A Statistical Theory of Sequence Alignment with Gaps

Dirk Drasdo
Max-Planck-Institut für Kolloid- und Grenzflächenforschung
Kantstr. 55, 14513 Teltow, Germany
present address: IMISE, Universität Leipzig
Liebigstr. 27, 04103 Leipzig, Germany
e-mail: drasdo@imise.uni-leipzig.de

Terence Hwa
Department of Physics
University of California at San Diego
La Jolla, CA 92093-0319
e-mail: hwa@ucsd.edu

Michael Lässig
Max-Planck-Institut für Kolloid- und Grenzflächenforschung
Kantstr. 55, 14513 Teltow, Germany
e-mail: lassig@mpikg-teltow.mpg.de

Abstract

A statistical theory of local alignment algorithms with gaps is presented. Both the linear and logarithmic phases, as well as the phase transition separating the two phases, are described in a quantitative way. Markov sequences without mutual correlations are shown to have *scale-invariant* alignment statistics. Deviations from scale invariance indicate the presence of mutual correlations detectable by alignment algorithms. Conditions are obtained for the optimal detection of a class of mutual sequence correlations.

Introduction

Sequence alignment is an important tool in molecular biology (Waterman 1994, Doolittle 1996). Alignment algorithms are designed to detect mutual correlations between DNA or protein sequences; such correlations are often indicative of functional and evolutionary relationships. Given two sequences, the so-called local alignment algorithms identify pairs of putatively correlated elements in two contiguous subsequences. The powerful algorithm of Smith and Waterman (1981) produces alignments with gaps (i.e., unpaired elements) to account for the occurrence of local insertions and deletions in molecular evolution. Mutual correlations between subsequences are detected by means of a scoring function: Based on the number of matches, mismatches, and gaps, a score is assigned to each alignment of the sequences compared. Maximization of this score is then used to select the optimal alignment, taken as a measure of the mutual correlations between the sequences. However, it is well known that the optimal alignment of a given pair of sequences strongly depends on the scoring parameters used, and so does its *fidelity*, that is, the extent to which it recovers the mutual correlations. The key problems of alignment statistics are to quantify the degree of sequence similarity based on alignment data (e.g., the score), to find the scoring parameters producing the alignment of highest fidelity, and to assess the significance of the results obtained.

This communication reports recent progress in the statistical theory of alignments with gaps. We show that such alignments can be understood using the concept of *scale invariance* familiar from the physics of phase transitions. Scale invariance is observed for long pairs of mutually uncorrelated Markov sequences aligned in their entirety, so-called global alignments (Needleman and Wunsch 1970). It manifests itself in a series of nontrivial power laws. For example, the score variance for such sequences grows with the power 2/3 of the sequence length (Hwa and Lässig 1996; Drasdo, Hwa, and Lässig 1997, 1998). Local alignments, however, show a phase transition separating two different phases of scoring parameters (Arratia and Waterman 1994). This introduces a finite characteristic length scale t_s, i.e., a scale independent of the sequence lengths. It can be defined in the so-called logarithmic phase as the average length of the aligned subsequences ending at a given pair of elements. Mutual correlations between sequences generate a second characteristic scale, the correlation length t_c (Hwa and Lässig 1996; Drasdo, Hwa, and Lässig 1998). This is the scale above which the optimal global alignment of the correlated sequences becomes significantly different from alignments of uncorrelated sequences at the same scoring parameters. The scales t_s and t_c have a strong dependence on the scoring parameters whose functional form has been studied in detail (Drasdo, Hwa and Lässig 1998; Hwa and Lässig 1998) and is summarized below. We find that high-fidelity alignments are obtained when t_s and t_c are of the same order of magnitude and are jointly minimized. This condition can be used to select the scoring parameters for optimal similarity detection by local alignment.

Review of alignment algorithms

We study local alignments of pairs of Markov sequences $Q = \{Q_i\}$ and $Q' = \{Q'_j\}$ with an approximately equal number of elements $\sim N/2$. Each element Q_i or Q'_j is chosen with equal probability from a set of c different alphabets. We mostly take $c = 4$ as appropriate for nucleotide sequences, although the results can be easily generalized to arbitary values of c. An alignment is defined as an ordered set of pairings (Q_i, Q'_j) and of gaps $(Q_i, -)$ and $(-, Q'_j)$ 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). We define the length of an alignment as the total number of aligned elements of both sequences, $L \equiv i_2 - i_1 + j_2 - j_1$.

A given alignment is conveniently represented as a *directed path* on a two-dimensional grid as shown in Fig. 1(b) (Needleman and Wunsch 1970). Using the rotated coordinates $r \equiv i - j$ and $t \equiv i + j$, this path is described by a single-valued function $r(t)$ measuring the "displacement" of the path from the diagonal of the alignment grid. The length L of the alignment equals the projected length of its path onto the diagonal.

