evaluated via Bayes rule :

$$P(\{\underline{A}^m\}|\{\underline{X}^n\}) \propto P(\{\underline{X}^n\}|\{\underline{A}^m\}) * P(\{\underline{A}^m\}) \quad (1)$$

Furthermore, it is common to assume an uninformative prior distribution for sequences  $\{\underline{A}^m\}$  by taking  $P(\{\underline{A}^m\})$  uniform, and the inference becomes the calculation of the *Maximum Likelihood Estimate* (MLE) given by

$$\{\underline{A}^m\}^* = \max_{\{\underline{A}^m\}} \{P(\{\underline{X}^m\}|\{\underline{A}^m\})\} \quad (2)$$

Most efficient methods to compute the likelihood function on a tree are based on a dynamic programming algorithm called Felsenstein's pruning algorithm [2], which exploits the recursion equation

$$P^n(\{\underline{X}^m\}|\underline{A}^n) = \prod_{m \in C(n)} \left[ \sum_{\underline{A}^m} P^m(\{\underline{X}^m\}|\underline{A}^m) P(\underline{A}^m|\underline{A}^n) \right] \quad (3)$$

where  $P^n$  represent the conditional probability of observing all existing data sequences that share node  $n$  as an ancestor given that the sequence of this ancestor is  $\underline{A}^n$ . The term  $C(n)$  denotes all children nodes of node  $n$ , and  $P(\underline{A}^m|\underline{A}^n)$  is the evolutionary propagator. Then, to compute the likelihood of the observed data, as a function of the sequence configuration at each internal node, the expression (3) is evaluated starting from the leaves, where  $P^n(\underline{A}^n) = \delta_{\underline{A}^n, \underline{X}^n}$ , to the root of the tree.

Since, for systems of realistic sizes, the number of possible sequences at the internal nodes of the tree is huge, the evaluation of the likelihood via the recursion relation (3) is intractable, and therefore the solution sought through an optimization scheme impossible. To reduce the phase space, the standard assumption is to consider an independent-site

approximation, where each site of the sequences evolves independently of all others. This allows to write equation (3) for each sequence site, being the observed data at leaves of the tree a single column of the aligned sequences. Therefore,  $P(\underline{A}^j|\underline{A}^i, \Delta t_{ij})$  is factorized leading to a probabilistic reversible model given by character states frequencies  $P_a(A_a)$  and propagator  $P_a(A_a^i|A_a^j, \Delta t_{ij})$  which describe the replacement of character in the  $a$  position of the sequence at node  $j$ ,  $A_a^j$ , by a character in the same position in the sequence at node  $i$ ,  $A_a^i$ , after an evolutionary time  $\Delta t_{ij}$ . The selection of the  $P_a(A_a^i|A_a^j, \Delta t_{ij})$  depends on the nature of the problem, and may reflect extra biological information introduced in the model.

However, as we mentioned in the introduction, co-evolutionary processes could be relevant both in proteins families and in genome sequences. In this work we are going to avoid the assumption of independent-site evolution, and use a Bayesian formalism, considering instead that each sequence is better described by a pairwise Potts model :

$$P(\underline{X}) = \frac{1}{Z} \exp \left\{ \sum_{1 \leq i < j \leq L} \lambda_{ij}(X_i, X_j) \right\} \quad (4)$$

where statistical couplings  $\lambda_{ij}$  between sites encode the epistatics signal of the system. This model has been widely used in biological systems [15], and it is the least biased statistical model that reproduces the empirical frequencies of characters by site and by site pairs, which are the most common statistical observables for biological data.

### III. MULTIVARIATE ORNSTEIN-UHLENBECK DYNAMICS

All phylogeny based method assume that a single evolutionary history underlies the sample of the sequences under study, then recombination and gene flow are ignored because these may give rise to graphs that are no longer trees but networks. Therefore the forces of evolution consistent with our framework are those that act on the single genotype level as selection, mutations and genetic drift. In this context we propose to use a multivariate OU dynamics, which can take into account selection when the potential is considered to be a fitness proxy, but not changes in selection rules on time, because this would invalidate the assumption of stationary evolution. However, it may consider changes in the mutation rate when times are measured in terms of a molecular clock rather than in physical time.

Although the discrete nature of biological evolution is an admitted fact since the discovery of the genetic code, there are good reasons to study AR problems in continuous variable

models as the multivariate OU dynamics. First, correlation between traits could be present also at the phenotypic level, where characters are real valued quantities, as body mass. Second, even at the gene level, there are some tricks that help to turn discrete into continuous variables [27], thus rendering the continuous approach applicable to the discrete case. And, finally, continuous variables may simplify the problem enough to allow for a precise analytical description, which can illustrate the relevance of the parameters of the problem.

### A. Formalism

We will first study a model of phylogenetic tree; one in which each specie is described by a continuous vector  $\vec{x} \in \mathbb{R}^L$  in  $L$  dimensions. Furthermore, we assume that the evolution of these characters follow an OU process (see details in A 1) . For this, continuous degrees of freedom evolves under a potential  $V(\vec{x}) = \frac{1}{2}\vec{x}^T \mathbf{C}^{-1} \vec{x}$  leading to the stationary distribution:

$$P^0(\vec{x}) \propto \exp -\frac{1}{2}\{\vec{x}^T \mathbf{C}^{-1} \vec{x}\} \quad (5)$$

For and OU process like this the corresponding propagator, i.e. the solution of the Fokker-Planck equation, is given by :

$$P^0(\vec{x}_j|\vec{x}_i; \Delta t_{ij}) \propto \exp -\frac{1}{2}\{\vec{x}_j^T \mathbf{\Sigma}_{ij}^{-1} \vec{x}_j + \vec{x}_i^T \mathbf{\Lambda}_{ij}^2 \mathbf{\Sigma}_{ij}^{-1} \vec{x}_i - 2\vec{x}_i^T \mathbf{\Lambda}_{ij} \mathbf{\Sigma}_{ij}^{-1} \vec{x}_j\} \quad (6)$$

where

$$\mathbf{\Sigma}_{ij} = \mathbf{C} - \mathbf{\Lambda}_{ij} \mathbf{C} \mathbf{\Lambda}_{ij}, \quad \mathbf{\Lambda}_{ij} = e^{-\mathbf{C}^{-1} \gamma \Delta t_{ij}}$$

and where  $\mathbf{C}$  is the correlation matrix,  $\Delta t_{ij}$  is the time distance between sequences  $\vec{x}_i$  and  $\vec{x}_j$  and  $\gamma$  is the characteristic time-scale governing the dynamics.

The evolution model is schematized in Figure 1. It is worth clarifying the conventions used for the notation. We will denote by:

- lowercase vector variables (i.e.  $\vec{x}, \vec{x}_i, \vec{x}_j, \vec{x}_1^h, \dots$ ), the real valued vector of dimension  $L$  defining the state of the given species.
- uppercase vector variables (i.e.  $\vec{X}, \vec{A}, \vec{D}, \dots$ ), the concatenation of many state vectors, for instance as the collection of all sequences  $\vec{X} = (\vec{x}_1, \dots, \vec{x}_N)$ .
- bold uppercase (i.e.  $\mathbf{\Sigma}_{ij}, \mathbf{C}, \mathbf{\Lambda}_{ij}$ )  $L \times L$  matrices acting at species level, i.e. over variables  $\vec{x}_i$ .



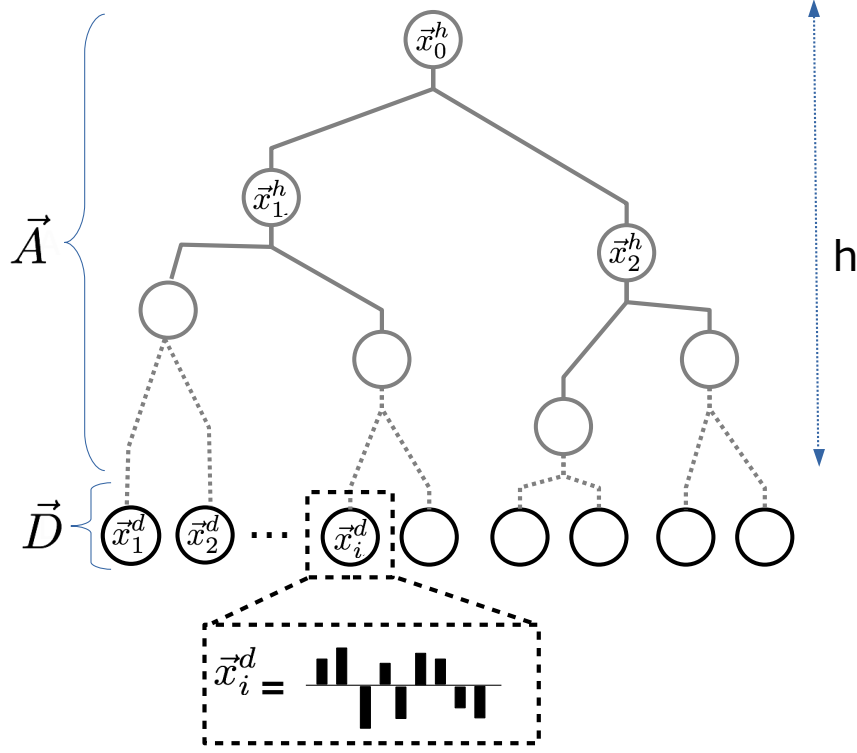


FIG. 1. Schematic representation of the evolutionary process in a binary and balanced tree  $\mathcal{T}$  with height  $h$ . The process starts at the root node 0 with a configuration  $\vec{x}_0$  sampled from  $P^0(\vec{x}_0)$ . The dynamic consists in independent realizations of the OU process on all branches from ancestral nodes to its descendant over times corresponding to the branch length. The observable data only consist of configurations of the leaf nodes ( $\vec{D}$ ), while configurations of ancestral nodes ( $\vec{A}$ ), remain unknown and must be inferred.

- uppercase blackboard bold matrices (i.e.  $\mathbb{K}, \mathbb{G}, \dots$ ) as block matrices at tree level, made of the composition of many species level matrices.
- indices  $i, j$  to run over the tree nodes, i.e. over the species.
- indices  $a, b$  to run over entries of the state vector or concatenated vectors.

