Small-coupling expansion for multiple sequence alignment

Louise Budzynski[1,2,*] and Andrea Pagnani[1,2,3]

[1]*DISAT, Politecnico di Torino, Corso Duca degli Abruzzi, 24, I-10129, Torino, Italy*
[2]*Italian Institute for Genomic Medicine, IRCCS Candiolo, SP-142, I-10060, Candiolo, Italy*
[3]*INFN, Sezione di Torino, Torino, Via Pietro Giuria, 1 10125 Torino Italy*

(Received 21 October 2022; accepted 27 March 2023; published 25 April 2023)

The alignment of biological sequences such as DNA, RNA, and proteins, is one of the basic tools that allow to detect evolutionary patterns, as well as functional or structural characterizations between homologous sequences in different organisms. Typically, state-of-the-art bioinformatics tools are based on profile models that assume the statistical independence of the different sites of the sequences. Over the last years, it has become increasingly clear that homologous sequences show complex patterns of long-range correlations over the primary sequence as a consequence of the natural evolution process that selects genetic variants under the constraint of preserving the functional or structural determinants of the sequence. Here, we present an alignment algorithm based on message passing techniques that overcomes the limitations of profile models. Our method is based on a perturbative small-coupling expansion of the free energy of the model that assumes a linear chain approximation as the zeroth-order of the expansion. We test the potentiality of the algorithm against standard competing strategies on several biological sequences.

DOI: 10.1103/PhysRevE.107.044125

I. INTRODUCTION

The evolution of biological molecules such as proteins is an ongoing, highly nontrivial dynamical process spanning over billions of years, constrained by the maintenance of relevant structural and functional determinants. One of the most striking features of natural evolution is how different evolutionary pathways produce an ensemble of molecules characterized by an extremely heterogeneous amino acid sequence, often with a sequence identity lower than 30%, but with virtually identical three-dimensional native structures. Thanks to the shrewd use of this structural similarity, it is nowadays possible to classify the entire set of known protein sequences into disjoint classes of sequences originating from a common ancestral sequence. Sequences belonging to the same class are called homologous.

Homologous sequences are best compared using sequence alignments [1]. Depending on the number of sequences to align, there are three possible options. (i) *Pairwise alignments* aim at casting two sequences into the same framework. The available algorithms are typically based on some versions of dynamic programming and scale linearly with the length of the sequences [2,3]. (ii) *Multiple sequence alignments* (MSA) maximize the global similarity of more than two sequences [4]. Dynamic programming techniques can be generalized to more than two sequences, but with a computational cost that scales exponentially with the number of sequences to be aligned. Producing MSAs of more than 10^3 sequences remains an open computational challenge. (iii) To align a larger number of homologous sequences, one first selects a representative subset called *seed* for which the use of MSA is computationally feasible. Every single homolog eventually is aligned to the *seed* MSA. In this way one can easily align up to 10^6 sequences [5–7].

Standard alignment methods are based on the *independent site evolution* assumption [1], i.e., the probability of observing a sequence is factorized among the different sites. From a statistical mechanics perspective, such an approximation corresponds to a noninteracting 21 colors (20 amino acids + 1 gap symbol) Potts model. Profile-hidden Markov models [6], for instance, are of that type. The computational complexity of profile models is polynomial. However, profile models neglect long-range correlations, although they are an important statistical feature of homologous proteins. This well-known phenomenon is at the basis of what biologists call *epistasis* (i.e., how genetic variation depends on the genetic context of the sequence). Recently, epistasis received renewed attention from the statistical mechanics' community [8]. Given an MSA of a specific protein family, one could ask what is the best statistical description of such an ensemble of sequences. Summary statistics such as one-site frequency count $f_i(a)$ (i.e., the empirically observed frequency of observing amino acid a at position i in the MSA), two-site frequency count $f_{ij}(a,b)$ (i.e., the frequency of observing the amino acid realization a, b at position i and j, respectively), and in principle higher-order correlations, could be used to inverse statistical modeling of the entire MSA. One can assume that each a sequence in the MSA is independently drawn from a multivariate distribution $P(a_1,\dots,a_L)$ constrained to reproduce the multibody empirical frequency counts of the MSA. The use of the maximum-entropy principle is equivalent to assume a Boltzmann-Gibbs probability measure for P. The related Hamiltonian is a 21-colors generalized

[*] Present address: Dipartimento di Fisica, Università La Sapienza, P.le A. Moro 5, 00185, Rome, Italy; louise.budzynski@polito.it

Potts model characterized by two sets of parameters: local fields $H_i(a)$ and epistatic two-site interaction terms $J_{i,j}(a, b)$. Such parameters can be learned more or less efficiently using the so-called direct coupling analysis (DCA) [9]. This method has found many interesting applications ranging from the prediction of protein structures [10,11], protein-protein interaction [12–14], prediction of mutational effects [15–18], and so on. Inherent to this strategy, there is the counterintuitive step of constructing an MSA based on a statistical independence of sites assumption, which is used, in turn, to predict long-range correlations. To solve this loophole, we propose a mean-field message-passing strategy to align sequences to a reference Potts model. To do so, we consider a first-order perturbative expansion *a la* Plefka [19], setting as zeroth order of the expansion the linear chain approximation. Recently, other strategies were proposed which take into account long-range correlations: the search for remote homology [20], a simplified version of the message-passing strategy presented here [21], the alignment of two Potts models [22], and a more machine-learning-inspired method based on transformers [23].

