Finite-temperature local protein sequence alignment: Percolation and free-energy distribution

S. Wolfsheimer[*]

Department of Applied Mathematics, Université Paris Descartes, 45 rue des Saint-Pères, F-75270 Paris Cedex 06, France

O. Melchert and A. K. Hartmann

Institut für Physik, Carl von Ossietzky Universität Oldenburg, D-26111 Oldenburg, Germany
(Received 13 July 2009; published 17 December 2009)

Sequence alignment is a tool in bioinformatics that is used to find homological relationships in large molecular databases. It can be mapped on the physical model of directed polymers in random media. We consider the finite-temperature version of local sequence alignment for proteins and study the transition between the linear phase and the biologically relevant logarithmic phase, where the free energy grows linearly or logarithmically with the sequence length. By means of numerical simulations and finite-size-scaling analysis, we determine the phase diagram in the plane that is spanned by the gap costs and the temperature. We use the most frequently used parameter set for protein alignment. The critical exponents that describe the parameter-driven transition are found to be explicitly temperature dependent. Furthermore, we study the shape of the (free-) energy distribution close to the transition by rare-event simulations down to probabilities on the order 10^{-64}. It is well known that in the logarithmic region, the optimal score distribution ($T=0$) is described by a modified Gumbel distribution. We confirm that this also applies for the free-energy distribution ($T>0$). However, in the linear phase, the distribution crosses over to a modified Gaussian distribution.

I. INTRODUCTION

Biological sequence analysis is an interdisciplinary scientific field which uses concepts from statistics, computer science, and molecular biology. Some approaches used in the context of sequence analysis are, from a conceptional point of view, related to models in statistical mechanics of disordered systems. One of the most fundamental tools in the area of sequence analysis is *sequence alignment* (see, for example, [1,2]). It is used to quantify similarities between two (or more) biological sequences, such as DNA, proteins, or RNA. Modern search tools for large databases, such as BLAST [3] or FASTA [4], heavily rely on sequence alignment algorithms.

In this paper, we consider algorithms for pairwise local protein alignment, which aims at finding "conserved" regions of two input protein sequences. The most prominent example is the Smith-Waterman algorithm [5]. The algorithm finds optimal alignments (OAs) according to an objective function. Each alignment is assigned a score, which is maximal for optimal alignments. The optimal alignment score serves as a scalar measure of similarity of the input sequences. Since alignments have a geometrical interpretation as directed paths [6], the problem of finding an optimal alignment is directly related to the ground state of directed paths in random media (DPRM) [7–11] in 1+1 dimensions. From this point of view, the alignment score corresponds to the negative energy and the optimal alignment to the ground state of the system.

However, in some cases optimal alignments are not desirable and one is interested in ensembles of probabilistic align-ments. This is particularly the case when one wishes to compare so-called weak homologs, i.e., sequences that are related on a relatively long evolutionary time scale. In the literature, some examples can be found, where probabilistic alignments clearly outperform optimal alignments [12–15]. From the physical perspective, a natural generalization to probabilistic alignments can be achieved by introducing a temperature and considering canonical ensembles of alignments for each pair of input sequences instead of the ground state alone [12,16,17]. The finite-temperature approach provides also interesting applications when one wishes to assess the reliability of alignments by so-called posterior probabilities [1,16].

For both approaches, for OA and for finite-temperature alignments (FTAs), the choice of the algorithmic parameters remains ambiguous. In particular, the choice of the so-called gap costs (see below) requires some heuristic experimentation. Interestingly, this question can be approached by the theory of critical phenomena. The study of sequence alignment from that perspective yields interesting results that have improved the optimal choice of parameters of sequence alignment [17–19]. The *linear-logarithmic phase transition* [20–22] is the most important aspect regarding this issue. The name stems from the fact that there is a continuous parameter-driven transition between phases, where the average score grows linearly or logarithmically with sequence length, respectively. There is much empirical evidence that the optimal choice of scoring parameters is close to the phase boundary on the logarithmic side [23,24]. The underlying reason is that the transition is driven by the balance between the *score matrix* that measures the similarity between letters of the underlying alphabet (i.e., amino acids in the case of protein alignment) and the gap costs. The latter ones control how strong insertions or deletions of subsequences are to be penalized. Hence, one would like to identify similar regions, which means to try to avoid gaps, giving them a penalty. On the other hand, one would like to ignore small local evolu-

---

*Formerly at Institut für Physik, Carl von Ossietzky Universität Oldenburg, D-26111 Oldenburg, Germany.

tionary changes to the sequences, which means one should not make the gap penalty too strong. This leads to an optimum choice of the gap penalty parameters at "intermediate values."

At $T=0$, i.e., for OA, Hwa and Lässig studied the transition by looking at the dynamic growth of the local (and global) score when advancing in the search space [24]. Later, the critical values were studied analytically by a self-consistent equation [22] or numerically by a finite-size-scaling analysis [25]. Both studies rely on a simple scoring model with a single mismatch parameter. In the latter procedure, the problem was approached by considering the linear-logarithmic phase transition as a percolation phenomenon [26].

The aim of our study is to go beyond the models that have been considered so far. In particular, we studied the most widely used protein alignment model, i.e., local alignment with the scoring matrix blosum62 [27] and affine gap costs (see below), where the linear-logarithmic phase transition is of actual relevance to the database queries or alignment analysis of protein sequences.

We considered the geometrical interpretation of alignments and studied numerically the percolation properties of OA and FTA. This allowed us to determine critical exponents that describe the parameter-driven linear-logarithmic phase transition. Furthermore, we determined the phase diagram in the plane that is spanned by the temperature and the gap costs. Finally, we studied the distribution of the optimal score and the free energy close to the transitions.

In Sec. II, we review the model and algorithms to compute the partition functions and methods to sample alignments from the canonical ensemble. The main results for different observables and the (free-)energy distributions are presented in Sec. III followed by a discussion in Sec. IV.