With these assumptions, the probability of a configuration  $\vec{X}$  of the full system (internal nodes configurations and leaves) can be constructed by using the stationary distribution (5) and the propagator (6). For instance, for the case of a tree with binary topology,

$$P^0(\vec{X}) = P^0(\vec{x}_0)P^0(\vec{x}_1|\vec{x}_0)P^0(\vec{x}_2|\vec{x}_0)P^0(\vec{x}_3|\vec{x}_1)P^0(\vec{x}_4|\vec{x}_1)\dots$$

For an arbitrary tree, the probability  $P^0(\vec{X})$  can be shaped as

$$P^0(\vec{X}) = \frac{e^{-\mathcal{H}^0(\vec{X})}}{Z^0} \quad (7)$$

invoking the Boltzmann distribution of a system with a pairwise Hamiltonian

$$\mathcal{H}^0 = - \sum_{i < j} \vec{x}_i^T \mathbf{J}_{ij} \vec{x}_j - \sum_{i=1}^N \vec{x}_i^T \mathbf{H}_i \vec{x}_i. \quad (8)$$

The interaction terms are  $L \times L$  matrices

$$\mathbf{J}_{ij} = \begin{cases} \mathbf{\Lambda}_{ij} \mathbf{\Sigma}_{ij}^{-1} & \text{if } i \text{ and } j \text{ are in contact,} \\ \mathbf{0} & \text{otherwise.} \end{cases} \quad (9)$$

and there is a Gaussian local interaction given by

$$\mathbf{H}_i = -\frac{1}{2} \cdot \begin{cases} \mathbf{C}^{-1} \left[ 1 + \sum_{j \in c(0)} \frac{\mathbf{\Lambda}_{0j}^2}{1 - \mathbf{\Lambda}_{0j}^2} \right] & \text{if } i = 0, \\ \mathbf{\Sigma}_{a(i)i}^{-1} & \text{if } i \text{ is a leaf,} \\ \mathbf{C}^{-1} \left[ \frac{1}{1 - \mathbf{\Lambda}_{a(i)i}^2} + \sum_{j \in c(i)} \frac{\mathbf{\Lambda}_{ij}^2}{1 - \mathbf{\Lambda}_{ij}^2} \right] & \text{otherwise.} \end{cases} \quad (10)$$

where  $c(i)$  and  $a(i)$  refers to child and ancestral nodes of nodes  $i$ .

Not surprisingly, the Gaussian nature of the Ornstein-Uhlenbeck survives for the whole tree, and the equilibrium distribution of the concatenated variables  $\vec{X}$  is also a Gaussian  $P^0(\vec{X}) \propto \exp(-\frac{1}{2} \vec{X}^T \mathbb{H} \vec{X})$ , with inverse covariance matrix

$$\mathbb{H} = 2 \begin{pmatrix} \mathbf{H}_1 & \frac{1}{2} \mathbf{J}_{12} & \dots & \frac{1}{2} \mathbf{J}_{1N} \\ \frac{1}{2} \mathbf{J}_{21} & \mathbf{H}_2 & \dots & \frac{1}{2} \mathbf{J}_{2N} \\ & & \dots & \\ \frac{1}{2} \mathbf{J}_{N1} & \frac{1}{2} \mathbf{J}_{N2} & \dots & \mathbf{H}_N \end{pmatrix} \quad (11)$$

A Gaussian distribution could be considered all bout non-problematic, at this point. However, there are a couple of good reasons to keep the system as an additive pairwise interaction model. First, although treatable, the matrix  $\mathbb{H}$  could run uncomfortably large, since its size is  $(L \times N)^2$ ,  $L$  being the size of the sequences, and  $N$  the number of species in the tree. Notice that for discrete models  $L$  could be particularly large, since a 21 one hot encoding is normally used to codify discrete amino acid alphabet into continuous variables, implying that  $L$  is 21 times larger than the real amino acid sequence. The second reason is that dealing with a pairwise model brings to bear the statistical mechanics toolbox to compute exactly or approximate the inference problem.

## B. Inference problem

In the context of ASR, tree-nodes are decomposed in two groups, leaves and internal nodes:  $\vec{X} = \{\vec{A}, \vec{D}\}$ , with  $\vec{A}$  being the internal (ancestral) nodes configurations and  $\vec{D}$  being the leaves (data) configurations (see Fig. 1). The MAP estimate of ancestral sequences configurations can be computed by maximizing the posterior distribution

$$P(\vec{A}|\vec{D}) = \frac{P^0(\vec{X})}{P(\vec{D})} \propto P^0(\vec{A}, \vec{D}) \propto \exp\left\{-\mathcal{H}^d(\vec{A})\right\} \quad (12)$$

over internal sequences  $\vec{A}$ , where  $P^0(\vec{A}, \vec{D})$  from eq. (7) is the equilibrium distribution for the whole system, but with the leaves evaluated at the observed values  $\vec{D}$ .

The term in the exponent, therefore, is

$$\mathcal{H}^d(\vec{A}) = - \sum_{1 \leq i < j \leq N_h} \vec{x}_i^T \mathbf{J}_{ij} \vec{x}_j - \sum_{i=1}^{N_h} \left( \vec{x}_i^T \mathbf{H}_i \vec{x}_i + \vec{h}_i \vec{x}_i \right), \quad (13)$$

and the fields and couplings  $\{\mathbf{J}_{ij}, \mathbf{H}_i\}$  are equal to those of  $\mathcal{H}^0$ , and

$$\vec{h}_i = \begin{cases} \sum_{j \in c(i)} \mathbf{\Lambda}_{ij} \mathbf{\Sigma}_{ij}^{-1} \vec{x}_j & \text{if } i \text{ is in contact with leaf } \vec{x}_j, \\ \vec{0} & \text{otherwise.} \end{cases} \quad (14)$$

To maximize the posterior distribution over hidden sequences is equivalent to finding the mode of the distribution  $P(\vec{A}|\vec{D})$ . The posterior (12) can be rewritten as:

$$P(\vec{A}|\vec{D}) \propto \exp\left\{\vec{A}^T \mathbb{K} \vec{A} + \vec{S}^T \vec{A}\right\} \quad (15)$$

where  $\mathbb{K}$  is a block matrix, corresponding to the part of the matrix  $\mathbb{H}$  (eq. (11)) that acts over the hidden nodes, with elements

$$\mathbf{K}_{ij} = \begin{cases} \frac{1}{2} \mathbf{J}_{ij} & \text{if } i \neq j, \\ \mathbf{H}_i & \text{if } i = j. \end{cases} \quad (16)$$

and  $\vec{S}$  is a concatenation of vectors  $\vec{h}_i$  defined in (14).

As (15) is a Gaussian distribution it's mode match with its mean and is given by:

$$\hat{\mu} = -\frac{\mathbb{K}^{-1} \vec{S}^T}{2} \quad (17)$$

i.e., the mode of the posterior distribution can be computed directly from the expression (17). Unfortunately, for practical applications it implies the inversion of a matrix  $\mathbb{K}$  of very high dimensions  $\dim(\mathbb{K}) = L \times N_h$ , with  $L$  the length of the sequences and  $N_h$  the number of internal nodes. Gaussian message-passing algorithm (GaMP) [29–31] could be adapted to overcome this issue, which correctly solves the problem on trees, reducing optimization to compute max-marginals

$$M_i(\vec{\mu}^i) = \max_{\vec{A}} \left\{ P^d(\vec{A}) : \vec{x}_i = \vec{\mu}^i \right\} \quad (18)$$

Then, the set of ancestral configurations that jointly maximizes the posterior distribution (15) is given by  $\{\vec{\mu}^1, \vec{\mu}^2, \dots, \vec{\mu}^{N_h}\}$ . Adaptation of GaMP update rules to compute of max-marginals yields to max-product update rules [31], details of the algorithm obtained are shown in appendix (A 2). In what follows, we use equation (17) to compute the accuracy of this estimator analytically, and (18) through the GaMP algorithm, to efficiently evaluate (17) without inverting the  $\mathbb{K}$  matrix.

### C. Evaluating estimator accuracy via mean square error

It is important to check whether the inference process defined above provides a solution that does not only maximize the posterior distribution, but that actually defines a set of inferred sequences which are similar to the actual one. We measure this similarity by the distance

$$\hat{d}(\mathbf{C}, \mathbf{C}_0) = \left\langle \|\vec{M} - \vec{A}\|^2 \right\rangle_{\vec{M}, \vec{A}} = \left\langle \sum_b^{L \times N_h} (M_b - A_b)^2 \right\rangle_{\vec{M}, \vec{A}} \quad (19)$$

where  $\mathbf{C}$  encodes the information about the actual co-evolutionary process, and therefore the joint statistics of the real ancestral sequences  $\vec{A}$  and the observed ones  $\vec{D}$  defined by equation (7) while  $\mathbf{C}_0$  has a structure that is defined by the researcher before the start of the inference and is not necessarily equal to  $\mathbf{C}$  as schematized in Figure 2.  $\vec{A} = [\vec{x}_1^h, \dots, \vec{x}_{N_h}^h]$  is the concatenated vector of true ancestral sequences and  $\vec{M} = [\vec{\mu}^1 \dots \vec{\mu}^{N_h}]$  are the concatenated inferred configurations at internal nodes, for a realization of the data  $\vec{D}$ .

The estimator (17) can be rewritten as

$$\hat{\mu}_0(\mathbf{C}_0, \vec{D}) = -\frac{\mathbb{K}_0^{-1} * \mathbb{A}_0 \vec{D}^T}{2}$$

to make explicit its dependence on the data  $\vec{D}$  and where  $\mathbb{A}_0$  is a block matrix dependent of the correlation matrix  $\mathbf{C}_0$  and the tree topology

$$\mathbf{A}_{i,j}^0 = \begin{cases} \Lambda_{ij}(\mathbf{C}_0) * \Sigma_{ij}^{-1}(\mathbf{C}_0) & \text{if } i \text{ is an external hidden node in contact with leaves node } j . \\ \mathbf{0} & \text{otherwise} \end{cases} \quad (20)$$

We exploit different  $\mathbf{C}_0$  to analyze the effect on the estimator accuracy when, for instance, a diagonal approximation of  $\mathbf{C}$  is used for the reconstruction, as is the case of the independent site evolutionary models used by Pupko's algorithm [6]. In what follows, the block matrices with subscript 0 indicate that they are a function of the correlation matrix  $\mathbf{C}_0$ .