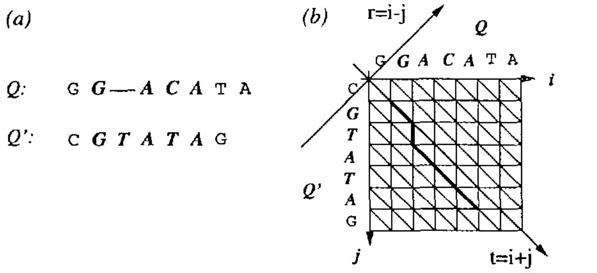

Figure 1: (a) One possible local alignment of two nucleotide sequences, $Q = \{GGACATA...\}$ and $Q' = \{CGTATAG...\}$. The aligned subsequences are shown in boldface, with 4 pairings (three matches, one mismatch) and one gap. The length L of an alignment is the total number of elements participating in that alignment. For the example shown, $L = 9$. (b) The alignment in (a) can be represented uniquely as a *directed* path (the thick path) directed along the diagonal of the alignment grid; each vertical (horizontal) bond of the path corresponds to a gap in sequence Q (Q'). L equals the projected length of the directed path onto the diagonal.

Each alignment is assigned a score S, maximization of which defines the optimal alignment. We use here the simplest class of *linear* scoring functions, with S given by the total number N_+ of matches ($Q_i = Q'_j$), the total number N_- of mismatches ($Q_i \neq Q'_j$), and the total number N_g of gaps. The most general such function involves three scoring parameters:

$$S = \sigma_+ N_+ + \sigma_- N_- + \sigma_g N_g \ , \qquad (1)$$

with σ_+, σ_-, and σ_g denoting the score of a match, mismatch, and gap, respectively. However, the *optimal* alignment configuration of a given sequence pair Q and Q' is left invariant if all three scoring parameters in (1) are multiplied by the same factor. Without loss of generality, we can therefore use the scoring function

$$S = \sigma L + \sqrt{c-1}\, N_+ - \frac{1}{\sqrt{c-1}} N_- - \gamma N_g \ , \qquad (2)$$

which is normalized in such a way that a pairing of two independent elements has the average score 2σ and the variance 1. Here $L = 2N_+ + 2N_- + N_g$ denotes again the length of the alignment defined above, and c is the size of the alphabet set. The two scoring parameters entering Eq. (2) have a simple interpretation: σ is an overall score gain for each element aligned, and γ is the cost of each gap. Hence σ controls the length L of the optimal alignment, while changing γ affects its number of gaps, i.e., the displacement of the optimal alignment path [1]. Note that for global alignment, we can take $L = N$ fixed. For this case, the first term in (2) becomes an overall additive constant which does not affect the alignment. Hence, the statistics of global alignments depends only on the single parameter γ.

The dynamic programming algorithm obtains optimal alignment paths from the "score landscape" $S(r,t)$, where $S(r,t)$ denotes the optimal score for the set of all alignment paths ending at the point (r,t). The score landscape for local alignments is computed by the Smith-Waterman (1981) recursion relation

$$S(r,t) = \max \left\{ \begin{array}{l} S(r-1,t-1) + \sigma - \gamma \\ S(r+1,t-1) + \sigma - \gamma \\ S(r,t-2) + s(r,t) + 2\sigma \\ 0 \end{array} \right\} \qquad (3)$$

with

$$s(r,t) = \left\{ \begin{array}{ll} \sqrt{c-1} & \text{if} \quad Q_{(r+t)/2} = Q'_{(t-r)/2} \\ -\frac{1}{\sqrt{c-1}} & \text{if} \quad Q_{(r+t)/2} \neq Q'_{(t-r)/2} \end{array} \right. . \qquad (4)$$

Evaluation of this relation starts with the initial condition $S(r, t = 0) = 0$ and stops at $t = N$. (We have used various versions of the algorithm and boundary conditions; see Appendix B of Drasdo, Hwa, and Lässig (1998) for a detailed discussion.)

The score landscape $S(r,t)$ has the absolute maximum

$$\Sigma \equiv \max_{r,t} S(r,t) \ . \qquad (5)$$

[1] In statistical physics, such a path is known as a *directed polymer*; see Krug and Spohn (1991), Halpin-Healy and Zhang (1995) for recent reviews. The scoring parameters γ and σ can be interpreted as the *line tension* (governing the displacements) and the *chemical potential* (governing the length) of the polymer, respectively.

