Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Bulletin of Mathematical Biology 67 (2005) 169–191

Bulletin of Mathematical Biology

www.elsevier.com/locate/ybulm

Toward an accurate statistics of gapped alignments

Maik Kschischo[a], Michael Lässig[b], Yi-Kuo Yu[c,d,*]

[a]University of Applied Sciences Koblenz, RheinAhrCampus Remagen, Südallee 2, 53424 Remagen, Germany

[b]Institut für Theoretische Physik, Universität zu Köln, Zülpicher Str. 71, 50937 Köln, Germany

[c]National Center for Biotechnology Information, National Library of Medicine, National Institute of Health, Bethesda, MD 20894, USA

[d]Department of Physics, Florida Atlantic University, Boca Raton, FL 33431-0991, USA

Received 15 September 2003; accepted 13 July 2004

Abstract

Sequence alignment has been an invaluable tool for finding homologous sequences. The significance of the homology found is often quantified statistically by p-values. Theory for computing p-values exists for gapless alignments [Karlin, S., Altschul, S.F., 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA 87, 2264–2268; Karlin, S., Dembo A., 1992. Limit distributions of maximal segmental score among Markov-dependent partial sums. Adv. Appl. Probab. 24, 13–140], but a full generalization to alignments with gaps is not yet complete. We present a unified statistical analysis of two common sequence comparison algorithms: maximum-score (Smith–Waterman) alignments and their generalized probabilistic counterparts, including maximum-likelihood alignments and hidden Markov models. The most important statistical characteristic of these algorithms is the distribution function of the maximum score $S_{\max}$, resp. the maximum free energy $F_{\max}$, for mutually uncorrelated random sequences. This distribution is known empirically to be of the Gumbel form with an exponential tail $P(S_{\max} > x) \sim \exp(-\lambda x)$ for maximum-score alignment and $P(F_{\max} > x) \sim \exp(-\lambda x)$ for some classes of probabilistic alignment. We derive an exact expression for λ for particular probabilistic alignments. This result is then used to obtain accurate λ values for generic probabilistic and maximum-score alignments. Although the result

[*] Corresponding author.

E-mail addresses: kschischo@rheinahrcampus.de (M. Kschischo), lassig@thp.Uni-Koeln.de (M. Lässig), yyu@fau.edu (Y.-K. Yu).

0092-8240/$30 © 2004 Society for Mathematical Biology. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.bulm.2004.07.001

demonstrated uses a simple match–mismatch scoring system, it is expected to be a good starting point for more general scoring functions.

© 2004 Society for Mathematical Biology. Published by Elsevier Ltd. All rights reserved.

1. Introduction

Alignment algorithms remain important in the analysis of biological sequences. In database searches, local similarities between sequences have to be distinguished from random matches. In at least two ways, this problem has become more challenging in recent years. With the increasing size of databases, random matches become more likely, and this effect decreases the confidence level of the sequence similarities found (Spang and Vingron, 2000).

The degree of similarity between two or more sequences is often measured by the alignment score. The common algorithms like BLAST and FASTA find a definite alignment of maximal score $S_{\max}$ for a given pair of sequences (Altschul et al., 1990; Pearson, 1988). Its statistical significance can be characterized by the so-called p-value, i.e., by the probability $P(S_{\max} > x)$ that a score value $S_{\max} > x$ occurs in alignments of uncorrelated random sequences. The underlying probability distribution function is known to be of Gumbel form (Gumbel, 1958), $P(S_{\max} > x) = 1 - \exp(-\kappa \exp(-\lambda x))$, for alignments without gaps and it is widely believed that the same functional form also applies to gapped alignments. The Gumbel parameters λ and κ, however, are known analytically only for the special case of gapless alignments (Karlin and Altschul, 1990; Karlin and Dembo, 1992) and have to be obtained by simulation otherwise. Known analytical approximations are restricted to the case of very large gap cost (Siegmund and Yakir, 2000; Metzler, 2002) or employ heuristics using a greedy approximation to the original Smith–Waterman algorithm (Mott and Tribe, 1999).

A somewhat different alignment approach utilizes the concept of likelihood or hidden Markov models; examples include HMMer (Eddy, 1998) and SAM (Karplus et al., 1998). These produce a probability distribution over alignments which is inferred from an underlying stochastic model of sequence evolution. The well known forward–backward algorithm serves for the computation of the likelihood and the most probable alignment can be filtered out by the Viterbi algorithm (see e.g., Durbin et al. (1998)). Again, there is to date no complete statistical theory to assess the significance of the results.

The alignment problem has an interesting connection to the statistical physics of disordered systems. This has been exploited to develop the *scaling theory* of gapped alignments discussed in a number of recent publications. Along these lines, Bundschuh (2002) has obtained the Gumbel parameter λ for a particular limit of gapped alignment called the longest common subsequence problem. Using the forward–backward version, Kschischo and Lässig (2000) have generalized the scaling theory to probabilistic alignments and identified the maximal free energy $F_{\max}$ as the relevant quantity for significance estimates.[1] Yu and Hwa (2001) have established that the probability distribution of $F_{\max}$ is of the Gumbel form, have provided a general criterion for determining the parameter λ, and have obtained λ exactly for a family of probabilistic alignments.

In this paper, we derive an accurate approximation for the Gumbel parameter λ of local maximum-score alignments with gaps. We refer to this approximation as the *cooling map*. Accurate λ values can be obtained without the need of extensive numerical simulations of random sequences. We use the idea that maximum score alignments can be obtained as the limit case of probabilistic alignments. The limit is governed by a variable τ which we call the temperature. The well known phase transition (Arratia and Waterman, 1994) of Smith–Waterman alignment (Smith and Waterman, 1981) was shown to exhibit scale invariance (Drasdo et al., 1998; Kschischo and Lässig, 2000), which allows one to compare alignments with different parameters. This leads to a scaling formula for λ as a function of the parameters in the alignment. To fix the scale of this scaling function, it is sufficient to analyze two special cases of probabilistic alignment with exact Gumbel parameter λ. We then use scaling theory to extract the Gumbel parameter λ of generic alignments from these 'solvable' families. Although the procedure is illustrated with a match–mismatch scoring function, it should be valid for position-independent scoring functions like the popular PAM matrices (Dayhoff et al., 1978). Currently, the theory is restricted to linear gap scores. We believe that it provides a good starting point for more general scoring functions.

This paper is organized as follows. In Section 2 we give a brief introduction to sequence alignment and define the most important quantities. The main results of this paper including the cooling map for the calculation of λ are summarized in Section 3. More detailed explanations are given afterwards. We conclude with a discussion of the perspectives and limitations of the method.

2. Review of sequence alignment

We give a brief introduction to sequence alignment and describe a simple scoring scheme for local alignments. We describe probabilistic alignment algorithms and the limiting procedure from probabilistic to maximum-score alignment. The phase diagram separating the local and global regimes of local probabilistic alignment is introduced.

2.1. Definition of alignments

A local alignment of two sequences $\mathbf{a} = \{a_i\}$ $(i = 1, \ldots, M)$ and $\mathbf{b} = \{b_j\}$ $(j = 1, \ldots, N)$ is defined as an ordered set of pairings (i, j) and of gaps $(i, -)$ and $(-, j)$ involving the elements of two contiguous subsequences $\{a_{m'}, \ldots, a_m\}$ and $\{b_{n'}, \ldots, b_n\}$; see Fig. 1(a). Its length is defined as the total number of aligned elements, $L \equiv m - m' + n - n' < M + N$.

[1] Notice that in Kschischo and Lässig (2000) the free energy considered contains both the forward and backward contributions, while the maximal free energy defined in this paper only contains the forward part.

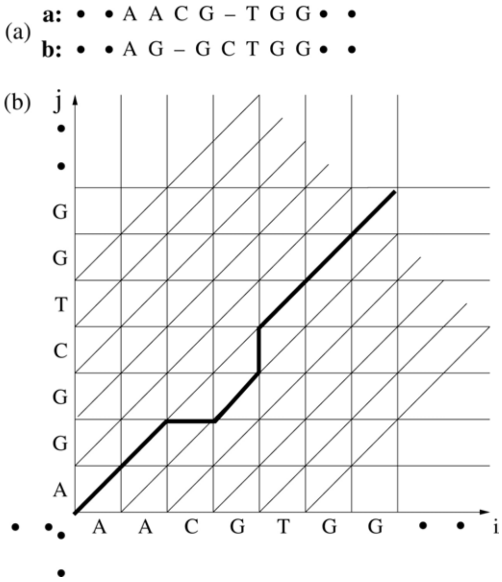

Fig. 1. (a) One possible local alignment of two sequences **a** and **b** with elements taken from a 4-letter alphabet. In the grid figure, the mth element of sequence **a** has its i coordinate equal to $m - 1/2$, and similarly the nth element of sequence **b** has its j coordinate equal to $n - 1/2$. Only the aligned subsequences are shown, with 6 pairings (five matches, one mismatch) and two gaps. (b) Unique representation of this alignment as directed path **A** (thick line) on an alignment grid.