II. PARTITION FUNCTION CALCULATION AND SAMPLING

An alignment relates letters from one sequence $\mathbf{a} = a_1,\ldots,a_L \in \Sigma^L$ to a second one $\mathbf{b} = b_1,\ldots,b_M \in \Sigma^M$, where Σ denotes the underlying alphabet. Here, we consider protein sequences wherein Σ is given by the 20 letter amino acid alphabet. Given the pair $\mathbf{a}$ and $\mathbf{b}$, an alignment $\mathcal{A}$ is an ordered set of pairings $\{(i_1,j_1),\ldots,(i_{N_m},j_{N_m})\}$ with $1 \leq i_k < i_{k+1} \leq L$, $1 \leq j_k < j_{k+1} \leq M$. If $a_{i_k} = b_{j_k}$, the pair (i_k,j_k) is called a match, otherwise, a mismatch. Consequently, we will refer to N_m as the number of matches plus mismatches. The state space of all alignments of $\mathbf{a}$ and $\mathbf{b}$ shall be written as $\chi_{\mathbf{a},\mathbf{b}}$.

When comparing sequences, one has to account for so-called insertions or deletions of subsequences that occur in evolutionary processes. Regarding alignments, these processes are represented by gaps, which are defined as follows. If $i_{k+1} = i_k + 1$ and $j_{k+1} = j_k + 1 + l$ with $l > 0$ and $(i_k,j_k),(i_{k+1},j_{k+1}) \in \mathcal{A}$, then $\mathbf{b}$ is said to contain a *gap of length l* between j_k and j_{k+1} and likewise for the sequence $\mathbf{a}$. If $j_1 = l + 1 \geq 2$, then $\mathbf{b}$ is said to have a *gap of length l* at the beginning, if $j_N = M - l < M$, then $\mathbf{b}$ has a *gap of length l* at the end and likewise for the sequence $\mathbf{a}$.

For the comparison of sequences, it is relevant to give a measure for the similarity or the degree of conservation between the sequences or regions of the sequences under consideration. The classical way to accomplish this is to assign a *score* for each alignment via an *objective function* $\mathcal{S}: \chi_{\mathbf{a},\mathbf{b}} \to \mathbb{R}$ and then maximizing $\mathcal{S}$ among all alignments

$$\mathbf{S}_0(\mathbf{a},\mathbf{b}) = \max_{\mathcal{A}} \mathcal{S}(\mathcal{A};\mathbf{a},\mathbf{b}),$$

$$\mathcal{A}^{\mathrm{opt}} = \mathrm{argmax}\, \mathcal{S}(\mathcal{A};\mathbf{a},\mathbf{b}). \tag{1}$$

For the choice of the objective function and its parameters, it is necessary to decide (i) whether we are interested in a locally conserved region or whether the entire sequences should be considered, (ii) how matches and mismatches should be evaluated, and (iii) how gaps should influence the alignment or how a gap penalty should affect the overall score.

To address the first issue, there are in principle two types of objective functions available, namely, *optimal local alignment scores* $\mathbf{S}_0^{\mathrm{local}}$ and *optimal global alignment scores* $\mathbf{S}_0^{\mathrm{global}}$. Optimal global alignment scores involve contributions from all matches, mismatches, and gaps. Based on this, the optimal local alignment score is the optimum of all global alignments of all possible contiguous subsequences of $\mathbf{a}$ and $\mathbf{b}$,

$$\mathbf{S}_0^{\mathrm{local}}(\mathbf{a},\mathbf{b}) = \max_{\substack{1 \leq i' < i \leq L \\ 1 \leq j' < j \leq M}} \mathbf{S}_0^{\mathrm{global}}(a_{i'} \ldots a_i, b_{j'} \ldots b_j). \tag{2}$$

The second issue requires the knowledge of a relationship between the letters of the underlying alphabet. This is usually realized by *substitution or score matrices* that assign each pair of letters a number $\sigma(a,b)$. Here, we use the most frequently used matrix blosum62 [27]. In most cases, scoring matrices are derived from biological data by the scaled log-odd ratio $\Lambda \frac{P_{a,b}}{f_a f_b}$, where $P_{a,b}$ is the probability of observing the pair of letters a and b, f_a and f_b denote the background frequencies of observing the letters a and b independently, and Λ defines a scale [1]. The entries are usually rounded to integers.

Regarding the gaps, one compromises between a computational feasible and a biological evident penalty function g. That means, each gap Γ of length l_Γ yields a negative contribution of $-g(l_\Gamma)$ to the overall score, which is then defined as

$$\mathcal{S}(\mathcal{A};\mathbf{a},\mathbf{b}) = \sum_{(i,j) \in \mathcal{A}} \sigma(a_i,b_j) - \sum_{\Gamma} g(l_\Gamma), \tag{3}$$

g is usually a monotonously increasing function of the gap length. The alignment algorithms for gaped alignments with arbitrary gap penalties exhibit a cubic time complexity $\{\mathcal{O}[\max(L,M)^2 \min(L,M)]\}$. In practice, *affine gap* cost functions

$$g(l_\Gamma) = \alpha^{\mathrm{open}} + \beta^{\mathrm{ext}}(l_\Gamma - 1), \quad \text{with } \alpha^{\mathrm{open}} \geq \beta^{\mathrm{ext}} > 0 \tag{4}$$

are commonly used because the computational complexity then reduces to $\mathcal{O}(LM)$ [28]. The parameters α^{open} and β^{ext} are called gap-open penalty and gap extension penalty, re-
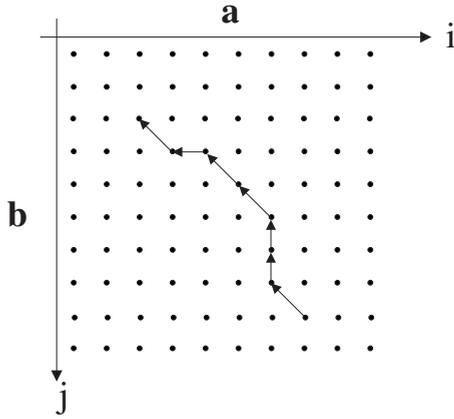
FIG. 1. Representation of an alignment as a directed path in quenched disorder. The disorder is realized by random sequences.

spectively. With the above choice, one mimics the biological observations that (i) longer gaps appear less frequently than shorter ones and (ii) opening a gap is less likely than extending an existing one.

There is some evidence that this form describes the natural process of insertions and deletions quite well [29–31].

An alignment can be represented as a directed path on a lattice of size $L \times M$ (see Fig. 1). The path is given by the set of matches and mismatches and gaps in the alignment $\mathcal{A}$. Due to the conditions $i_k < i_{k+1}$ and $j_k < j_{k+1}$, the path is directed. By convention, we say that each path element may be orientated in $(-1, -1)$, $(-1, 0)$, or $(0, -1)$ direction. Diagonal elements denote matches or mismatches and l_Γ consecutive vertical or horizontal elements correspond to gaps of length l_Γ in one of the sequences.