The optimal alignment path ends at the point (r_2, t_2), where $S(r_2, t_2) = \Sigma$. The path is found by back tracking [2] from this endpoint to the initial point (r_1, t_1) given by $S(r_1, t_1) = 0$. The length of the optimal path is $L = t_2 - t_1$. For global alignments, one uses the simpler recursion relation (Needleman and Wunsch, 1970)

$$S^G(r,t) = \max \left\{ \begin{array}{l} S^G(r-1,t-1) + \sigma - \gamma \\ S^G(r+1,t-1) + \sigma - \gamma \\ S^G(r,t-2) + s(r,t) + 2\sigma \end{array} \right\} . \quad (6)$$

The optimal global alignment path ends at the point $(r_2 = 0, t_2 = N)$ and is tracked back to the initial point $(r_1 = 0, t_1 = 0)$.

Alignment of Uncorrelated Sequences

The cornerstone of the theory of alignment with gaps is the *global* alignment statistics of mutually uncorrelated Markov sequences. (We distinguish their score data by the subscript 0 from those of mutually correlated sequences to be discussed below.) Consider the average score $\overline{S_0^G}(t)$ obtained from global alignment of long sequence pairs ($N \gg 1$). We can take this quantity to be either $\overline{S_0^G}(r = 0, t)$ or $\overline{\max_r S^G(r,t)}$. The overbar denotes an ensemble average over sequence pairs, although in practice, ensemble averaged quantities can often be obtained from a *single* sequence pair (Hwa and Lässig, 1998; Drasdo, Hwa, and Lässig 1998). For large values of t, it is easy to show that $\overline{S_0^G}(t)$ is asymptotically linear in t (Arratia and Waterman 1994), with

$$\overline{S_0^G}(t) \simeq (\sigma + E_0(\gamma))\, t \, . \quad (7)$$

From the definition of the scoring function (2), it is clear that the prefactor has a nontrivial dependence only on γ. The function $E_0(\gamma)$ is monotonically decreasing and can be calculated asymptotically for large γ, with the result $E_0(\gamma) \sim 1/\gamma$ (Hwa and Lässig, unpublished). Numerically, we find this function to be well approximated by the form $E_0(\gamma) \propto 1/(\gamma + \text{const})$ over the interval $\gamma > \gamma_0 \equiv 1/(2\sqrt{c-1})$ (Drasdo, Hwa and, Lässig 1998) [3].

A number of other quantities are governed by nontrivial power laws, e.g., the variance of the score landscape

$$(\Delta S_0^G(t))^2 \equiv \overline{(S_0^G)^2}(t) - (\overline{S_0^G}(t))^2 = B^2(\gamma)\, t^{2/3} \quad (8)$$

and the mean square displacement of between two points t_1 and t_2 of the optimal alignment path,

$$\overline{(r_0(t_2) - r_0(t_1))^2} = A^2(\gamma)\, |t_2 - t_1|^{4/3} \, . \quad (9)$$

These power laws reflect the statistical scale invariance of global alignments without inter-sequence correlations. The exponents 2/3 and 4/3 governing the t dependence are "universal", that is, the dependence on the scoring parameters is contained entirely in the amplitude functions $B(\gamma)$ and $A(\gamma)$. These functions are well approximated by the form $A^{3/4}(\gamma) \sim B^{-3}(\gamma) \sim E_0(\gamma)$. The scaling laws (8) and (9) are believed to be exact for the closely related problem of first-passage percolation (Licea, Newman, and Piza 1996; Licea and Newman 1996; see also Krug and Spohn (1991) and Halpin-Healy and Zhang (1995) for reviews on recent progress from the statistical physics perspective.) That the same scaling laws apply to global alignment of uncorrelated sequences [4] was conjectured only recently (Hwa and Lässig 1996) and has since been verified by extensive numerical simulations (Drasdo, Hwa, and Lässig 1998). The same scaling has been found for pairs of unrelated cDNA sequences; an example is shown in Fig. 2 (Drasdo, Hwa, and Lässig 1998).

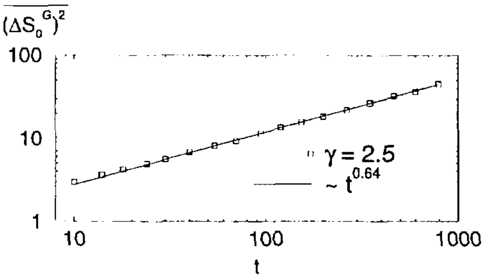

Figure 2: Score fluctuations for a pair of unrelated cDNA seqences (P.lividius cDNA for COLL2alpha gene with $N = 5511$ (Exposito *et al.* 1995) and Drosophila melanogaster (cDNA1) protein 4.1 homologue (coracle) mRNA, complete cds. with $N = 5921$ (Fehon, Dawson, and Artavanis-Tsakonas 1994)). The straight line is a least mean square fit to the data which is in very good agreement to the expected power law given by Eq. (8).