In contrast to local alignments which allow for unpaired regions to both sides of the aligned subsequences, global alignments align the two sequences from head to toe and thus have length $L = M + N$. An alignment can be uniquely represented as a *directed path* **A** on the two-dimensional grid of Fig. 1(b).

2.2. Scoring of alignments

The elements of the sequences come from an alphabet χ of size c. For DNA sequences this will be the four bases A, C, G, T and for protein sequences the 20 amino acids. Each letter a occurs with frequency $p(a)$; we have $\sum_{a\in\chi} p(a) = 1$. The score of an alignment is defined as the sum of the scores of its pairings and gaps. A scoring system has to specify the substitution scores $s(a, b)$ for all pairings (a, b) and the gap score s_g. (Here one often distinguishes further between gaps following a pairing and gaps following another gap; this is called the affine gap cost.) It will prove convenient to normalize the scoring function in such a way that random pairings have a specified score average 2σ and variance 1,

$$\begin{aligned} &\sum_{a,b\in\chi} p(a)p(b)s(a,b) = 2\sigma, \\ &\sum_{a,b\in\chi} p(a)p(b)(s(a,b) - 2\sigma)^2 = 1. \end{aligned} \tag{1}$$

In the remainder of this paper, we use random sequences composed of equally distributed letters ($p(a) = 1/c$ for $a \in \chi$) with $c = 4$ and a simple scoring system that distinguishes only between matches ($s(a, b) = s_+$ for $a = b$), mismatches ($s(a, b) = s_-$ for $a \neq b$), and gaps. Normalized according to (1), the scores can be written in the form

$$\begin{aligned} s_+ &= \sqrt{c-1} + 2\sigma \\ s_- &= -1/\sqrt{c-1} + 2\sigma \\ s_g &= -\gamma + \sigma \end{aligned} \tag{2}$$

with two adjustable parameters, the gap cost γ and the score gain per aligned element, σ. The results below are generalizable to position-independent scoring matrices $s(a, b)$ (such as the PAM matrices (Dayhoff et al., 1978) for amino acid pairings). Eq. (1) can be fulfilled by a simple rescaling of the matrix entries.

2.3. Alignment algorithms

The celebrated Smith–Waterman (Smith and Waterman, 1981) algorithm finds the local score maxima $S_{i,j}$ for all points of the alignment grid. $S_{i,j}$ is defined as the maximum score over the set of alignment paths ending at the point (i, j),

$$S_{i,j} \equiv \max_{\mathbf{A}|i,j} S(\mathbf{A}). \tag{3}$$

Here, the score is given as the sum over pairings and gaps in the alignment $\mathbf{A}$

$$\begin{aligned} S(\mathbf{A}) &= \sum_{\textit{pairings } a,b} s(a, b) + \sum_{\textit{gaps}} s_g \\ &= \sigma L + \sum_{\textit{pairings } a,b} (s(a, b) - 2\sigma) - \sum_{\textit{gaps}} \gamma. \end{aligned} \tag{4}$$

The Smith–Waterman dynamic programming algorithm reads

$$S_{i,j} = \max\{0, S_{i-1,j} + s_g, S_{i,j-1} + s_g, S_{i-1,j-1} + s(a_i, b_j)\}, \tag{5}$$

where $s(a_i, b_j)$ denotes the score for a pairing of a_i and b_j. The total maximum score for a given pair of sequences is then simply $S_{\max} = \max_{i,j} S_{i,j}$. The lower cutoff score 0 is essential for local alignment and is absent from the corresponding algorithm for global alignment.

A probabilistic alignment takes into account alignment paths of arbitrary score. Each path $\mathbf{A}$ is associated with a weight factor $\exp[S(\mathbf{A})/\tau]$ given in terms of its score and the additional parameter $\tau > 0$. That is, the maximum (3) is replaced by the sum

$$Z_{i,j} \equiv \sum_{\mathbf{A}|i,j} \exp[S(\mathbf{A})/\tau] \tag{6}$$

over *all* alignment paths ending at the point (i, j). The exponential weighting of different paths is motivated by the additivity of the alignment score: when the alignment is composed of two pieces, the sum of the corresponding scores is just the score of the whole alignment.

A similar dynamic programming (see, e.g., Kschischo and Lässig (2000), Yu and Hwa (2001)) is available for probabilistic alignment. The recursion relation reads

$$Z_{i,j} = 1 + \nu[Z_{i-1,j} + Z_{i,j-1}] + v(a_i, b_j) Z_{i-1,j-1}, \tag{7}$$

where

$$v(a, b) = \exp[s(a, b)/\tau], \quad \nu = \exp(s_g/\tau) \tag{8}$$

are the weights of pairings and gaps, respectively. The $+1$ term corresponds to the lower cutoff score 0 in (5) and is only present in the case of local alignment.

Apart from the scoring matrix (2), probabilistic alignments have three parameters. The average gap frequency and length of the paths are controlled by γ and σ, respectively, while τ governs the relative weight of paths with different scores. Note, that $Z_{i,j}$ denotes a weight and not a probability. This important difference from the forward algorithm of hidden Markov models will be used later.

Probabilistic alignments are related to standard Smith–Waterman alignments in a simple way. From (6), we obtain

$$S_{i,j} = \lim_{\tau \to 0} F_{i,j}, \tag{9}$$

where $F_{i,j} \equiv \tau \ln Z_{i,j}$. Therefore, the total maximum $F_{\max} \equiv \max_{i,j} F_{i,j}$ is the finite-temperature counterpart of the score maximum $S_{\max}$. Indeed, Yu and Hwa (2001) have shown that $F_{\max}$ obeys Gumbel statistics for independent random sequences.

In analogy with statistical physics, the parameter τ is called the *temperature*; the quantities Z_{ij} are the local *partition function*. The total partition function Z is the sum over paths analogous to (6) without constrained end point. It can be computed as $\sum_{i,j} Z_{i,j}$. The quantities $F_{i,j}$ and $F \equiv \tau \log Z$ are called the local and total *free energies*, respectively. This connection has been used by Zhang and Marr (1995), Miyazawa (1996) and Hwa and Lässig (1998).

2.4. Normalization of probabilistic alignments

We now turn to specific probabilistic alignments given, for example, by a hidden Markov model producing *correlated* sequence pairs with a joint probability distribution $Q[\mathbf{a}, \mathbf{b}]$. In this case, the total partition function $Z[\mathbf{a}, \mathbf{b}] := Z$ for a given pair of sequences has the important interpretation as the ratio of their probability in the Markov model and their 'null probability' $Q_0[\mathbf{a}, \mathbf{b}]$ without evolutionary correlations,

$$Z[\mathbf{a}, \mathbf{b}] = \frac{Q[\mathbf{a}, \mathbf{b}]}{Q_0[\mathbf{a}, \mathbf{b}]}, \tag{10}$$

see Kschischo and Lässig (2000) and Yu et al. (2002). Here $Q_0[\mathbf{a}, \mathbf{b}]$ is given by

$$Q_0[\mathbf{a}, \mathbf{b}] = \prod_{a \in \mathbf{a}} p(a) \prod_{b \in \mathbf{b}} p(b). \tag{11}$$

The free energy is the appropriate generalization of the log odds score to gapped alignment. The alignment weights (8) are then determined by the mutation and insertion/deletion probabilities of the underlying Markov model for sequence evolution. We do not need the details of this mapping here. It is important to note that the Markov model imposes a normalization condition[2] on the weights (8). Then, the parameters $s(a, b)$ and s_g are no longer independent. In the parametrization (1), we can express σ in terms of the other scoring parameters. For the scoring system (2), we write

$$\sigma = \sigma_1(\gamma, \tau). \tag{12}$$

As can be inferred[3] from Yu et al. (2002), even under *position-specific* scoring functions this family of alignments has $\lambda\tau = 1$. In Section 4.2, we show that there exists a second family of 'solvable' alignments which has $\lambda\tau = 2$. This family is given by another normalization condition, which reads for the scoring system (2)

$$\sigma = \sigma_2(\gamma, \tau). \tag{13}$$

The functions $\sigma_{1,2}(\gamma, \tau)$ are roots of a quadratic and a quartic equation, respectively; see Section 4.2.

2.5. *The phase diagram of local alignment*

For local alignments, there are two different regimes, the *local* and the *global* regimes. In the following, we discuss the statistics of alignments over an ensemble of random sequences without mutual correlations, which is the 'null model' for significance estimates used in Eq. (10). Averages over this null ensemble are denoted by $\langle \cdots \rangle_0$. Of particular importance for understanding the local and global regimes are the average free energy

$$\langle F \rangle_0 = \sum_{\mathbf{a},\mathbf{b}} Q_0[\mathbf{a}, \mathbf{b}]\tau \ln Z[\mathbf{a}, \mathbf{b}] \tag{14}$$

and its local counterparts $\langle F_{i,j} \rangle_0$. The properties of these quantities are determined by which paths contribute most to the local partition sums $Z_{i,j}$ for typical sequence pairs. The contribution of a given path having length L is determined by the score gain per aligned element σ, leading to a score term σL; see Eq. (4). Consequently, long paths dominate for sufficiently large σ but are strongly suppressed for small σ.

