Scaling Laws and Similarity Detection in Sequence Alignment with Gaps

Dirk Drasdo[1,3], Terence Hwa[2], and Michael Lässig[1]†

[1] Max-Planck Institut für Kolloid- und
Grenzflächenforschung,
Kantstr. 55, 14513 Teltow, Germany

[2] Department of Physics
University of California at San Diego
La Jolla, CA 92093-0319

[3] Institute for Medical Informatics, Statistics and Epidemiology
University of Leipzig
Liebigstr. 27, 04103 Leipzig, Germany

Abstract

We study the problem of similarity detection by sequence alignment with gaps, using a recently established theoretical framework based on the morphology of alignment paths. Alignments of sequences without mutual correlations are found to have scale-invariant statistics. This is the basis for a scaling theory of alignments of correlated sequences. Using a simple Markov model of evolution, we generate sequences with well-defined mutual correlations and quantify the *fidelity* of an alignment in an unambiguous way. The scaling theory predicts the dependence of the fidelity on the alignment parameters and on the statistical evolution parameters characterizing the sequence correlations. Specific criteria for the optimal choice of alignment parameters emerge from this theory. The results are verified by extensive numerical simulations.

Key words: sequence comparison; alignment algorithm; homology; evolution model; longest common subsequence

Related (p)reprints available at *http://matisse.ucsd.edu/∼hwa/pub.html*.

† *Corresponding author.* Email: lassig@mpikg-teltow.mpg.de. Fax: +49 3328 46215.

1 Introduction

Sequence alignment has been one of the most valuable computational tools in molecular biology. It has been used extensively in discovering and understanding functional and evolutionary relationships among genes and proteins. There are two basic types of alignment algorithms: algorithms without gaps such as the original BLAST (Altschul *et al.*, 1990), and algorithms with gaps, for example, variants of the Smith-Waterman local alignment algorithm (Smith and Waterman, 1981) as implemented in the current generation of BLAST and FASTA. Gapless alignment is widely used in database searches because the algorithms are fast (computational time scales linearly with sequence length) and the results depend very weakly on the choice of scoring systems (Altschul *et al.*, 1990; Altschul, 1993). However, gapless alignment is not sensitive to weak sequence similarities (Pearson, 1991). For detailed similarity analysis, algorithms with gaps are therefore needed (Waterman, 1989; 1994).

At present, there are two main obstacles to the wider application of the more powerful gapped alignment algorithms. Firstly, they require substantially longer computational time than gapless alignments (depending quadratically on the sequence length). More importantly, gapped alignments lack a quantitative theory assessing the statistical significance of the results obtained. It is this second issue we address in the present paper.

In a typical pairwise alignment, one assigns a score to each alignment of two sequences. The score is based on the degree of match/mismatch for each pair of aligned elements, and on the number of gaps used. 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 can depend strongly on the scoring parameters used. The same is true for the *fidelity* of the optimal alignment, that is, the extent to which mutual correlations are recovered. Hence, the key problem of alignment statistics is to quantify the degree of sequence similarity based on attainable alignment data, and to find the scoring parameters producing alignments of the highest fidelity. Optimal scoring parameters have been chosen mostly by trial and error so far, although there have been systematic efforts to establish a more solid empirical footing (Benner, 1993; Vingron and Waterman, 1994; Koretke *et al.*, 1996). The statistical theory presented here gives a systematic way to find optimal alignment parameters, and to understand their dependence on the inter-sequence correlations. It is expected to be most useful in the alignment of weakly homologous sequences, where a judicious choice of scoring parameters is critical.

To guide the choice of scoring parameters, a quantitative measure of sequence similarity is necessary. The most widely used measure is the p-value, which expresses the likelihood that a given alignment score is obtained by chance. To compute the p-value, it is necessary to understand quantitatively the score distribution, particularly the large-score tail of the distribution, given the vast number of sequences in the database. While there is an exact theory to compute the asymptotic distribution for arbitrary scoring parameters in gapless alignment (Karlin & Altschul, 1990; 1993), no theory is available for alignment with gaps. Direct numerical simulation using shuffled sequences has been used instead. The shuffling

method is very time consuming, as tens of thousands of shuffles are typically needed to reconstruct the tail of the distribution[1].

In this paper, we shall adopt a different approach which is not based on the p-value. We develop a general *scaling theory* relating the fidelity of the alignment (which is unobservable for unknown homology) to alignment score data which are observable. The theory is motivated by knowledge obtained from related problems of statistical physics, and is supported by extensive numerical simulations on synthetic sequences. One outstanding virtue of our approach is that the statistical significance of an alignment can be estimated based on the alignment score data of a *single* sequence, without the need of shuffling. This general approach can also be extended to estimate statistical significance via the p-value, as demonstrated recently by Olsen *et al.* (1999b).

