Finite-temperature Sequence Alignment

MAIK KSCHISCHO(1) and MICHAEL LÄSSIG(2)

(1) *Max-Planck Institut für Kolloid- und Grenzflächenforschung*
14424 Potsdam, Germany
(2) *Institut für theoretische Physik, Universität zu Köln*
Zülpicher Str. 77, 50937 Köln, Germany

We develop a statistical theory of probabilistic sequence alignments derived from a 'thermodynamic' partition function at finite temperature. Such alignments are a generalization of those obtained from information-theoretic approaches. Finite-temperature statistics can be used to characterize the significance of an alignment and the reliability of its single element pairs.

1 Introduction

The standard algorithms of Needleman-Wunsch [1] and of Smith-Waterman [2] align sequences by maximizing a score function that favors matching element pairs over mismatches and gaps. Maximum-score alignments are an accurate measure of similarity for closely related sequences. With increasing evolutionary distance, however, they tend to become sensitive to the choice of scoring parameters and therefore less reliable. One is thus lead to ask: How 'likely' is the maximum-score alignment compared to its alternatives?

In this paper, we develop the theory of probabilistic alignments derived from a thermodynamic partition function. These are called *finite-temperature* alignments. The theory is a generalization of the scaling approach to maximum-score alignments discussed in a number of recent publications [3,4,5,6,7]. The partition function formalism has also been used in other recent work [8,9]. Finite-temperature alignments have two key applications: (i) estimating the reliability of the single element pairs in an alignment and (ii) assessing the relative significance of different high-score local alignments. The latter are found efficiently using a new algorithm called the *ridge path* algorithm.

A probabilistic notion of alignment is also inherent to information-theoretic approaches. (The conceptual relationship between Bayesian statistics and statistical mechanics has recently been discussed in the related context of protein potentials [10].) The known maximum-likelihood [11,12,13] and Bayesian minimum message length alignments [14,15] can be regarded as special cases of finite-temperature alignments, which are obtained by minimization of a suitable entropy function. It is shown, however, that minimum-entropy alignments do not have maximal accuracy (in a sense to be defined below). This caveat to-

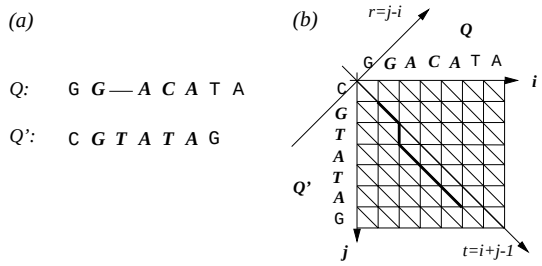

Figure 1: (a) One possible local alignment of two sequences Q and Q' with elements taken from a 4-letter alphabet and numbered by i and j, respectively. The aligned subsequences are shown in boldface, with 4 pairings (three matches, one mismatch) and one gap. (b) Unique representation of this alignment as directed path $\mathcal{A}$ (thick line) on an alignment grid with coordinates $t = i + j - 1$ and $r = j - i$. The vertices of this graph are labeled by even values of $t + r$. The diagonal bonds represent site pairs (i, j) and are labeled by odd values of $t + r$. The alignment length is $L = 9$.

wards the application of Bayesian statistics is expected to be important in a wider context of related problems.

2 The definition of finite-temperature alignments

A local alignment of two sequences $Q = \{Q_i\}$ $(i = 1, \ldots, N)$ and $Q' = \{Q'_j\}$ $(j = 1, \ldots, N')$ is defined as an ordered set of pairings (i, j) and of gaps $(i, -)$ and $(-, j)$ involving the elements of two contiguous subsequences $\{Q_{i_1}, \ldots, Q_{i_2}\}$ and $\{Q'_{j_1}, \ldots, Q'_{j_2}\}$; see Fig. 1(a). Its length is defined as the total number of aligned elements, $L \equiv i_2 - i_1 + j_2 - j_1 < N + N'$. An alignment can be uniquely represented as a *directed path* $\mathcal{A}$ on the two-dimensional grid of Fig. 1(b) [1]. Using the coordinates $r \equiv i - j$ and $t \equiv i + j - 1$, this path is the graph of a single-valued function $r(t)$ for $t = i_1 + i_2, i_1 + i_2 + 1, \ldots, i_2 + j_2$. An alignment $\mathcal{A}$ with N_+ matches $(Q_i = Q'_j)$, N_- mismatches $(Q_i \neq Q'_j)$, and N_g gaps has length $L = N_+ + N_- + 2N_g$. The simplest scoring functions are linear in N_+, N_-, and N_g. Any such function can be written in the normal form [7,5]

$$S(\mathcal{A}) = \sigma L + \sqrt{c-1}\, N_+ - \frac{1}{\sqrt{c-1}}\, N_- - \gamma N_g \qquad (1)$$

with two adjustable parameters, the gap cost γ and the score gain per aligned element, σ. (The prefactors of N_+ and N_- are such that for two letters Q_i and Q'_j chosen randomly from a c-letter alphabet, the pairing (Q_i, Q'_j) has score average 2σ and variance 1.)