For sequence pairs of long sequences and $i \approx j \to \infty$ the asymptotic behavior of the ensemble-averaged local free energy $\langle F_{i,j} \rangle_0$ is given by[4]

$$\begin{array}{ll} \langle F_{i,j} \rangle_0 \simeq [\sigma - \sigma_c(\gamma, \tau)] \cdot (i + j) & \text{for } \sigma > \sigma_c \\ \langle F_{i,j} \rangle_0 \to F_0(\gamma, \sigma, \tau) & \text{for } \sigma < \sigma_c \end{array} \tag{15}$$

with a parameter-dependent threshold value $\sigma_c(\gamma, \tau) < 0$ (Kschischo and Lässig, 2000). Both regimes can be characterized by the average length $L = \partial \langle F \rangle_0 / \partial \sigma$ of a local alignment (compare Eqs. (4) and (6)). In the global alignment regime ($\sigma > \sigma_c(\gamma, \tau)$), the entire sequences are aligned, i.e., $L \simeq 2N$. In the local alignment regime ($\sigma < \sigma_c(\gamma, \tau)$), L

[2] For hidden Markov models this condition corresponds to the conservation of transition probabilities.

[3] In Yu et al. (2002), the substitution scores are rescaled to keep $\tau = 1$, and the result $\lambda = 1$ in this normalization. Here we allow for a general temperature τ without rescaling the substitution scores. It is easy to see the result then reads $\lambda\tau = 1$ regardless of the value of τ.

[4] We use $\simeq$ to indicate asymptotic equality and $\sim$ for asymptotic proportionality. The symbol $\approx$ indicates approximate equality of two numerical values.

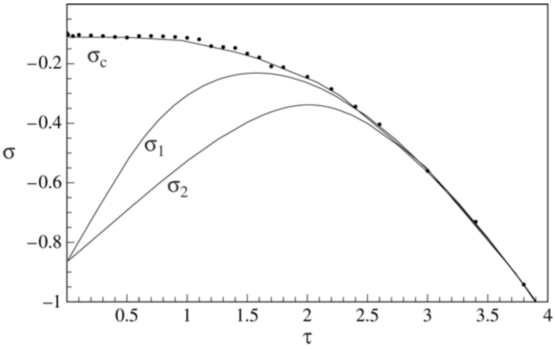

Fig. 2. Phase diagram of probabilistic sequence alignment for $\gamma = 5.5$. The critical line $\sigma_c(\gamma, \tau)$ separates the local ($\sigma < \sigma_c$) and the global ($\sigma > \sigma_c$) alignment regimes. The dots show numerical simulations which compare favorably with the curve obtained from (31) in Section 3.2. The curves $\sigma_1(\gamma, \tau)$ and $\sigma_2(\gamma, \tau)$ correspond to the two families of solvable cases; see Section 4.2 for details.

reaches a finite limit $L_0(\gamma, \sigma, \tau) \equiv \partial F_0(\gamma, \sigma, \tau)/\partial\sigma$. The two regimes are separated by a phase transition. In the zero-temperature limit, this is the well known transition (Arratia and Waterman, 1994) of maximum-score alignments (Smith and Waterman, 1981). This transition persists for probabilistic alignments, but the transition point $\sigma_c(\gamma, \tau)$ changes with temperature. Fig. 2 shows the temperature dependence of $\sigma_c(\gamma, \tau)$ for a given value of γ. The numerical data are in good agreement with the transition curve which we compute in Section 3.2 below; see Eqs. (31) and (32). The two lines $\sigma_1(\gamma, \tau)$ and $\sigma_2(\gamma, \tau)$ of 'solvable' alignments are also shown in Fig. 2.

3. Approximating the Gumbel parameter λ

The cooling map allows for a rapid and accurate evaluation of the Gumbel parameter λ for local maximum score alignments with gaps. In this section we collect the two main results underlying the cooling map and detail our practical implementation for the calculation of λ.

The two basic results are:

(1) The parameter dependence of the Gumbel parameter has the functional form

$$\lambda(\gamma, \sigma, \tau) = b(\gamma, \tau)\,\Phi(\sigma - \sigma_c, \gamma). \tag{16}$$

The function Φ contains the universal singularity

$$\Phi \simeq |\sigma - \sigma_c(\gamma, \tau)|^{1/2}, \tag{17}$$

as the phase transition is approached from the local regime. The exponent 1/2 is independent of the parameter values. The explicit temperature dependence of λ is determined by the prefactor $b(\gamma, \tau)$. This nonuniversal amplitude is directly related to the amplitude of F_0 (see Fig. 5 and Eq. (36)). It is seen to be approximately independent of temperature in the range $0 < \tau < 0.5$ and increases strongly for $\tau > 1$ as more and more paths contribute to the partition function.

(2) There are functions $\sigma_1(\gamma, \tau)$ and $\sigma_2(\gamma, \tau)$ with

$$\sigma = \sigma_1(\gamma, \tau) \Longleftrightarrow \lambda\tau = 1 \quad \text{for } \tau > 0 \tag{18}$$

and

$$\sigma = \sigma_2(\gamma, \tau) \Longleftrightarrow \lambda\tau = 2 \quad \text{for } \tau > 0. \tag{19}$$

In the following discussion we keep the gap parameter γ at a fixed value as was done in the phase diagram in Fig. 2. In Section 4.1 we derive Eqs. (16) and (17) from scaling theory. The second result determines the Gumbel parameter along the lines $\sigma = \sigma_1(\gamma, \tau)$ and $\sigma = \sigma_2(\gamma, \tau)$ in the phase diagram. Analytical results for both functions can be found in Section 4.2.

3.1. *The cooling map*

Eq. (16) allows us to relate the λ values at two different points $(\sigma^{(0)}, \tau^{(0)})$ and $(\sigma^{(1)}, \tau^{(1)})$ in the local alignment regime of the phase diagram. We find

$$\frac{\lambda(\gamma, \sigma^{(0)}, \tau^{(0)})}{\lambda(\gamma, \sigma^{(1)}, \tau^{(1)})} = \frac{b(\gamma, \tau^{(0)})}{b(\gamma, \tau^{(1)})} \frac{\Phi(\sigma^{(0)} - \sigma_c(\gamma, \tau^{(0)}), \gamma)}{\Phi(\sigma^{(1)} - \sigma_c(\gamma, \tau^{(1)}), \gamma)}. \tag{20}$$

In the special case, when the two points have equal distances $|\sigma^{(0)} - \sigma_c(\gamma, \tau^{(0)})| = |\sigma^{(1)} - \sigma_c(\gamma, \tau^{(1)})|$ from the critical line σ_c, only the ratio $b(\gamma, \tau^{(0)})/b(\gamma, \tau^{(1)})$ of the prefactors is important. To fix this ratio, we use the solvable cases (18) and (19). Placing the two points $\sigma^{(1)} = \sigma_1(\gamma, \tau^{(1)})$ and $\sigma^{(0)} = \sigma_2(\gamma, \tau^{(0)})$ on the solvable lines and adjusting the temperature values $\tau^{(0)}$ and $\tau^{(1)}$ according to the condition

$$|\sigma_1(\gamma, \tau^{(1)}) - \sigma_c(\gamma, \tau^{(1)})| = |\sigma_2(\gamma, \tau^{(0)}) - \sigma_c(\gamma, \tau^{(0)})|, \tag{21}$$

we find with (18) and (19)

$$\frac{b(\gamma, \tau^{(1)})}{b(\gamma, \tau^{(0)})} = \frac{\tau^{(0)}}{2\tau^{(1)}}. \tag{22}$$

The map $\tau^{(1)} = R(\tau^{(0)})$ under the condition (21) is illustrated graphically in Fig. 3. Since $R(\tau)$ is always smaller than τ we refer to the map as the *cooling map*.We define it here more formally as

$$|\sigma_1(\gamma, R(\tau)) - \sigma_c(\gamma, R(\tau))| = |\sigma_2(\gamma, \tau) - \sigma_c(\gamma, \tau)|, \tag{23}$$

and rewrite Eq. (22) as

$$\frac{b(\gamma, R(\tau))}{b(\gamma, \tau)} = \frac{\tau}{2R(\tau)}. \tag{24}$$

3.2. *λ for maximum score alignment*

We now turn to the calculation of λ in the limit $\tau \to 0$. Consider two points (σ, τ) and $(\sigma + \rho, R(\tau))$ not necessarily on the solvable lines. If both points have equal distances