In order to keep the path representation for local alignment unique, we require that (i) the first and the last path element always points in $(-1, -1)$ direction, hence, gaps at the beginning and end of the alignment never occur, and (ii) a gap in the sequence **b** is not allowed to directly follow a gap in sequence **a** (see [12]).

The optimal alignment can be computed by the dynamic programming algorithm (like a transfer-matrix method) by Smith and Waterman [5]. For affine gap costs, it requires three $(L+1) \times (M+1)$ matrices $D_{i,j}$, $P_{i,j}$, and $Q_{i,j}$ that are computed iteratively [28]. The matrix element $D_{i,j}$ is the optimal local alignment score of the subsequences $a_1 \ldots a_i$ and $b_1 \ldots b_j$ given that a_i and b_j are paired. $P_{i,j}$ and $Q_{i,j}$ are auxiliary matrices storing the optimal alignment score of the subsequences $a_1 \ldots a_i$ and $b_1 \ldots b_j$ given that the alignment ends in a gap in either sequence. The recursion relation to compute these matrices reads as

$$D_{i,j} = \sigma(a_i, b_j) + \max \begin{cases} 0 \\ D_{i-1,j-1} \\ P_{i-1,j-1} \\ Q_{i-1,j-1} \end{cases}$$

$$P_{i,j} = \max \begin{cases} D_{i-1,j} - \alpha^{\mathrm{open}} \\ Q_{i-1,j} - \alpha^{\mathrm{open}} \\ P_{i-1,j} - \beta^{\mathrm{ext}} \end{cases}$$

$$Q_{i,j} = \max \begin{cases} D_{i,j-1} - \alpha^{\mathrm{open}} \\ Q_{i,j-1} - \beta^{\mathrm{ext}} \end{cases} \tag{5}$$

with the boundary conditions

$$D_{i,0} = P_{i,0} = Q_{i,0} = -\infty \quad \text{for} \ i = 0, \ldots, L,$$

$$D_{0,j} = P_{0,j} = Q_{0,j} = -\infty \quad \text{for} \ j = 0, \ldots, M.$$

In a physical interpretation, $\sigma(a_i, b_j)$ plays the role of a random chemical potential with quenched disorder and the gap cost function $g(l_\Gamma)$ describes a line tension that forces the alignment path on a straight diagonal line.

The optimal alignment score, which in physical terms corresponds to the negative ground-state energy, is given by $\mathbf{S}_0^{\mathrm{local}}(\mathbf{a}, \mathbf{b}) = \max(\max_{i,j}\{D_{i,j}\}, 0)$. If $\mathbf{S}_0^{\mathrm{local}}(\mathbf{a}, \mathbf{b}) = 0$, the optimal alignment is the empty alignment (a path of length 0). Otherwise, the alignment starts at the position of the maximum of $D_{i,j}$. The optimal alignment (ground state) can be determined by a backtrace procedure. Given that the current state at position (i, j) is a match or mismatch, the path is extended in diagonal direction if $D_{i,j} = \sigma(a_i, a_j) + D_{i-1,j-1}$ and, accordingly, in vertical or horizontal direction if $D_{i,j} = \sigma(a_i, a_j) + P_{i-1,j-1}$ or $D_{i,j} = \sigma(a_i, a_j) + Q_{i-1,j-1}$. Similar conditions appear, if the current state is a gap in either sequence. This is repeated until a match/mismatch with $D_{i,j} = \sigma(a_i, b_j)$ is met. In the linear phase, the ground state might be highly degenerate. For this reason, one should include additional matrices $N_{i,j}^{(D)}$, $N_{i,j}^{(P)}$, $N_{i,j}^{(Q)}$ that account for the degeneracies of the alignments that end at (i, j). With help of these matrices, it is possible to sample all ground states uniformly.

In the picture of DPRM, one usually uses a temporal and a spatial coordinate, which are defined as

$$t = \frac{1}{2}(i + j) \quad \text{and} \quad x = \frac{1}{2}(i - j).$$

A local alignment problem is seen as a dynamical growth process, which starts at space-time event (t_0, x_0), where the dynamic programming matrix $D_{i,j}$ is maximal. In each time step described, the path is extended by one diagonal, vertical, or horizontal element. The spatial variable $x - x_0$ describes the deviation from a straight diagonal line for each time step. The stopping condition $D_{i,j} = \sigma(a_i, b_j)$ given that the current state is a match defines the final point (t_1, x_1) in the space time. We define the roughness of the path as the maximal deviation from a straight diagonal line, i.e., $\Delta = \max_{t_1 \le t \le t_0}|x(t) - x_0|$. Note that this definition refers only to the local alignment path and is "time independent" in contrast to Ref. [24].

Next, we describe the generalization of the alignment problem to a canonical ensemble of alignments. Let us consider the canonical ensemble of all alignments $\mathcal{A}$ for a quenched pair of sequences. The partition function at temperature T is given by

$$Z_T = \sum_{\mathcal{A}} \exp[\mathcal{S}(\mathcal{A}; \mathbf{a}, \mathbf{b})/T].$$

This sum can be computed by a generalization of Eq. (5) [16],

$$Z_{i,j}^D = (1 + Z_{i-1,j-1}^D + Z_{i-1,j-1}^P + Z_{i-1,j-1}^Q)e^{\sigma(a_i,b_j)/T},$$

$$Z_{i,j}^P = (Z_{i-1,j}^D + Z_{i-1,j}^Q) \cdot e^{-\alpha^{\mathrm{open}}/T} + Z_{i-1,j}^P \cdot e^{-\beta^{\mathrm{ext}}/T},$$

$$Z_{i,j}^Q = Z_{i,j-1}^D \cdot e^{-\alpha^{\mathrm{open}}/T} + Z_{i,j-1}^Q \cdot e^{-\beta^{\mathrm{ext}}/T}, \qquad (6)$$

with the boundary conditions

$$Z_{i,0}^D = Z_{i,0}^P = Z_{i,0}^Q = 0 \quad \text{for } i = 0, \ldots, L,$$

$$Z_{0,j}^D = Z_{0,j}^P = Z_{0,j}^Q = 0 \quad \text{for } j = 0, \ldots, M.$$

Since an alignment may start anywhere and may also include the empty alignment, the full partition function is given by