Since the algorithm is designed to detect residual similarities between sequences in a divergent evolution, it is clear that the fidelity measure has to emerge from the underlying evolution process. We use a simple probabilistic evolution model to generate daughter sequences from ancestor sequences by local substitutions, insertions, and deletions. The model is certainly too simple to describe realistic evolution processes, but it allows an unambiguous identification of inherited mutual similarities between sequences. The fidelity of an alignment is then simply the fraction of the inherited similarities recovered by it. Maximization of the fidelity is used as a criterion to select optimal scoring parameters. These depend, of course, on the parameters of the primary evolution process. A link between evolution parameters and scoring parameters is also inherent to maximum-likelihood methods (Bishop and Thompson, 1986; Thorne *et al.*, 1991, 1992). It has been found, however, that maximum-likelihood methods do not maximize the fidelity as defined above (Kschischo and Lässig, 1999).

The existing theory of gapless alignments has been used successfully to describe local alignments with *few gaps* (in a sense to be made precise below). The theory of this paper describes the opposite limit of alignments with *many gaps*. The statistics of such alignments differs significantly from — but is shown to be consistent with — the gapless limit. We focus on global alignments of long sequences obtained by the Needleman-Wunsch (1970) algorithm, which inherently have many gaps. An important special case of this theory is the problem of the longest common subsequence (LCS), for which a number of conjectures and bounds exist. Additionally, we have shown in two recent communications (Hwa and Lässig, 1998; Drasdo *et al.* 1998) that results on local alignments close to the phase transition to global alignment (Waterman *et al.*, 1987; Arratia and Waterman, 1994) can also be described by this theory. This regime of the Smith-Waterman algorithm is important for biological applications since it has been found empirically to produce "good" alignments (Vingron and Waterman, 1994). The phase transition, in particular, is found to differ qualitatively from the corresponding transition for gapless alignments.

The statistical theory of gapped alignments presented here is based on a *geometrical*

[1]Waterman and Vingron (1994) proposed a declumping method which required only $\sim$ 10 shuffles for random amino acid sequences. However, the declumping algorithm itself is rather time consuming, and the direct estimate by simulation is recommended over the declumping method (Hardy and Waterman, 1997).

approach introduced recently by two of us (Hwa and Lässig, 1996). This approach focuses on the morphology of the optimal *alignment paths*. The notion of an alignment path (recalled below) provides a very fruitful link to various well-studied problems of statistical mechanics (Kardar, 1987; Fisher and Huse, 1991; Hwa and Fisher, 1994) as has also been noticed by Zhang and Marr (1995). The important statistical properties of alignment paths are described by a number of *scaling laws* (Hwa and Lässig, 1996; Drasdo *et al.*, 1997) explained in detail below. Their validity for sequence alignment is supported by extensive numerical evidence. The resulting scaling theory of alignment has three main virtues:

(i) It distinguishes clearly between *universal* (parameter-independent) properties of alignments and those depending on the scoring parameters (and hence governing their optimal choice). We find generic alignments with gaps and LCS alignments share the same universal properties, which differ from those of gapless alignments.

(ii) It relates score data of alignments to their fidelity and to the underlying evolutionary parameters characterizing the similarities of the sequences compared.

(iii) Its key statistical *averages* turn out to be significant for the alignment of *single* sequence pairs that are sufficiently long.

These scaling laws are important for the statistics of *uncorrelated* and *correlated* sequences as we show in detail below. They lead to a systematic score-based parameter optimization for global (Needleman-Wunsch) alignments as well as for local (Smith-Waterman) alignments (Olsen *et al.*, 1999a). Statistical scaling theories have also been developed for related optimization problems in structural biology, notably protein folding (Wang *et al.*, 1996; Onuchic *et al.*, 1997).

This paper is organized as follows. In Section 2, we define the evolution process, recall the global alignment algorithm used throughout this paper, and discuss the qualitative aspects of the geometrical approach. The quantitative theory of alignment starts in Section 3, where we give a detailed description of the alignment statistics for uncorrelated random sequences, and present the power laws governing alignment paths and scores. In Section 4, we turn to sequences with mutual correlations inherited by a realization of our evolution process. We establish a scaling theory that explains the parameter dependence of alignments in a quantitative way. Hence we derive optimal alignment parameters as a function of the evolution parameters. Furthermore, we show how the evolutionary parameters and the optimal alignment of a given pair of sequences can be deduced from its score data.

2 The Geometrical Approach to Sequence Alignment

Evolution model

The evolution process used in this study has as its input an "ancestor" sequence $Q = \{Q_1, \ldots, Q_i, \ldots, Q_N\}$ of length $N \gg 1$. Each element Q_i is chosen from a set of c different letters with equal probability $1/c$, independently of the elements at other positions. Hence,

the ancestor sequence is a Markov random sequence. The numerical results presented below are for the case $c = 4$ as appropriate for nucleotide sequences, but for some derivations, it is useful to consider general c-letter alphabets.