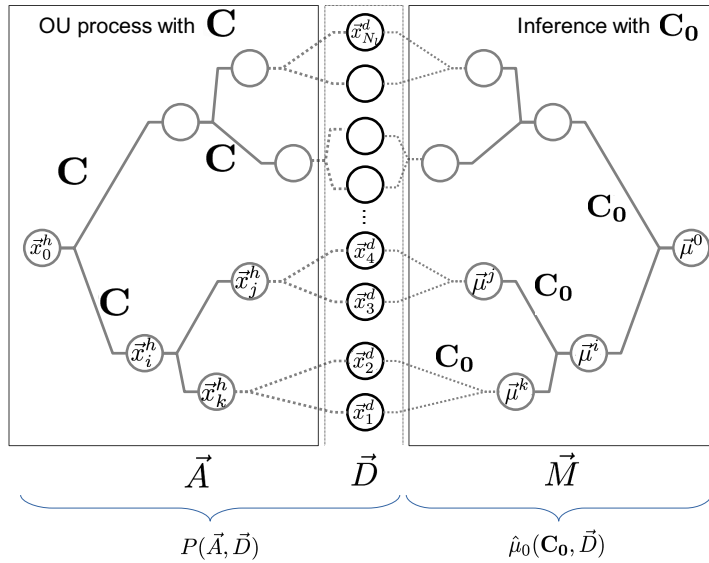


FIG. 2. Schematic representation of the inference process. On the left, ancestral sequences are generated by an Ornstein-Uhlenbeck process with equilibrium covariance  $\mathbf{C}$ . On the right, the reconstruction of the tree is obtained by maximizing the likelihood of a similar process, albeit with a potentially different covariance  $\mathbf{C}_0$ . For a perfect reconstruction  $\vec{M} = \vec{A}$ .

In equation (19) the sum goes over each of the  $L \times N_h$  elements of the concatenated vector and the average is done by the joint distribution  $P(\vec{A}, \vec{M})$  :

$$\hat{d}(\mathbf{C}, \mathbf{C}_0) = \sum_b \int (M_b - A_b)^2 P(\vec{A}, \vec{M}) d\vec{A} d\vec{M}$$

The joint distribution  $P(\vec{A}, \vec{M})$  can be obtained from:

$$P(\vec{A}, \vec{M}) = \int d\vec{D} \delta[\hat{\mu}_0(\vec{D}) - \vec{M}] P(\vec{A}, \vec{D}) \quad (21)$$

with  $P(\vec{A}, \vec{D})$  as the full probability of all sequences: ancestral ( $\vec{A}$ ) and leaves ( $\vec{D}$ ) concatenated vectors.  $P(\vec{A}, \vec{D})$  was defined in (7) and can be rewritten as

$$P^0(\vec{X}) = P(\vec{A}, \vec{D}) \propto \exp \left\{ \vec{A} \mathbb{K} \vec{A}^T + \vec{A} * \vec{S}^T - \frac{1}{2} \vec{D} \mathbb{G} \vec{D}^T \right\} \quad (22)$$

where the block matrix  $\mathbb{K}$  is defined in (16) and vector  $\vec{S}$  is a concatenation of vectors  $\vec{h}_i$  defined in (14). Finally, the elements of the block matrix  $\mathbb{G}$  are given by

$$\mathbf{G}_{i,j} = \begin{cases} \Sigma_{a(i),i}^{-1} & \text{if } i = j. \\ \mathbf{0} & \text{otherwise} \end{cases} \quad (23)$$

Then, to evaluate equation (21) we use expression (22) and the exponential representation of Dirac's delta function:

$$\begin{aligned} P(\vec{A}, \vec{M}) &= \int d\vec{D} \delta[\hat{\mu}_0(\vec{D}) - \vec{M}] P(\vec{A}, \vec{D}) \\ &= \int d\vec{D} \int d\vec{q} \exp \left\{ i * \vec{q} * (\hat{\mu}_0(\vec{D}) - \vec{M})^T \right\} * P(\vec{A}, \vec{D}) \\ &= \exp \left\{ \vec{A} \mathbb{K} \vec{A}^T \right\} \int d\vec{q} \exp \left\{ -i * \vec{q} * \vec{M}^T \right\} \int d\vec{D} \exp \left\{ -\frac{1}{2} \vec{D} \mathbb{G} \vec{D}^T + \vec{D} \mathbb{A}^T \vec{A}^T - \frac{i}{2} \vec{q} * \mathbb{K}_0^{-1} \mathbb{A}_0 \vec{D}^T \right\} \end{aligned} \quad (24)$$

such that (see details in A 3):

$$P(\vec{A}, \vec{M}) \propto \exp \left\{ -\frac{1}{2} \left[ -2 \vec{A} \mathbb{K} \vec{A}^T + 4 \vec{A} \mathbb{Q}_1^T \mathbb{Q}_0^{-1} \mathbb{K}_0 \vec{M}^T + 4 * \vec{M} \mathbb{K}_0 \mathbb{Q}_0^{-1} \mathbb{K}_0 \vec{M}^T \right] \right\} \quad (25)$$

where  $\mathbb{Q} = \mathbb{A} \mathbb{G}^{-1} \mathbb{A}^T$ ,  $\mathbb{Q}_0 = \mathbb{A}_0 \mathbb{G}^{-1} \mathbb{A}_0^T$  and  $\mathbb{Q}_1 = \mathbb{A}_0 \mathbb{G}^{-1} \mathbb{A}^T$ .

The equation (25) can be resumed as

$$P(\vec{A}, \vec{M}) \propto \exp \left\{ -\frac{1}{2} \vec{Z} \mathbb{V}^{-1} \vec{Z}^T \right\} \quad (26)$$

with  $\vec{Z} = [\vec{A}, \vec{M}]$  and

$$\mathbb{V}^{-1} = \begin{pmatrix} -2\mathbb{K} & 2\mathbb{Q}_1^T \mathbb{Q}_0^{-1} \mathbb{K}_0 \\ 2\mathbb{K}_0 \mathbb{Q}_0^{-1} \mathbb{Q}_1 & 4\mathbb{K}_0 \mathbb{Q}_0^{-1} \mathbb{K}_0 \end{pmatrix} \quad (27)$$

Equation (26) can be seen as a bivariate normal distribution where the variables are given by the pair of concatenated vectors  $[\vec{A}, \vec{M}]$ . The covariance matrix for this distribution  $\mathbb{V}_{ij} = \langle \vec{Z}_i, \vec{Z}_j \rangle$  allow us to compute terms in the mean square error defined by the equation (19) if we know  $\mathbb{V}_{11} = \langle \vec{A} \cdot \vec{A}^T \rangle$ ,  $\mathbb{V}_{22} = \langle \vec{M} \cdot \vec{M}^T \rangle$  and  $\mathbb{V}_{12} = \langle \vec{A} \cdot \vec{M}^T \rangle$ . For this we invert  $\mathbb{V}^{-1}$  obtaining:

$$\mathbb{V} = \begin{pmatrix} -[2\mathbb{K} + \mathbb{Q}]^{-1} & \frac{1}{2}[2\mathbb{K} + \mathbb{Q}]^{-1} * \mathbb{Q}_1^T \mathbb{K}_0^{-1} \\ \frac{1}{2}\mathbb{K}_0^{-1} \mathbb{Q}_1 [2\mathbb{K} + \mathbb{Q}]^{-1} & \frac{1}{4}\mathbb{K}_0^{-1} [\mathbb{Q}_0 - \mathbb{Q}_1 [2\mathbb{K} + \mathbb{Q}]^{-1} \mathbb{Q}_1^T] \mathbb{K}_0^{-1} \end{pmatrix} \quad (28)$$

Then, we simply evaluate the expression

$$\hat{d}(\mathbf{C}, \mathbf{C}_0) = \sum_b^{L \times N_h} ([\mathbb{V}_{11}]_{bb} - 2 * [\mathbb{V}_{12}]_{bb} + [\mathbb{V}_{22}]_{bb}) \quad (29)$$

Notice that, the mean square error from equation (29) is a function of both, the correlation matrix  $\mathbf{C}$  that defines the potential in the direct OU process and the matrix  $\mathbf{C}_0$  which represents the approximate correlation matrix used for the inference. If  $\mathbf{C}_0$  is diagonal, we have an independent site approximation. On the other hand, if  $\mathbf{C}_0 \equiv \mathbf{C}$  we are inferring with the actual correlation for the evolutionary process.

#### IV. COMPARISON WITH NUMERICAL EXPERIMENTS

In order to test our results, we extract the direct process correlation  $\mathbf{C}$  randomly from a Wishart distribution:

$$\mathbf{C}_{L \times L} \sim W(k, \mathbf{P}) \propto |\mathbf{C}|^{(k-L-1)/2} \exp \left\{ -\frac{1}{2} \text{Tr}(\mathbf{P}^{-1} \mathbf{C}) \right\}$$

where parameters  $k > L - 1$  and  $\mathbf{P}$  stand for the number of degrees of freedom and the scale matrix respectively. The expected value for  $\mathbf{C}$  is  $E(\mathbf{C}) = k * \mathbf{P}$  and the variance of its

elements is given by

$$\text{Var}(C_{ij}) = k * (p_{ij}^2 + p_{ii} * p_{jj})$$

If we set  $\mathbf{P} = \mathbb{I}$ , where  $\mathbb{I}$  is the identity matrix, then  $\mathbf{C}/k \rightarrow \mathbb{I}$  for  $k \gg L$ . This makes it possible to sample correlation matrices with different levels of covariance by tuning the ratio  $k/L$ , which allows us to understand when it becomes relevant to go beyond the independent site approximation (zero off-diagonal correlation matrix).

In order to assess the relevance of neglecting the covariance of the characters in the ancestral reconstruction process, we will evaluate  $\hat{d}(\mathbf{C}, \mathbf{C}_0)$  at two extreme cases:

$$\hat{d}_1 = \hat{d}(\mathbf{C}, \mathbf{C}) \quad \hat{d}_0 = \hat{d}(\mathbf{C}, \mathbf{C}_0 = \text{diag}(\mathbf{C}))$$

where in the second case the inference process is carried out with a factorized assumption on the distribution of the characters, and therefore with a diagonal  $\mathbf{C}_0$ . For a fixed  $\mathbf{C}$ , both  $\hat{d}_1$  and  $\hat{d}_0$  are monotonically increasing functions of the speed of the evolutionary process  $\gamma$  (not shown in figures), starting from  $\hat{d} = 0$  when  $\gamma = 0$  up to the average distance between two uncorrelated equilibrium configurations when  $\gamma = \infty$ , with  $\hat{d}_0 > \hat{d}_1$  at every  $\gamma$ , as expected.