II. SETUP OF THE PROBLEM

Although here we will focus on proteins, the method can be extended to other biological sequences, such as RNA and DNA. Let $\mathbf{A} = (A_1, \dots, A_N)$ be an unaligned amino acid sequence of length N, containing a protein domain $\mathbf{S} = (S_1, \dots, S_L)$ of a known protein family. While $\mathbf{A}$ contains only amino acids (represented as upper-case letters from the amino acid alphabet), $\mathbf{S}$ might also contain gaps that are used to indicate the deletion of an amino acid in the sequence $\mathbf{A}$. We assume that the protein family is described by a Potts Hamiltonian

$$\mathcal{H}_{\rm DCA}(\mathbf{S}) = -\sum_{i=1}^{L} H_i(S_i) - \sum_{i<j} J_{ij}(S_i, S_j). \qquad (1)$$

The couplings J_{ij} and external fields H_i are learned from the seed MSA in a preprocessing step, using DCA, and the subsequence $\mathbf{S}$ is assumed to have the same length L as the seed. The energy $\mathcal{H}_{\rm DCA}$ is considered as a score for the subsequence $\mathbf{S}$ to belong to the protein family. In this setting, our problem consists in finding a subsequence $\mathbf{S}$ with the lowest energy (i.e., with the highest score). Contrary to profile models, the Hamiltonian $\mathcal{H}_{\rm DCA}$ also includes pairwise interactions related to residue coevolution, hopefully leading to more accurate alignments in cases where the conservation of single residues is not sufficient to describe the protein family. The Hamiltonian in Eq. (1) does not model the insertions statistics because the parameters J_{ij} and H_i are learned from the seed MSA in which all columns containing inserts are removed. Therefore, as in [21], we added the insertion cost $\mathcal{H}_{\rm ins}$, whichis learned from the insertion statistics contained in the full seed alignment. Similarly to [21], we also added an additional gap cost $\mathcal{H}_{\rm gap}$ to correct the gap statistics learned in $\mathcal{H}_{\rm DCA}$ (that deeply depends on how the seed is constructed). In this setting, the alignment problem corresponds to finding a subsequence $\mathbf{S} = (S_1, \dots, S_L)$ of the original sequence $\mathbf{A} = (A_1, \dots, A_N)$, such that the following conditions hold.

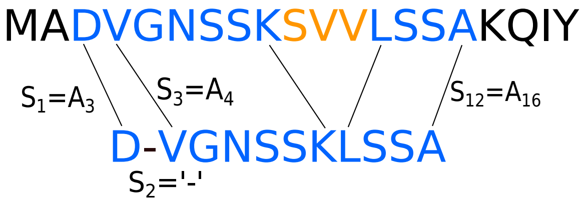

FIG. 1. Example of alignment. Top: Original sequence $\mathbf{A}$ of length $N = 20$. Bottom: Aligned sequence $\mathbf{S}$ of length $L = 12$. Match states are enlightened in (dark gray) blue. There is one gap at position 2 in the subsequence $\mathbf{S}$. Three amino acids are skipped in the original sequence [in (light gray) orange]: they are interpreted as insertions.

(1) $\mathbf{S}$ is an ordered list of amino acids in $\mathbf{A}$ (called *match* states), with the possibility of adding gaps states denoted "—" between two consecutive positions, and of skipping some amino acids of $\mathbf{A}$ (i.e., interpreting them as insertions).

(2) The subsequence $\mathbf{S}$ minimizes the total energy $\mathcal{H} = \mathcal{H}_{\rm DCA} + \mathcal{H}_{\rm ins} + \mathcal{H}_{\rm gap}$.

An example of a sequence $\mathbf{A}$ and its alignment $\mathbf{S}$ is illustrated in Fig. 1. To formulate this problem as a statistical physics model, we introduce for each position $i = 1, \dots, L$ a pair of variables $y_i = (x_i, n_i)$, where $x_i \in \{0, 1\}$ is a binary variable, and $n_i \in \{0, 1, \dots, N, N+1\}$ is a pointer. The variable x_i indicates whether position i is a gap "$-$" ($x_i = 0$) or a match state ($x_i = 1$). When i is a match, the pointer n_i indicates the position of the match state in the full-length sequence $\mathbf{A}$. When i is a gap, the pointer keeps track of the last match state before position i. Note that we added pointer values $n = 0$ and $n = N + 1$. These value are used for gap states at the beginning and at the end of the aligned sequence: if matched symbols start to appear only from a position $i > 1$, we fill the previous positions $j < i$ with gaps having pointer $n_j = 0$. Similarly, if the last matched state appears at position $i < L$, we fill the next positions $j > i$ with gaps having pointers $n_j = N + 1$. The Potts Hamiltonian rewritten in terms of the variables $\mathbf{y} = (y_1, \dots, y_L)$ is

$$\mathcal{H}_{\rm DCA}(\mathbf{y}) = -\sum_{i=1}^{L} H_i(A_{x_i.n_i}) - \sum_{i<j} J_{ij}(A_{x_i.n_i}, A_{x_j.n_j}),$$

where $A_0 = -$ is the gap state. We will use shorthand notations $H_i(y_i) \equiv H_i(A_{x_i.n_i})$ and $J_{ij}(y_i, y_j) \equiv J_{ij}(A_{x_i.n_i}, A_{x_j.n_j})$ in the rest of the paper. The insertion cost $\mathcal{H}_{\rm ins}$ and the gap cost $\mathcal{H}_{\rm gap}$ take the form introduced in [21]. In particular, for the insertion cost we have

$$\mathcal{H}_{\rm ins}(\mathbf{y}) = \sum_{i=2}^{L} \varphi_i(n_i - n_{i-1} - 1),$$

with $\varphi_i(\Delta n) = (1 - \delta_{\Delta n,0})[\lambda_o^i + \lambda_e^i(\Delta n - 1)]$, and $\Delta n_i = n_i - n_{i-1} - 1$ the number of skipped amino acids between position $i - 1$ and i. The parameters $\{\lambda_o^i, \lambda_e^i\}$ are inferred from the insertion statistics (see [21] Sec. IV B). In addition, for the gap cost we have

$$\mathcal{H}_{\rm gap}(\mathbf{y}) = \sum_{i=1}^{L} \mu(x_i, n_i),$$

with $\mu(1, n) = 0$ for match states, $\mu(0, 0) = \mu(0, N+1) = \mu_{\text{ext}}$ for the external gaps, and $\mu(0, n) = \mu_{\text{int}}$ for the internal gaps (with $0 < n < N+1$). The values of μ_{int}, and μ_{ext} are chosen according to the procedure described in [21], Sec. IV C: one realigns sequences of the seed MSA using several values of $\mu_{\text{int}}, \mu_{\text{ext}}$, and picks the ones minimizing the Hamming distance between the realigned seed and the original seed.