The evolution process generates a daughter sequence $Q' = \{Q'_1, \ldots, Q'_j, \ldots, Q'_{N'}\}$ from the ancestor sequence Q. This process involves local insertions and deletions of random elements with the same probability $\tilde{p}$, and point substitutions by a random element with probability p. Insertion, deletion, and substitution events at one point of the sequence are independent of the events at other points. The evolution process can thus be formulated as a Markov process along the sequence (Bishop and Thompson, 1986, Thorne *et al.*, 1991; Hwa and Lässig, 1996). This Markov process models evolution in time with *cumulative* mutation probabilities p and $\tilde{p}$, which are related to the PAM distance of the sequences. These parameters should not be confused with the mutation *rates* per unit time. The precise evolution rules used in this paper are given in Appendix A. They are chosen such that the average length of the daughter sequence equals the length N of the ancestor sequence.

A specific realization of this Markov process defines a unique evolutionary relation between the sequences Q and Q'; see Fig. 1(a). Of course, the same pair of sequences can be linked by many different evolutionary relations. For a given relation, there is a well-defined set of conserved elements, i.e., elements that are neither deleted nor substituted. We call these conserved pairs of elements ($Q_i = Q'_j$) *native pairs*. The average fraction of ancestor elements Q_i conserved in the daughter sequence Q' is

$$U(p, q) = (1 - p)\,(1 - q)\ , \tag{1}$$

where

$$q = \frac{\tilde{p}}{1 - \tilde{p}} \tag{2}$$

is the effective insertion/deletion rate (see Appendix A). $U(p, q)$ quantifies the mutual similarity between sequences. Identical sequences have $U = 1$; mutually uncorrelated sequences are obtained for $p = 1$, i.e., $U = 0$. In the remainder of this paper, we take U and q as the basic parameters characterizing the evolution process. The primary goals of sequence alignment are to identify the native pairs and to estimate the mutual similarity U.

Alignment and Scoring Scheme

We align the sequences $Q = \{Q_1, \ldots, Q_i, \ldots, Q_N\}$ and $Q' = \{Q'_1, \ldots, Q'_j, \ldots, Q'_{N'}\}$ using the simplest version of the global alignment algorithm by Needleman and Wunsch (1970). A global alignment of two sequences is defined as an ordered set of pairings (Q_i, Q'_j) (matches or mismatches) and of gaps $(Q_i, -)$ and $(-, Q'_j)$, each element Q_i and Q'_j belonging to exactly one pairing or gap (see Fig. 1(b)). A special case is alignments without mismatches. These produce always an LCS of Q and Q', defined as a sequence $Q'' = \{Q''_1, \ldots, Q''_k, \ldots, Q''_L\}$ of maximal length L with $Q''_k = Q_{i_k} = Q'_{j_k}$, $i_1 < \ldots < i_L$, $j_1 < \ldots < j_L$.

Any alignment of Q and Q' is assigned a score S, maximization of which defines the optimal alignment. We use here the simplest *linear* gap function, with the alignment score