A finite-temperature alignment is a probability distribution

$$P(\mathcal{A}) = \frac{1}{Z} \exp[S(\mathcal{A})/\tau] \; . \tag{2}$$

over *all* alignment paths $\mathcal{A}$. The form of Eq. (2) is fixed by the requirement that the probability of disjoint pieces of an alignment should be multiplicative while the score is additive. The normalization factor $Z = \sum_{\mathcal{A}} \exp[S(\mathcal{A})/\tau]$ is called the *partition function* of alignment [8,9,3], and $F \equiv \tau \log Z$ is the *free energy*. A finite-temperature alignment thus has 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. (In the language of statistical mechanics, $P(\mathcal{A})$ defines a Gibbs ensemble at temperature τ for directed paths $\mathcal{A}$ with line tension γ and chemical potential σ.)

A finite-temperature alignment contains *any* element pair (Q_i, Q'_j) with a finite probability $\rho(r = j - i, t = i + j - 1)$. This is determined by the sum $Z(r,t) \equiv \sum_{\mathcal{A}:(r,t)\in\mathcal{A}} \exp[S(\mathcal{A})/\tau]$ over all paths passing through the point (r,t). The normalization of the alignment probabilities is somewhat arbitrary since we are interested only in the relative importance of two different site pairs (r,t) and (r',t'), which is given by the ratio $\rho(r,t)/\rho(r',t')$. The local free energy $F(r,t) \equiv \tau \log Z(r,t)$ can be computed by a simple generalization of the Smith-Waterman dynamic programming algorithm; see Appendix A. The limit value $S(r,t) \equiv \lim_{\tau\to 0} F(r,t)$ is the maximum score of any path containing the point (r,t). Finite-temperature alignment thus reduces to the usual Smith-Waterman alignment for $\tau \to 0$; see also Appendix A.

3 The statistics of finite-temperature alignments

Alignment statistics describes averages (denoted by overbars) over an ensemble of sequence pairs (Q, Q') with well-defined mutual correlations. This ensemble should not be confused with the Gibbs ensemble $P(\mathcal{A})$ defining a finite-temperature alignment for a *given* sequence pair. The finite-temperature statistics of this paper involves a double average over sequence pairs and alignment paths. We have performed extensive numerical work for various sequence ensembles and alignment parameters. Here we summarize our main findings; details can be found elsewhere [16].

Consider first finite-temperature alignments for pairs of Markov sequences Q and Q' without mutual correlations (i.e, each letter Q_i and Q'_j is drawn independently from a c-letter alphabet). The properties of these alignments are determined by which paths contribute most to the partition function Z, or equivalently, to the local sums $Z(r,t)$. Due to the 'chemical potential' term

σL in the scoring function (1), we expect longer paths dominate over shorter ones for large σ, but are exponentially suppressed for sufficiently small σ. This is indeed the case. For pairs of sequences of equal length $N \to \infty$, we find the asymptotic behavior of the average local free energy

$$\begin{array}{lcll} \overline{F}(r,t) & \simeq & [\sigma - \sigma_c(\gamma,\tau)] \cdot 2N & \text{for } \sigma > \sigma_c(\gamma,\tau) \\ \overline{F}(r,t) & \to & F_0(\gamma,\sigma,\tau) & \text{for } \sigma < \sigma_c(\gamma,\tau) \end{array} \tag{3}$$

with a parameter-dependent threshold value $\sigma_c(\gamma,\tau) < 0$. This determines the average length of alignment paths, $\overline{L} = \partial\overline{F}/\partial\sigma$. In the *global* alignment regime ($\sigma > \sigma_c(\gamma,\tau)$), the entire sequences are aligned, i.e., $\overline{L} \simeq 2N$. In the *local* alignment regime ($\sigma < \sigma_c(\gamma,\tau)$), $\overline{L}$ reaches a finite limit $L_0(\gamma,\sigma,\tau) \equiv \partial F_0(\gamma,\sigma,\tau)/\partial\sigma$. There is a continuous phase transition between the two regimes; that is, F_0 and L_0 diverge as σ approaches the threshold value σ_c from below. The relations (3) have a finite zero-temperature limit describing the well-known phases of maximum-score Smith-Waterman alignments [2,4,5,6]. The local alignment regime has the characteristic score $S_0(\gamma,\sigma) \equiv \lim_{\tau\to 0} F_0(\gamma,\sigma,\tau)$.