We finally introduce the Boltzmann probability law over the set of possible alignments

$$P(\mathbf{y}) = \frac{\chi_{\text{in}}(y_1)\prod_{i=2}^{L}\chi_{\text{sr}}(y_{i-1}, y_i)\chi_{\text{end}}(y_L)}{Z(\beta)} e^{-\beta\mathcal{H}(\mathbf{y})}, \tag{2}$$

where χ_{in}, χ_{sr}, and χ_{end} are Boolean functions ensuring that the ordering constraints are satisfied. The constraint for $\mathbf{S}$ to be an ordered list of amino acids is $\mathbf{A}$ can indeed be encoded with the function $\chi_{\text{sr}}(x_{i-1}, n_{i-1}, x_i, n_i)$ between two consecutive positions

$$\begin{aligned}
\chi_{\text{sr}}(0, n_{i-1}, 0, n_i) &= \mathbb{I}[n_{i-1} = n_i],\\
\chi_{\text{sr}}(1, n_{i-1}, 0, n_i) &= \mathbb{I}[n_{i-1} = n_i \vee n_i = N+1],\\
\chi_{\text{sr}}(0, n_{i-1}, 1, n_i) &= \mathbb{I}[0 \leqslant n_{i-1} < n_i < N+1],\\
\chi_{\text{sr}}(1, n_{i-1}, 1, n_i) &= \mathbb{I}[0 < n_{i-1} < n_i < N+1],
\end{aligned}$$

and with additional constraints imposed in the first and last position

$$\begin{aligned}
\chi_{\text{in}}(x_1, n_1) &= \delta_{x_1,0}\delta_{n_1,0} + \delta_{x_1,1}\mathbb{I}[0 < n_1 < N+1],\\
\chi_{\text{end}}(x_L, n_L) &= \delta_{x_L,0}\delta_{n_L,N+1} + \delta_{x_L,1}\mathbb{I}[0 < n_L < N+1].
\end{aligned}$$

Configurations $\mathbf{y}$ violating the ordering constraints have zero probability. The parameter β plays the role of an inverse temperature: by increasing β, the distribution concentrates on the allowed configurations achieving the smallest energy, i.e., on the best alignments.

III. SMALL COUPLING EXPANSION

An efficient strategy for approaching this constrained optimization problem is to use BELIEF-PROPAGATION (BP). BP is a message-passing method to approximate probability distributions of the form of Eq. (2). In particular, it allows to compute marginal probabilities on any small subset of variables, as well as the partition function $Z(\beta)$. BP is exact when the factor graph representing interactions between variables is a tree and is used as a heuristic for sparse graphs. In our case, however, the set of couplings J_{ij} is defined for all pairs (i, j), resulting in a fully connected factor graph, as shown in the left panel of Fig. 2. This makes the problem difficult for BP. However, although the interactions are very dense (all couplings are nonzero), they are typically weak for distant sites. Conversely, interactions between two neighboring sites are typically stronger as they encode the one-dimensional structure of the amino acid sequence.

Therefore, in this work we develop an approximation method where long-range couplings are treated perturbatively. More precisely, we perform a small-coupling expansion of the free-energy $F = -\frac{1}{\beta}\log Z(\beta)$ associated with the Boltzmann distribution in Eq. (2), where the zeroth order corresponds to the model defined on the one-dimensional chain, i.e., with long-range couplings set to zero: $J_{ij} = 0$ for $|i - j| > 1$. Higher orders take into account the contribution of long-range couplings in a perturbative way. We perform the expansion up to the first-order term and let the computation of higher orders for future work. This pertubative expansion is similar to a Plefka expansion to obtain the Thouless-Anderson-Palmer (TAP) equations [19,24,25]. The main difference is that in the Plefka expansion, the zeroth order is the mean-field model (i.e., including only external fields H_i) and all couplings J_{ij} are treated perturbatively, while in our approach the zeroth order includes also the short-range couplings $J_{i,i+1}$. We then study the stationary points of the perturbed free-energy with respect to single-sites and nearest-neighbors sites marginal probabilities $P_i(y_i)$ and $P_{i,i+1}(y_i, y_{i+1})$ to obtain a set of approximate BP equations. The technical details of this small-coupling expansion are given in the Supplemental Material [26], Secs. II and III. In the rest of the paper we refer to these approximate BP equations as the small coupling expansion (SCE) equations.

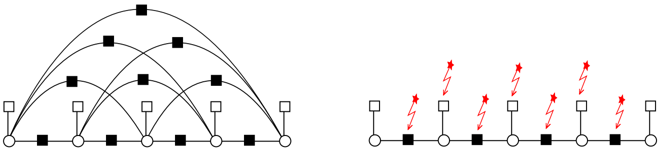

FIG. 2. Left panel: Fully connected factor graph associated to the probability Eq. (2) with $L = 5$. Variables y_i are represented by white dots, external fields H_i by white squares, and couplings J_{ij} by black squares. Right panel: Factor graph obtained after the perturbative expansion. External fields $H_2, \dots, H_{L-1}$ and short-range couplings $J_{i,i+1}$, $i \in \{1, \dots, L-1\}$ are modified according to Eq. (3) [illustrated by red (light gray) stars].

This set of SCE equations can be seen as BP equations whose associated factor graph is a linear chain, as represented in the right panel of Fig. 2, or equivalently to the equations obtained with the transfer matrix method (or dynamic programming or forward-backward algorithm [1]). The contribution of the long-range couplings J_{ij}, $|i - j| > 1$ results into a modification of the external fields H_i and short-range couplings $J_{i,i+1}$:

$$\begin{aligned}
\widetilde{H}_i &= H_i + f_i \quad \text{for} \quad i \in \{2, \dots, L-1\},\\
\widetilde{J}_{i,i+1} &= J_{i,i+1} + g_i \quad \text{for} \quad i \in \{1, \dots, L-1\}.
\end{aligned} \tag{3}$$