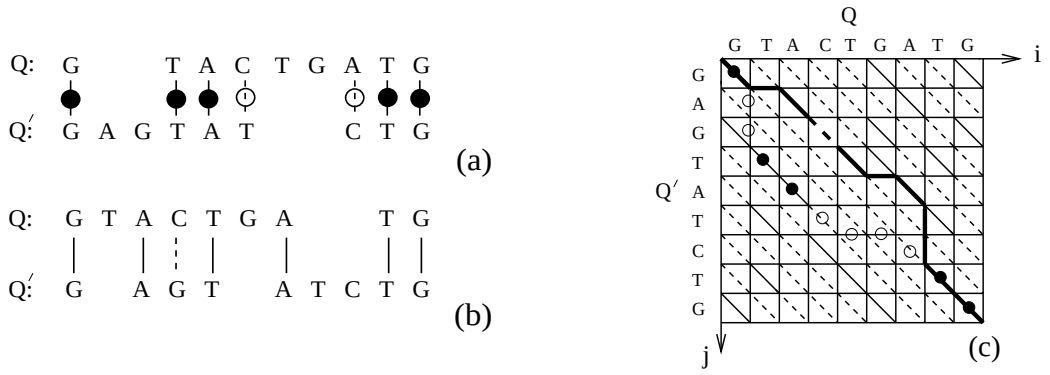

Fig. 1: (a) An evolutionary relation linking the ancestor sequence $Q = \{G, T, A, C, T, G, A, T, G\}$ to the daughter sequence $Q' = \{G, A, G, T, A, T, C, T, G\}$. Native pairs are marked by bonds with full circles, substitutions by bonds with empty circles. The unpaired letters Q_i are deleted, the unpaired letters Q'_j are inserted. (b) A possible alignment between Q and Q' with matches ($Q_i = Q'_j$) (full lines), mismatches ($Q_i \neq Q_j$) (dashed lines) and gaps (unpaired letters). (c) Lattice representation. The evolution path $R(t)$ corresponding to (a) is marked by circles; there are five native bonds (full circles). The alignment path corresponding to (b) appears as thick line whose solid (dashed) diagonal bonds are matches (mismatches) and whose horizontal and vertical bonds are gaps. It covers three of the five native bonds, producing the fidelity $\mathcal{F} = 3/5$.

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 used. Hence, the most general such function involves three scoring parameters:

$$S = \mu_+ N_+ + \mu_- N_- + \mu_g N_g \; . \tag{3}$$

To find the optimal global alignment, we can use without loss of generality (see Appendix B) the simpler scoring function

$$S = \sqrt{c-1}\, N_+ - \frac{1}{\sqrt{c-1}}\, N_- - \gamma N_g \; , \tag{4}$$

which has only a single scoring parameter, the effective gap cost γ. As a function of γ, we can distinguish different alignment regimes:

(i) For $\gamma \to \infty$, the optimal alignment becomes gapless. The match/mismatch scores in (4) are chosen such that gapless alignments of uncorrelated random sequences produce a score of mean 0 and variance 1 per element pair.

(ii) For $\gamma \geq \gamma_0$, the optimal alignment contains matches, mismatches, and gaps. This regime is the most interesting for biological sequences and the focus of this paper. Optimal values of γ are typically of order 1.

(iii) For $\gamma < \gamma_0 \equiv 1/(2\sqrt{c-1})$, the score cost of a mismatch is higher than that of two gaps. Hence, the optimal alignments contain only matches and gaps. They are independent of γ in this regime, producing always longest common subsequences of the sequences aligned. An LCS of two sequences of length N has a length $L = N_+$ which is related to the score of the corresponding alignment,

$$S = \sqrt{c-1}\, L - 2\gamma(N - L) \; . \tag{5}$$

The fidelity of an alignment

As discussed above, mutual correlations between the sequences $Q = \{Q_i\}$ and $Q' = \{Q_j\}$ arise from the set of native pairs ($Q_i = Q'_j$). The *fidelity* $\mathcal{F}$ of an alignment can be quantified as the fraction of correctly matched native pairs, see Fig. 1(b). This is an unambiguous measure of the goodness of an alignment, and it will be used below to find optimal alignment parameters. To evaluate $\mathcal{F}$ directly, the native pairs have to be distinguished from random matches ($Q_i = Q'_j$) involving mutated elements. Hence, the fidelity defined in this way depends not only on the sequences Q and Q' but also on the evolution path linking them. Of course, the evolution path is not known in actual applications of sequence alignment. However, the scaling theory discussed below relates statistical properties of $\mathcal{F}$ to *observable* alignment data, making it a useful and measurable quantity.

Lattice representation

Any alignment of two sequences $\{Q_i\}$ and $\{Q'_j\}$ is conveniently represented on a two-dimensional $N \times N'$ grid as in Fig. 1(c) (Needleman and Wunsch, 1970). The cells of this grid are labeled by the index pair (i, j). The diagonal bond in cell (i, j) represents the pairing of the elements (Q_i, Q'_j). The horizontal bond between cells (i, j) and $(i, j+1)$ represents a gap $(Q_i, -)$ located on sequence Q' between the elements Q'_j and Q'_{j+1}. The vertical bond between cells (i, j) and $(i+1, j)$ represents a gap located on sequence Q between the elements Q_i and Q_{i+1}. In this way, any alignment defines a unique *directed path* on the grid. Using the rotated coordinates $r \equiv j - i$ 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 Needleman-Wunsch dynamic programming algorithm obtains optimal alignments (denoted by the subscript $\times$) from the "score landscape" $S(r, t)$ computed recursively for all lattice points. Here $S(r, t)$ denotes the maximum score of all paths ending at the point (r, t). The recursion relation requires boundary conditions. We mostly use boundary conditions corresponding to *rooted* alignment paths starting at the point $(r = 0, t = 0)$, but some statistical quantities are conveniently evaluated for *unrooted* paths starting at an arbitrary point $(r, t = 0)$. The precise form of the algorithm and of the boundary conditions used in this paper are detailed in Appendix C. For given T, the maximum of the score landscape $S(x, T) = S_\times(T) \equiv \max_r S(r, T)$ determines the endpoint $x = r_\times(T)$; the entire path $r_\times(t)$ is then found by back-tracing. Of course, optimal paths defined in this way are not unique

since (i) the maximum score $S_\times(T)$ may be attained at different points x and (ii) for given x, the back-tracing may produce more than one path $r_\times(t)$. It can be shown that with probability 1, the resulting ambiguities for the displacement $r_\times(t)$ are only of the order of a single lattice spacing. For more precise formulations of this 'macroscopic' uniqueness of the optimal path, see Fisher and Huse (1991), Hwa and Fisher (1994), Kinzelbach and Lässig (1995). The 'microscopic' ambiguities do not affect any of the results reported below.