Figs. 2(a,b) show an example of the free energy 'landscape' $F(r,t)$ and of the corresponding alignment probabilities $\rho(r,t) \sim \exp[F(r,t)/\tau]$ in the local alignment regime at finite τ. Both $F(r,t)$ and $\rho(r,t)$ are seen to be small (of order F_0) at many points of the alignment grid, with disconnected *islands* of larger values forming around locally significant paths. Free-energy islands persist for $\tau = 0$. In this limit, they can be defined by the condition $S(r,t) > 0$, and their statistics has recently been characterized in detail [17].

We now turn to the detection of sequence similarity by finite-temperature alignments. Consider pairs of Markov sequences Q, Q' with mutually correlated subsequences $\hat{Q}$ and $\hat{Q}'$ of approximately equal length $\hat{N} = \hat{N}'$; the remainder of Q and Q' has no correlations. The 'daughter' sequence $\hat{Q}'$ is obtained from its 'ancestor' $\hat{Q}$ by a simple Markov evolution process [7] with substitution probability p and insertion/deletion probability $\widetilde{p}$. A particular outcome of this process is uniquely represented by the sequences Q and Q' together with a specific path [3] on the alignment grid called the *evolution path* $\mathcal{E}$. This path is of length $L_{\mathcal{E}} = \hat{N} + \hat{N}'$. The ancestor elements Q_i that are neither deleted nor substituted define, together with the corresponding daughter elements Q'_j, the *conserved pairs* (Q_i, Q'_j). These are contained in the path $\mathcal{E}$.

In Figs. 2(c,d), we show a finite-temperature alignment for this case. The free energy landscape develops a larger island around the evolution path $\mathcal{E}$. Consequently, the alignment probability $\rho(r,t)$ is now concentrated around that path, and so are the points of maximal local free energy, $F(r,t) = F_{\max}$. Clearly, this 'correlation island' can only be detected if the alignment parameters are chosen in the local alignment regime such that the randomly generated

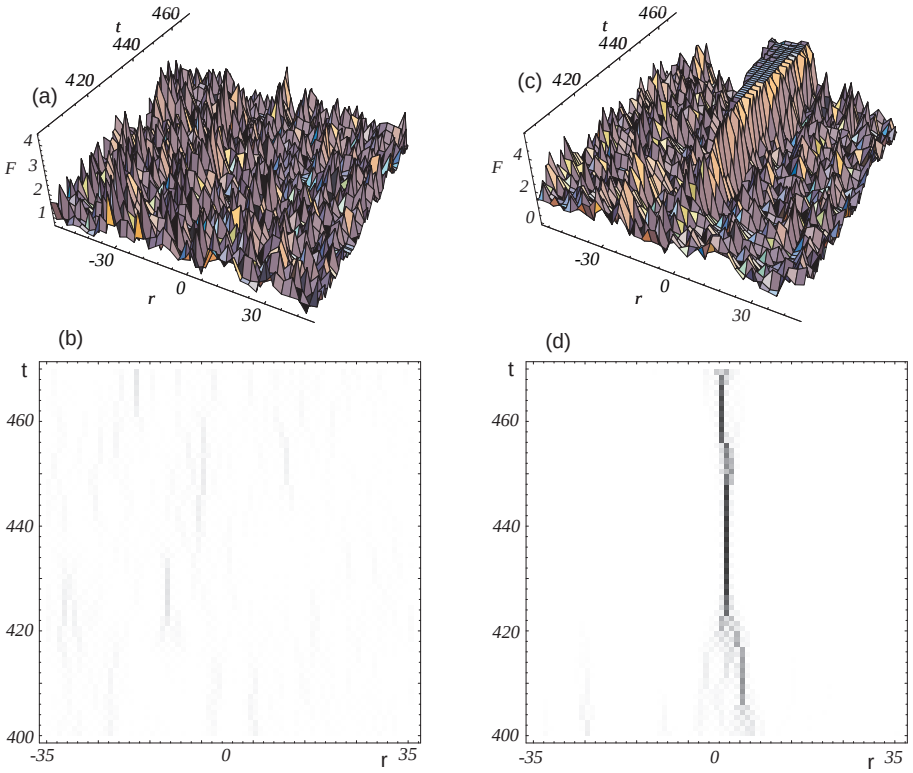

Figure 2: (a,b) The free energy landscape $F(r,t)$ and the alignment probability $\rho(r,t)$ for a pair of mutually uncorrelated sequences in the local alignment regime, shown only in a part of the alignment grid. Typical values of $F(r,t)$ are of order F_0. (c,d) The same for a pair of sequences with mutual correlations. The leading score 'island' lies around the evolution path and has a free energy maximum $F_{\max} \gg F_0$.