$$Z = 1 + \sum_{i=1}^{L} \sum_{j=1}^{M} Z^D i, j.$$

Note that contributions from Z^P and Z^Q are explicitly excluded because they are auxiliary only and contain noncanonical alignments. In the limit $T \to 0$, Eqs. (6) reduce to the recursion relation of the original Smith-Waterman algorithms (5). Once the transfer matrices $Z_{i,j}^D$, $Z_{i,j}^P$, and $Z_{i,j}^Q$ are determined, it is possible to directly draw alignments from the canonical distribution $P_T[\mathcal{A}] = \exp[\mathcal{S}(\mathcal{A}; \mathbf{a}, \mathbf{b})/T]/Z_T$ with zero autocorrelation. This direct Monte Carlo algorithm was proposed by Mückstein *et al.* [12] for local alignment. A general description of such methods is presented in the textbook of Durbin *et al.* [1].

III. RESULTS

To study properties of the linear-logarithmic phase transition, we generated ensembles of n^{sample} random sequences, which were drawn from the distribution $P(\mathbf{a}) = \prod_{i=1}^{L} f_{a_i}$, where f are the amino acid frequencies that were derived together with blosum62 [27]. Furthermore, we only consider pairs of sequences of equal length, i.e., $L = M$, between $L = 40$ and $L = 5120$. It turned out that for the finite-size-scaling analysis, that is, discussed in the following, only system sizes with $L \geq 480$ yield consistent results. The number n^{sample} of samples varied between 6400 for $L = 480$ and 800 for the largest system.

For each sample, $n^{\mathrm{align}} = 100$ alignments were drawn from the canonical ensemble at various temperatures T using the backtracing procedure, as described above. We used different gap-open parameters α^{open} and temperatures T between $T = 0$ and $T = 4$. The gap extension parameter β^{ext} was set to 1 throughout.

The case $T = 0$ corresponds to optimal alignments (ground states). Note that for small gap costs (i.e., in the linear phase), the ground-state degeneracy grows exponentially with the system size, whereas in the logarithmic phase usually only a few optimal alignments are observed. Thermal averages and averages over ground states over a fixed realization of a sequence pairs will be denoted as $\langle \cdot \rangle_T$ and $\langle \cdot \rangle_0$, respectively. Averages over realizations of random sequence pairs will be written as $[\cdot]$ in the following. In statistical mechanics of disordered systems, the latter one is often called the average over the disorder.
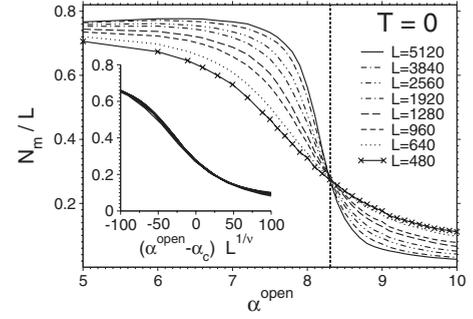
FIG. 2. Results for the number N_m of aligned letters (matches plus mismatches) as a function of gap opening penalty α^{open}. Curves for different sequence lengths intersect at the critical parameter α_c. For a more clear presentation, single data points are only shown for one system size. Inset: after rescaling the abscissa with appropriate values for α_c and ν, the data listed in Table I collapse on a single master curve.

So as to describe the linear-logarithmic transition, we considered different observables described in the subsequent sections.

A. Geometric properties

Here, we describe the results for the number N_m of paired letters (matches plus mismatches). This quantity turned out to be an adequate quantity to extract properties of the phase transition, such as the critical gap costs $\alpha_c(T)$ and the scaling behavior close to criticality.

We consider the averaged number $[\langle N_m \rangle_T]/L$ of paired letters per sequence length as a function of gap opening penalty α^{open} and temperature T. Hence, it is the fraction of matches/mismatches with respect to the maximal possible number of pairs. This observable corresponds to the percolation probability (the probability that a geometrical object spans the entire lattice) in standard percolation theory. Usually, a crossover from 1 to 0 is observed when passing the phase boundary. Here, a perfectly percolating local alignment $N_m = L$ is hardly found even in the percolating phase. Nevertheless, we applied the same finite-size-scaling analysis as for the usual percolation probability because $[\langle N_m \rangle_T]/L$ is in the order of unity in the linear regime and vanishes in the logarithmic regime.

Figure 2 displays $[\langle N_m \rangle_T]/L$ as a function of the gap-open penalty α^{open} for different lengths L and zero temperature. The curves for different sequence lengths intersect at the critical value α_c as expected for a second-order phase transition. Using finite-size-scaling theory [26], we may extrapolate data from finite sequence lengths to the thermodynamic limit $L \to \infty$. In this limit, the observable $[\langle N_m \rangle_T]/L$ is on the order of 1 below the threshold, i.e., $\alpha^{\mathrm{open}} < \alpha_c$, and it jumps to 0 exactly at α_c. In finite systems, $L < \infty$, the crossover extends over a range $\sim L^{-1/\nu}$ as can be seen in Fig. 2. Scaling theory leads us to expect that the behavior of $[\langle N_m \rangle_T](\alpha^{\mathrm{open}}; T; L)$ close to criticality is described by

$$[\langle N_m \rangle_T](\alpha^{\mathrm{open}}; T; L)/L = f[(\alpha^{\mathrm{open}} - \alpha_c)L^{1/\nu}], \qquad (7)$$

where f is an universal scaling function and the exponent ν describes the divergence of the "correlation length" at the
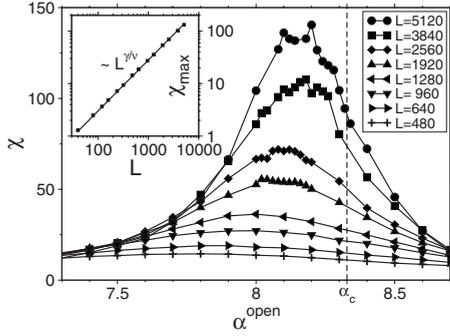
FIG. 3. Critical fluctuations of N_m. The positions of the peaks approach the critical value α_c and their heights diverge like $L^{\gamma/\nu}$. Inset: fit of $\chi_{\max}(L)$ to the scaling form $\sim L^{\gamma/\nu}$.