The evolutionary relation linking the sequences Q and Q' can also be represented as a directed path $R(t)$ on the alignment grid, called the *evolution path* (Hwa and Lässig, 1996). On this path, horizontal and vertical bonds represent deleted and inserted elements, respectively. For a given realization of the evolution process, the resulting path $R(t)$ is unique. A fraction U of the bonds along the evolution path are *native* bonds representing the native pairs ($Q_i = Q'_j$). The fidelity of an alignment is then simply the fraction of overlap between the trajectories of the optimal alignment path $r_\times(t)$ and the evolution path $R(t)$; see Fig. 1(c).

Alignment morphology

Alignment algorithms are designed to trace the mutual correlations between sequences. As it becomes clear from Figs. 2, the presence of such correlations affects both the morphology of the optimal alignment path $r_\times(t)$ and the associated score statistics. Fig. 2(a) shows the path $r_\times(t)$ for a pair of mutually uncorrelated random sequences. This path is seen to be intrinsically rough; i.e., the displacement has large variations. This "wandering" is caused by random agglomerations of matches in different regions of the alignment grid. Fig. 2(b) shows the corresponding score landscape $S(r,t)$ at a given value of t. The maximum score value occurs at the point $x = r_\times(t)$ and is seen to be not very pronounced; near-optimal score values occur also at distant points such as x_1. The statistics of alignment paths and scores for uncorrelated sequences are discussed in detail in Section 3 below.

The optimal alignment path for a pair of mutually correlated sequences (obtained from the evolution process described above) behaves quite differently, as shown in Fig. 2(c). Its wandering is essentially restricted to a "corridor" of finite width centered around around the evolution path $R(t)$. In this way, the path $r_\times(t)$ covers a finite fraction $\mathcal{F}$ of the native bonds. The corresponding score landscape is shown in Fig. 2(d). The maximum at $r_\times(t)$ is now very pronounced; all paths ending at points far from $r_\times(t)$ have substantially lower scores than the optimal path. The alignment statistics of mutually correlated sequence pairs is described in Section 4.

The morphology of the optimal alignment path depends strongly on the choice of the scoring parameter γ. As an example, Fig. 3 shows the optimal paths $r_\times(t)$ (dashed lines) for the *same* pair of correlated sequences with the same underlying evolution path $R(t)$ (the solid line), and for three different values of γ: At small γ, the path $r_\times(t)$ follows the evolution path only on large scales. On small scales, variations in the displacement $r_\times(t)$ are seen to be larger than those of $R(t)$ (Fig. 3(a)). The intrinsic roughness of the optimal alignment path limits its overlap with the evolution path, hence suppressing the fidelity. The fidelity