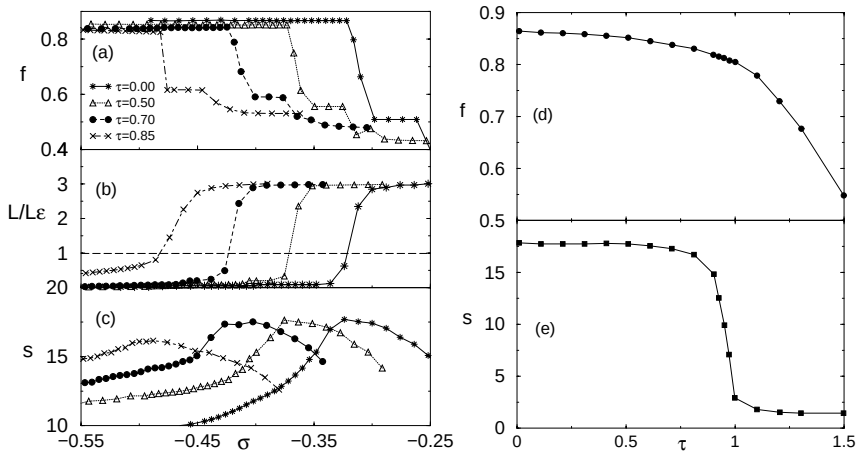

Figure 3: Parameter dependence of the alignment accuracy for a pair of mutually correlated sequences with evolution parameters $p = 0.1, \widetilde{p} = 0.1$ and evolution paths of length $L_{\mathcal{E}} = 950$. (a) The fidelity f, (b) the relative length $L/L_{\mathcal{E}}$, and (c) the average significance ratio s as a function of σ for various values of τ and $\gamma = \gamma^*(\tau)$. (d,e) The relative maxima $f^*(\tau)$ and $s^*(\tau)$.

islands are significantly smaller. This is measured by the *significance ratio*

$$s \equiv F_{\max}/F_0 \ . \tag{4}$$

How can the accuracy of such alignments be quantified and optimized? For a given element pair ($Q_i = Q'_j$), we can ask whether it is a conserved pair or a random match. For the alignment distribution P, we define the weighed *fidelity* $f \equiv \sum_c \rho(r,t) / \sum_m \rho(r,t)$, where $\sum_m$ runs over all matches (i,j) of the alignment grid and $\sum_c$ over the conserved matches only. This definition is statistically equivalent to other fidelity measures used previously for maximum-score alignments [3,6].

The alignment data have a strong parameter dependence. In Figs. 3(a,b), this is exemplified for the dependence of the fidelity f and the length L on σ at fixed γ and τ. For small σ, f is at its relative maximum and $L < L_{\mathcal{E}}$; typical alignment paths are close to the evolution path but are too short. For larger σ, f is small and $L > L_{\mathcal{E}}$, indicating that the paths are too long. There is a unique intermediate point where f is still maximal and $L = L_{\mathcal{E}}$, i.e., where the alignment and the the evolution path match best. This point is very close to the point where the significance ratio s reaches its relative maximum; see Fig. 3(c). The same is true for the relative maxima $f^*(\tau)$ and

$s^*(\tau)$ optimized over both σ and γ. This defines optimal parameter values $\gamma^*(\tau)$ and $\sigma^*(\tau)$ (which, of course, also depend on the sequence characteristics given here by p and $\widetilde{p}$). $f^*(\tau)$ and $s^*(\tau)$ are always found to be decreasing functions of τ as shown in Figs. 3(d,e). Hence, the global maxima are attained at parameter values $\tau = 0$, $\gamma^* \equiv \gamma^*(\tau = 0)$, $\sigma^* \equiv \sigma^*(\tau = 0)$. We draw two important conclusions: (a) *Alignments can be optimized by maximization of the significance ratio s.* (b) *The optimal alignment is always a zero-temperature (maximum-score) alignment.* For $\tau = 0$, the significance ratio becomes $s = S_{\max}/S_0$. The zero-temperature optimization has been discussed in detail in a previous publication [6]. The scaling theory of alignment explains theoretically why the alignment accuracy and s have a common parameter dependence.