In Figure 3 is shown the ratio between  $\hat{d}_0 = \hat{d}(\mathbf{C}, \mathbf{C}_0 = \text{diag}(\mathbf{C}))$  and  $\hat{d}_1 = \hat{d}(\mathbf{C}, \mathbf{C}_0 = \mathbf{C})$  as function of the time-scale parameter  $\gamma$  for a binary tree with height  $h = 7$  and for a direct potential  $\mathbf{C}$  sampled from a Wishart distributions where  $k = L = 10$ . The plot was also reproduced numerically using GaMP from 100 simulations of the evolution-inference process for each  $\gamma$  value.

Not surprisingly, extreme cases for the typical time-scale  $\gamma = 0$  and  $\gamma = \infty$  produce no difference ( $\hat{d}_1 = \hat{d}_0$ ) in the accuracy of the estimator using either  $\mathbf{C}$  or  $\mathbf{C}_0$ . One of the cases  $\gamma \rightarrow 0$  means that the process is too slow to produce any changes in the sequences along the tree, and all observed leaves and all hidden nodes have the same value and can be equally (trivially and perfectly) reconstructed independently of the covariance assumed. The other case,  $\gamma \rightarrow \infty$  corresponds to an extremely fast evolutionary process, such that all nodes are equilibrium samples, and there is no information whatsoever in the observed data. In this limit both methods are equally bad.

We can find effective values for the parameter  $\gamma$  where the two regimes described above start to be noticeable. We know that the correlation between two sequences linked by an OU process is given by:



$$\mathbf{\Lambda}^{\Delta t} \mathbf{C} = \sum_a^L \rho_a^{-1} e^{-\rho_a \Delta t * \gamma} |s_a\rangle \langle s_a| = \sum_a^L \rho_a^{-1} e^{-\Delta t / \tau_a} |s_a\rangle \langle s_a|$$

with  $\{\rho_a, \langle s_a|\}$  eigenstates of  $\mathbf{C}^{-1}$  and  $\tau_a = (\gamma * \rho_a)^{-1}$ , then we can consider that two configurations are *uncorrelated* when  $\Delta t \gg \max(\tau_a) = \frac{1}{\gamma * \rho_{min}}$  or  $\gamma \gg \frac{1}{\Delta t \rho_{min}}$ . Inversely an *strongly correlated* regime occur when  $\Delta t \ll \min(\tau_a) = \frac{1}{\gamma * \rho_{max}}$  or  $\gamma \ll \frac{1}{\Delta t \rho_{max}}$ . These criteria could be generalized for a tree if we set  $\Delta t = \Delta t_{av}$  where  $\Delta t_{av}$  is the average time between connected nodes in the tree. Then we have  $\gamma \gg \frac{1}{\Delta t_{av} \rho_{min}} = \gamma_d$  for *uncorrelated* regime and  $\gamma \ll \frac{1}{\Delta t_{av} \rho_{max}} = \gamma_c$  for the *strongly correlated* regime.

The interval where phylogenetic-based inference methods are relevant is

$$\frac{1}{\Delta t_{av} \rho_{max}} \equiv \gamma_c \ll \gamma \ll \gamma_d \equiv \frac{1}{\Delta t_{av} \rho_{min}}$$

In this regime, the data is neither too correlated (*strongly correlated* regime) around the tree as to make inference unnecessary or too uncorrelated, making the phylogeny irrelevant (*uncorrelated* regime). In Figure 3 this interval is signaled with vertical lines, and it coincides with the region where the full correlated nature of the process more significantly outperforms the factorized inference  $\frac{\hat{d}_0}{\hat{d}_1} > 1$ .

Notice that the ratio has two maximums that are originated by different mechanisms. The first one reflects the value of  $\gamma$  at which the system instantly recognizes the difference between the two matrices in the expressions pre-multiplied by the factor  $\gamma$  as  $\mathbf{\Lambda} = \exp -\gamma \mathbf{C}^{-1}$  and  $\mathbf{\Lambda}_0 = \exp -\gamma \mathbf{C}_0^{-1}$ . This effect has different proportions at each level in the tree as is shown in Figure 4 where we plot the ratio between the inference errors but for different levels of the tree  $\hat{d}_0^l$  y  $\hat{d}_1^l$ . These errors are computed as in equation (29), but using in place of  $\mathbb{V}$  a matrix formed only by the blocks associated with the corresponding level  $l$ . For internal nodes, the effective value of  $\gamma$  is higher as soon as the node is more distant from the leaves.

At larger values of  $\gamma$ , the fraction  $\frac{\hat{d}_0}{\hat{d}_1}$  starts to grow with  $\gamma$  because the differences between  $\mathbf{\Lambda}$  and  $\mathbf{\Lambda}_0$  becomes more important. This growth saturates at the second peak of the plot, from this point the system moves toward the *uncorrelated* regime and  $\mathbf{\Lambda} \approx \mathbf{\Lambda}_0 \rightarrow 0$ . Again, this exponential decay from the maximum emerges first for more internal levels of the tree.

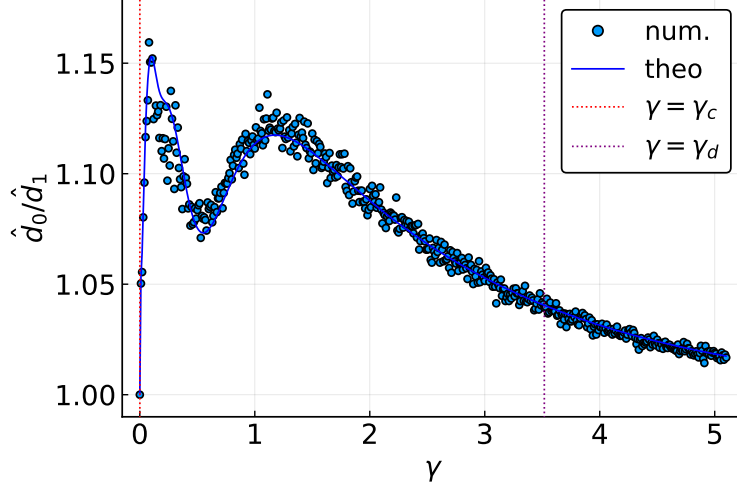


FIG. 3. Ratio between  $\hat{d}_0$  and  $\hat{d}_1$  as function of the parameter  $\gamma$  for a binary tree and homogeneous topology with  $\Delta t = 1.0$  and height  $h = 7$ . The OU potential  $\mathbf{C}$  is sampled from a Wishart distribution with  $\mathbf{P} = \mathbb{I}$  and  $k = L = 10$ . Solid line corresponds to the evaluation of the equation (29), whereas points are the result of averaging 100 numerical simulations of the evolution and inference (using GaMP) for each value of  $\gamma$ . The vertical dashed lines represent the borders of the interval for time-scale parameter where the inference results non-trivial.

## V. SEQUENCES WITH DISCRETE CHARACTERS

Unfortunately, when dealing with sequences with discrete characters, we don't have a global co-evolutionary propagator for the statistical model of equation (4). Then, it's impossible to derive an explicit form of the joint probability of ancestral configurations given the contemporary sequences of equation (12). However, instead of the common factorization of the propagator, we may keep the global nature and tractability of the problem transforming the Potts model over discrete variables (4) into a Gaussian distribution, and to assume an OU dynamics as an approach to the evolutionary process. In this way, we may exploit the same inference scheme described for continuous traits in III B. In what follows, we expand on this idea.

We know that it is possible to represent a Potts model (4) unambiguously in the space of the frequencies from the correlation matrix  $\{\mathbf{C}\}$ . In this space we can also define a multivariate Gaussian model that describes the same statistics of the discrete model but for continuous variables [27]. In practice, from a set of  $M$  sequences sampled by the pairwise

Potts model, we first transform each sequence of length  $L$  and alphabet  $\{1, \dots, q\}$  toward an array of length  $q * L$  where each element take values on binary alphabet  $\{0, 1\}$ . Then, each original sequence site is mapped to  $q$  binary variables, taking value 1 if the position corresponds with the original state of the residue and 0 for the rest of the positions. The result is a new alignment with ones and zeros keeping the same structure from the original alignment but now with dimensions  $M \times (q * L)$ . These binary variables could be approximated by real valued variables, which allow the computation of the covariance matrix  $\mathbf{C}$  needed to parametrize the multivariate Gaussian distribution by

$$P^0(\vec{x}) \propto \exp -\frac{1}{2} \{ \vec{x}^T \mathbf{C}^{-1} \vec{x} \} \quad (30)$$

The Gaussian distribution obtained can be easily propagated using an Ornstein-Uhlenbeck process. That is, we can build from the potential  $\mathbf{C}$ , the propagator  $P^0(\vec{x}_j | \vec{x}_i; \Delta t_{ij})$  given by the equation (6), and use it as the propagator of the Potts model. This propagated statistics is expected to be comparable to the one obtained from simulating the original Potts model with a Monte Carlo (MCMC) procedure.

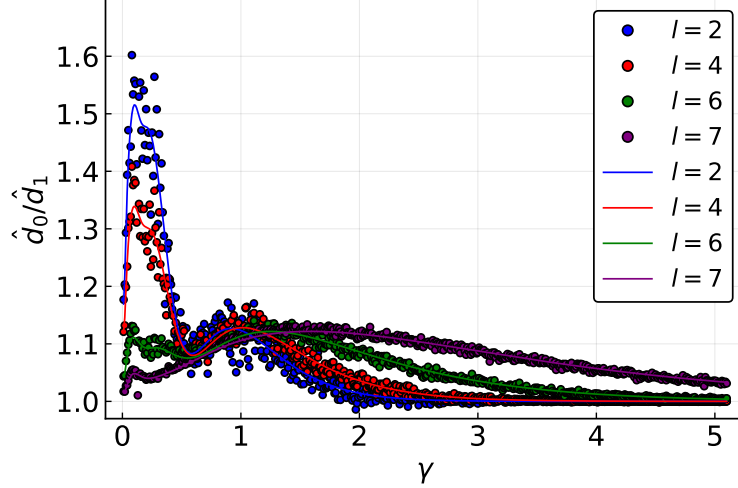


FIG. 4. Ratio between  $d_0^l$  and  $d_1^l$  as function of the parameter  $\gamma$  for different levels  $l$ , for a tree with binary and homogeneous topology with  $\Delta t = 1.0$  and height  $h = 7$ . The OU potential  $\mathbf{C}$  is sampled from a Wishart distribution with  $\mathbf{P} = \mathbb{I}$  and  $k = L = 10$ . Solid line correspond to the evaluation of the equation (29) but using in place of  $\mathbb{V}$  a matrix formed only by blocks associated with the level  $l$ . Points are the result of averaging 100 numerical simulations of the evolution and inference (using GaMP) for each vale of  $\gamma$ .