Single-site fields f_i and nearest-neighbors pairwise fields g_i are computed explicitly from the set of conditional probabilities $P(y_i|y_j)$ for any i, j with $|i - j| > 1$:

$$f_l(y_l) = -\sum_{i=1}^{l-1}\sum_{j=l+1}^{L}\sum_{y_i,y_j} J_{ij}(y_i, y_j)P_i(y_i|y_l)P_j(y_j|y_l), \tag{4}$$

and

$$g_l(y_l, y_{l+1}) = \sum_{i=1}^{l}\sum_{j=\zeta_i^l}^{L}\sum_{y_i,y_j} J_{ij}(y_i, y_j)P_i(y_i|y_l)P_j(y_j|y_{l+1}), \tag{5}$$

with $\zeta_i^l = \max(l+1, i+2)$. The SCE equations are recursive equations for a set of *forward* messages $F_i(y_i)$, $\widehat{F}_i(y_i)$ and *backward* messages $B_i(y_i)$, $\widehat{B}_i(y_i)$, defined on the edges of the one-dimensional chain, as shown in Fig. 3. We give here the

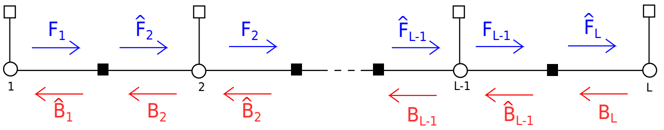

FIG. 3. BP messages defined on the one-dimensional chain. In blue (top arrows): The set of forward messages F_i, $\widehat{F}_i$, and in red (bottom arrows) the set of backward messages B_i, $\widehat{B}_i$.

exact form of the approximate BP equations, their derivation is given in the Supplemental Material [26], Sec. III. For the forward messages we have

$$\begin{aligned}
F_1(y_1) &= \frac{1}{z_{1\to e_1}} e^{\beta H_1(y_1)},\\
F_i(y_i) &= \frac{1}{z_{i\to e_i}} e^{\beta \widetilde{H}_i(y_i)} \widehat{F}_i(y_i),\\
&\text{for}\quad i \geqslant 2,\\
\widehat{F}_{i+1}(y_i) &= \frac{1}{\widehat{z}_{e_i\to i+1}} \sum_{y_i} e^{\beta \widetilde{J}_{e_i}(y_i,y_{i+1})} F_i(y_i),
\end{aligned} \tag{6}$$

where F_i is defined for $i \in \{1,\dots,L-1\}$ and $\widehat{F}_i$ for $i \in \{2,\dots,L-1\}$, and $z_{i\to e_i}$, $\widehat{z}_{e_i\to i+1}$ are normalization factors ensuring that the BP messages are normalized to 1. For the backward messages we have

$$\begin{aligned}
B_L(y_L) &= \frac{1}{z_{L\to e_{L-1}}} e^{\beta H_L(y_L)},\\
B_i(y_i) &= \frac{1}{z_{i\to e_{i-1}}} e^{\beta \widetilde{H}_i(y_i)} \widehat{B}_i(y_i), \quad \text{for}\quad i \leqslant L,\\
\widehat{B}_i(y_i) &= \frac{1}{\widehat{z}_{e_i\to i}} \sum_{y_{i+1}} e^{\beta \widetilde{J}_{e_i}(y_i,y_{i+1})} B_{i+1}(y_{i+1}),
\end{aligned} \tag{7}$$

where B_i is defined for $i \in \{2,\dots,L\}$ and $\widehat{B}_i$ for $i \in \{1,\dots,L-2\}$, and $z_{i\to e_{i-1}}, \widehat{z}_{e_i\to i}$ are normalization constants. Single-site and nearest-neighbors marginal probabilities $P_i(y_i)$ and $P_{i,i+1}(y_i, y_{i+1})$ can be expressed in terms of the BP messages

$$\begin{aligned}
P_1(y_1) &= \frac{1}{z_1} e^{\beta H_1(y_1)} \widehat{B}_1(y_1),\\
P_i(y_i) &= \frac{1}{z_i} e^{\beta \widetilde{H}_i(y_i)} \widehat{F}_i(y_i) \widehat{B}_i(y_i), \; 2 \leqslant i \leqslant L-1,\\
P_L(y_L) &= \frac{1}{z_L} e^{\beta H_L(y_L)} \widehat{F}_L(y_L),
\end{aligned} \tag{8}$$

and for $i \in \{1,\dots,L-1\}$:

$$P_{i,i+1}(y_i, y_{i+1}) = \frac{e^{\beta \widetilde{J}_{i,i+1}(y_i,y_{i+1})}}{z_{i,i+1}} F_i(y_i) B_{i+1}(y_{i+1}). \tag{9}$$

From the set of marginal probabilities, one finally computes the conditional probabilities $P_i(y_i|y_j)$, for all i, j with $|i-j| > 1$, from the chain rule, which is valid when long-range couplings are neglected

$$P_i(y_i|y_l) = \sum_{y_{i-1}} P_{i-1}(y_{i-1}|y_l) P_i(y_i|y_{i-1}) \quad \text{if } i > l+1, \tag{10}$$

with a similar expression similarly when $i < l-1$. A solution of the SCE equations can be found iteratively (see Supplemental Material [26] Sec. III C for a complete description of the algorithm). From a random initialization of the BP messages, the algorithm first computes the marginals P_i, $P_{i,i+1}$ from Eqs. (8) and (9), then updates the set of conditional probabilities $P_i(y_i|y_j)$ from Eq. (10), and finally computes the long-range fields f_i, g_i using Eqs. (4) and (5). BP messages are then updated using the new value of f_i, g_i, and these steps are repeated until convergence. Each iteration has complexity $O(L^3Q^4)$, with Q the size of the state space for variable y_i [in our case $Q = 2(N+2)$], the bottleneck being the computation of fields f_i, g_i. Although this algorithm is slower than DCALIGN, the approximate BP algorithm derived in [21], it has the advantage to derive the small coupling expansion in a rigorous way, which in turns allows to compute thermodynamic quantities such as free-energy and entropy (see Supplemental Material [26], Sec. V for their explicit expression) that were not available with the previous approach [21]. The free-energy could be used to optimize the Hamiltonian's parameters (in particular, the gap costs $\mu_{\text{int}}, \mu_{\text{ext}}$ defined in $\mathcal{H}_{\text{gap}}$). We leave this for future work. Note that DCALIGN equations [21] can be recovered from this perturbative expansion, at the cost of assuming the factorization $P_{ij}(y_i, y_j) \simeq P_i(y_i)P_j(y_j)$ for $|i-j| > 1$ in the first-order term of the free-energy (see Supplemental Material [26] Sec. III D for an explicit derivation).