4 Minimum-entropy alignments and information theory

In this section, we compare the significance optimization of alignments with alternative approaches grounded upon Bayes' principle [14,15]. For simplicity, we limit ourselves to sequences with mutual correlations over the entire length (i.e., $\hat{Q} = Q$ and $\hat{Q}' = Q'$) and $N = N' \gg 1$. We can then choose global alignments with an arbitrary value of $\sigma > \sigma_c(\gamma, \tau)$ (e.g., $\sigma = 0$) and have to optimize only γ and τ. In the framework of Bayesian statistics, the Markov evolution process is a probabilistic machine producing sequence pairs (Q, Q') as data with a frequency distribution $\mathcal{Z}_e(Q, Q')$. This distribution is characterized by the evolution parameters, here p and $\widetilde{p}$. Finite-temperature alignments with different γ, τ are regarded as hypotheses about this process. These are gauged by their *message length*, which is defined as (minus) the joint log-probability of data (Q, Q') and hypothesis (γ, τ). For a parameter-independent 'prior' probability of the hypothesis, the relevant part of the message length is (minus) the log-probability of the data under a given hypothesis. This probability (or likelihood) is given by a suitably normalized alignment partition function $\mathcal{Z}_{\gamma,\tau}(Q, Q') \equiv Z_{\gamma,\tau}(Q, Q')/\mathcal{N}$ (details can be found elsewhere [16]).

The Bayesian statistical analysis results in a 'posterior' distribution over the hypotheses, in this case, over γ and τ. In particular, we can define 'best' parameters (γ_e, τ_e) from a 'minimum entropy' principle for the *ensemble average* $\overline{\log \mathcal{Z}_{\gamma,\tau}} = \sum_{Q,Q'} \mathcal{Z}_e(Q, Q') \log \mathcal{Z}_{\gamma,\tau}(Q, Q')$. Indeed, the difference

$$-\overline{\log \mathcal{Z}_{\gamma,\tau}} + \overline{\log \mathcal{Z}_e} = \sum_{Q,Q'} \mathcal{Z}_e(Q, Q') \log \left(\frac{\mathcal{Z}_e(Q, Q')}{\mathcal{Z}_{\gamma,\tau}(Q, Q')} \right) \tag{5}$$

is just the *relative entropy* or Kullback-Leibler divergence of the probability distributions $\mathcal{Z}_{\gamma,\tau}$ and $\mathcal{Z}_e$. It reaches its minimum 0 if and only if the two

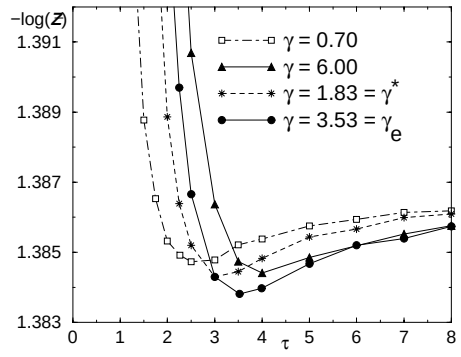

Figure 4: Parameter dependence of the likelihood $\mathcal{Z}_{\gamma,\tau}(Q,Q')$ for a given pair of long sequences. The sequences are generated by a Markov evolution process with parameters $p = 0.80$ and $\widetilde{p} = 0.072$ and alphabet size $c = 4$. The function $-\log \mathcal{Z}_{\gamma,\tau}$ has its unique minimum at the point $(\gamma_e = 3.5, \tau_e = 3.6)$ given by (6).

distributions are equal: $\mathcal{Z}_{\gamma_e,\tau_e}(Q,Q') = \mathcal{Z}_e(Q,Q')$ for all Q, Q'. This fixes the local weights $\exp(s_\pm/\tau_e)$ and $\exp(-\gamma_e/\tau_e)$ of matches, mismatches, and gaps in terms of the substitution and insertion/deletion frequencies, i.e., in terms of the evolution parameters $p, \tilde{p}$ (cf. Appendix A). The relations can be written in the form

$$\begin{aligned} \gamma_e &= \tau_e \log\Big[\frac{1-2\widetilde{p}}{\widetilde{p}}\big(1-p+\frac{p}{c}\big)\Big] - \sqrt{c-1} \\ \tau_e &= \frac{c}{\sqrt{c-1}}\left[\log\Big[\frac{c(1-p)}{p}+1\Big]\right]^{-1}. \end{aligned} \tag{6}$$