Then, we could ask whether it is possible to reconstruct the discrete configurations of the internal nodes of the tree if the direct evolutionary process is carried out by a Monte Carlo dynamics for a Potts' model (4). Pott's parameters  $\lambda$ , for sequences with length  $L$  and alphabet size  $q$ , are chosen randomly from a Gaussian distribution with mean  $\mu_\lambda > 0$  and standard deviation  $\sigma_\lambda \ll \mu_\lambda$ . The connectivity  $c$  between sites of the sequences could take different values in order to simulate regimes with different covariance relevance.

Given a tree structure, the direct evolutionary process starts sampling the root configuration from (4) via MCMC. Then, each bifurcation event in the tree starts from the configuration at the ancient node and on average  $\Delta t * L * \delta$  MCMC changes are proposed with Metropolis acceptance rate, leading to a new configuration at the child node. Here  $\Delta t$  is the tree branch length between two nodes and  $\delta$  is the time-scale for the MCMC simulation from which we regulate the phylogeny. From sampling tree-nodes via MCMC we obtain at leaves the discrete data configurations  $\{\underline{X}^m\}$  and as internal nodes sequences which we would like to predict  $\{\underline{A}^m\}$ .

As we already discussed in the introduction, most ASR algorithms of the literature assume that sequences follow a single-site evolutionary model. Within this class, we will use an efficient dynamic programming algorithm developed by Pupko et al.[6] (included in FastML program) as a benchmark to compare our results. For this, we use the independent site model of evolution, introduced by Felsenstein [2], with constant mutation rate  $\mu$  given by

$$P_i(A_1|A_2, \Delta t) = e^{-\mu\Delta t}\delta_{A_1, A_2} + (1 - e^{-\mu\Delta t})P_i(A_1) \quad (31)$$

This model describes, in a time interval  $\Delta t$ , no mutations with probability  $e^{-\mu\Delta t}$  and one or more mutations with probability  $1 - e^{-\mu\Delta t}$ . In this last case, a specific character states is selected according  $P_i(A_1)$ . Details of our implementation of the algorithm appear in appendix (A 4 a). In the next section, we are going to show results for the reconstruction process carried out with both, the FastML algorithm and the continuous OU approximation discussed above.

Notice that the OU propagator needs both  $\mathbf{C}$  and  $\gamma$ .  $\mathbf{C}$  is obtained from an i.i.d sample of the Potts model, which is transformed to its continuous version as described above. On the other hand, the time-scale parameter  $\gamma$  is inferred by maximizing the likelihood of the data points at leaves of the tree as is described in appendix (A 4 b).

## A. Numerical Results

We generate data via MCMC simulations, as was described in the previous section. For this, we consider a tree of binary and homogeneous topology with  $h = 9$  bifurcations events. For a system with  $L = 10$  and  $q = 2$ , a Potts model of connectivity  $c = 3$  is designed with ferromagnetic couplings Gaussian distributed with mean  $\mu_J = 0.8$  and standard deviation  $\sigma_J = 0.2$ . For simulations, we use different time-scale  $\delta$  in the range  $[0.2, 11.2]$ , which allows us to explore different regimes of correlation between nodes configurations. To avoid statistical noise, we repeat the sampling procedure 100 times for each value of  $\delta$ . Then we applied two different reconstructions strategies: continuous OU approximation with either the full covariance matrix  $\mathbf{C}$  (MP1) or with a matrix resulting of neglect sites covariance  $\text{diag}(\mathbf{C})$  (MP0) and Pupko’s ML strategy with Felsenstein evolutionary model of equation (31) (FastML).

In Figure 5 we present results of the different strategies. We use the hamming distance between real and inferred ancestral sequences to evaluate the performance of the predictions. As can be seen, the performance between the different no-correlated approximations (MP0 and FastML) almost coincide, while MP1 is consistently better than both approximations. The ratios between the hamming distances obtained from methods MP0 or FastML with hamming distances from MP1 are shown in the inset plot. This illustrates that the major gain in the prediction came from using MP1 instead of FastML and at intermediate values of  $\delta$  where node configurations have enough divergence but are still correlated.

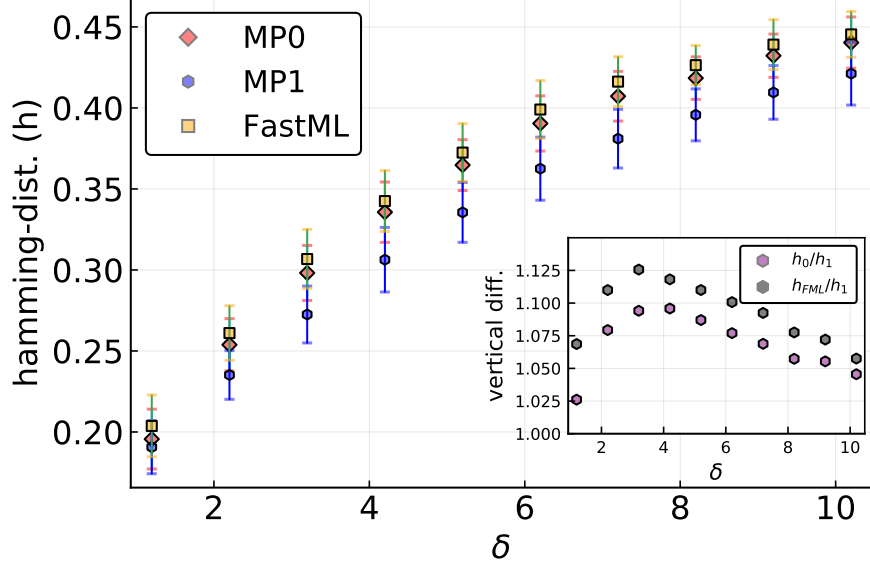


FIG. 5. Mean hamming distance between all true sequences and those obtained after the inference method. Each point correspond to 100 sampling-inference process averaged. The length of the sequence vectors was  $L = 10$  and the alphabet size  $q = 2$ , the average site connectivity is  $c = 3$ , the number of ancestral sequences  $M = 511$  as consequence of a binary non-homogeneous tree with  $K = 9$  bifurcations processes. Label MP0 correspond to the case when inference is carried out using continuous approximation neglecting covariance between sites, MP1 to the case when the inference is done using the full correlation matrix  $\mathbf{C}$  and FastML is the reference algorithm for maximum-likelihood reconstruction developed by Pupko [6]. The inserted plot indicates the ratio between hamming distances from MP0 or FastML with hamming distances from MP1 approach.

## VI. CONCLUSIONS

In this work, we have studied the reconstruction of ancestral sequences described by global co-evolutionary models and study the impact of the intra-species correlations in the performance of the inference process. We studied two types of sequences. Sequences of continuous variables that evolved according to an Ornstein-Uhlenbeck dynamics on a finite tree, and sequences of discrete variables defined by a Potts model and sampled on trees via Monte Carlo simulations. Exploiting the Gaussian character of the Ornstein-Uhlenbeck process, we were able to design a fast algorithm that provides a MAP estimate of the ancestral sequences and takes into account the covariance between variables, i.e. our

algorithm goes beyond the standard independent site approximation. We were also able to analytically quantify the precision of our reconstruction and showed that this analytical computation describes correctly the results from the algorithm in artificial data sets. Then, exploiting a known projection of the Potts model on a multivariate Gaussian distribution, we tested our algorithm in artificial discrete data sets. Also, for this kind of models, our results support the idea that there is a wide range of parameters for which the intra-species correlation significantly affect the inference process. We finally show that in this regime, our algorithm outperforms a traditional reconstruction method based on the independent site approximation.

A different, but connected issue, is the ability to distinguish, from a sequence alignment, which correlations are originated by phylogeny and which ones are originated by epistasis. In the context of Inverse Statistical Physics of biological sequences, global co-evolutionary models are used to describe sequence variability in ensembles of homologous sequences. This allows to unveil statistical constraints acting on this variability and relate them to biological features. Potts models in Direct Coupling Analysis (DCA) have found widespread application in protein-structure prediction from sequences [32, 33]. However, one of the basic assumptions of this global statistical modeling is that sequences form an, at least, approximately independent sample of an unknown probability distribution, which is to be learned from data. In the case of protein families, this assumption is obviously violated by phylogenetic relations between protein sequences. In [28] it was shown that phylogenetic correlations between sequences lead to a changed residue-residue correlation structure, represented by a fat tail in the eigenvalue spectrum of the data covariance matrix. Furthermore, the phylogeny has a global influence on all the parameters of the model and it could impact in those applications that use DCA as a sequence model.

Notice that, the equilibrium distribution reached by the OU process is the Gaussian version of the DCA model, where the covariance matrix  $\mathbf{C}$  is a quadratic potential that represents selection forces. Then, in the light of results of [28], we must consider that when we model biological sequences, the phylogenetic correlations interfere in the covariance signal empirically estimated from data and this not properly represent the phenotypic constraints resulting from natural selection. It is therefore necessary to disentangle covariances in the data corresponding to their multivariate Gaussian equilibrium distribution from those resulting from the historical correlations. To solve this problem [26] developed a methodology

which leads to a clear gain in accuracy in the inferred equilibrium distribution. However, in our ancestral reconstruction method, we assume that this covariance matrix between residues is known. A future step is to couple both inference methods.

**Acknowledgments:** We acknowledge interesting discussions with Martin Weigt and Pierre Barrat-Charlaix. This project has received funding from the European Union’s Horizon 2020 research and innovation program MSCA-RISE-2016 under Grant Agreement No. 734439 INFERNET. This research has also been funded by the Office for the management of funds and projects of the Ministry of Science, Technology and Environment of the Republic of Cuba within the project PN223LH010-015.