The score statistics for *local* alignments of mutually uncorrelated Markov sequences can be inferred from that of global alignments. Of foremost importance is the existence of a phase transition (Arratia and Waterman 1994) at $\sigma + E_0(\gamma) = 0$. This defines the phase transition line $\sigma_c(\gamma) = -E_0(\gamma) \propto 1/(\gamma + \text{const})$. The two phases are distinguished by the asymptotics of the average optimal score $\overline{\Sigma_0(N)}$ for long sequences, which is of order N for $\sigma > \sigma_c$ and of order $\log N$ for $\sigma < \sigma_c$

[2] There is an exponential number of degenerate paths having the same score. Since most of these paths overlap each other very closely we can resolve the degeneracies by random choices during back tracking.

[3] For $\gamma < \gamma_0$, it is always favorable to replace a mismatch by two gaps, and the algorithm becomes biologically irrelevant.

[4] Note that the global alignment problem differs significantly from the usual first-passage percolation problem since the former contains $O(N)$ random numbers while the latter contains $O(N^2)$ random numbers; see also Arratia and Waterman (1994). This difference is, however, irrelevant for the asymptotic scaling behavior (Hwa and Lässig, unpublished). A detailed heuristic discussion addressing the correspondence of these two problems is given by Cule and Hwa (1998) in the context of a number of closely related physics problems.

(Arratia and Waterman 1994). A comprehensive understanding [5] of the score statistics in the vicinity of the phase transition line follows from the scaling laws (7) and (8), as we now show.

Consider an optimal path of local alignment ending at a point (r, t). Let the average score be $\overline{S_0}(t)$ and the average alignment length be $\overline{L_0}(t) \leq t$. We discuss first the linear phase where $\delta\sigma \equiv \sigma - \sigma_c > 0$. For large t, we have $\overline{S_0}(t) \simeq \delta\sigma \cdot t$ and $\overline{L_0}(t) \simeq t$ as in global alignment, since the typical optimal score $S_0(t) \sim \overline{S_0}(t) \pm \Delta S_0^G(t)$ becomes large for large t and is thus unaffected by the constraint $S > 0$ special to local alignment. But this condition can be violated at small t if $S_0(t) < \Delta S_0^G(t)$. Using the score variance in Eq. (8), we find the linear behavior to hold only for t exceeding a characteristic length scale

$$t_s(\sigma, \gamma) \sim B^{3/2}(\gamma)|\delta\sigma|^{-3/2}. \qquad (10)$$

Note that t_s diverges as the phase transition line is approached, i.e., as $\delta\sigma \rightarrow 0$. This indicates that the preasymptotics for $t < t_s$ is given by the "critical" behavior right along the transition line where $\delta\sigma = 0$. There, global alignments yield score in the range $-\Delta S_0^G(t)$ to $\Delta S_0^G(t)$. For local alignment, the averages are dominated by the *finite* fraction of sequence pairs with $S_0^G(t) \geq 0$ for all t. Since such sequences have scores of the order $\Delta S_0^G(t)$, we obtain the important result

$$\overline{S_0^c}(t) \sim B(\gamma) t^{1/3} \qquad (11)$$

describing the average alignment score right at the phase transition line.

Slightly on the other side of the phase transition line, i.e., for $\delta\sigma \lesssim 0$, the score experiences a small negative drift, $-|\delta\sigma|\ t$. As $\delta\sigma \rightarrow 0^-$, this effect is negligible for small t and the average score follows the critical behavior (11) until the negative drift "catches up", i.e., when $\overline{S_0^c}(t) \sim |\delta\sigma| t$. Hence, the average length $\overline{L_0}(t)$ saturates to a value $t_s \equiv \lim_{t\to\infty} \overline{L_0}(t)$ whose parameter dependence is given by Eq. (10). The corresponding saturation value of the score, $S_{\text{sat}} \equiv \lim_{t\to\infty} \overline{S_0}(t)$, is given by

$$S_{\text{sat}}(\sigma, \gamma) \sim \overline{S_0^c}(t_s) \sim B^{3/2}(\gamma)|\delta\sigma|^{-1/2}\,. \qquad (12)$$

This behavior of the average alignment length and score close to the phase transition can be written in the form (Hwa and Lässig 1998)

$$\frac{\overline{S_0}(t)}{S_{\text{sat}}} = \mathcal{S}_\pm\left(\frac{t}{t_s}\right)\,, \quad \frac{\overline{L_0}(t)}{t_s} = \mathcal{L}_\pm\left(\frac{t}{t_s}\right)\,. \qquad (13)$$

The scaling functions $\mathcal{S}_\pm$ and $\mathcal{L}_\pm$ are again universal; i.e., the entire dependence on the scoring parameters is contained in the constants (10) and (12). The subscript of the scaling functions refers to the sign of $\delta\sigma$; the two branches correspond to the linear and the logarithmic phase, respectively. The branches $\mathcal{S}_\pm$ share the same asymptotic behavior $\mathcal{S}_\pm(\tau) \sim \tau^{1/3}$ for $\tau \ll 1$. For $\tau \gg 1$, $\mathcal{S}_+(\tau) \sim \tau$ but $\mathcal{S}_-(\tau) \rightarrow O(1)$ signaling a finite saturation score. The scaling form (13) has been verified numerically (Hwa and Lässig 1998). Fig. 3(a) shows $\overline{S_0}(t)$ for various values of γ and σ close to the transition ($|\delta\sigma|/\sigma_c(\gamma) \leq 0.05$). Plotting $\overline{S_0}(t)/S_{\text{sat}}$ as a function of t/t_s shows a clear data collapse to a two-branched function $\mathcal{S}_\pm$ with the predicted asymptotics, see Fig. 3(b).