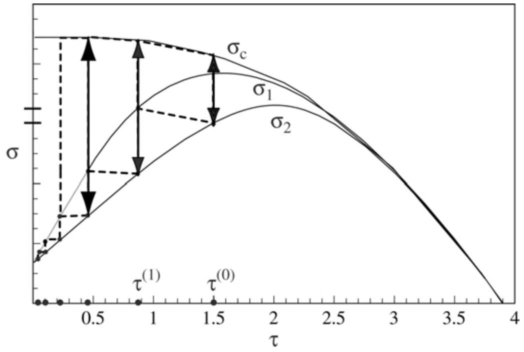

Fig. 3. Illustration of the cooling map. The initial temperature $\tau^{(0)}$ is mapped to $\tau^{(1)} = R(\tau^{(0)})$. The map preserves the distance from the phase transition line σ_c. The map can be iterated to lower and lower temperatures (dots).

from the phase transition line ($|\sigma + \rho - \sigma_c(\gamma, R(\tau))| = |\sigma - \sigma_c(\gamma, \tau)|$), then

$$\rho = \sigma_c(\gamma, R(\tau)) - \sigma_c(\gamma, \tau) = \sigma_1(\gamma, R(\tau)) - \sigma_2(\gamma, \tau), \tag{25}$$

see Eq. (23). The relation between the Gumbel parameters at both points is

$$\lambda(\gamma, \sigma + \rho, R(\tau)) = \frac{\tau}{2R(\tau)}\lambda(\gamma, \sigma, \tau). \tag{26}$$

By iterating the cooling map (see Fig. 3) from the initial point $\tau^{(0)} = \tau$, we obtain ($n = 1, 2, \ldots$)

$$\lambda(\gamma, \sigma + \rho^{(n)}, \tau^{(n)}) = B^{(n)}\lambda(\gamma, \sigma, \tau) \tag{27}$$

with

$$\begin{aligned} \tau^{(n)} &= R(\tau^{(n-1)}) \\ \rho^{(n)} &= \sum_{k=1}^{n}(\sigma_1(\gamma, \tau^{(k)}) - \sigma_2(\gamma, \tau^{(k-1)})) \\ B^{(n)} &= \prod_{k=1}^{n} \frac{\tau^{(k-1)}}{2\tau^{(k)}}. \end{aligned} \tag{28}$$

These sequences are rapidly converging. Their limit

$$\lambda(\gamma, \sigma + \rho^{(\infty)}(\gamma, \tau), 0) = B^{(\infty)}(\gamma, \tau)\lambda(\gamma, \sigma, \tau) \tag{29}$$

is practically reached after only a few iterations. With an initial point on the solvable line $\sigma = \sigma_1(\gamma, \tau)$, we obtain

$$\lambda(\gamma, \sigma_1(\gamma, \tau) + \rho^{(\infty)}(\gamma, \tau), 0) = B^{(\infty)}(\gamma, \tau)/\tau, \tag{30}$$

a parametric representation of the λ values for maximum-score alignment.

The computed values agree very well with those from numerical simulations, as shown in Fig. 4. The formula (30) also illustrates why the so-called Viterbi algorithm is often

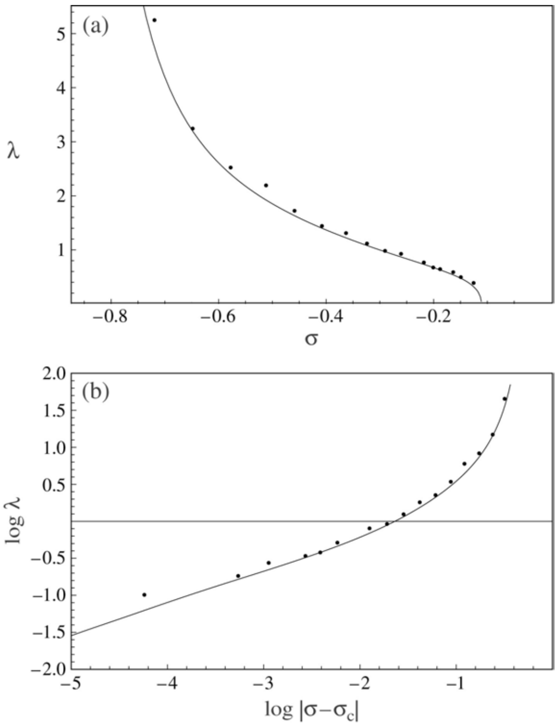

Fig. 4. (a) The Gumbel parameter λ for maximum-score alignment as a function of σ for $\gamma = 5.5$. Dots represent numerical results from pairs of independent random sequences of length $N = 1500$. The curve is obtained from (30). (b) The same values of λ as a function of the distance from the phase transition point $|\sigma - \sigma_c(\gamma, \tau = 0)|$ in a log–log plot. The asymptotic singularity is given by (17).

inaccurate. This is a zero-temperature alignment derived from a maximum-likelihood point; its parameters are $(\gamma, \sigma_1(\gamma, \tau), 0)$. However, its properties are not simply related to those at the maximum-likelihood point $(\gamma, \sigma_1(\gamma, \tau), \tau)$ since the correction terms $\rho^{(\infty)}$ and $B^{(\infty)}$ are neglected.

Since the function $\sigma_1(\tau)$ converges rapidly to $\sigma_c(\tau)$ for large values of τ, we can also compute the zero-temperature phase transition point from the same cooling map,

$$\sigma_c(\gamma, 0) = \lim_{\tau\to\infty} [\sigma_1(\gamma, \tau) + \rho^{(\infty)}(\gamma, \tau)]. \tag{31}$$

Using the fact that $\lambda(\gamma, \sigma_c(\gamma, \tau), \tau) = 0$ and comparing with (29), we have

$$\lambda(\gamma, \sigma_c(\gamma, \tau) + \rho^{(\infty)}(\gamma, \tau), 0) = 0 = \lambda(\gamma, \sigma_c(\gamma, 0), 0).$$

The entire phase transition curve $\sigma_c(\gamma, \tau)$ is then accurately expressed as

$$\sigma_c(\gamma, \tau) = \sigma_c(\gamma, 0) - \rho^{(\infty)}(\gamma, \tau), \tag{32}$$

see Fig. 2.

3.3. *Practical implementation*

A practical implementation requires knowledge of the cooling map (23). It is important to have an accurate low-temperature limit. With the scale of the temperature explicitly included, we find that the following low-temperature approximation works very well:

$$\frac{R(\tau)}{\tilde{\tau}} = \frac{\tau}{2\tilde{\tau}} + \exp\left(-c_1 \frac{\tilde{\tau}}{\tau} + c_2\right) \tag{33}$$

with

$$\tilde{\tau}(\gamma) = \arg\max_{\tau} \sigma_1(\gamma, \tau) \tag{34}$$

chosen as our scale for temperature. The two constants $c_1 = 3.4$ and $c_2 = 1.1$ were obtained from a fit to the data. A heuristic derivation of (33) will be given in Appendix C. Alternatively, the cooling map could be obtained from an iteration according to the definition (23). This requires the knowledge of $\sigma_c(\gamma, \tau)$, which can be rapidly computed (Kschischo and Lässig, 2000). Analytical expressions for $\sigma_1(\gamma, \tau)$ and $\sigma_2(\gamma, \tau)$ are given in Eqs. (39) and (43). In Algorithm 1 these calculations are used as subroutines.

```
**Require:** $\gamma, \tau_0$
  $\tau \leftarrow \tau_0$
  $\sigma \leftarrow \sigma_1(\gamma, \tau)$ {*From Eq.* (39)}
  {*Computation of Eq.* (28)}
  **repeat**
    $\tau' \leftarrow R(\tau)$ {*see text for R*}
    $\rho \leftarrow \rho + \sigma_1(\tau') - \sigma_2(\tau)$ {*From Eq.* (43)}
    $B \leftarrow B \times \tau/(2\tau')$
  **until** convergence is reached
  $\lambda(\sigma + \rho, 0) \leftarrow B/\tau$
```

Algorithm 1. Pseudocode for the computation of λ for Smith–Waterman alignment. The algorithm requires the starting value τ_0 and the gap parameter γ as input. The arrows indicate the directions of assignments and × denotes multiplication. Comments are enclosed by curly brackets.

4. Background on the results

In this section we provide some background on the main results used in Section 3 for the approximation of the Gumbel parameter λ.