Decoding strategies

Once a solution to the SCE equations is found, an assignment can be computed from the marginals using a decoding strategy. We use and compare the performance of two strategies: (i) the *nucleation* already used in [21], (ii) and *Viterbi decoding* in which we use the nearest-neighbors pairwise marginals $P_{i,i+1}$ to compute the solution having the largest probability of being generated by a Markov chain using transition probabilities $P(y_{i+1}|y_i)$. Note that neither of the two strategies are guaranteed to produce an assignment achieving the largest probability w.r.t. Eq. (2): in principle, one should use a decimation strategy and recompute after each assignment of a variable the new marginals conditioned on the previous assignments. However, these two strategies are faster than decimation, and we see that they provide very good alignments. In particular, we show below that VITERBI decoding outperforms the nucleation strategy on protein families PF00684 and PF00035 taken from the PFAM database, see [27] [7]. More details on the decoding strategies are given in the Supplemental Material [26], Sec. IV.

IV. EPSILON COUPLING ANALYSIS

The SCE approach allows us to find a solution to the constrained optimization problem of finding the best alignment of the original sequence **A** to a seed MSA. We use this method to explore the energy landscape around a given optimal alignment found with our algorithm. We use a general technique called the epsilon coupling analysis, introduced in [28], see also [29] for its application to RNA secondary structures: starting from the optimal solution $\mathbf{y}^0$, we add a repulsive

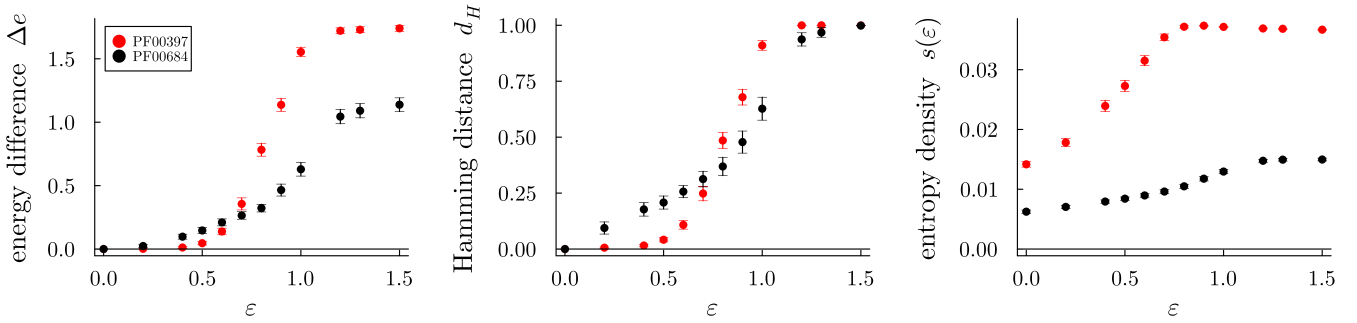

FIG. 4. Result for ϵ coupling. Red (light gray) points: On family PF00397, averaged over 100 sequences. Black points: On family PF00684, averaged over 40 sequences. Left: Difference between energy densities of the ground state: $\Delta e = [\mathcal{H}(\mathbf{y}^\epsilon) - \mathcal{H}(\mathbf{y}^0)]/L$. Middle: Hamming distance between the ground state at ϵ and at $\epsilon = 0$. Right: Entropy density $s(\epsilon)$. Results are obtained with SCE + VITERBI decoding, with an annealed scheme $\beta \in \{0, 0.05, \dots, 0.4\}$.

external field to the Hamiltonian $\mathcal{H}(\mathbf{y})$ that repels $\mathbf{y}^0$ with intensity ϵ:

$$\mathcal{H}'(\mathbf{y};\epsilon,\mathbf{y}^0) = \mathcal{H}(\mathbf{y}) + \epsilon \sum_{i=1}^{L} \delta_{y_i,y_i^0}. \tag{11}$$

This additional term [viz. the Hamming distance $d_H(\mathbf{y}, \mathbf{y}^0)$ between the optimal solution and a configuration $\mathbf{y}$] penalizes structures that are close to the ground state $\mathbf{y}^0$, allowing to explore other minima. One computes the optimal solution $\mathbf{y}^\epsilon$ of $\mathcal{H}'$ for many values of ϵ, using again the SCE + decoding strategy. For each value of ϵ, one compares the new ground state with the true one by computing their Hamming distance $d_H(\mathbf{y}^0, \mathbf{y}^\epsilon)$, and their difference in energy density $\Delta e = [\mathcal{H}(\mathbf{y}^\epsilon) - \mathcal{H}(\mathbf{y}^0)]/L$. We also compute, for each value of ϵ, the entropy density $s(\epsilon)$ associated with the perturbed model (11). Results are shown in Fig. 4 for two protein families (PF00397 and PF00684) selected from the PFAM database [7]. We restrict our analysis to short families ($L = 67$ for PF00684 and $L = 31$ for PF00397) to avoid a significant slowing down of the alignment algorithm. As ϵ increases, $\mathbf{y}^\epsilon$ starts to depart from $\mathbf{y}^0$ ($d_H > 0$) and simultaneously the difference in energy density Δe becomes positive. This indicates that we do not find other optimal solutions, instead we find solutions with higher energy ($\Delta e > 0$), but close in hamming distance to the true ground state, suggesting a landscape with a single minimum in a basin of attraction. This analysis is compatible with our computation of the entropy: we obtain for both families a rough estimate of the number of optimal configurations $e^{Ls(\epsilon)}$ between one and two configurations. At larger ϵ values, the energy density difference Δe, the Hamming distance d_H, and the entropy $s(\epsilon)$ reach a plateau at $\epsilon \simeq 1.0$ for both protein families. The solutions $\mathbf{y}^\epsilon$ found for these values of ϵ are mostly made of gaps, i.e., are not good alignments, which indicates that in this regime the free-energy landscape is substantially modified by the perturbation.