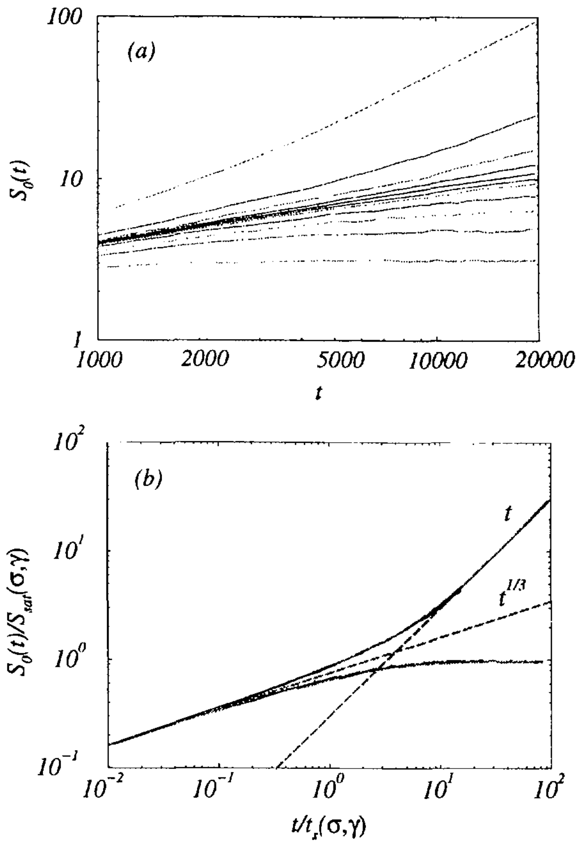

Fig. 3: Local alignment of Markov sequences without mutual correlations. (a) The score average $\overline{S_0}(t)$ for various values of γ and σ obtained from an ensemble of 1000 random sequence pairs of 10000 elements each. The curves correspond to $\gamma = 3.0$ and $\delta\sigma/\sigma_c(\gamma) = 0.05$ to -0.05 (top to bottom). (b) The same data plotted according to the scaling form of Eq. (13) exhibits the two branches of the scaling function $\mathcal{S}_\pm$. The asymptotic behavior is seen to follow the scaling theory; the expected power laws are indicated by dashed lines.

Given the saturation values t_s and S_{sat}, we can estimate the average optimal score $\overline{\Sigma_0}(N) = \overline{\max_{t<N} S_0(t)}$ and length $\overline{\Lambda_0}(N) = \overline{\max_{t<N} L_0(t)}$. The probability of the length $L_0(t \gg t_s)$ being *much larger* than $\overline{L_0} = t_s$ is expected to be Poisson distributed (Waterman and Vingron 1994b). For sequences of lengths

[5] For a discussion in the context of a related physics problem, see Muñoz and Hwa (1998).

of the order $N \gg t_s$, the probability of finding some $t_s < t < N$ such that $L_0(t) \gg t_s$ is of the order $(N/L_0)\exp(-\text{const}\cdot L_0/t_s)$. Since $\overline{\Lambda_0}$ is given by the largest L_0 for which this probability remains finite, we find asymptotically the result

$$\overline{\Lambda_0}(N) \sim t_s(\sigma,\gamma)\log(N/t_s) + O(\log\log(N/t_s))\ . \quad (14)$$

Similarly, the optimal score is

$$\overline{\Sigma_0}(N) \sim S_{\mathrm{sat}}(\sigma,\gamma)\log(N/t_s)+O(\log\log(N/t_s))\ . \quad (15)$$