- 
- [1] Joy JB, Liang RH, McCloskey RM, Nguyen T, and Poon AFY. Ancestral reconstruction. *PLoS Comput Biol*, 2016,12(7),e1004763.
  - [2] J Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *J. Mol. Evol*, 1981,17,368–376.
  - [3] Yang Z, Kumar S, and Nei M. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 1995,141(4),1641-50, PMID: 8601501.
  - [4] Koshi JM and Goldstein RA. Probabilistic reconstruction of ancestral protein sequences. *Journal of Molecular Evolution*, 1996,42(2),313-20, PMID: 8919883.
  - [5] Pagel M. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology*, 1999,48(3),612-622.
  - [6] JPupko T, Pe I, Shamir R, and Graur D. A. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular Biology and Evolution*, 2000,17(6),890-6. PMID:10833195.
  - [7] Yang Z. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 2007,24(8),1586-91, PMID: 17483113.
  - [8] Huelsenbeck JP and Bollback JP. Empirical and hierarchical bayesian estimation of ancestral states. *Systematic Biology*, 2001,50(3),351-66, PMID: ISI:000169823200006.
  - [9] Peter K. Vlasov Cedric Notredame Michael S. Breen, Carsten Kemena and Fyodor A. Kondrashov. Epistasis as the primary factor in molecular evolution. *Nature*, 2012,490.7421, 535–538, ISSN: 0028-0836, 1476-4687. doi: 10.1038/ nature11510.
  - [10] Michael J. Harms and Joseph W. Thornton. Evolutionary bio-chemistry: Revealing the historical and physical causes of protein properties. *Nature Reviews Genetics*, 2013, , 559–571, ISSN: 1471-0056, 1471-0064. doi: 10.1038/nrg3540.
  - [11] Anders Olson C., Nicholas C. Wu, and Ren Sun. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol*, 2014,24.22,2643–2651, ISSN: 0960-9822. doi: 10.1016/j.cub.2014.09.072.
  - [12] Frank J Poelwijk Michael A Stiffler Nicholas P Gauthier Chris Sander Nathan J Rollins, Kelly P Brock and Debora S Marks. 3d protein structure from genetic epistasis experiments. *Curr Biol*, 2018,24.22,2643–2651, doi: 10.1101/320721.
  - [13] Kimura M. Attainment of quasi linkage equilibrium when gene frequencies are changing by

- natural selection. *Genetics*, 1965, 52 , 875–890.
- [14] Angelo Vulpiani Hai-Jun Zhou Chen-Yi Gao, Fabio Cecconi and Erik Aurell. Dca for genome-wide epistasis analysis: the statistical genetics perspective. *Physical Biology*, 2019, 16 , 026002.
  - [15] Nguyen H. Chau, Zecchina R. N, and Berg J. Inverse statistical problems: from the inverse ising problem to data science. *Adv. Phys.*, 2017, 66, 97-261.
  - [16] Ronald M Levy, Allan Haldane, and William F Flynn. Potts hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Current Opinion in Structural Biology*, 2017, 43, 55-62.
  - [17] Matteo Figliuzzi Remi Monasson Simona Cocco, Christoph Feinauer and Martin Weigt. Inverse statistical physics of protein sequences: A key issues review. *Reports on Progress in Physics*, 2018, 81(3):032601.
  - [18] B. Lunt A. Bertolino-D. S. Marks C. Sander R. Zecchina J. N. Onuchic T. Hwa F. Morcos, A. Pagnani and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 2011, 108(49) ,E1293–E1301.
  - [19] Christian Stocker Pierre Barrat-Charlaix Michael Socol-ich Peter Kast Donald Hilvert Remi Monasson Simona Cocco Martin Weigt William P. Russ, Matteo Figliuzzi and Rama Ranganathan. An evolution-based model for designing chorismate mutase enzymes. *Science*, 2020, 369(6502) ,440–445.
  - [20] Zeng HL., Dichio V., Rodríguez-Horta E., and Aurell E. Thorell K. Global analysis of more than 50,000 sars-cov-2 genomes reveals epistasis between eight viral genes. *Proceedings of the National Academy of Sciences*, 2020, ,p. 202012331.
  - [21] Huelsenbeck JP and Nielsen R. Effect of nonindependent substitution on phylogenetic accuracy. *Systematic Biology*, 48:317–328, 1999.
  - [22] Nasrallah CA, Mathews DH, and Huelsenbeck JP. Quantifying the impact of dependent evolution among sites in phylogenetic inference. *Systematic Biology*, 60:60–73, 2011.
  - [23] Muntoni Anna Paola, Andrea Pagnani, Martin Weigt, and Francesco Zamponi. Aligning biological sequences by exploiting residue conservation and coevolution. *Phys. Rev. E*, 102, 2020.
  - [24] Krzysztof Bartoszek, Jason Pienaar, Petter Mostad, Staffan Andersson, and Thomas F. Hansen. A phylogenetic comparative method for studying multivariate adaptation. *Journal of Theoretical*

- Biology*, 314:204–215, 2012.
- [25] Venelin Mitov, Krzysztof Bartoszek, Georgios Asimomitis, and Tanja Stadler. Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts. *Theoretical Population Biology*, 131:66–78, 2020.
  - [26] Edwin Rodríguez Horta, Alejandro Lage-Castellanos, Pierre Barrat Charlaix, and Martin Weigt. Global multivariate model learning from hierarchically correlated data. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(7):073501, jul 2021.
  - [27] Baldassi C., Zamparo M., Feinauer C., Procaccini A., Zecchina R., Weigt M., and Pagnani A. A fast and accurate multivariate gaussian modeling of protein families: Predicting residue contacts and protein-interaction partners. *PLoS ONE*, 2014,9,e92721.
  - [28] Edwin Rodriguez Horta and Martin Weigt. On the effect of phylogenetic correlations in coevolution based contact prediction in proteins. *PLoS Comput Biol*, 17(5), 2021.
  - [29] Weiss Y. and Freeman W. T. Correctness of belief propagation in gaussian graphical models of arbitrary topology. *Neural Computation*, 13:2173–2200, 2001.
  - [30] Malioutov D.M., Johnson J.K., and Willsky A.S. Walk-sums and belief propagation in gaussian graphical models. *Machine Learning Research*, 7:2031–2064, 2006.
  - [31] Dany Bickson. Gaussian belief propagation: Theory and application. *arXiv*, 2009,0811.2518(cs).
  - [32] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.
  - [33] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Remi Monasson, and Martin Weigt. Inverse Statistical Physics of Protein Sequences: A Key Issues Review. *Reports on Progress in Physics*, 81(3):032601, 2018. arXiv: 1703.01222.
  - [34] C.W Gardiner. *Handbook of Stochastic Methods*. 2004.
  - [35] Rajesh Singh, Dipanjan Ghosh, and R. Adhikari. Fast bayesian inference of the multivariate ornstein-uhlenbeck process. *arxiv:1706.04961*, 2017.

## Appendix A: Description of technical details

### 1. Ornstein-Uhlenbeck dynamics

Let us consider a system characterized by  $L$  continuous degrees of freedom and whose state is fully described by an  $L$ -dimensional vector  $\vec{x} \in \mathbb{R}^L$ . We suppose that this system evolves under the potential  $V(\vec{x}) = \frac{1}{2}\{\vec{x}^T \mathbf{J} \vec{x}\}$  being  $\mathbf{J}$  a symmetric and positive definite coupling matrix, and according to the Langevin equation

$$\gamma^{-1} \frac{d\vec{x}}{dt} = -\mathbf{J}\vec{x} + \vec{\xi}(t) \quad (\text{A1})$$

which represents a multivariate Ornstein-Uhlenbeck process, where  $\vec{\xi}(t)$  is an stochastic term, and  $\gamma^{-1}$  is the characteristic timescale governing the dynamics. Modeling the stochastic term  $\vec{\xi}(t)$  as a uncorrelated white noise, we obtain the Ito stochastic differential equation for a multivariate OU process [34, 35]:

$$dx_i = - \sum_{j=1}^L \left( \gamma J_{ij} x_j dt + \left( \sqrt{2D} \right)_{ij} dW_j \right) \quad (\text{A2})$$

where:

- $dW_j = \xi_j(t)dt$  represent a stochastic Wiener process.
- $D_{ij} = \gamma \delta_{ij}$  is a matrix of diffusions coefficients.

It can be shown that the corresponding Fokker-Planck equation is

$$\partial_t P_{1|1} = \hat{L} P_{1|1} \quad (\text{A3})$$

where  $P_{1|1} \equiv P(\vec{x}, t | \vec{x}_0, t_0)$  is the probability density of displacement from  $\vec{x}_0$  at time  $t_0$  to  $\vec{x}$  at time  $t$ , and  $\hat{L}$  is the Fokker-Planck operator given by:

$$\hat{L}(\vec{x}) = -\gamma \left( \sum_{i,j=1}^L \frac{\partial}{\partial x_i} J_{ij} x_j + \sum_{i=1}^L \frac{\partial^2}{\partial x_i^2} \right) \quad (\text{A4})$$

The stationary solution for the Fokker-Planck equation  $\hat{L}P^0(\vec{x}) = 0$  is:

$$P^0(\vec{x}) = \frac{1}{\sqrt{(2\pi)^L |\mathbf{C}|}} \exp \left\{ -\frac{1}{2} \vec{x}^T \mathbf{C}^{-1} \vec{x} \right\} \quad (\text{A5})$$

which is a zero mean Gaussian distribution with  $\mathbf{C} = \langle \vec{x} \vec{x}^T \rangle$  as the covariance matrix. If from the beginning we add to the potential  $V(\vec{x})$  a linear term  $\vec{h} \cdot \vec{x}$  with  $\vec{h}$  a local field, we would obtain a Gaussian distribution with mean value shift from zero and covariance matrix  $\mathbf{C} = \langle \vec{x} \vec{x}^T \rangle - \langle \vec{x} \rangle \langle \vec{x} \rangle$ .

The solution of the Fokker-Planck equation is a multivariate normal distribution

$$P(\vec{x}|\vec{x}_0, \Delta t) = [(2\pi)^L \det \mathbf{\Sigma}]^{-1/2} \exp \left\{ -\frac{1}{2} (\vec{x} - \vec{\mu})^T \mathbf{\Sigma}^{-1} (\vec{x} - \vec{\mu}) \right\} \quad (\text{A6})$$

where

$$\mathbf{\Lambda} = e^{-\gamma \mathbf{J}}, \quad \vec{\mu} = \mathbf{\Lambda}^{\Delta t} \vec{x}_0, \quad \mathbf{\Sigma} = \mathbf{C} - \mathbf{\Lambda} \mathbf{C} \mathbf{\Lambda}, \quad \Delta t = t - t_0$$

Matrices  $\mathbf{J}$ ,  $\mathbf{C}$  and  $\mathbf{D}$  are not independent as they are related by Liapunov stationary condition :

$$\mathbf{J} \mathbf{C} + (\mathbf{C} \mathbf{J})^T = 2\mathbf{D}$$

from which we get

$$\mathbf{J} = \mathbf{C}^{-1} \quad (\text{A7})$$

Then we can conclude that the system evolves under a quadratic potential  $V(\vec{x}) = \frac{1}{2} \vec{x}^T \mathbf{C}^{-1} \vec{x} = \frac{1}{2} \sum_{i,j} J_{ij} x_i x_j$  and the evolutionary process leads to stationary distribution A5 when  $\gamma \Delta t \gg 1$ .