FIG. 4. Rescaled distributions of the observable N_m for $T=0$. Inset: scaling analysis of the moments. The first moment scales as $\mu_1 \sim L$. Higher-order moments increase slower than $\mu_k \sim L^k$.

critical point $\alpha^{\text{open}} = \alpha_c$. We used Eq. (7) to extract numerical values for the critical exponents ν and the critical gap costs α_c with high precession from all data for a fixed temperature simultaneously. The fit is performed by minimizing a weighted-χ^2-like objective function $Q(\alpha_c, \nu)$ [32] that measures the distance (in units of the standard error) of the data from the (a prior unknown) master curve. For the example in Fig. 2, i.e., $T=0$, we obtained $\alpha_c = 8.306(4)$ and $\nu = 1.58(5)$ with acceptable quality of $\hat{Q} \equiv Q(\alpha_c, \nu) = 2.2$ ($\hat{Q}$ should be on the order of 1 [32]). Statistical errors have been determined by bootstrapping [33,34]. Repeating this analysis for several temperatures, one may probe the critical line $\alpha_c(T)$ (see below).

We also tested a related quantity, which is defined as the number of matches/mismatches plus the number of gaped letters $N_m + N_g$, which results in the same critical exponent and critical point within error bars (not shown). Gaps seem to play only a marginal role close to the critical point. We observe that the roughness Δ of the alignment path, as defined above at the critical point, diverges only logarithmically with the sequence length. This supports the equivalence the description of the description of both variants of this quantity.

Next, we study the critical fluctuations of N_m. We define the susceptibilitylike quantity $\chi = ([\langle N_m^2 \rangle_T] - [\langle N_m \rangle_T]^2)/L$ as a function of α^{open} (see Fig. 3).

Close to the critical point, χ diverges like $\chi \sim L^{\gamma/\nu}$. In order to extract the height of the maxima from $\chi(\alpha^{\text{open}})$, we performed parabolic fits in the form $\chi(\alpha^{\text{open}}) = -C[\alpha_c(L) - \alpha^{\text{open}}]^2 + \chi_{\max}(L)$ for each system size L. The exponent γ itself is determined by a fit of $\chi_{\max}(L)$ to the scaling form $\sim L^{\gamma/\nu}$. For $T=0$, we obtain $\gamma/\nu = 0.95(1)$.

One may also use the scaling of $\alpha_c(L)$ to determine the critical value α_c and the critical exponent ν as a cross check via $\alpha_0(L) = \alpha_c - AL^{-\nu}$. When restricting the range to $L \geq 1280$, we obtain $\nu = 1.4(3)$ for $T=0$, which agrees within the error bars with the value stated above. Furthermore, the critical value α_c agrees within 1.5 standard deviations. We also checked that those values agree for FTA, $T=1$.

Alternatively, one can determine γ from the second moment of the averaged distribution $[P(N_m)]$ of N_m at $\alpha^{\text{open}} = \alpha_c$. Hence, we performed further simulations at criticality with a larger sample size (for the largest system size, $n^{\text{sample}} \approx 2 \times 10^4$ for $T=0$ and $n^{\text{sample}} = 1.6 \times 10^4$ for $T>0$).
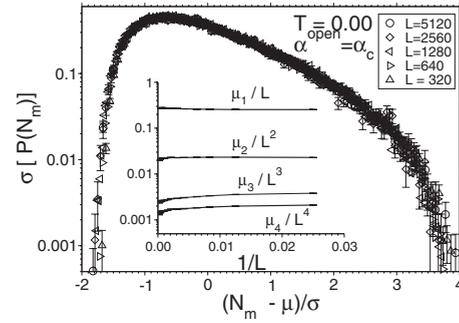
This allowed us to cross check the value of γ and to extract higher moments of the distribution.

Figure 4 displays distributions $[P(\langle N_m \rangle_T)]$ for $T=0$ (similar results have been obtained for $T>0$). The distributions have been rescaled to zero mean and unit variance. In all cases, we observe that the first moments μ_1 scale as $\mu_1 \sim L^{1+\epsilon}$ with $\epsilon = 0$ within the error bars. For the second moment, one would expect the scaling behavior $\mu_2 \equiv L\chi \sim L^{1+\gamma/\nu}$. Indeed, we find $\gamma/\nu = 0.955(8)$ for OA by a least-square fit. This is consistent with the numerical value obtained by the height of maxima. The third and the fourth moments scale as $\mu_3 \sim L^{2+\gamma_3}$ and $\mu_4 \sim L^{3+\gamma_4}$, respectively. For temperatures $T<2$, both exponents γ_3 and γ_4 agree within the statistical errors.

The resulting critical values α_c and critical exponents ν, γ, γ_3, γ_4 are summarized in Table I. All ratio of exponents γ/ν, γ_3/ν, and γ_4/ν are on the order of 1 and seem to increase with the temperature. Note that for a perfectly one-parameter scaling of the complete distribution with $\mu_1 \sim L$, one would expect $\gamma/\nu = \gamma_3/\nu = \gamma_4/\nu = \ldots = 1$. This property is only approximately fulfilled according to our data. This is shown in Fig. 5, where also the resulting phase diagram is displayed.

The standard order parameter in percolation problems is the relative size of the largest cluster. Since local sequence alignments (and its interpretation as DPRM) exhibit one spatial dimension, the observable N_m/L can also be interpreted as order parameter, which is one if the alignment covers the entire sequence. The usual finite-size-scaling ansatz for the relative size of the largest cluster order parameter reads as

$$[\langle N_m \rangle_T](\alpha^{\text{open}}; T; L)/L = L^{-\beta^{\text{geo}}/\nu} f[(\alpha^{\text{open}} - \alpha_c)L^{1/\nu}], \quad (8)$$

where the exponent β^{geo} describes the divergence of the largest cluster. By comparing this relation with Eq. (7), we may infer $\beta^{\text{geo}} = 0$ and verify that the scaling relation $\gamma + 2\beta^{\text{geo}} = d\nu$ [26] (with $d=1$ in our case) is again only approximately fulfilled. We confirmed that $\beta = 0$ within the error bars by considering β as free parameter in the finite-size-scaling analysis for N_m.