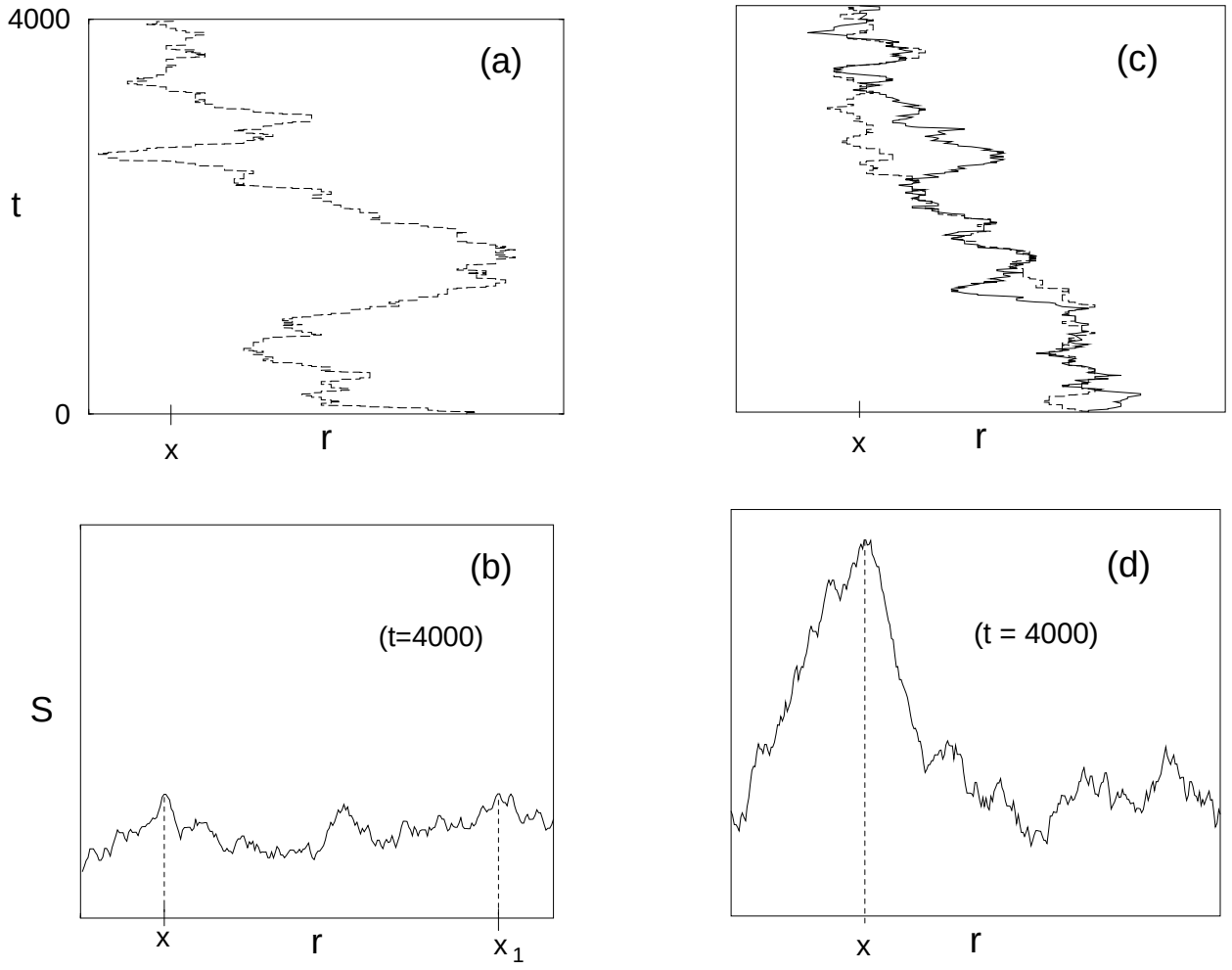

Fig. 2: (a) The optimal alignment path $r_\times(t)$ and (b) a slice of the score landscape $S(r, t = 4000)$ for a pair of mutually uncorrelated random sequences. The score maximum is at x, which defines the endpoint $x \equiv r_\times(t = 4000)$ of the optimal path. Similar score values occur also at distant points such as x_1. (c) The paths $r_\times(t)$ (dashed line), $R(t)$ (solid line) and (d) the score landscape $S(r, t = 4000)$ for a pair of sequences with mutual correlations. The score maximum at x is now pronounced; all distant points r have a substantially lower score. Hence the fluctuations of the alignment path $r_\times(t)$ are confined to a corridor around the evolution path $R(t)$.

is highest at some intermediate value γ^*, where the alignment path follows the target path most closely (Fig. 3(b)). At large γ, the alignment path contains large straight segments (Fig. 3(c)), which again reduces the fidelity.

A qualitative understanding of this parameter dependence may be gained from an analogy to random walks, regarding $r_\times(t)$ as the trajectory of a walker trying to follow a curvy path $R(t)$. The intrinsic properties of the walker are parametrized by γ. (In statistical mechanics, γ is called the effective line tension of the fluctuating path $r(t)$.) For small γ, the walker is "drunk" and cannot follow the path $R(t)$ without meandering to its left and right. This is the regime of Fig. 3(a), which we call the *random fluctuation* regime. For large values of γ, on the other hand, the walker is lazy and bypasses the larger turns of the path $R(t)$; this is the *shortcut* regime (Fig. 3(c)). From this analogy, it becomes plausible that a walker who is neither too drunk nor too lazy will follow the path $R(t)$ most closely and thereby achieve the highest fidelity (Fig. 3(b)). Such a criterion for the optimal parameter γ^* will indeed emerge from the quantitative theory described in the remainder of this paper.

3 Alignment of Uncorrelated Sequences

A statistical theory of alignment can hardly predict the optimal alignment for a specific pair of sequences. What can be characterized are quantities averaged over realizations of the evolution process for given parameters U and q. It will be shown, however, that these *ensemble averages* are also relevant for the alignment statistics of single pairs of "typical" sequences provided they are sufficiently long.