V. PERFORMANCE ANALYSIS

We assess the quality of MSAs generated by our SCE method and compare them to state-of-the art alignments provided by HMMER [6], on small protein families PF00397, PF00684, and PF00035 taken from PFAM [7] (with $L = 67$ for PF00035). As done in [21], we do not consider the entire sequences, whose length N is often much larger than L, but a "neighborhood" of the hit selected by HMMER. In practice, we add δ amino acids at the beginning and at the end of the hit resulting in a final length $N = \delta + L + \delta$ (with $\delta = 20$ for PF00397 and PF00684, and $\delta = 10$ for PF00035). We consider sequence-wise measures, also used in [21], to evaluate the similarity between two candidate MSAs (a "reference" and a "target" MSA). (i) THe **Hamming** distance between two alignments ($\mathbf{S}^{\text{ref}}$ and $\mathbf{S}^{\text{tar}}$) of the same sequence $\mathbf{A}$ in the reference and target MSAs, respectively. (ii) **Gap +**: The number of match states in $\mathbf{S}^{\text{ref}}$ that are replaced by a gap in $\mathbf{S}^{\text{tar}}$. (iii) **Gap −**: The number of gap states in $\mathbf{S}^{\text{ref}}$ that are replaced by a match state in $\mathbf{S}^{\text{tar}}$. (iv) **Mismatch**: The number of amino acid mismatches, i.e., the number of times we have a match state in both $\mathbf{S}^{\text{ref}}$ and $\mathbf{S}^{\text{tar}}$, but corresponding to different amino acids positions in the full sequence $\boldsymbol{A}$. All quantities are normalized by L, the length of the sequences.

In addition, we compare the quality of alignments by computing for each sequence of the MSAs the difference in energy density $\Delta e = [\mathcal{H}_{\text{DCA}}(\mathbf{S}^{\text{ref}}) - \mathcal{H}_{\text{DCA}}(\mathbf{S}^{\text{tar}})]/L$.

A. Comparison with HMMER

We first compare the MSA produced by our SCE algorithm (target MSA) with the MSA produced by HMMER (reference MSA), see Fig. 5. For each family we choose a random sample of sequences and compare the alignments produced by the two methods. The difference in energy density for each sequence (sorted in decreasing order) is plotted on the left panels. For the three families, we see that for a large fraction of the sample set, the energy $\mathcal{H}_{\text{DCA}}$ of the SCE alignment is lower than the one of HMMER, thus resulting in a better alignment found by SCE. For PF00397 and PF00684, for the rest of the sample set, the difference in energy is zero: both methods find the same alignment. The distribution of similarity metrics (Hamming distance, Gap±, Mismatch) are mostly concentrated on the first bins for both families, indicating that the alignments found by SCE and HMMER are close. For PF00035, it is only on a tiny fraction of the sample set that SCE finds a solution with either equal or slightly higher energy compared to HMMER. The distribution of similarity metrics is broader on this family, indicating that SCE and HMMER find substantially different alignments on a large fraction of samples.

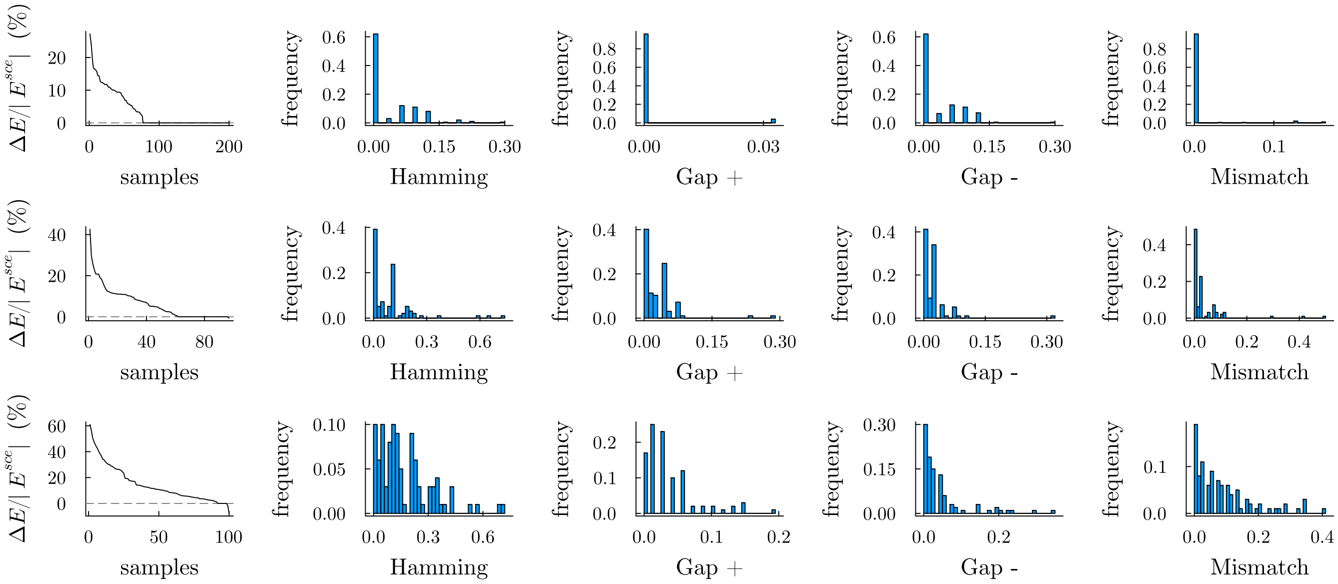

FIG. 5. Comparison of SCE with HMMER. Top: Protein family PF00397 (on a set of 200 sequences). Middle: Protein family PF00684 (on a set of 100 sequences). Bottom: Protein family PF00035 (on a set of 100 sequences). On left panels, we plot the difference in energy between the ground state found with HMMER $E^{\text{hmmer}} = \mathcal{H}_{\text{DCA}}(\mathbf{S}^{\text{hmmer}})$ and the ground state found with SCE $E^{\text{sce}} = \mathcal{H}_{\text{DCA}}(\mathbf{S}^{\text{sce}})$ (percent of the ground-state energy E^{sce} found with SCE). Positive $\Delta E = E^{\text{hmmer}} - E^{\text{sce}}$ means that SCE strategy has found a better (lower in energy) alignment than HMMER. Samples are sorted by decreasing values of $\Delta E/|E^{\text{sce}}|$. Then, from left to right, we plot the histograms of Hamming distances, Gap +, Gap −, and Mismatch.