4.1. *Scaling of local alignments*

The first result ((16), (17)) is derived from the scaling theory of alignment (Drasdo et al., 1998; Hwa and Lässig, 1998; Olsen et al., 1999; Drasdo et al., 2000). The score fluctuations of *global* alignment belong to the university class of directed polymers in a random medium; see Halpin-Healy and Zhang (1995) for a review. For probabilistic alignment,

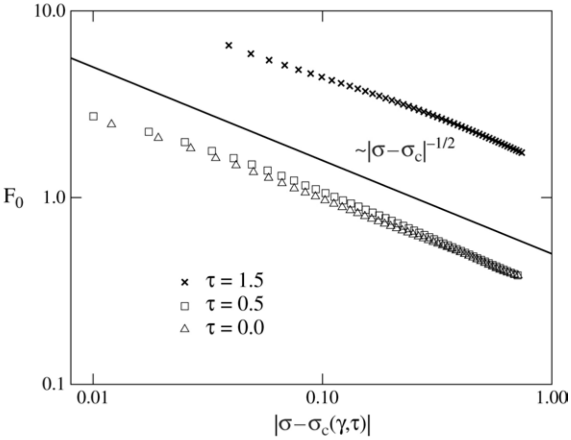

Fig. 5. The universal scaling of local alignments. The ensemble-averaged free energy obeys a power law $F_0 \sim |\sigma - \sigma_c(\gamma, \tau)|^{-1/2}$ with a universal (parameter-independent) exponent of 1/2. The prefactor, however, is not universal and changes strongly with τ and γ (here $\gamma = 1.7$).

the fluctuation of the free energy in the global alignment regime ($\sigma > \sigma_c$) is given by

$$\langle F_{i,j}^2 \rangle_0 - \langle F_{i,j} \rangle_0^2 \sim (i + j)^{2/3}. \tag{35}$$

This scaling is interesting even for local alignment. For a pair of random sequences, there are often high scoring islands. Within these islands, the alignment behaves as a global alignment and the free energy grows linearly ($F \simeq |\sigma - \sigma_c| L_0$) with the length L_0 of the island (compare with Eq. (15)). Note that these islands occurred through the upward fluctuations of the free energy. Therefore, the amplitude of the fluctuations also sets the bound for the length and the free energy of typical islands. Comparing the fluctuations (35) with the linear growth of the free energy within a typical island, one obtains the typical length of the island $L_0 \sim |\sigma - \sigma_c|^{-3/2}$. From this one deduces

$$F_0 \sim |\sigma - \sigma_c(\gamma, \tau)|^{-1/2} \tag{36}$$

as $\sigma_c(\gamma, \tau)$ is approached from below. The characteristic power law manifests the continuous phase transition between the local and global alignment regimes. The exponent 1/2 is a universal property of sequence alignments independent of alignment parameters (Drasdo et al., 1998) and in particular of the temperature (Kschischo and Lässig, 2000). Fig. 5 shows the singular behavior (36) for different values of the temperature. The exponent is seen to be temperature independent, while the prefactor varies with τ (as does the transition point σ_c).

The quantity λF_0 is dimensionless, thus Eq. (36) suggests scaling (16) and (17) for the Gumbel parameter λ.

4.2. Solvable probabilistic alignments

In this section, we discuss two cases in which the λ values can be obtained analytically. As described in Yu and Hwa (2001), the key to understanding the tail of the score distribution is in *global alignment*, since the value of λ is determined by the equation

$$\lim_{t\to\infty}\langle e^{\lambda\tau\ln(\widetilde{W}_t)}\rangle_0 = \lim_{t\to\infty}\langle \widetilde{W}_t^{\lambda\tau}\rangle_0 = 1, \tag{37}$$

where the quantity $\widetilde{W}_t$, defined by Eq. (A.3), depends only on global alignment weights Eq. (A.2). Details are described in Appendix A.

This normalization condition can be solved explicitly in the two cases $\lambda\tau = 1$ and $\lambda\tau = 2$. The solution to each of the two cases imposes a different relationship among the alignment parameters. For our simple scoring system, they are the Eqs. (12) and (13).

The first case $\langle\widetilde{W}_t\rangle_0 = 1$ has been elaborated by Yu et al. (2002). Here we work out the second case $\lim_{t\to\infty}\langle\widetilde{W}_t^2\rangle_0 = 1$. We use the (weak) additional assumption that substitution scores at different lattice points can be treated as independent. In fact, it has been shown numerically (Olsen et al., 1999; Bundschuh, 2002) that the effect of such an assumption is generally negligible. The calculation of our second case is much more difficult since the relevant partition function is that of an interacting system. Quite remarkably, the result can still be expressed as a simple relation between the alignment parameters. The relevant definitions and results are given below; some details of the derivations can be found in Appendices A and B.

4.2.1. *First condition*

As will be shown in Appendix A, to achieve $\langle\widetilde{W}_t\rangle_0 = 1$, we only need to satisfy

$$2\nu + v = 1, \tag{38}$$

where $v \equiv \langle v(a,b)\rangle_0 \equiv y f_s(\tau)$ is the average substitution weight, $y = \exp(2\sigma/\tau)$, and $f_s(\tau) \equiv \sum_{a,b}\exp[(s(a,b))/\tau]p(a)p(b)$ with $p(a)$ being the background frequency of character a. Similarly, we may also write the linear gap weight ν as $\nu \equiv y^{1/2}\exp(-\gamma/\tau) \equiv y^{1/2} f_\gamma(\tau)$. The condition (38) is therefore a quadratic equation in $y^{1/2}$ and can be readily solved to yield $\sigma_1(\gamma,\tau)$ as

$$\sigma_1(\gamma,\tau) = \tau\log\left(\frac{f_\gamma(\tau)}{f_s(\tau)} + \sqrt{\left(\frac{f_\gamma(\tau)}{f_s(\tau)}\right)^2 + \frac{1}{f_s(\tau)}}\right). \tag{39}$$

Here, and in Eq. (43), we suppress the γ dependence. Note that along the phase trajectory $\sigma_1(\gamma,\tau)$, the λ value is simply $1/\tau$.

Although the condition (38) was pointed out in Yu and Hwa (2001) and Yu et al. (2002), a formal mathematical derivation was omitted. In Appendix A, a formal derivation for the condition (38) is given. The basic mathematical structure used to obtain this result involves discrete Laplace transform and Fourier transform. After those are done, the quantity $\langle\widetilde{W}_t\rangle_0$ can be expressed as a contour integral (arising from the inverse Laplace transform) which gives us the value 1 for all $t \geq 1$ if and only if the condition (38) holds.

4.2.2. *Second condition*

To explain the condition for $\langle\widetilde{W}_t^2\rangle_0 = 1$, we define the variance Δ of the substitution weight,

$$\Delta \equiv \langle v^2(a,b)\rangle_0 - v^2. \tag{40}$$

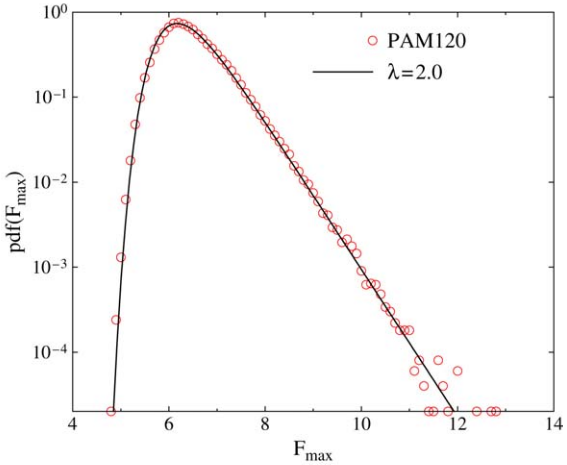

Fig. 6. The histogram and the Gumbel fit using the second condition at $\tau = 1$. The circles represent the alignment score histogram of 500,000 random sequence pairs using the PAM120 scoring matrix and linear gap cost $\gamma = 4.5$. Each random sequence generated has length $N = 600$. The solid line corresponds to a Gumbel fit with $\lambda = 2.0$ as expected, together with the other fitted parameter $\ln \kappa = \ln(KN^2) = 12.4$.

Using the same notation as before, we can write Δ as

$$\Delta = y^2[f_s(\tau/2) - f_s^2(\tau)]. \tag{41}$$

The end result of this calculation is the following condition:

$$(1+v)\sqrt{(1-v)^2 - 4\nu^2} = \Delta, \tag{42}$$

upon the satisfaction of which we can have $\lim_{t\to\infty}\langle \widetilde{W}_t^2\rangle_0 = 1$ and consequently $\lambda = 2/\tau$. Note that here we also need to have $2\nu + v < 1$, that is to say $\langle \widetilde{W}_t\rangle_0$ decays exponentially with t. Eq. (42) can be recast in terms of y, $f_s(\tau)$, $f_\gamma(\tau)$ and we then have a quartic equation in y. Among the four roots of y, we pick the real root r with range $0 < r < 1$. We can then call

$$\sigma_2(\gamma, \tau) = \frac{\tau}{2}\log(r). \tag{43}$$

The derivation is very similar to, but more involved than, that of Friedberg and Yu (1994) and Yu (1999) for a related problem. As in the first case, it involves again Fourier and Laplace transformations; see Appendix B for the sketch of the process. A complete derivation that also includes the affine gap functions will be provided in a separate publication (Yu, 2004).

Our prediction is tested by an extensive numerical simulation at $\tau = 1$ using the PAM120 scoring matrix and a linear gap cost $\gamma = 4.5$. Fig. 6 shows the score histogram obtained from aligning half a million pairs of random sequences of length $N = 600$ together with a Gumbel fit. The tail is given by the parameter $\lambda = 2.0 \pm 0.02$ as expected.