Another important property of the Ornstein-Uhlenbeck process is that the times correlation function obeys the linear regression theorem

$$\mathbf{G}(t_1 - t_2) \equiv \langle \vec{x}_1(t_1) \vec{x}_2(t_2) \rangle = \mathbf{\Lambda}^{\Delta t} \mathbf{C} \quad (\text{A8})$$

describing the covariance of configurations  $\vec{x}_1$  and  $\vec{x}_2$  separated in time by  $\Delta t = t_1 - t_2$ .

## 2. Message passing for continuous variables

Posterior probability distribution (12) can be written from the following pairwise factorization

$$P^d(\vec{A}|\vec{D}) \propto \prod_{i=1}^{N_h} \phi_i(\vec{x}_i) \prod_{1 \leq i < j \leq N_h} \psi_{ij}(\vec{x}_i, \vec{x}_j) \quad (\text{A9})$$

where  $\psi_{ij} = \exp(\vec{x}_i \mathbf{J}_{ij} \vec{x}_j)$  and  $\phi_i(\vec{x}_i) = \exp(\vec{x}_i \mathbf{H}_i \vec{x}_i + h_i \vec{x}_i)$  are edge potentials and self potential which define the tree-graph.

Message from  $i$  to  $j$  over their shared edge on the tree is given by,

$$m_{ij}(\vec{x}_j) \propto \int d\vec{x}_i \psi_{ij}(\vec{x}_i, \vec{x}_j) \phi_i(\vec{x}_i) \prod_{k \in \partial i \setminus j} m_{ki}(\vec{x}_i) \quad (\text{A10})$$

and marginals

$$M_i(\vec{x}_i) \propto \phi_i(\vec{x}_i) \prod_{k \in \partial i} m_{ki}(\vec{x}_i) \quad (\text{A11})$$

*a. Gaussian message passing algorithm [31]*

From previous expressions, we note that both binary and unary potentials are Gaussian :

$$\phi_i(\vec{x}_i) \propto N(\vec{\mu}_i = \frac{-\vec{h}_i}{2 * \mathbf{H}_i}, \boldsymbol{\sigma}_i = -\frac{1}{2 * \mathbf{H}_i}) \quad (\text{A12})$$

Since Gaussian densities' product over a common variable is, up to a constant factor, also a Gaussian density, we can write messages in the following way

$$m_{ij}(\vec{x}_j) \propto N(\vec{\mu}_{ij}, \boldsymbol{\sigma}_{ij}^{-1}) \quad (\text{A13})$$

and the product

$$\phi_i(\vec{x}_i) \prod_{k \in \partial i \setminus j} m_{ki}(\vec{x}_i) \propto N(\vec{\mu}_{i \setminus j}, \boldsymbol{\sigma}_{i \setminus j}^{-1}) \quad (\text{A14})$$

is also Gaussian with

$$\begin{aligned} \boldsymbol{\sigma}_{i \setminus j}^{-1} &= \boldsymbol{\sigma}_i^{-1} + \sum_{k \in \partial i \setminus j} \boldsymbol{\sigma}_{ki}^{-1} \\ \vec{\mu}_{i \setminus j} &= \boldsymbol{\sigma}_{i \setminus j} \left( \boldsymbol{\sigma}_i^{-1} \vec{\mu}_i + \sum_{k \in \partial i \setminus j} \boldsymbol{\sigma}_{ki}^{-1} \vec{\mu}_{ki} \right) \end{aligned} \quad (\text{A15})$$

Plugging this into the message defining equation A10, we obtain

$$\begin{aligned}
m_{ij}(\vec{x}_j) &\propto \int d\vec{x}_i \psi_{ij}(\vec{x}_i, \vec{x}_j) \phi_i(\vec{x}_i) \prod_{k \in \partial i \setminus j} m_{ki}(\vec{x}_i) \\
&\propto \int d\vec{x}_i \exp(\vec{x}_i \mathbf{J}_{ij} \vec{x}_j) \exp(-\frac{1}{2} \vec{x}_i \boldsymbol{\sigma}_{i \setminus j}^{-1} \vec{x}_i + \boldsymbol{\sigma}_{i \setminus j}^{-1} \vec{\mu}_{i \setminus j} \vec{x}_i) \\
&= \int d\vec{x}_i \exp(-\frac{1}{2} \vec{x}_i \boldsymbol{\sigma}_{i \setminus j}^{-1} \vec{x}_i + (\boldsymbol{\sigma}_{i \setminus j}^{-1} \vec{\mu}_{i \setminus j} + \mathbf{J}_{ij} \vec{x}_j) \vec{x}_i) \\
&\propto \exp((\boldsymbol{\sigma}_{i \setminus j}^{-1} \vec{\mu}_{i \setminus j} + \mathbf{J}_{ij} \vec{x}_j)^2 / (2 * \boldsymbol{\sigma}_{i \setminus j}^{-1})) \\
&\propto N(\vec{\mu}_{ij} = \boldsymbol{\sigma}_{ij} \mathbf{J}_{ij} \vec{\mu}_{i \setminus j}, \boldsymbol{\sigma}_{ij} = -\boldsymbol{\sigma}_{i \setminus j}^{-1} / (\mathbf{J}_{ij})^2)
\end{aligned} \tag{A16}$$

leading to the update rules for Gaussian parameters:

$$\begin{aligned}
\boldsymbol{\sigma}_{ij}^{-1} &= -\mathbf{J}_{ij}^2 / (\boldsymbol{\sigma}_i^{-1} + \sum_{k \in \partial i \setminus j} \boldsymbol{\sigma}_{ki}^{-1}) \\
\vec{\mu}_{ij} &= -(\vec{\mu}_i \boldsymbol{\sigma}_i^{-1} + \sum_{k \in \partial i \setminus j} \boldsymbol{\sigma}_{ki}^{-1} \vec{\mu}_{ki}) / \mathbf{J}_{ij}
\end{aligned} \tag{A17}$$

substituting A12 and A16 in A11 we obtain the marginals as a Gaussian density  $M_i(\vec{x}_i) \propto N(\vec{\eta}_i, \boldsymbol{\kappa}_i)$  with

$$\begin{aligned}
\vec{\eta}_i &= (\boldsymbol{\sigma}_i^{-1} \vec{\mu}_i + \sum_{k \in \partial i} \boldsymbol{\sigma}_{ki}^{-1} \vec{\mu}_{ki}) / (\boldsymbol{\sigma}_i^{-1} + \sum_{k \in \partial i} \boldsymbol{\sigma}_{ki}^{-1}) \\
\boldsymbol{\kappa}_i^{-1} &= \boldsymbol{\sigma}_i^{-1} + \sum_{k \in \partial i} \boldsymbol{\sigma}_{ki}^{-1}
\end{aligned} \tag{A18}$$

Update rules equation in A18 match with ones associated to the max-product algorithm, as we will show below. This means that our solution in (18) can be obtained as

$$\vec{x}_i^* = \vec{\eta}_i$$

#### b. Max-Product rule

A continuous version of max-product algorithm could be obtained, replacing the integral-product rule by

$$m_{ij}(\vec{x}_j) \propto \arg \max_{\vec{x}_i} \psi_{ij}(\vec{x}_i, \vec{x}_j) \phi_i(\vec{x}_i) \prod_{k \in \partial i \setminus j} m_{ki}(\vec{x}_i) \tag{A19}$$

similar to A16 we get

$$\begin{aligned}
m_{ij}(\vec{x}_j) &\propto \arg \max_{\vec{x}_i} \exp(\vec{x}_i \mathbf{J}_{ij} \vec{x}_j) \exp(-\frac{1}{2} \vec{x}_i \boldsymbol{\sigma}_{i \setminus j}^{-1} \vec{x}_i + \boldsymbol{\sigma}_{i \setminus j}^{-1} \vec{\mu}_{i \setminus j} \vec{x}_i) \\
&= \arg \max_{\vec{x}_i} \exp(-\frac{1}{2} \vec{x}_i \boldsymbol{\sigma}_{i \setminus j}^{-1} \vec{x}_i + (\boldsymbol{\sigma}_{i \setminus j}^{-1} \vec{\mu}_{i \setminus j} + \mathbf{J}_{ij} \vec{x}_j) \vec{x}_i)
\end{aligned} \tag{A20}$$

deriving and equating to zero the exponential term, we find

$$\vec{x}_i^{max} = \frac{\boldsymbol{\sigma}_{i \setminus j}^{-1} \vec{\mu}_{i \setminus j} + \mathbf{J}_{ij} \vec{x}_j}{\boldsymbol{\sigma}_{i \setminus j}^{-1}} \tag{A21}$$

substituting  $\vec{x}_i^{max}$  back in A20 we get,

$$\begin{aligned}
m_{ij}(\vec{x}_j) &\propto \exp((\boldsymbol{\sigma}_{i \setminus j}^{-1} \vec{\mu}_{i \setminus j} + \mathbf{J}_{ij} \vec{x}_j)^2 / (2 * \boldsymbol{\sigma}_{i \setminus j}^{-1})) \\
&\propto N(\vec{\mu}_{ij} = \boldsymbol{\sigma}_{ij} \mathbf{J}_{ij} \vec{\mu}_{i \setminus j}, \boldsymbol{\sigma}_{ij} = -\boldsymbol{\sigma}_{i \setminus j}^{-1} / (\mathbf{J}_{ij})^2)
\end{aligned} \tag{A22}$$

which is identical to the messages derived for the sum-product case, then as intuitively we could guess the rules obtained to find the marginals (A17 and A18) leads to max marginals for the Gaussian version of Message Passing and then as marginals are Gaussian the maximum value correspond to the mean.