It is instructive to compare the score statistics described above with that of *gapless* local alignments widely used in large-scale database searches (Altschul et. al. 1990). In gapless alignments ($\gamma \to \infty$), the alignment paths are constrained to a single value of r. Since the score value for each pairing is an independent random variable with average 2σ and variance 1, we have $\overline{S_0^G}(t) = \sigma t$ and $\Delta S_0^G(t) = t^{1/2}$ for the "global version" of the gapless alignment. Using these expressions in place of Eqs. (7) and (8), and repeating the above analysis, we obtain the following properties for gapless local alignment: There is a phase transition at $\sigma = \sigma_c = 0$ where $\overline{S_0^c} \sim t^{1/2}$. For $\sigma < 0$, the average length and score of the optimal alignment are still of the form $\overline{\Lambda_0}(N) \sim t_s\log(N/t_s)$ and $\overline{\Sigma_0}(N) \sim S_{\mathrm{sat}}\log(N/t_s)$, respectively, but with a different dependence on σ given by $t_s \sim |\sigma|^{-2}$ and $S_{\mathrm{sat}} \sim |\sigma|^{-1}$. These results are of course well known (Karlin and Altschul 1990; Karlin, Dembo, and Kawabata 1990; Dembo and Karlin 1991)[6]. While gapless local alignment is sufficiently simple so that even the complete distribution function $P(\Sigma_0)$ is known, the inclusion of gaps greatly complicates the problem. Indeed, even the first moment $\overline{\Sigma_0}$ (i.e., the form of the coefficient $S_{\mathrm{sat}}(\sigma,\gamma)$) has not been analyzed systematically prior to this work. Knowledge of the leading moments can be used to construct an effective description for $P(\Sigma_0)$ (Bundschuh *et al.*, unpublished); this is a direction currently being pursued by many groups (Waterman and Vingron 1994a, 1994b; Altschul and Gish 1996).

Alignment of Correlated Sequences

Evolution acts on DNA by local substitutions, insertions, and deletions of nucleotides, as well as by rearrangements of large segments of the sequence. Hence DNA sequences in different organisms can have subsequences that differ only by local mutations, with many pairs of conserved elements (i.e., elements that are neither deleted nor substituted at any point of the evolution process) inherited from a common ancestor. We model such mutations by a simple Markov process, allowing for local deletions and insertions (of random elements) at an average frequency q, as well as random point substitutions with probability p per element; see Drasdo, Hwa, and Lässig (1998) for details. The average fraction $U = (1-p)(1-q)$ of ancestor elements conserved in the daughter sequence quantifies the degree of correlations between the two sequences. In the sequel, we consider pairs of Markov sequences Q and Q' with mutually correlated subsequences $\hat{Q}$ and $\hat{Q}'$ of approximately equal length $\hat{N}/2 \ll N$; the remainder of the sequences Q and Q' has no mutual correlations. The subsequences $\hat{Q}$ and $\hat{Q}'$ are related by a realization of the above Markov process characterized by the parameters U and q. The pairs of conserved elements $(Q_i, Q'_j) \in \hat{Q}\times\hat{Q}'$ are to be identified by local alignment; the fraction $\mathcal{F}$ of correctly detected conserved pairs defines the *fidelity* of an alignment.

The local alignment statistics of correlated sequences is again based on the properties of global alignments. Hence consider first the optimal global alignment of the sequences $\hat{Q}$ and $\hat{Q}'$ for given values of γ and σ. If the alignment covers a finite fraction $\mathcal{F}$ of the conserved pairs, it will have an increased number of matches and hence a higher score than alignments of uncorrelated sequences. Indeed, the average score $\overline{S^G}(t)$ has the asymptotic form

$$\overline{S^G}(t) \simeq (\sigma + E(\gamma;U,q))\,t = (\delta\sigma + \delta E(\gamma;U,q))\,t \quad (16)$$

for large values of t, with a finite score gain $\delta E(\gamma;U,q) \equiv E(\gamma;U,q) - E_0(\gamma) \geq 0$ per unit of t over uncorrelated sequences. Mutual correlations between sequences introduce a length scale into global alignment, the *correlation length* t_c (Hwa and Lässig. 1996; Drasdo, Hwa, and Lässig, 1998). For $t > t_c$, the alignment becomes statistically different from the global alignment of mutually uncorrelated sequences. Hence, t_c is the threshold length for similarity detection by global alignment. Its value can be estimated by equating the score gain $\overline{S^G}(t_c) - \overline{S_0^G}(t_c) = \delta E\cdot t_c$ with the r.m.s. score for random sequences, $\Delta S_0^G(t_c)$, given by Eq. (8). We obtain

$$t_c \sim B^{3/2}(\gamma)\,(\delta E)^{-3/2}\ , \quad (17)$$

which should be compared with (10) for the saturation length of local alignments.

The scaling theory of global alignments (Drasdo, Hwa, and Lässig, 1998) establishes the parameter dependence of the score gain δE. This is found to have the approximate scaling form $\delta E(\gamma;U,q)/U = \delta\mathcal{E}(x,y)$, where $x \equiv C(\gamma)/U$, $y \equiv q/U^2$, and $C(\gamma) \approx E_0(\gamma)$ is another amplitude function. Similar scaling forms are found for the correlation length t_c and for the average fidelity $\overline{\mathcal{F}}$. Fig. 4 shows numerical data for the scaled score gain $\delta\mathcal{E}(x,y)$ obtained from single sequence pairs with various values of U, q and γ. As predicted by the scaling theory, the curves of Fig. 4 have clear maxima, which turn out to be close to the maxima of $\overline{\mathcal{F}}$ and to the minima of t_c (Drasdo, Hwa, and Lässig, 1998). We conclude that *global alignments can be optimized efficiently by maximization of the score gain δE.* Notice

[6] Note that the parameter σ here plays the role of the important parameter λ in gapless local alignment (Karlin and Altschul 1990). It is straightforward to verify that in the vicinity of the phase transition ($\sigma \to 0^-$), $|\sigma| \propto \lambda$.