As mentioned above, the roughness only grows logarithmically with the system size. This implies that the fractal dimension d_r of the alignment path equals the topological

TABLE I. Critical gap-open penalty α_c and critical exponents ν, γ, γ_3, γ_4 for the observable N_m.

T	α_c	ν	γ/ν	γ_3/ν	γ_4/ν
0.00	8.306(4)	1.58(5)	0.955(8)	0.920(5)	0.903(9)
0.50	8.450(2)	1.54(1)	0.947(6)	0.92(1)	0.91(1)
1.00	8.871(4)	1.55(2)	0.946(6)	0.92(1)	0.91(1)
1.33	9.339(3)	1.53(2)	0.948(8)	0.92(1)	0.91(1)
2.00	10.791(3)	1.51(2)	0.931(9)	0.92(1)	0.89(1)
2.50	12.296(4)	1.50(4)	0.921(9)	0.92(1)	0.88(1)
3.50	16.227(1)	1.46(1)	0.87(1)	0.870(8)	0.80(1)
4.00	18.557(2)	1.38(5)	0.84(2)	0.924(5)	0.76(2)

dimension $d=1$, which is in agreement with the scaling relation $d_r = d - \beta^{\mathrm{geo}}/\nu$ in a trivial way.

B. Energetic properties

As mentioned above, the size of the largest cluster is usually regarded as the order parameter in percolation problems. In the nonpercolating phase, the size of the largest cluster typically grows logarithmically with the system size, whereas in the percolating phase its extension is comparable to the system size [26]. The average score of local alignments exhibits the same crossover when crossing the linear-logarithmic boundary. For this reason, we regard the average score $[S]/L$ (OA) or free energy $[F_T]/L$ (FTA) per length as a second-order parameter, as in [25]. Note that there is no direct geometrical interpretation for this quantity.

Scaling theory states that the order parameter scales as

$$[F_T]/L = L^{-\beta/\nu} f[(\alpha - \alpha_c)L^{1/\nu}], \quad (9)$$

with some universal scaling function f. This allows one to extract the critical value α_c and exponents ν and β from the data with the same method as described above. Here, we fixed α_c and ν with the values that have been obtained from the data collapse for the observable N_m and regard β/ν as a free parameter. The result for $T=0$ is shown in Fig. 6. The quality of these fits varied between $\hat{Q}=1.58$ for $T=1.00$ and $\hat{Q}=7.49$ for $T=4.00$.

By regarding ν and β as free parameters, we also used the scaling form (9) as a second cross check for the exponent ν.

Within the error bars, it is comparable with the results of the finite-size-scaling analysis for the observable N_m [for example, $\nu=1.50(7)$ for $T=0$]. For larger temperatures ($T \gtrsim 2$), only system sizes $L \geq 1280$ led to convincing results for this kind of check.

As can be seen in the second column of Table II, the free energy per length $[F_T]/L$ decreases like $\sim L^{-\beta/\nu}$, where β/ν increases monotonously with the temperature from 0.6324(7) to 0.7958(5). For small temperatures, i.e., $T<2$, the exponent $\beta \approx 1$ is not temperature dependent. Hence, the phase behavior regarding the exponent β is not universal any more when exceeding $T=2$.

C. Free-energy distributions close to criticality

In analogy to the distribution of the observable N_m in Fig. 4, the resulting rescaled score distributions right at $\alpha^{\mathrm{open}} = \alpha_c$ are shown in Fig. 7. Simulations for FTA yield comparable results. We performed an analysis of the moments of $P[S]$ and $P[F]$, respectively. The first moment scales as $\mu_1 \sim L^{1-\beta/\nu}$, as expected from the finite-size analysis shown in Fig. 6. We checked that the fit parameters agree with those from finite-size scaling.

Regarding the second moment, no divergence was observed. Its scaling behavior is given by $\mu_2 \sim L^{1-\omega}$ with $\omega>0$. The resulting fit parameters are listed in Table II. In the limit $\alpha^{\mathrm{open}} \to \infty$ and $L \to \infty$, the score distribution is predicted to follow a Gumbel distribution

$$P_{\mathrm{gumbel}}(S) = \lambda \exp\{-\lambda(S-S_0) - \exp[-\lambda(S-S_0)]\} \quad (10)$$

(according to the Karlin-Altschul-Dembo theory [35–37]).

FIG. 5. Results for FTA. Left: phase diagram for FTA. The linear phase is located below the critical line. Right: critical exponents as a function of the temperature.
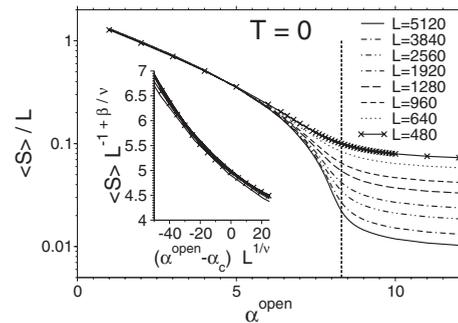
FIG. 6. Finite-size-scaling analysis for the average score per length.

TABLE II. Critical exponents for the average score/free energy per length. β/ν was obtained from finite-size scaling (see Fig. 6) and cross checked via the scaling of the first moment of the score distribution at criticality $\alpha = \alpha_c$. The exponent ω describes the fluctuations of the score distribution at $\alpha = \alpha_c$.

T	β/ν	ω
0.00	0.6324(07)	0.386(5)
0.50	0.6406(10)	0.391(5)
1.00	0.6443(11)	0.387(6)
1.33	0.6564(09)	0.380(9)
2.00	0.6835(10)	0.36(1)
2.50	0.7081(08)	0.33(1)
3.50	0.7691(06)	0.23(2)
4.00	0.7958(05)	0.20(2)

In the linear phase, the conditions of this theory are not valid any more. Interestingly, right at the critical point, the shape of the distributions is well described by a Gumbel distribution, at least in the high probability region (down to $P[S] \sim 10^{-4}$). In previous studies, we observed parabolic corrections to Eq. (10) that occur in the far right tail of the optimal score distribution [38,39]. The corrected distribution is empirically well described by

$$P(S) = \frac{1}{z'} P_{\text{gumbel}}(S) \exp[-\lambda_2 (S - S_0)^2], \qquad (11)$$

where λ_2 is a correction parameter and the normalization constant z' is indistinguishable from 1. We found evidence that λ_2 vanishes for $L \rightarrow \infty$ but persists even for gapless alignment $\alpha^{\text{open}} = \infty$ [39].