### 3. Integration of equation 24

We must first evaluate:

$$\begin{aligned}
I_3 &= \int d\vec{D} \exp \left\{ -\frac{1}{2} \vec{D} \mathbb{G} \vec{D}^T + \vec{D} \mathbb{A}^T \vec{A}^T - \frac{i}{2} \vec{q} * \mathbb{K}_0^{-1} \mathbb{A}_0 \vec{D}^T \right\} \\
&= cte * \exp \left\{ \frac{1}{2} \left( \vec{A} \mathbb{A} - \frac{i * \vec{q} \mathbb{K}_0^{-1} \mathbb{A}_0}{2} \right) \mathbb{G}^{-1} \left( \vec{A} \mathbb{A} - \frac{i * \vec{q} \mathbb{K}_0^{-1} \mathbb{A}_0}{2} \right)^T \right\}
\end{aligned} \tag{A23}$$

defining  $\mathbb{Q} = \mathbb{A} \mathbb{G}^{-1} \mathbb{A}^T$ ,  $\mathbb{Q}_0 = \mathbb{A}_0 \mathbb{G}^{-1} \mathbb{A}_0^T$  and  $\mathbb{Q}_1 = \mathbb{A}_0 \mathbb{G}^{-1} \mathbb{A}^T$  we obtain:

$$I_3 = cte * \exp \left\{ \frac{1}{2} \left[ \vec{A} \mathbb{Q} \vec{A}^T - i * \vec{A} (\mathbb{Q}_1^T \mathbb{K}_0^{-1}) \vec{q}^T - \frac{1}{4} \vec{q} \mathbb{K}_0^{-1} \mathbb{Q}_0 \mathbb{K}_0^{-1} \vec{q}^T \right] \right\} \tag{A24}$$

Substituting  $I_3$  in 24 we get:

$$P(\vec{A}, \vec{M}) \propto \exp \left\{ \frac{1}{2} \left[ \vec{A} \mathbb{Q} \vec{A}^T + 2 \vec{A} \mathbb{K} \vec{A}^T \right] \right\} * I_4 \tag{A25}$$

where

$$\begin{aligned}
I_4 &= \int d\vec{q} \exp \left\{ -\frac{1}{2} \vec{q} \frac{\mathbb{K}_0^{-1} \mathbb{Q}_0 \mathbb{K}_0^{-1}}{4} \vec{q}^T - \frac{i}{2} \left( \vec{A} \mathbb{Q}_1^T \mathbb{K}_0^{-1} + 2 \vec{M} \right) * \vec{q}^T \right\} \\
&= cte * \exp \left\{ -\frac{1}{2} \left( \vec{A} \mathbb{Q}_1^T \mathbb{K}_0^{-1} + 2 \vec{M} \right) \mathbb{K}_0 \mathbb{Q}_0^{-1} \mathbb{K}_0 \left( \vec{A} \mathbb{Q}_1^T \mathbb{K}_0^{-1} + 2 \vec{M} \right)^T \right\}
\end{aligned} \tag{A26}$$

and replacing  $I_4$  in A25 we finally get:

$$\begin{aligned}
P(\vec{A}, \vec{M}) &\propto \exp \left\{ -\frac{1}{2} \left[ -2 \vec{A} \mathbb{K} \vec{A}^T + 4 \vec{A} \mathbb{Q}_1^T \mathbb{Q}_0^{-1} \mathbb{K}_0 \vec{M}^T + 4 * \vec{M} \mathbb{K}_0 \mathbb{Q}_0^{-1} \mathbb{K}_0 \vec{M}^T \right] \right\} \\
&\propto \exp \left\{ -\frac{1}{2} \vec{z} \mathbb{V}^{-1} \vec{z}^T \right\}
\end{aligned} \tag{A27}$$

with  $\vec{z} = \begin{bmatrix} \vec{A} \\ \vec{M} \end{bmatrix}$  and

$$\mathbb{V}^{-1} = \begin{pmatrix} -2\mathbb{K} & 2\mathbb{Q}_1^T \mathbb{Q}_0^{-1} \mathbb{K}_0 \\ 2\mathbb{K}_0 \mathbb{Q}_0^{-1} \mathbb{Q}_1 & 4\mathbb{K}_0 \mathbb{Q}_0^{-1} \mathbb{K}_0 \end{pmatrix} \tag{A28}$$

inverting  $\mathbb{V}^{-1}$  we get:

$$\mathbb{V} = \begin{pmatrix} -[2\mathbb{K} + \mathbb{Q}]^{-1} & \frac{1}{2}[2\mathbb{K} + \mathbb{Q}]^{-1} * \mathbb{Q}_1^T \mathbb{K}_0^{-1} \\ \frac{1}{2}\mathbb{K}_0^{-1} \mathbb{Q}_1 [2\mathbb{K} + \mathbb{Q}]^{-1} & \frac{1}{4}\mathbb{K}_0^{-1} [\mathbb{Q}_0 - \mathbb{Q}_1 [2\mathbb{K} + \mathbb{Q}]^{-1} \mathbb{Q}_1^T] \mathbb{K}_0^{-1} \end{pmatrix} \quad (\text{A29})$$

which allows to compute the error as:

$$\hat{d}(\mathbf{C}, \mathbf{C}_0) = \sum_i ([\mathbb{V}_{11}]_{ii} - 2 * [\mathbb{V}_{12}]_{ii} + [\mathbb{V}_{22}]_{ii}) \quad (\text{A30})$$

#### 4. Initializing parameters

##### a. Time scale parameter $\mu$ for FastML inference

The mutation rate parameter  $\mu$  in the Felsenstein model 31 is typically unknown, then it must be inferred from the data. Note that under model 31 the average Hamming distance between two equilibrium sequences at evolutionary time distance  $\Delta t$  can be computed as

$$\bar{d}_H(\Delta t) = (1 - e^{-\mu \Delta t}) \bar{d}_H(\infty) \quad (\text{A31})$$

where average  $\bar{d}_H(\infty)$  is the Hamming distance between two independent equilibrium sequences in the independent-site model. Therefore, we can take any two sequences at leaves of the tree, calculate their Hamming distance together with their time separation on the phylogenetic tree by adding all branch lengths along their connecting path, and use the result as an instance of  $d_H(\Delta t)$ . Taking all pairs of sequences from the alignment, we can bin the observed times, calculate average Hamming distances for each time bin, and fit the functional form of equation A31 to obtain the desired value of  $\mu$ .

As proof of concept, we show our implementation of this algorithm for data generated with single site model 31 on a homogeneous and binary tree with  $H = 9$  bifurcation events. Figure 6 show results of fitting equation A31 and Figure 7 show the distance between inferred and true ancestral sequences for different values of the mutation rates. From figures, it is possible to note that reconstruction get worse for higher mutations rate and for deepest internal nodes, an expected feature for this problem.

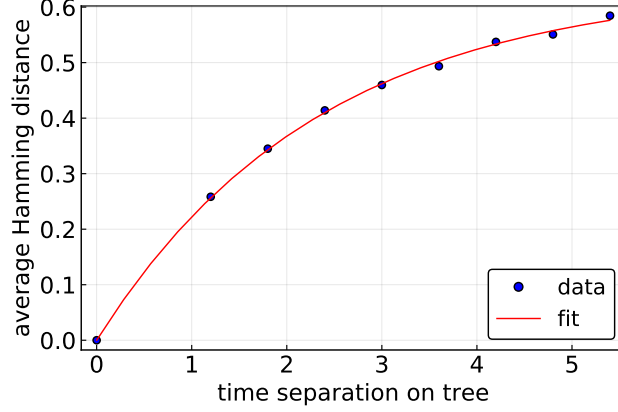


FIG. 6. Fitting  $\mu$

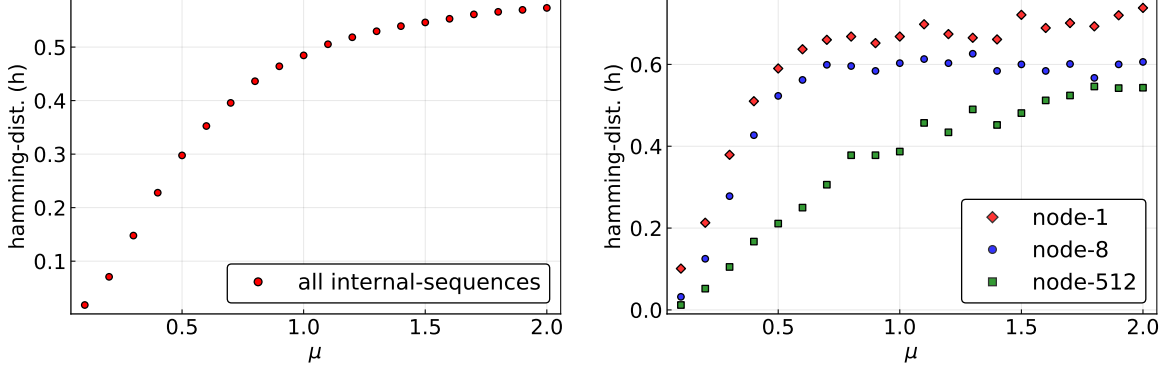


FIG. 7. Ancestral reconstruction via our implementation of FastML algorithm. **left:** Distance between inferred and true full set of ancestral sequences. **right:** Distance between inferred and true sequences at different internal nodes in the tree.

*b. Time scale parameter  $\gamma$*

The timescale parameter  $\gamma$  is a priori unknown and must be inferred from data as an initial step before the ancestral reconstruction algorithm. Since the process is always Gaussian, the distribution of the leaves has to be Gaussian itself. Furthermore, we know the covariance between any two elements. Within one leaf, the covariance is the equilibrium covariance  $\mathbf{C}$ , and among leaves it has to be  $\Lambda^{\Delta t_{ij}} \mathbf{C}$  with  $\Lambda = \exp(-\mathbf{C}^{-1} * \gamma)$  and  $\Delta t_{ij}$  is the path time between nodes  $i$  and  $j$  along the branches of the tree.

Then the leaves distribution is given by

$$P(\vec{D} = [\vec{x}_l^1, \dots, \vec{x}_l^{N_l}]) = \frac{1}{\sqrt{2\pi^{N_l * L} \det \mathbb{G}}} \times \exp\left(-\frac{1}{2} \vec{D} \mathbb{G}^{-1} \vec{D}^T\right)$$

where  $\mathbb{G}$  is a block matrix whose structure is induced by the phylogenetic tree and its elements are given by:

$$\mathbf{G}_{i,j} = \begin{cases} \mathbf{C} & i = j. \\ \Lambda^{\Delta t_{ij}} \mathbf{C} & \text{otherwise} \end{cases} \quad (\text{A32})$$

Then the likelihood is given by

$$L(\vec{D} | \mathbf{C}, \gamma) = \frac{1}{2} \log \det \mathbb{G}^{-1} - \frac{1}{2} \vec{D} \mathbb{G}^{-1} \vec{D}^T \quad (\text{A33})$$

depending only on the tree as well as on  $\mathbf{C}$  and  $\gamma$ . As we know the tree and the covariance matrix  $\mathbf{C}$  we can obtain the  $\gamma$  parameter by maximizing the likelihood function.