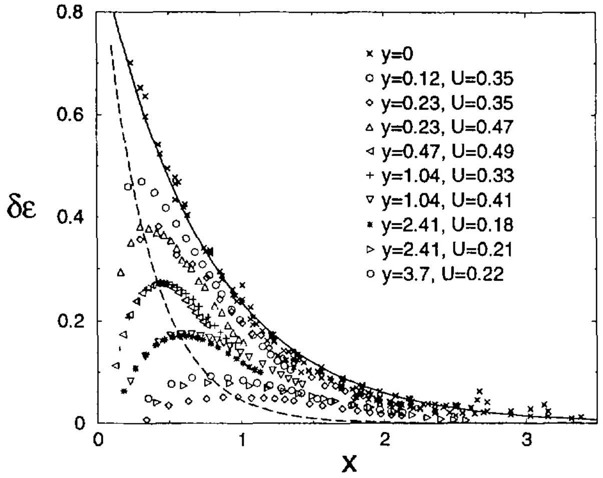

Fig. 4: Global alignment of Markov sequences with mutual correlations. The score gain $\delta\mathcal{E}(x,y)$ obtained from single sequence pairs with various evolution parameters U, q and alignment parameters γ. The data for different (U,q,γ) corresponding to the same values of (x,y) collapse approximately, as predicted by the scaling theory. The lines are the theoretical loci of the maxima (dashed) and the theoretical limit curve $\delta\mathcal{E}(x,0)$ (solid). See Drasdo, Hwa, and Lässig (1998) for details.

that this is quite different from maximizing of the total score, which is frequently (but erroneously) used in applications. δE can be extracted directly from alignment data. An efficient method based on the score landscape approach is described by Hwa and Lässig (1998).

The optimal global alignment of the subsequences $\hat{Q}$ and $\hat{Q}'$ is to be reproduced as a local alignment of the entire sequences Q and Q', at least approximately. Therefore its score given by (16) must be the absolute score maximum, i.e.,

$$(\delta\sigma + \delta E)\,\hat{N} > \overline{\Sigma_0}(N)\ . \qquad (18)$$

This requires (i) $\delta\sigma > -\delta E$ so that the l.h.s. of (18) is positive and (ii) $\delta\sigma < 0$ so that the r.h.s. is small, i.e., only of order $\log N$ as given by Eq. (15). Hence *weak correlations ($\delta E \to 0$) can only be detected with scoring parameters set in the logarithmic phase close to the transition line.*

According to Eq. (18), detection is possible if $\hat{N}$ exceeds the threshold length $N_0 = \overline{\Sigma_0}(N)/(\delta\sigma + \delta E)$. Minimizing N_0 by using (15) and (12) determines the optimal value of σ for given γ,

$$\delta\sigma^*(\gamma; U, q) = -\tfrac{1}{3}\delta E(\gamma; U, q)\ . \qquad (19)$$

By Eqs. (10) and (17), this is equivalent to the condition $t_s \sim t_c$, producing an optimal detection threshold $N_0 \sim t_c \log(N/t_c)$. The optimal value of γ is then determined as for global alignments (Drasdo, Hwa, and Lässig 1998). Hence *local alignments are efficiently optimized by maximization of the score gain δE while keeping $\delta\sigma = -\delta E/3$.*

Discussion

The scaling theory of alignment with gaps is based on the scale-invariant statistics of global alignments of mutually uncorrelated Markov sequences. In local alignments of correlated sequences, this scale invariance is broken by the simultaneous presence of two length scales: the saturation length t_s and the correlation length t_c. The theory presented here provides a coherent description of local alignments both in the linear and the logarithmic phase, including a quantitative understanding of the phase transition. This is important for similarity detection since scoring parameters suitable for the analysis of weak correlations are found to be close to the phase transition line in the logarithmic phase. We show that minimizing the length scale t_c and keeping t_s of the order t_c produces alignments with high fidelity and low detection threshold $N_0 \sim t_c \log(N/t_c)$. These conditions can be turned into an optimization procedure for local alignments based on score data. A crucial question is, of course, whether these findings carry over to the mutation statistics of real sequences and to the algorithmic variants commonly used (which have scoring functions with more than two parameters). An important empirical result indicates that this may well be the case: As pointed out by Vingron and Waterman (1994), optimal alignments of weakly correlated sequences are indeed found in the vicinity of the phase transition line, just as predicted by this theory.