Here, we extend this study to finite-temperature alignment, i.e., to the free-energy distribution as a generalization of the score distribution. We employed generalized ensemble Monte Carlo simulations combined with Wang-Landau sampling [40] in the sequence space (details can be found elsewhere [41]). The (production) run for $L = 120$ employed 4.8×10^7 Monte Carlo steps for each distribution over the disorder.

In the following, we use the phase diagram as a guide to study the free-energy distribution for various temperatures. We kept the gap costs fixed ($\alpha^{\text{open}} = 12$, $\beta^{\text{ext}} = 1$) and only
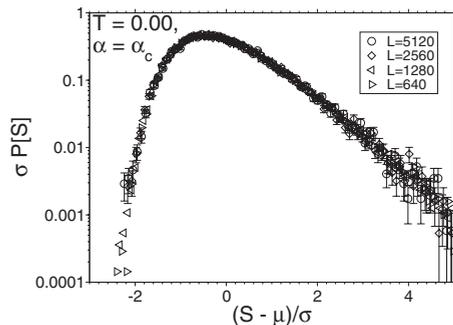
FIG. 7. Rescaled score distributions $T = 0$. The free-energy distributions of FTA look comparable.
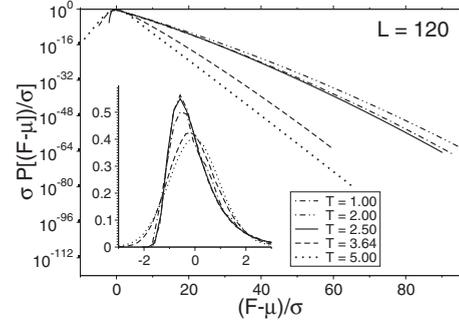
FIG. 8. Rescaled free-energy distribution of finite-temperature alignments. At $T = 2.50$ and below, the data are well described by a modified Gumbel distribution. For large temperatures, an exponential tail is observed. Inset: the same data shown with a linear ordinate. In the high probability region, the data for $T = 5.00$ are well described by a Gaussian distribution.

varied the temperature (between $T = 0$ and $T = 5$). The interpolating points are indicated by stars in the phase diagram in Fig. 5.

In the logarithmic regime ($T = 0, 1, 2, 2.5$), the free-energy distribution is well described by the modified Gumbel distribution (11) (see Fig. 8). Note that we have again rescaled the distributions to unit variance and zero mean. The fit parameters only change slightly with the temperature (see Table III).

The crossover from the logarithmic to the linear regime comes along with a change in the skewness, as can be seen in the inset of Fig. 8. In the high probability region, for a large value $T = 5.00$, a Gaussian distribution describes the data well. This was confirmed by a Kolmogorov-Smirnov test that yielded a p value of 0.14. For $T = 1/0.275 \approx 3.64$, the evidence for a Gaussian distribution is much smaller (a p value of 2×10^{-11}). We also checked that the change in the shape is accompanied by a change from logarithmic to linear growth of typical free energies (the position of the maximum) with the sequence length, i.e., the free energy becomes extensive (not shown here). This result can be understood in the following way. The partition functions that appear in the transfer-matrix calculations (6) become (more or less) independent and hence factorize in the linear phase. The total free energy decomposes into a sum of independent contributions and the central limit theorem applies.

When considering the rare-event tail at higher temperatures, the free-energy distribution is rather exponential than Gaussian, as can be seen in the main plot of Fig. 8. Hence,

TABLE III. Fit parameters of least χ^2 fits of the free-energy distributions to the modified Gumbel distribution (11) for $L = 120$ in the logarithmic phase.

T	λ	$10^4 \lambda_2$	s_0
0.00	0.2966(4)	3.182(1)	37.4(1)
1.00	0.2924(1)	2.900(5)	24.6(1)
2.00	0.2907(2)	3.122(7)	31.56(6)
2.50	0.2980(2)	3.16(1)	38.29(7)

we observe a crossover from a Gaussian distribution in the high probability region to the characteristic exponential tail of the Gumbel distribution. With the same argumentation as for the optimal alignment, sequence pairs appearing in the tail feature high similarities. The overall free energy is dominated by the ground state. This was confirmed by looking at the difference between the free energy and the ground-state energy for those sequences that occur in the tail of the distribution. The summation in the transfer matrix is virtually replaced by maximization, yielding an exponential tail. The finite-size effect that is responsible for the curvature of the optimal alignment statistics seems to be of marginal order in this case.

IV. DISCUSSION

We presented a finite-size-scaling analysis of the linear-logarithmic phase transition of finite-temperature protein sequence alignment. This phase transition is crucial to determine the set of parameters, where the alignment is of highest sensitivity. We have used the blosum62 scoring matrix together with affine gap costs, which is the most frequently used scoring system for actual database queries. This goes much beyond previous studies, which have investigated only simple scoring systems.

Two order parameters were studied in detail: the number of matches (i.e., the alignment length) and the average free energy per length. We have analyzed the phase transition using finite-size-scaling techniques. Using sophisticated algorithms, large systems could be studied, such that corrections to finite-size scaling are negligible. The resulting critical line $a_c(T)$ in the range $T=0-4$ provides a guide for biological applications, where suboptimal alignments play an important role.

Numerical values of the critical exponents ν, γ, and β suggest that the percolation transition is not universal with respect to different temperature values.

Under a different scoring model, the critical points will change. Since the critical exponents are not universal along the transition line, they are likely to change as well. For example, blosum80 is more sensitive to strongly related proteins than blosum62 (i.e., it features a stronger penalty for mismatches and has a smaller expected score). This implies that the critical line shown in Fig. 5 is expected to run below the one for blosum62 (i.e., located at smaller values of α^{open} for given T) because mismatches have to be circumvented with lesser costs so as to archive longer alignments.

For practical applications, the location of the critical point is most important, while the knowledge of the critical exponents is not so crucial. This means, unfortunately, that one has to determine individually the critical point for each scoring system used. Nevertheless, since we cross checked the critical points and exponents by considering different observables, the methodology is readily applicable to other scoring systems, like other matrices from the blosum or pam family. Since the number of matches turned out to be a simple and stable observable that exhibits the behavior of a percolation probability, we suggest to use this value in future studies targeted toward practical applications [42].