B. Comparison with the seed

To explore further the differences between our method and HMMER, we compare the alignments found with the two methods and the seed MSA. More precisely we realign each sequence of the seed MSA (reference MSA) with our method and with HMMER to obtain a new MSA (target MSAs). The results, given in the Supplemental Material [26], Sec. I, Fig. 1. (for the protein family PF00397), show that the MSA obtained with SCE is closer to the seed MSA than the one obtained with HMMER, suggesting that SCE performs better than the alignment task.

C. Comparison of decoding methods

We compare the performances of two decoding methods: *nucleation* and *Viterbi* (see Supplemental Material [26], Sec. IV). For each family, we compare the two decoding methods used on the set of marginal probabilities computed from our SCE algorithm. Results are shown in the Supplemental Material [26], Fig. 2 (for families PF00397, PF00684, and PF00035). While for family PF00397, both decoding methods find essentially the same alignment, the situation is different for families PF00684 and PF00035: although for a large fraction of the sequences, both decoding methods find the same alignment, we can clearly see that VITERBI finds a better solution on a nonnegligible fraction of the sequences, with a significantly lower energy, and nucleation leads to a better alignment only for a few sequences.

D. Remote homology detection

We test the performance of our SCE algorithm on homology search for the RNA family RF00162 taken from RFAM database [30]. The goal of homology detection is to determine whether a sequence is evolutionary related (i.e., homologous) to a family of sequences. It is common that homology search fails at identifying distantly related sequences [31]. As a testing ground, we use the SAM riboswitch seed alignment from the RFAM family [30] RF00162 (which have length $L = 108$). This dataset was proposed in [20] as a stress test for alignment algorithms. Following this setup, the MSA is divided into a training set and a test set. Sequences in the test set are selected to be distant to the training set and distant from each other (see [20] for details). In addition, a set of nonhomologous decoy sequences is randomly generated as follows: each character is drawn i.i.d. from the nucleotide composition of the positive test sequences, with a length matching a randomly selected positive test sequence [20]. To wrap up, we have three mutually nonoverlapping set of sequences: (i) training: from which we learn the parameters of our model; (ii) test: a set of homologous sequences; and (iii) decoy: a randomly generated set of nonaligned sequences.

For each sequence in the test set and for the 11 sequences randomly extracted from the set of decoy sequences, we compute the alignment found with SCE + VITERBI decoding. The parameters are learned from the training set: the parameters of the Potts model are trained with a Boltzmann machine DCA learning algorithm and the parameters of the insertion cost $\mathcal{H}_{\text{ins}}$ are learned from the insertion statistics (see [21] Sec. IV B). The parameters of the gap cost μ_{int}, μ_{ext} are taken from [21], Table II.

To score the alignments, we compare their energy density $e = \mathcal{H}(\mathbf{S})/L$. We also compute, for each alignment $\mathbf{S}$ found by our algorithm, its Hamming distance w.r.t. each aligned sequence in the training set. We then collect the minimum attained value. Results are given in Fig. 6. and show that our

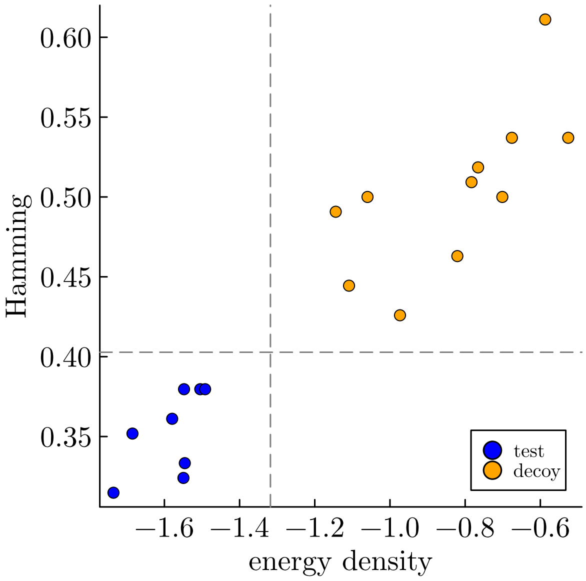

FIG. 6. Remote homology detection on the RNA family RF00162 [20]. Alignments found by SCE + Viterbi decoding, on a set of eight test sequences [in blue (dark gray)] and 11 decoy sequences [in orange (light gray)]. x axis: energy density $e = \mathcal{H}(\mathbf{S})/L$ of the alignment. y axis: Hamming distance from the solution to closest aligned sequence in the training set. Vertical (or horizontal) dashed line shows the average between the right-most (or highest) blue point and the left-most (or lowest) orange point, indicating that the two sets can be separated with both observables.

method is able to disentangle between decoy sequences and true sequences belonging to RF00162: the alignments found for the test set have smaller energy and are closer to the training set.

VI. CONCLUSION

We proposed an alternative method based on a perturbative expansion of the model around the linear chain and obtained a set of approximate message-passing equations that we used to find optimal alignments. We tested the potentiality of our algorithm on protein families taken from the PFAM database [7]. The results obtained on these families suggest that including long-range correlations is crucial for the alignment task and it is a promising direction to go beyond current state-of-the-art bioinformatics tools based on profile models, which, from a statistical mechanics standpoint, are assuming statistical independence of sites. Additionally, we compare the performances of two different decoding strategies, and show that for two of the protein families studied in this paper, the VITERBI decoding algorithm outperforms the nucleation strategy presented in [21]. We test the performance of our method on remote homology search, for the RNA family RF00162 taken from the RFAM database [30], and obtain promising results suggesting that our method is able to detect distant homologs. The method proposed in this paper treats perturbatively the contribution of long-range couplings J_{ij}, with $|i - j| > 1$ using a small coupling expansion *a la* Plefka [19]. While this assumption might not be justified as some of the couplings might not be in the perturbative regime, our approach is an initial step to include them to go beyond the independent-site assumption. Moreover, in the context of DCA [11], it was empirically shown that the first-order approximation of the Plefka expansion is enough to capture relevant structural and functional features of the protein family.