Acknowledgments

The authors have benefited from discussions with S.F. Altschul, R. Bundschuh, R. Durbin, M.A. Muñoz, M. Vingron, and M.S. Waterman. T.H. acknowledges the financial support of a research fellowship by the A.P. Sloan Foundation, and an young investigator award from the Arnold and Mabel Beckman Foundation.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic Local Alignment Search Tool. *J. Mol. Biol.* 215, 403-410.

Altschul, S.F. and Gish, W. 1996. Local Alignment Statistics. *Methods in Enzymology* 266, 460-480.

Arratia, R. and Waterman, M.S. 1994. A Phase Transition for the Score in Matching Random Sequences Allowing Deletions. *Ann. Appl. Prob.* 4, 200-225.

Cule, D. and Hwa, T. 1998. Static and Dynamic Properties of Inhomogeneous Elastic Media on Disordered Substrate. *Phys. Rev. B.* in press.

Dembo, A. and Karlin, S. 1991. Strong Limit Theorems of Empirical Functionals for Large Exceedances of Partial Sums of IID Variables. *Ann. Prob.* 19, 1737-1755.

Doolittle, R.F. 1996. *Methods in Enzymology* 266. Academic Press, San Diego.

Drasdo, D., Hwa, T. and Lässig, M. 1997. DNA Sequence Alignment and Critical Phenomena. *Mat. Res. Soc. Symp. Proc.* 263, 75-80.

Drasdo D., Hwa, T. and Lässig, M. 1998. Scaling Laws and Similarity Detection in Sequence Alignment with Gaps. Los Alamos e-print archive: physics/9802023.

Exposito J.Y., Boute N., Deleage G., Garrone R. 1995. Characterization of two Genes Coding for a Similar Four-Cysteine Motif of the Amino-Terminal Propeptide of a Sea Urchin Fibrillar Collagen. *Eur. J. Biochem.* 234:59-65.

Fehon R.G., Dawson I.A., Artavanis-Tsakonas S. 1994. A Drosophila Homologue of Membrane-Skeleton Protein 4.1 is Associated with Septate Junctions and is Encoded by the Coracle Gene. *Development* 120:545-557.

Halpin-Healy, T. and Zhang, Y.-C. 1995. Kinetic Roughening Phenomena, Stochastic Growth, Directed Polymers and All That: Aspects of Multidisciplinary Statistical Mechanics. *Phys. Rep.* 254, 215-414.

Hwa, T. and Lässig, M. 1996. Similarity Detection and Localization. *Phys. Rev. Lett.* 76, 2591-2594.

Hwa, T. and Lässig, M. 1998. Optimal Detection of Sequence Similarity by Local Alignment. *Proceedings of the Second Annual International Conference on Computational Molecular Biology (RECOMB98)*, in press.

Karlin S. and Altschul, S.F. 1990. Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes. *Proc. Natl. Acad. Sci. U.S.A.* 87, 2264-2268.

Karlin, S., Dembo, A., and Kawabata, T. 1990. Statistical Composition of High-Scoring Segments from Molecular Sequences. *Ann. Stat.* 18, 571-581.

Krug, J. and Spohn, H. 1991. Kinetic Roughening of Growing Interfaces. In *Solids far from equilibrium: Growth, Morphology, and Defects*, C. Godreche ed. Cambridge University Press.

Licea, C., Newman, C.M., and Piza, M.S.T. 1996. Superdiffusitivity in First-Passage Percolation. *Prob. Theory Relat. Fields* 106 559-591.

Licea, C. and Newman. C.M. 1996. Geodesics in Two-Dimensional First-Passage Percolation. *Ann. Prob.* 24 399-410.

Muñoz, M.A. and Hwa, T. 1998. On Nonlinear Diffusion with Multiplicative Noise. *Europhys. Lett.* 41, 147-152.

Needleman, S.B. and Wunsch, C.D. 1970. A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of two Proteins. *J. Mol. Biol.*, 48, 443-453.

Smith, T.F. and Waterman, M.S. 1981. Identification of Common Molecular Subsequences. *J. Mol. Biol.*, 147, 195-197.

Vingron, M. and Waterman, M.S. 1994. Sequence Alignment and Penalty Choice. Review of Concepts, Case Studies and Implications. *J. Mol. Biol* 235, 1-12.

Waterman, M.S. 1994. *Introduction to Computational Biology*, Chapman & Hall.

Waterman, M.S., and Vingron, M. 1994a. Rapid and Accurate Estimates of Statistical Significance for Sequence Data Base Searches. *Proc. Natl. Acad. Sci. U.S.A.* 91, 4625-4628.

Waterman, M.S. and Vingron, M. 1994b. Sequence Comparison Significance and Poisson Approximation. *Stat. Sci.* 9, 367-381.