The free-energy distribution, which can be seen as a generalization of the score distribution over random sequences, crosses over from a modified Gumbel distribution with a parabolic correction in the tail given by Eq. (11) in the logarithmic phase to a modified Gaussian distribution with a linear rare-event tail in the linear phase. This is another example showing that the large-deviation properties of order-parameter distributions change significantly close to phase transitions.

Finite-temperature alignment, or in general probabilistic alignment, provides a variety of analysis methods in bioinformatics. The possibility of *directly* translating score-based alignment into a probabilistic description with only one parameter, namely, the temperature, is an advantage over pair HMM, which contain a huge set of parameters. Miyazawa pointed out that there is a "canonical temperature" $T=1$, for which the finite-temperature alignment achieves a more probabilistic meaning [16]. In more recent terms, this means that using the value $T=1$ finite-temperature alignment approximates a pair HMM. On the other side, by varying the temperature, a single scalar parameter, one is able to put more weight on suboptimal alignments and possibly detect alternative alignments. The knowledge of the phase diagram, which is provided in this study, prevents us from overdispersions by giving totally meaningless alignments too much weight.

To apply the methods to biological alignment problems, where one expects a large number of suboptimal alignments, one would replace the optimal score by the free energy. Evidence of the emergence of suboptimal alignments will then be signaled by a significantly larger free energy in comparison to the optimal score. To calibrate the free energy, one may use normalized values with respect to mean and standard deviation of the free-energy distribution of random sequences (also termed z score) at the same temperature.

[1] A. Krogh, R. Durbin, S. Eddy, and G. Mitchison, *Biological Sequence Analysis* (Cambridge University Press, New York, 1998).

[2] P. Clote and R. Backofen, *Computational Molecular Biology: An Introduction* (John Wiley & Sons, Chichester, 2005).

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, J. Mol. Biol. **215**, 403 (1990).

[4] W. R. Pearson and D. J. Lipman, PNAS USA **85**, 2444 (1988).

[5] T. F. Smith and M. S. Waterman, J. Mol. Biol. **147**, 195 (1981).

[6] M. Q. Zhang and T. G. Marr, J. Theor. Biol. **174**, 119 (1995).

[7] D. A. Huse and C. L. Henley, Phys. Rev. Lett. **54**, 2708 (1985).

[8] M. Kardar, Nucl. Phys. B **290**, 582 (1987).

[9] M. Mezard, J. Phys. (France) **51**, 1831 (1990).

[10] D. S. Fisher and D. A. Huse, Phys. Rev. B **43**, 10728 (1991).

[11] M. Kardar, in *Directed Paths in Random Media*, edited by F. D. P. Ginzparg and J. Zinn-Justin, Proceedings of the Les Houches Summer School of Fluctuating Geometries in Statistical Mechanics and Field Theory, LXII, 1994 (Elsevier, Amsterdam, 1996).

[12] U. Mückstein, I. L. Hofacker, and P. F. Stadler, Bioinformatics **18**, 153 (2002).

[13] L. Jaroszewski, W. Li, and A. Godzik, Protein Sci. **11**, 1702 (2002).

[14] R. Koike, K. Kinoshita, and A. Kidera, Proteins **56**, 157 (2004).

[15] R. Koike, K. Kinoshita, and A. Kidera, Proteins **66**, 655 (2007).

[16] S. Miyazawa, Protein Eng. **8**, 999 (1995).

[17] M. Kschischo and M. Lässig, in *Proceedings of the Pacific Symposium on Biocomputing, Honolulu, 2000*, edited by R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein (World Scientific, Singapore, 2000), pp. 621–632.

[18] R. Olsen, R. Bundschuh, and T. Hwa, in *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, edited by T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H.-W. Mewes, and R. Zimmer (AAAI Press, Menlo Park, CA, 1999), Vol. 270, p. 211.

[19] D. Drasdo, T. Hwa, and M. Lässig, J. Comput. Biol. **7**, 115 (2000).

[20] M. S. Waterman and M. Eggert, J. Mol. Biol. **197**, 723 (1987).

[21] R. Arratia and M. S. Waterman, Ann. Appl. Probab. **4**, 200 (1994).

[22] R. Bundschuh and T. Hwa, Discrete Appl. Math. **104**, 113 (2000).

[23] M. Vingron and M. S. Waterman, J. Mol. Biol. **235**, 1 (1994).

[24] T. Hwa and M. Lässig, in *Proceedings of the Second Annual International Conference on Computational Molecular Biology (RECOMB98)*, edited by S. Istrail, P. Pevzner, and M. S. Waterman (ACM, New York, 1998), p. 109.

[25] M. E. Sardiu, G. Alves, and Y. Yu, Phys. Rev. E **72**, 061917 (2005).

[26] D. Stauffer and A. Aharony, *Introduction to Percolation Theory* (Taylor & Francis, London, 1994).

[27] S. Henikoff and J. G. Henikoff, PNAS USA **89**, 10915 (1992).

[28] O. Gotoh, J. Mol. Biol. **162**, 705 (1982).

[29] G. H. Gonnet, M. A. Cohen, and S. A. Benner, Science **256**, 1443 (1992).

[30] X. Gu and W.-H. Li, J. Mol. Evol. **40**, 464 (1995).

[31] R. Cartwright, BMC Bioinf. **7**, 527 (2006).

[32] J. Houdayer and A. K. Hartmann, Phys. Rev. B **70**, 014418 (2004).

[33] B. Efron, Ann. Stat. **7**, 1 (1979).

[34] A. K. Hartmann, *Practical Guide to Computer Simulations* (World Scientific, Singapore, 2009).

[35] S. Karlin and S. F. Altschul, PNAS USA **87**, 2264 (1990).

[36] S. Karlin and A. Dembo, Adv. Appl. Probab. **24**, 113 (1992).

[37] A. Dembo, S. Karlin, and O. Zeitouni, Ann. Probab. **22**, 2022 (1994).

[38] A. K. Hartmann, Phys. Rev. E **65**, 056102 (2002).

[39] S. Wolfsheimer, B. Burghardt, and A. K. Hartmann, Algor. Mol. Biol. **2**, 9 (2007).

[40] F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).

[41] S. Wolfsheimer, I. Herms, S. Rahmann, and A. K. Hartmann (unpublished).

[42] In addition, one may avoid the computational effort of backward sampling by introducing auxiliary variables, if one is interested in the expected number of matches alone.