5. Discussion

We have discussed here a unified statistical analysis of probabilistic and maximum-score alignments with gaps. In particular, we have shown how exact results on particular alignment families and scaling can be combined to infer the Gumbel parameter λ accurately.

To compute p-values, the second Gumbel parameter κ has to be determined as well. While this was not the primary focus of our study, we are currently working in this direction. However, a precise λ value can facilitate a rapid estimate of κ either from a single but large-size pairwise alignment using the island method (Olsen et al., 1999) or from the average score of aligning a few random sequence pairs as described in Yu et al. (2002).

We would like to emphasize that our method should be readily applicable to position independent scoring functions, since only the score average σ of random pairings enters the theory. Suitable modifications analogous to those in Yu et al. (2002), could possibly render our method applicable to position specific scoring functions. This will be of particular importance for profile searches (see e.g., Eddy (1998)).

The implications of our work are two-fold. Conceptually, probabilistic and maximum-score alignments have long been regarded as rather different statistical entities, linked only by ad hoc procedures like the Viterbi algorithm. Regarding maximum-score alignments as the zero-temperature limit of probabilistic alignments opens a new avenue to understand the mathematics of the former. The statistics of Smith–Waterman alignments may be better understood beyond the heuristic level.

From a practical point of view, maximum-score alignments retain their importance since they are easier to interpret than their probabilistic counterparts, and their fidelity (i.e., the fraction of correctly aligned element pairs) tends to be higher (Kschischo and Lässig, 2000). On the other hand, a recent work (Yu et al., 2002) indicates comparable performance between maximum-score alignment and probabilistic alignment when tested on a real biological database. Thus we expect the alignment tools of the future will be a judicious combination of probabilistic and maximum-score parts.

Appendix A

In this appendix, we demonstrate that condition (38) is indeed all we need for the first case. Recall that the quantity $Z_{i,j}$, which consists of weights of all paths terminating at point (i, j) regardless of the starting points, satisfies the recursion relation (7). In a simpler context, we may consider the quantity $w_{i,j}$ that sums the weights of all paths starting at the origin and terminating at point (i, j). The quantity $w_{i,j}$ is also the global alignment weight between subsequences $\{a_1, a_2, \ldots, a_i\}$ and $\{b_1, b_2, \ldots, b_j\}$. The recursion relation for $w_{i,j}$ is simply

$$w_{i,j} = \nu[w_{i-1,j} + w_{i,j-1}] + v(a_i, b_j)w_{i-1,j-1}. \tag{A.1}$$

To reveal the mathematical structure involved, it is convenient to introduce a new set of coordinates $(x = i - j, t = i + j)$ on the alignment lattice as shown in Fig. 7. The

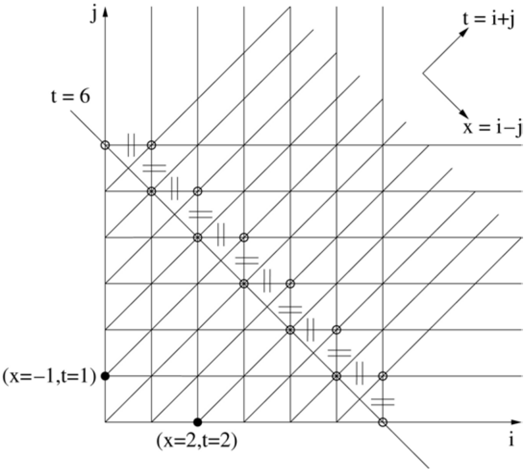

Fig. 7. The alignment lattice.

recursion (A.1) now reads

$$w(x, t+1) = \nu[w(x+1, t) + w(x-1, t)] + v(x, t)w(x, t-1) \qquad \text{(A.2)}$$

where $v(x, t-1) \equiv v(a_i, b_j)$ is introduced to reflect that the pairing (a_i, b_j) is located at $(x, t-1)$. The initial conditions for (A.2) are $w(x, t=0) = \delta_{x,0}$ and $w(x, t<0) = 0$. From this point on, the procedure to solve this problem is very similar to that of Friedberg and Yu (1994) and Yu (1999). Note that at a fixed location (i, j) the substitution weight $v(a_i, b_j)$, and consequently $v(x, t-1)$, changes when a new pair of sequences is considered.

We are now ready to write down the precise definition of $\widetilde{W}_t$:

$$\widetilde{W}_t = \sum_x w(x, t) + \sum_{x'} v(x', t)w(x', t-1). \qquad \text{(A.3)}$$

Note that in (A.3), if x is summed over even integers, then x' will be summed over odd integers, and vice versa. An explicit example is given in Fig. 7, where the open circles indicate the vertices whose weights are summed over at time $t = 6$. And the double slash on the bonds indicate that no weight flow through those bonds should be included. Basically, the quantity $\widetilde{W}_t$ sums all the global alignment weights arriving at the time slice t in the alignment lattice.

Let us first emphasize that $\langle v(x, t)w(x, t-1)\rangle_0 = \langle v(x, t)\rangle_0 \langle w(x, t-1)\rangle_0$. This is exact because $v(x, t) = v(a_{(t+1+x)/2}, b_{(t+1-x)/2})$ while the alignment weight at $w(x, t-1)$ depends only on subsequences $\{a_1, \ldots, a_{(t-1+x)/2}\}$ and $\{b_1, \ldots, b_{(t-1-x)/2}\}$. Denoting $\langle w(x, t)\rangle_0$ by $\phi(x, t)$, we may then write down easily the corresponding iterative equation

$$\phi(x, t+1) = v\phi(x, t-1) + \nu[\phi(x+1, t) + \phi(x-1, t)] \qquad \text{(A.4)}$$

whereas the quantity $\langle \widetilde{W}_t \rangle_0$ is obtained by

$$\langle \widetilde{W}_t \rangle_0 = \sum_x \phi(x,t) + v \sum_{x'} \phi(x', t-1). \tag{A.5}$$

Eq. (A.4) can be easily solved by going through a discrete Laplace transform and a discrete Fourier transform. Defining

$$\phi_z(x) \equiv \sum_{t=0}^{\infty} z^t \phi(x,t) \tag{A.6}$$

$$\phi_z^k \equiv \sum_x \mathrm{e}^{-\mathrm{i}kx} \phi_z(x) \tag{A.7}$$

we obtain

$$\phi_z(x) - \delta_{x,0} = z^2 v \phi_z(x) + z\nu[\phi_z(x+1) + \phi_z(x-1)]$$
$$\phi_z^k - 1 = [z^2 v + 2z\nu\cos(k)]\phi_z^k.$$

Apparently, the quantity of interest $\langle \widetilde{W}_t \rangle_0$ in (A.5) consists of only the zero-momentum mode. To be explicit, we may write

$$\langle \widetilde{W}_t \rangle_0 = \phi^{k=0}(t) + v\phi^{k=0}(t-1),$$

and consequently

$$\langle \widetilde{W}_t \rangle_0 = \oint \frac{\mathrm{d}z}{2\pi\mathrm{i}} \left[\frac{\phi_z^{k=0}}{z^{t+1}} + v \frac{\phi_z^{k=0}}{z^t} \right] = \oint \frac{\mathrm{d}z}{2\pi\mathrm{i}} \frac{1}{z^{t+1}} \frac{1+vz}{1-z^2v-2\nu z}. \tag{A.8}$$

When $\nu = (1-v)/2$, we may rewrite $vz^2 + 2\nu z - 1$ as $vz^2 + z - vz - 1 = (vz+1)(z-1)$. And therefore the contour integral becomes

$$\langle \widetilde{W}_t \rangle_0 = \oint \frac{\mathrm{d}z}{2\pi\mathrm{i}} \frac{1}{z^{t+1}} \frac{1}{1-z} = 1.$$

Appendix B

We now sketch the derivation of condition (42). Recall that the quantity Δ is defined as

$$\langle v(x,t)v(x',t')\rangle_0 - v^2 = \Delta\delta_{x,x'}\delta_{t,t'}. \tag{B.1}$$

Before computing $\widetilde{W}_t^2$, let us first define the following quantities:

$$\phi(x_1,x_2,t) \equiv \langle w(x_1,t)w(x_2,t)\rangle_0, \tag{B.2}$$

$$\phi^>(x_1,x_2,t) \equiv \langle w(x_1+1,t+1)w(x_2,t)\rangle_0, \tag{B.3}$$

$$\phi^<(x_1,x_2,t) \equiv \langle w(x_1,t)w(x_2+1,t+1)\rangle_0. \tag{B.4}$$

It has not escaped our attention that $\phi^>(x_1,x_2,t) = \phi^<(x_2,x_1,t)$, but we find it more convenient to have both $\phi^>$ and $\phi^<$ introduced. After some calculation, we can write $\langle \widetilde{W}_t^2 \rangle_0$ as

$$\langle \widetilde{W}_t^2 \rangle_0 = \sum_{x_1,x_2} \phi(x_1,x_2,t)$$

$$+ \sum_{x_1,x_2} [\Delta\delta_{x_1,x_2} + v^2]\phi(x_1, x_2, t-1)$$
$$+ v \sum_{x_1,x_2} [\phi^{>}(x_1, x_2, t-1) + \phi^{<}(x_1, x_2, t-1)].$$