Minimization of the message length can, in principle, be used to infer a priori unknown evolution parameters from a *single* sequence pair (Q,Q') (see also the discussion in previous papers [11,12,13]). The reason is that $\log \mathcal{Z}_{\gamma,\tau}(Q,Q')$ is *self-averaging*, i.e., it converges to the ensemble average $\overline{\log \mathcal{Z}_{\gamma,\tau}}$ with probability 1 in the limit of long sequences. This is illustrated in Fig. 4, where $-\log \mathcal{Z}_{\gamma,\tau}(Q,Q')$ is shown for a given pair of sequences generated with evolution parameters $p, \widetilde{p}$. There is indeed a unique minimum at the point (γ_e, τ_e) given by (6). The procedure can be extended to infer also the type of mutations [14,15].

Bayesian statistics thus focuses on the evolution process rather than on the conserved element pairs. The posterior distribution over alignment parameters, however, does not reproduce the point $(\gamma^* < \gamma_e, \tau = 0)$ of maximal fidelity and significance ratio. In particular, the minimum message length alignment is not the most accurate one. *Reconstructing the evolution characteristics is not equivalent to finding sequence similarities.* Indeed, the two extremization principles are quite different. Maximizing $\log \mathcal{Z} \sim F/\tau$ fixes the *local* Boltzmann

factors, while maximizing $s = F/F_0$ involves the *nonlocal* statistics of the random free energy islands contained in F_0. How the latter can be incorporated into the framework of information theory is still an open problem.

5 Local reliability and ridge scores

While finite-temperature alignments do not improve the overall accuracy of the optimal maximum-score alignment $\mathcal{A}^*$, they are very useful in quantifying the reliability of its site pairs, as we discuss qualitatively in the sequel. A number of different approaches to this problem are discussed in the literature [18,19,20,21].

In an alignment with parameters γ^*, σ^*, and $\tau > 0$, the dominant paths are $\mathcal{A}^*$ and subleading paths with small random deviations from $\mathcal{A}^*$. However, since $\mathcal{A}^*$ itself has small random deviations from the evolution path $\mathcal{E}$, the leading suboptimal paths are nearly as good approximations to $\mathcal{E}$. The reliability of a site pair $(r,t) \in \mathcal{A}^*$ depends on the number of such alternative paths passing through different points (r',t). This in turn is related to the alignment probability at some temperature τ_c, the characteristic scale for the decrease of $f^*(\tau)$ and $s^*(\tau)$ to significantly below the zero-temperature values. We find $\tau_c \approx S_0(\gamma^*, \sigma^*)$ in accordance with scaling theory [16].

As an example of such a reliability estimate, consider two correlated sequences $\hat{Q}$ and $\hat{Q}'$ with $\hat{Q}'$ containing a repeat (i.e., a subsequence and an adjacent copy of it) of length n. The conserved element pairs then fall in two disconnected groups as shown in Fig. 5(a). For n not too large compared to the length of the groups, the optimal path $\mathcal{A}^*$ interpolates between both groups and thus contains spurious matches (Fig. 5(b)). The finite-temperature alignment at $\tau = \tau_c$ clearly exhibits this region of unreliable site pairs; see Fig. 5(c).

In other cases, the optimal path $\mathcal{A}^*$ may contain only one group of the conserved pairs, while the optimal path $\mathcal{A}'$ covering the other group has a much smaller score; see the example of Figs. 5(d,e). The second and possible further groups are then usually found by *declumping* [22,23]. This procedure involves partially rerunning the dynamic programming algorithm for each subleading alignment to be found. Here we proceed differently, noting that both $\mathcal{A}^*$ and $\mathcal{A}'$ satisfy the *ridge condition* $S(r,t) = \max(S(r-1,t), S(r,t), S(r+1,t))$ for all of their points (r,t). More generally, we define for arbitrary alignent paths $\mathcal{A}$ the *ridge score*

$$R(\mathcal{A}) \equiv \begin{cases} S(\mathcal{A}) & \text{if } S(r,t) = \max(S(r-1,t), S(r,t), S(r+1,t)) \\ & \quad \text{for all } (r,t) \in \mathcal{A} \text{ with } t+r \text{ even} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The local score maxima $R(r,t) \equiv \max_{\mathcal{A}:(r,t)\in\mathcal{A}} R(\mathcal{A})$ can be computed by a