Our approach provides a self-consistent derivation of the mean-field approximation used in [21], which, in turn, being variational, allows us to compute approximated thermodynamic potentials. We use this strategy to explore the free-energy landscape of this constrained optimization problem, obtaining, for the protein families studied in this paper, the global picture of a unique solution surrounded by a basin of attraction. The main limitation of our method is an increase of computational complexity with respect to the mean-field method of [21]: indeed, our SCE algorithm has an $O(L^3N^4)$ complexity (with L the length of the alignment $\mathbf{S}$ and N the length of sequence $\mathbf{A}$ to be aligned), compared to the $O(L^2N^2)$ complexity for the DCALIGN algorithm designed in [21]. Further investigations could be to use our approximation of the free-energy for developing methods to simultaneously optimize the model's parameters and find the optimal alignment, using, for instance, strategies based on expectation-maximization. Note finally that the method developed in this paper is not restricted to the alignment problem and could be used in other problems that have the structure of a one-dimensional chain with additional fully connected weak couplings.

ACKNOWLEDGMENTS

The authors thank Anna Paola Muntoni for interesting discussions and for sharing with us part of the code needed for data processing. A.P. acknowledges funding by the EU H2020 research and innovation program MSCA-RISE-2016 under Grant Agreement No. 734439 INFERNET, as well as for financial support from Future Artificial Intelligence Research (FAIR) and Centro Nazionale di Ricerca in High-Performance Computing, Big Data, and Quantum Computing (ICSC) founded by European Union-NextGenerationEU.

[1] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, Cambridge, England, 1998).

[2] S. B. Needleman and C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. **48**, 443 (1970).

[3] T. F. Smith and M. S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. **147**, 195 (1981).

[4] R. C. Edgar and S. Batzoglou, Multiple sequence alignment, Curr. Opin. Struct. Biol. **16**, 368 (2006).

[5] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. **25**, 3389 (1997).

[6] S. R. Eddy, Accelerated profile HMM searches, PLoS Comput Biol **7**, e1002195 (2011).

[7] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart *et al.*, The Pfam protein families database in 2019, Nucleic Acids Res. **47**, D427 (2019).

[8] D. de Juan, F. Pazos, and A. Valencia, Emerging methods in protein co-evolution, Nat. Rev. Genet. **14**, 249 (2013).

[9] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Inverse statistical physics of protein sequences: a key issues review, Rep. Prog. Phys. **81**, 032601 (2018).

[10] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, Protein 3D structure computed from evolutionary sequence variation, PLoS ONE **6**, 1 (2011).

[11] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. USA **108**, E1293 (2011).

[12] A. Procaccini, B. Lunt, H. Szurmant, T. Hwa, and M. Weigt, Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: Orphans and crosstalks, PLoS ONE **6**, 1 (2011).

[13] C. Baldassi, M. Zamparo, C. Feinauer, A. Procaccini, R. Zecchina, M. Weigt, and A. Pagnani, Fast and accurate multivariate gaussian modeling of protein families: Predicting residue contacts and protein-interaction partners, PLoS ONE **9**, 1 (2014).

[14] C. Feinauer, H. Szurmant, M. Weigt, and A. Pagnani, Interprotein sequence co-evolution predicts known physical interactions in bacterial ribosomes and the Trp operon, PLOS ONE **11**, 1 (2016).

[15] M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, and M. Weigt, Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1, Mol. Biol. Evol. **33**, 268 (2016).

[16] R. R. Cheng, O. Nordesjö, R. L. Hayes, H. Levine, S. C. Flores, J. N. Onuchic, and F. Morcos, Connecting the sequence-space of bacterial signaling proteins to phenotypes using coevolutionary landscapes, Mol. Biol. Evol. **33**, 3054 (2016).

[17] T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. Schärfe, M. Springer, C. Sander, and D. S. Marks, Mutation effects predicted from sequence co-variation, Nat. Biotechnol. **35**, 128 (2017).

[18] J. Trinquier, G. Uguzzoni, A. Pagnani, F. Zamponi, and M. Weigt, Efficient generative modeling of protein sequences using simple autoregressive models, Nat. Commun. **12**, 5800 (2021).

[19] T. Plefka, Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model, J. Phys. A: Math. Gen. **15**, 1971 (1982).

[20] G. W. Wilburn and S. R. Eddy, Remote homology search with hidden Potts models, PLoS Comput. Biol. **16**, 1 (2020).

[21] A. P. Muntoni, A. Pagnani, M. Weigt, and F. Zamponi, Aligning biological sequences by exploiting residue conservation and coevolution, Phys. Rev. E **102**, 062409 (2020).

[22] H. Talibart and F. Coste, PPalign: optimal alignment of Potts models representing proteins with direct coupling information, BMC Bioinf. **22**, 317 (2021).

[23] S. Petti, N. Bhattacharya, R. Rao, J. Dauparas, N. Thomas, J. Zhou, A. M. Rush, P. Koo, and S. Ovchinnikov, End-to-end learning of multiple sequence alignments with differentiable Smith-Waterman, Bioinformatics, **39**, btac724 (2023).

[24] A. Georges and J. S. Yedidia, How to expand around mean-field theory using high-temperature expansions, J. Phys. A: Math. Gen. **24**, 2173 (1991).

[25] M. Opper and D. Saad, *From Naive Mean Field Theory to the TAP Equations* (MIT Press, Cambridge, MA, 2001).

[26] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.107.044125 for additional figures and technical details on the method.

[27] https://www.ebi.ac.uk/interpro/ release 32.0.

[28] A. Pagnani, G. Parisi, and M. Ratiéville, Near-optimal configurations in mean-field disordered systems, Phys. Rev. E **68**, 046706 (2003).

[29] E. Marinari, A. Pagnani, and F. Ricci-Tersenghi, Zero-temperature properties of RNA secondary structures, Phys. Rev. E **65**, 041919 (2002).

[30] I. Kalvari, J. Argasinska, N. Quinones-Olvera, E. P. Nawrocki, E. Rivas, S. R. Eddy, A. Bateman, R. D. Finn, and A. I. Petrov, Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families, Nucleic Acids Res. **46**, D335 (2018).

[31] C. M. Weisman, A. W. Murray, and S. R. Eddy, Many, but not all, lineage-specific genes can be explained by homology detection failure, PLOS Biology **18**, 1 (2020).