In the absence of mutual correlations (i.e., for $U = 0$), the statistics of alignments is determined by a balance between the loss in score due to gaps and the gain in score due to an excess number of random matches. As discussed by Hwa and Lässig (1996), the corresponding alignment paths belong to a class of systems known in statistical mechanics as *directed polymers* in a random medium. The statistical properties of directed polymers have been characterized in detail, treating r and t as continuous variables (Kardar, 1987; Huse and Fisher, 1991; Hwa and Fisher, 1994; see also the recent review by Lässig, 1998). They take the form of *scaling laws* governing the large-distance asymptotics of *ensemble averages* over the random potential. A number of scaling properties can also be proved for discrete models closely related to the alignment problem (Gwa and Spohn, 1991). Licea *et al.* (1994, 1996) have studied these scaling laws in the context of first passage percolation.

For the alignment problem proper, the scaling properties are presented as Conjectures 1 to 4. These are supported by extensive numerical evidence as discussed below. The main difference of the alignment problem to the percolation problem lies in the statistics of the match/mismatch score $s(r,t)$ (see Appendix C): On an alignment grid of size $N \times N$, there are N^2 such variables, indicating whether the pairing of elements (Q_i, Q'_j) produces a match or a mismatch. Since these variables are determined by the $2N$ sequence elements, they have mutual correlations. In the analogous percolation problem, however, the $s(r,t)$ are *independent* random variables. We find this difference in the statistics of the random variables does not affect the scaling properties of Conjectures 1 to 4, which take the same form as for

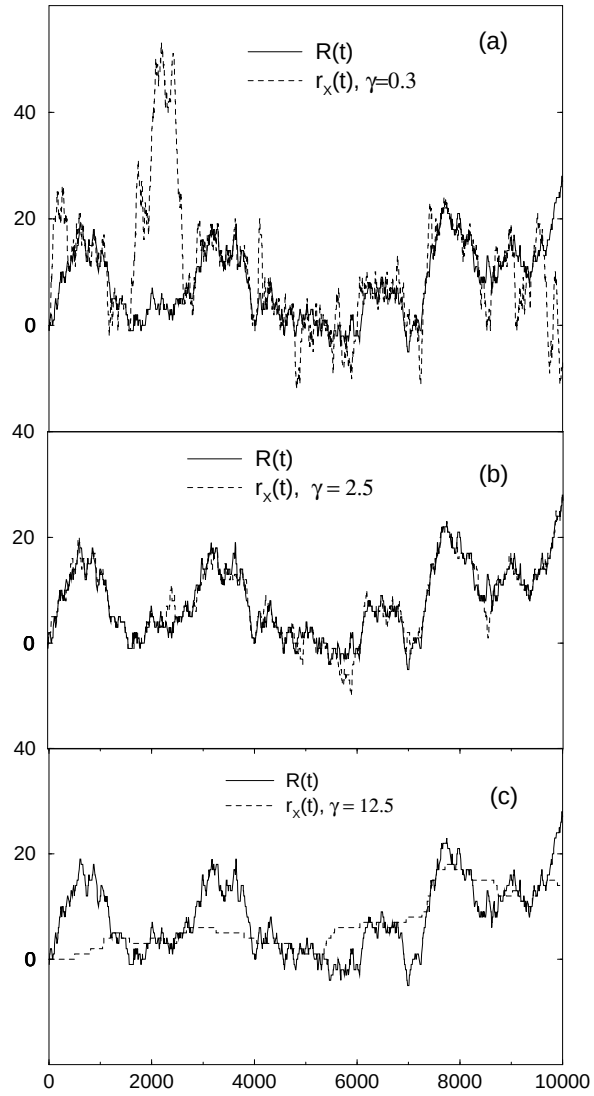

Fig. 3: Optimal alignment paths $r_\times(t)$ for the same pair of correlated sequences and three different values of γ. The evolution path $R(t)$ (solid lines) is the same in all three cases, while the optimal alignment paths $r_\times(t)$ (dashed lines) differ. (a) Random fluctuation regime ($\gamma < \gamma^*$). The path $r_\times(t)$ has strong fluctuations since the gap cost is low. (b) Optimal alignment parameter $\gamma = \gamma^*$. The fluctuations of the paths $r_\times(t)$ and $R(t)$ are of the same order of magnitude. (c) Shortcut regime ($\gamma > \gamma^*$). At high gap cost, the fluctuations of $R(t)$ are dominant, while $r_\times(t)$ contains large straight segments.

the percolation problem. The correlations between the variables $s(r,t)$ are observable in other alignment characteristics but these effects are always numerically small (see Appendix D).

Alignment path and score statistics

The scaling laws of Conjectures 1 and 2 describe the mean square displacement of the optimal alignment path from the diagonal, $\Delta_r^2(t) \equiv \overline{r_\times^2(t)}$, and related mean square score differences. These are obtained by averaging over an ensemble of mutually uncorrelated sequence pairs. Ensemble averages are denoted by overbars. These scaling laws are valid in the asymptotic limit of large t, i.e., for alignments with a large total number of gaps. (Below we denote by '$\simeq$' asymptotic equality and by '$\propto$' asymptotic proportionality up to a γ-independent factor of order 1.) Alignments in this limit have statistical properties qualitatively different from gapless (or nearly gapless) alignments. The statistical consistency of these alignment regimes is discussed at the end of this Section.