Following the same route as in Friedberg and Yu (1994), Yu (1999) and similar to what we did for the first condition, we perform a discrete Laplace transform as well as Fourier transforms similar to Eqs. (A.6) and (A.7)

$$\phi_z(x_1, x_2) \equiv \sum_{t=0}^{\infty} z^t \phi(x_1, x_2, t) \tag{B.5}$$

$$\phi_z^k(y) \equiv \sum_{x_1,x_2} \mathrm{e}^{-\mathrm{i}k(x_1+x_2)} \delta_{x_1-x_2,2y} \phi_z(x_1, x_2) \tag{B.6}$$

$$\phi_z^{k,l} \equiv \sum_y \mathrm{e}^{-2\mathrm{i}ly} \phi_z^k(y). \tag{B.7}$$

The quantity of interest, $\langle \widetilde{W}_t^2 \rangle_0$, is now rewritten as follows:

$$\langle \widetilde{W}_t^2 \rangle_0 = \oint \frac{\mathrm{d}z}{2\pi \mathrm{i}} \left\{ \frac{1 + zv^2}{z^{t+1}} \phi_z^{k=0,l=0} + \Delta \frac{\phi_z^{k=0}(y=0)}{z^t} \right.$$
$$\left. + v \, \frac{\phi^{>k=0,l=0} + \phi_z^{<k=0,l=0}}{z^t} \right\}, \tag{B.8}$$

where $\phi_z^k(y = 0)$ is nothing but taking value $y = 0$ in Eq. (B.6).

The evolution equations, similar to Eq. (A.4), of ϕ, $\phi^{>}$ and $\phi^{<}$ can be derived using the fundamental relation Eq. (A.2). After some tedious calculations, one arrives at

$$\begin{aligned} \phi_z^{k,l} &= 1 + z^2\Delta\phi_z^k(y = 0) + [z^2 v^2 + 2z\nu^2(\cos 2k + \cos 2l)]\phi_z^{k,l} \\ &\quad + z^2 \nu v(\mathrm{e}^{-2\mathrm{i}k+2\mathrm{i}l} + 1)\phi_z^{<k,l} + z^2 \nu v(\mathrm{e}^{-2\mathrm{i}k-2\mathrm{i}l} + 1)\phi_z^{>k,l} \end{aligned} \tag{B.9}$$

$$\phi_z^{>k,l} = zv\mathrm{e}^{2\mathrm{i}l} \phi_z^{<k,l} + \nu(\mathrm{e}^{2\mathrm{i}k+2\mathrm{i}l} + 1)\phi_z^{k,l} \tag{B.10}$$

$$\phi_z^{<k,l} = zv\mathrm{e}^{-2\mathrm{i}l} \phi_z^{>k,l} + \nu(\mathrm{e}^{2\mathrm{i}k-2\mathrm{i}l} + 1)\phi_z^{k,l}. \tag{B.11}$$

From Eq. (B.8), we see that the calculation can be simplified by setting $k = 0$ first. To lighten the notation, we only retain the variable l. Thus, $\phi_z^{k=0,l}$ becomes ϕ^l and $\phi_z^k(y = 0)$ becomes $\phi(y = 0)$. After some algebra and calculus, we obtain

$$\phi(y = 0) = \frac{1}{(1 + zv)\sqrt{(1 - zv)^2 - 4z\nu^2 - \Delta z^2}}. \tag{B.12}$$

Eq. (B.12) can then be substituted into Eq. (B.9) to obtain a complete expression for $\phi_z^{k=0,l}$.

In order to ensure $\lim_{t\to\infty} \langle \widetilde{W}_t^2 \rangle_0 = \text{constant}$, we analyze the pole locations in (B.8) and obtain the following conditions: (1) $\phi(y = 0)$ must not have any z pole such that $|z| < 1$; (2) $\phi(y = 0)$ must have a pole at $z = 1$ under the fact that $2\nu + v < 1$. Therefore, for a given Δ, we need to make sure that poles with $|z| < 1$ do not occur. To accomplish this, let us first observe that if $z \to 0$, then the denominator in $\phi(y = 0)$ becomes $1 - 0$. Therefore, what we want is that $(1 + zv)\sqrt{(1 - zv)^2 - 4zv^2} \geq \Delta z^2$ over the range $0 < z < 1$ and with equality holds at $z = 1$. Careful analysis then yields the condition (42).

Appendix C

In this appendix, we will provide the heuristic derivation of the asymptotic behavior (33) of the cooling map. To achieve this goal, we need to analyze our conditions for solvability in more detail. To demonstrate this procedure, let us focus on our simple match–mismatch scoring scheme.

Under our simple scoring scheme, we have

$$\begin{aligned} f_s(\tau) = \sum_{a,b} \exp[s(a,b)/\tau] &= \frac{1}{c}\exp[\sqrt{c-1}/\tau] \\ &\quad + \frac{c-1}{c}\exp[-1/(\tau\sqrt{c-1})]. \end{aligned} \tag{C.1}$$

As $\tau \to 0$, $f_s(\tau)$ diverges as $\frac{1}{c}\exp[\sqrt{c-1}/\tau]$, and the quantity $\exp[-\gamma/\tau]$ becomes vanishingly small. Under the first solvability condition (38), we consider the low-temperature limit. Recall that both $\nu = \exp[-(\gamma + \sigma)/\tau]$ and $v \equiv \exp[2\sigma/\tau] f_s(\tau)$ are positive. As $\tau \to 0$, we must have $v \to 1$ since otherwise the condition (38) breaks down. This then implies, as $\tau \to 0$,

$$\sigma_1(\gamma, \tau) \approx \frac{\tau}{2} \ln\left[\frac{1}{f_s(\tau)}\right] = -\frac{\sqrt{c-1}}{2} + \frac{\tau}{2}\ln(c) + \mathcal{O}(\mathrm{e}^{-\mathrm{const}/\tau}). \tag{C.2}$$

That is, it is very close to a straight line with slope $\ln(c)/2$ and with a correction term proportional to $\exp[-\mathrm{const}/\tau]$, which vanishes exponentially fast when $\tau \to 0$.

In a similar fashion, we analyze the case for $\sigma_2(\gamma, \tau)$. Under the condition (42) and remembering that $2\nu + v < 1$ for the second solvable class, we conclude $v < 1$ and consequently $\exp[2\sigma_2(\gamma, \tau)/\tau] f_s(\tau) < 1$. If for $\tau \to 0$, $\exp[2\sigma_2(\gamma, \tau)/\tau] f_s(\tau)$ becomes zero, the left hand side of (42) approaches 1 while the right hand side of (42) reaches 0. Therefore, as $\tau \to 0$, $\exp[2\sigma_2(\gamma, \tau)/\tau] f_s(\tau)$ must approach a finite constant smaller than one but larger than zero. Let us call this limiting number α. Let us calculate the leading term of $f_s(\tau/2) - f_s^2(\tau)$ as $\tau \to 0$,

$$f_s(\tau/2) - f_s^2(\tau) = (c-1)\left[\frac{\exp[\sqrt{c-1}/\tau]}{c}\right]^2 [1 + \mathcal{O}(\mathrm{e}^{-\mathrm{const}/\tau})].$$

When the zero-temperature limit of $\exp[2\sigma_2(\gamma, \tau)/\tau] f_s(\tau)$ approaches a finite positive constant smaller than one, ν must approach zero, and we have from (42)

$$1 - \alpha^2 = (c-1)\alpha^2.$$

We therefore have $\alpha = \sqrt{1/c}$. Consequently, we have, as $\tau \to 0$,

$$\sigma_2(\gamma, \tau) \approx \frac{\tau}{2} \ln\left[\frac{1/\sqrt{c}}{f_s(\tau)}\right] = -\frac{\sqrt{c-1}}{2} + \frac{\tau}{4}\ln(c) + \mathcal{O}(\mathrm{e}^{-\mathrm{const}/\tau}). \tag{C.3}$$

Now let us make the plausible assumption that the phase transition line, see Fig. 2, stays horizontal over a certain range of low temperature. Then it is easy to see how the $\tau/(2\tilde{\tau})$ emerges in the cooling map. Our definition of the cooling map requires

$$-\frac{\sqrt{c-1}}{2}+\frac{R(\tau)}{2}\ln(c)+\mathcal{O}(\mathrm{e}^{-\mathrm{const}/\tau})+[\sigma_c(\gamma,R(\tau))-\sigma_1(\gamma,R(\tau))]$$
$$=-\frac{\sqrt{c-1}}{2}+\frac{\tau}{4}\ln(c)+\mathcal{O}(\mathrm{e}^{-\mathrm{const}/\tau})+[\sigma_c(\gamma,\tau)-\sigma_2(\gamma,\tau)] \qquad \text{(C.4)}$$