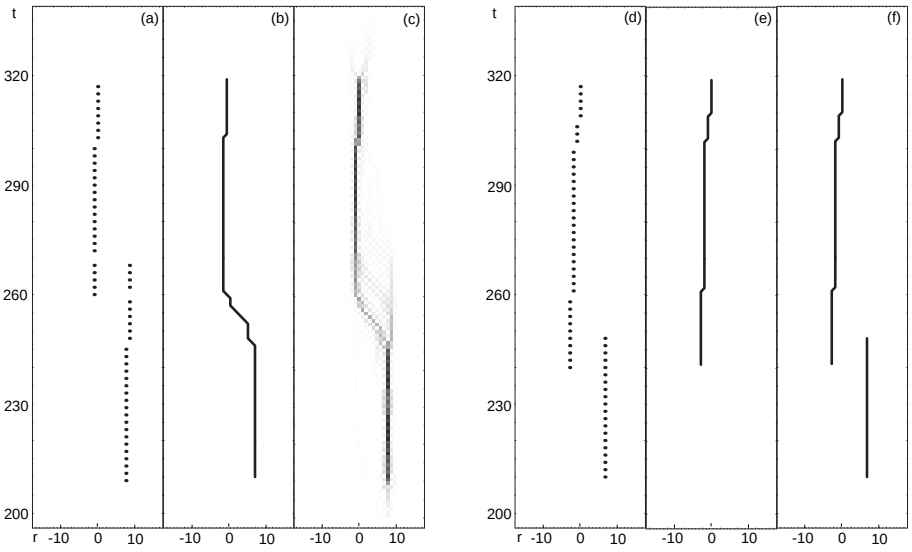

Figure 5: Local reliability and ridge paths for a pair of mutually correlated sequences $\hat{Q}$ and $\hat{Q}'$ with a repeat of length $n = 10$ in $\hat{Q}'$. (a) The positions of the conserved element pairs (Q_i, Q'_j) fall into two disconnected groups. The repeat elements are located at the lower end of the left group and at the upper end of the right group. (b) The optimal alignment path $\mathcal{A}^*$ interpolates between both groups. (c) The finite-temperature alignment with parameters $\tau = 1.0 \approx \tau_0$, $\gamma = 1.0 \approx \gamma^*(\tau)$, $\sigma = -0.44 \approx \sigma^*(\tau)$ exhibits the unreliable site pairs between the two groups of conserved elements. (d) A case similar to (a). (e) Here the optimal alignment $\mathcal{A}^*$ covers only one group of the conserved pairs. (f) The ridge path algorithm finds both groups.

modified dynamic programming algorithm in a single run, as described in Appendix B. We find they give directly all significant disjoint paths [16]. In the example shown, both $\mathcal{A}^*$ and $\mathcal{A}'$ are found; see Fig. 5(f).

We conclude that finite-temperature alignments together with the ridge path algorithm are useful methods to identify all significant local alignments of given sequences. These in turn are the building blocks to construct statistically well-defined longer or global alignments.

Acknowledgments

The authors have benefited from discussions with D. Drasdo, R. Durbin, T. Hwa, and M. Waterman.

Appendix A: The finite-temperature algorithm

The local free energy $F(r,t)$ can be decomposed as

$$F(r,t) = \begin{cases} F'(r,t) + F''(r,t) & \text{if } t+r \text{ even} \\ F'(r,t-1) + F''(r,t+1) + 2\sigma + s(r,t) & \text{if } t+r \text{ odd} \end{cases} . \quad (8)$$

Even values of $t+r$ belong to vertices of the alignment grid and odd values to site pairs (i,j). The restricted free energies $F'(r,t)$ and $F''(r,t)$ are defined at vertices. $F'(r,t) = \log Z'(r,t)$ is given by the partition function $Z'(r,t)$ of all paths with initial point at some $t' \le t$ and endpoint (r,t), and $F''(r,t)$ refers to the paths with initial point (r,t) and endpoint at some $t'' \ge t$. The match/mismatch scores

$$s(r,t) = \begin{cases} s_+ \equiv \sqrt{c-1} & \text{if} \quad Q_{(r+t+1)/2} = Q'_{(t-r)/2} \\ s_- \equiv -1/\sqrt{c-1} & \text{if} \quad Q_{(r+t+1)/2} \neq Q'_{(t-r)/2} \end{cases} \quad (9)$$

are defined for site pairs. $F'(r,t)$ can be obtained from the 'forward' recursion

$$F'(r,t) = \Omega_\tau(F'(r-1,t-1)+\sigma-\gamma, F'(r+1,t-1)+\sigma-\gamma, F'(r,t-2)+2\sigma+s(r,t)) \quad (10)$$

with

$$\Omega_\tau(x_1,x_2,x_3) = \tau \log\left(e^{x_1/\tau} + e^{x_2/\tau} + e^{x_3/\tau} + 1\right) , \quad (11)$$

and $F''(r,t)$ from the analogous 'backward' recursion. The last term on the r.h.s. of (11) is the contribution of a path starting at (r,t) and having length $L = 0$.