Because our map moves with $[\sigma_c(\gamma, R(\tau)) - \sigma_1(\gamma, R(\tau))] = [\sigma_c(\gamma, \tau) - \sigma_2(\gamma, \tau)]$, we obtain that

$$R(\tau)=\frac{\tau}{2}+\mathrm{Const}\ \mathrm{e}^{-\mathrm{const}/\tau}. \qquad \text{(C.5)}$$

Let us further argue that the deviation of the phase transition line from a horizontal line is of order $\exp[-\mathrm{const}/\tau]$ at low temperature. This is simply because at zero temperature, the partition function has contributions only from lowest energy paths, while at low but finite temperatures, suboptimal paths contribute as well. When one sums the Boltzmann weights from the lowest energy path and from the leading suboptimal paths, we have a free energy expression given by the lowest energy plus a correction term of order $\mathcal{O}(\exp[-\mathrm{const}/\tau])$. This increase in the average free energy indicates that one only needs to lower the average score gain per unit length by about the same amount to keep the system in the log phase. That is to say, as $\tau \to 0$, we have

$$\sigma_c(\gamma,\tau)\approx\sigma_c(\gamma,\tau=0)-C_1\mathrm{e}^{-C_2/\tau}.$$

With this extra correction considered, it will add an exponential correction term to the right hand side of (C.4). However, it will still lead to the same form (C.5) for the low-temperature part. This concludes our heuristic derivation of (33).

List of symbols

Symbol	Description	Section
λ, κ	Gumbel parameters	1
c	Number of letters in the sequences (e.g., $c = 4$ for nucleotides)	2.1
χ	Sequence alphabet of c letters	2.1
$\mathbf{a}, \mathbf{b}$	Sequences of nucleotides or amino acids	2.1
$\mathbf{A}$	Alignment (path)	2.1
L	Length of an alignment	2.1
$s(a, b)$	Score of pairing two letters $a, b \in \chi$	2.2
s_+, s_-, s_g	Match, mismatch and gap score	2.2
σ	Average score per paired element	2.2
γ	Gap parameter, $\gamma = s_g - \sigma$	2.2
τ	Temperature	2.2

Symbol	Description	Section
$S_{i,j}$	Alignment score for subsequences up to position i, j	2.3
$Z_{i,j}$	Partition function for subsequences up to position i, j	2.3
$F_{i,j}$	Free energy, $F_{i,j} = \tau \ln Z_{i,j}$	2.3
$S_{\max}$	Maximum score, $S_{\max} = \max_{i,j} S_{i,j}$	2.3
$F_{\max}$	Maximum score, $F_{\max} = \max_{i,j} F_{i,j}$	2.3
Z	Total partition function, $Z = \sum_{i,j} Z_{i,j}$	2.3
$v(a, b), \nu$	Weight for pairing letters a, b and gap weight	2.3
$Q_0[\mathbf{a}, \mathbf{b}]$	Joint probability of random sequences **a**, **b**	2.4
$Q[\mathbf{a}, \mathbf{b}]$	Joint probability of the sequence pair **a**, **b**	2.4
$\sigma_1(\gamma, \tau), \sigma_2(\gamma, \tau)$	The two solvable cases	2.4, 3, 4.2
$\langle\cdot\rangle_0$	Average with respect to $Q_0[\mathbf{a}, \mathbf{b}]$	2.5
F_0	Characteristic limit of free energy (local regime)	2.5
$b(\gamma, \tau)\Phi(\sigma - \sigma_c, \gamma)$	Functional form of λ	3, 4.1
R	Cooling map	3.1
ρ	Shift in σ	3.2
$\rho^{(n)}$	Iteration of ρ under the cooling map R	3.2
$\tau^{(n)}$	Iteration of τ under the cooling map R	3.2
$B^{(n)}, B^{(\infty)}$	Iteration of the prefactor b under R	3.2
$\tilde{\tau}(\gamma)$	Characteristic scale, maximum of $\sigma_1(\gamma, \tau)$ w.r.t τ	3.3
$\widetilde{W}_t$	Total global alignment weight arriving at the grid border	Appendix A, 4.2
y	$y = \exp(2\sigma/\tau)$	4.2
$f_s(\tau)$	$f_s(\tau) = \sum_{a,b} \exp(s(a, b)/\tau)$	4.2
$f_\gamma(\tau)$	$f_\gamma(\tau) = \exp(-\gamma/\tau)$	4.2
Δ	Variance of substitution weight	4.2
x, t	Rotated coordinates in the alignment grid, $x = i - j, t = i + j$	Appendix A
$w(x, t)$	Global alignment weight	Appendix A
$\phi(x, t)$	$\phi(x, t) = \langle w(x, t)\rangle_0$	Appendix A
$\phi_z(x), \phi_z^k$	Laplace and Laplace–Fourier transform of $\phi(x, t)$	Appendix A

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Arratia, R., Waterman, M.S., 1994. A phase transition for the score in matching random sequences allowing deletions. Ann. Appl. Probab. 4, 200–225.

Bundschuh, R., 2002. Asymmetric exclusion process and extremal statistics of random sequences. Phys. Rev. E 65, 031911.

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978. A model of evolutionary change in proteins. In: Dayhoff, M.O., Eck, R.V. (Eds.), Atlas of Protein Sequence and Structure 5 supp. Natl. Biomed. Res. Found, vol. 3. pp. 345–358.

Drasdo, D., Hwa, T., Lässig, M. 1998. A scaling theory of sequence alignment with gaps. In: Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology. ISMB98, pp. 52–58.

Drasdo, D., Hwa, T., Lässig, M., 2000. Scaling laws and similarity detection in sequence alignment with gaps. J. Comput. Biol. 7, 115–141.

Durbin, R., Eddy, S., Krogh, A., Mitchinson, G., 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, U.K.

Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics 14, 755–763.

Friedberg, R., Yu, Y.-K., 1994. Directed waves in random media: an analytical calculation. Phys. Rev. E. 49, 5755–5762.

Gumbel, E.J., 1958. Statistics of Extremes. Columbia University Press, New York, NY.

Halpin-Healy, T., Zhang, Y.C., 1995. Kinetic roughening phenomena, stochastic growth, directed polymers and all that. Phys. Rep. 254, 215–414.

Hwa, T., Lässig, M., 1998. Optimal detection of sequence similarity by local alignment. In: RECOMB98. pp. 109–116.

Karlin, S., Altschul, S.F., 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA 87, 2264–2268.

Karlin, S., Dembo, A., 1992. Limit distributions of maximal segmental score among Markov-dependent partial sums. Adv. Appl. Probab. 24, 13–140.

Karplus, K., Barrett, C., Hughey, R., 1998. Hidden Markov models for detecting remote protein homologies. Bioinformatics 14, 846–856.

Kschischo, M., Lässig, M., 2000. Finite-temperature sequence alignment. Pac. Symp. Biocomput. 5, 621–632.

Metzler, D., 2002. A Poisson model for gapped local alignments. Stat. Prob. Lett. 60, 91–100.

Miyazawa, S., 1996. A reliable sequence alignment method based on probabilities of residue correspondences. Protein Eng. 8, 999–1009.

Mott, R., Tribe, R., 1999. Approximate statistics of gapped alignment. J. Comput. Biol. 6, 91–112.

Olsen, R., Bundschuh, R., Hwa, T., 1999. Rapid assessment of extremal statistics for gapped local alignment. In: Lengauer, T. et al. (Eds.), Proceedings of The Seventh International Conference on Intelligent Systems for Molecular Biology. ISMB99, AAAI Press, Menlo Park, pp. 211–222.

Pearson, W.R., 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85, 2444–2448.

Siegmund, D., Yakir, B., 2000. Approximate p-value for local sequence alignments. Ann. Statist. 28, 657–680.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.

Spang, R., Vingron, M., 2000. Limits of Homology detection by pairwise sequence comparison. Bioinformatics 17, 338–342.

Yu, Y.-K., 1999. Calculation of wave center deflection and multifractal analysis of directed waves through the study of su(1,1) ferromagnets. In: Batchelor, M.T., Wille, L.T. (Eds.), Statistical Physics on the Eve of the 21st Century. World Scientific, NJ.

Yu, Y.-K., 2004. Replica model for an unusual directed polymer in $1+1$ dimensions and prediction of the extremal parameter of gapped sequence alignment statistics. Phys. Rev. E. 69, 061904.

Yu, Y.-K., Bundschuh, R., Hwa, T., 2002. Hybrid alignment: high performance with universal statistics. Bioinformatics 18, 864–872.

Yu, Y.-K., Hwa, T., 2001. Statistical significance of probabilistic sequence alignment and related local hidden Markov models. J. Comput. Biol. 8, 249–282.

Zhang, M.Q., Marr, T.G., 1995. Alignment of molecular sequences seen as random path analysis. J. Theor. Biol. 174, 119–129.