In the zero-temperature limit, the local free energy reduces to the local score maximum; $S(r,t) \equiv \lim_{\tau\to 0} F(r,t) = \max_{\mathcal{A}:(r,t)\in\mathcal{A}} S(\mathcal{A})$. The forward/backward decomposition is still of the form (8). The forward recursion (10) for $S'(r,t) \equiv \lim_{\tau\to 0} F'(r,t)$ reduces to the usual Smith-Waterman algorithm, i.e.,

$$\Omega_0(x_1,x_2,x_3) = \max(x_1,x_2,x_3,0) . \quad (12)$$

Appendix B: The ridge path algorithm

The local ridge score maximum $R(r,t)$ has a forward/backward decomposition of the form (8), with the forward score $R'(r,t)$ defined as the maximal ridge score (7) of all paths starting at some $t' \le t$ and ending at (r,t). We use the recursion

$$R'(r,t) = \quad (13)$$
$$\begin{cases} \Omega_0(R'(r-1,t-1)+\sigma-\gamma, R'(r+1,t-1)+\sigma-\gamma, R'(r,t-2)+2\sigma+s(r,t)) \\ \quad \text{if } S(r,t) = \max(S(r-1,t), S(r,t), S(r+1,t)) \\ 0 \quad \text{otherwise,} \end{cases}$$

which requires previous computation of the scores $S(r,t)$ according to Appendix A. The backward score $R''(t)$ is obtained in an analogous way.

References

1. Needleman, S.B. and Wunsch, C.D., *J. Mol. Biol.* **48**, 443-453, (1970).
2. Smith, T.F. and Waterman, M.S., *J. Mol. Biol.* **147**, 195-197, (1981).
3. Hwa, T. and Lässig, M., *Phys. Rev. Lett.* **76**, 2591-2594, (1996).
4. Hwa, T. and Lässig, M., *Proceedings of the Second Annual International Conference on Computational Molecular Biology (RECOMB98)*, 109-116 (ACM Press, New York, 1998).
5. Drasdo D., Hwa, T. and Lässig, M., *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology*, 52-58 (AAAI Press, Menlo Park, 1998).
6. R. Olsen, T. Hwa, and Lässig, M., *Pacific Symposium on Biocomputing* **4**, 302 (1999).
7. Drasdo D., Hwa, T. and Lässig, M., Preprint available at http://xxx.lanl.gov/abs/cond-mat/9802023.
8. Zhang, M.Q. and Marr, T.G., *J. Theo. Biol.* **174**, 119 - 29, (1995).
9. Miyazawa, S., *Protein Eng.* **8**, 999-1009, (1996).
10. Dewey, G., *Pacific Symposium on Biocomputing* **4**, 266 - 77 (1999).
11. Bishop, M.J. and Thompson, E.A., *J. Mol. Biol.* **190**, 159 - 65, (1986).
12. Thorne, J.L., Kishino, H. and Felsenstein, J., *J. Mol. Evol.* **33**, 114 - 24, (1991).
13. Thorne, J.L., Kishino, H., and Felsenstein, J., *J. Mol. Evol.* **34**, 3 - 16, (1992).
14. Allison, L., Wallace, C.S., and Yee, C.N., *J. Mol. Evol.* **35**, 77 - 90, (1991).
15. Zhu J., Liu J.S. and Lawrence E., *Bioinformatics* **14**, 25-39, (1998).
16. Kschischo M. and Lässig M., preprint (1999).
17. Olsen R., Bundschuh R., Hwa T., *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology* (1999).
18. Waterman, M., *Proc. Natl. Acad. Sci. USA* **80**, 3123-3124, (1983).
19. Waterman, M.S. and Eggert, M., *J. Mol. Biol.* **4**, 723 - 8, (1987).
20. Mevissen H.T. and Vingron M., *Protein Eng.* **9**, 127-132, (1996).
21. Vingron M., *Current Opinion in Structural Biology* **6**, 346-352, (1996).
22. Waterman M. and Vingron M., *Stat. Sci.* **9**, 367-81.
23. Waterman M. and Vingron M., *Proc. Natl. Acad. Sci. USA* **91**, 4625-8 (1994).