Conjecture 1 *For mutually uncorrelated sequences, the mean square displacement of the optimal alignment path has the asymptotic form*

$$\Delta_r^2(t) \simeq A^2(\gamma)\, t^{4/3} \;, \tag{6}$$

which is valid for $t \gg t_0(\gamma) \equiv A^{-3/2}(\gamma)$.

Remarks to Conjecture 1:

(i) The asymptotic law is valid for $\Delta_r^2(t) \gg 1$, i.e., $t \gg t_0(\gamma)$. For large γ, $t_0(\gamma)$ is the average distance between gaps. For $\gamma \to \infty$, this distance is found to diverge. Hence, the alignment becomes gapless in this limit for any given sequence pair.

(ii) The relation (6) says that the exponent 4/3 is a robust feature of the optimal alignment of uncorrelated random sequences, independent of the scoring parameter(s) or even scoring schemes used. A large gap cost efficiently suppresses the displacement only for the limited range of scales $t < t_0(\gamma)$. On larger scales, the cost of gaps is always outweighed by the gain in score from regions of the alignment grid with an excess number of random matches, leading to the power law (6) with a "universal" exponent. The dependence of the mean square displacement on the scoring parameters (γ in this case) is contained entirely in the coefficient $A(\gamma)$, which will be discussed below.

(iii) $\Delta_r^2(t)$ also describes the *auto-correlation* function of the optimal alignment path for a *single* sequence pair,

$$\Delta_r^2(t) \simeq C_r(t) \equiv T^{-1} \sum_{t_1=1}^{T} \left(r_\times(t_1+t) - r_\times(t_1)\right)^2 . \tag{7}$$

In this sense, the ensemble average is equivalent to averaging over initial points t_1, in the asymptotic limit $T \to \infty$.

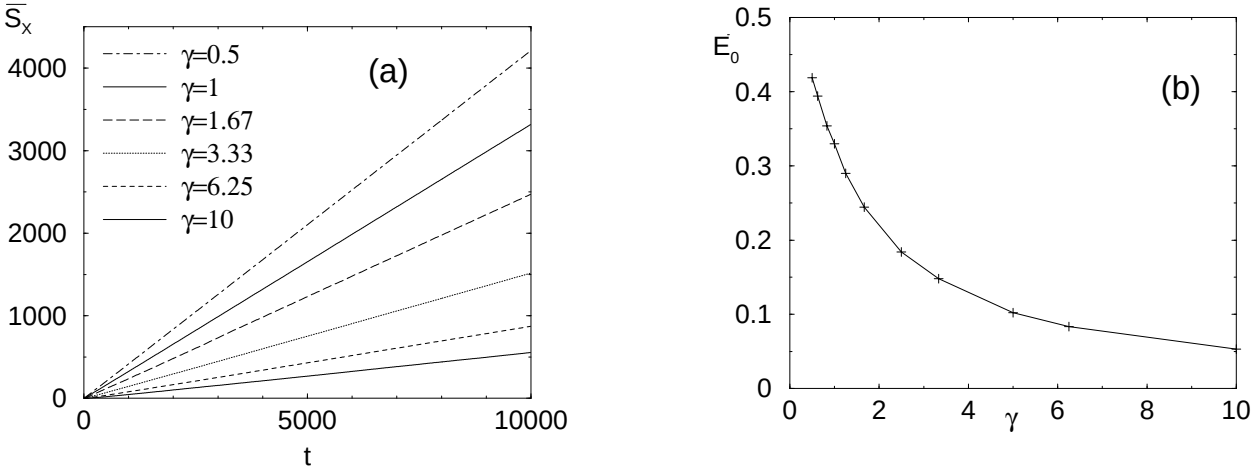

Fig. 4: (a) The average optimal score $\overline{S_\times(t)} \simeq E_0(\gamma)t$ as a function of t for several values of γ. The average has been obtained from an ensemble of 200 pairs of independent random sequences. (b) The asymptotic score per aligned element, $E_0(\gamma)$, obtained from the asymptotic slope of the lines in (a).

(iv) The higher moments of the displacement follow similar scaling laws, $\overline{r_\times^{2k}(t)} \propto \Delta_r^{2k}(t)$. Hence, the whole probability distribution for the variable $x = r_\times(t)$ can be written in scaling form, $P(x,t) \simeq t^{-2/3}\mathcal{P}(xt^{-2/3})$, where the scaling function $\mathcal{P}$ has an exponential tail for large values of its argument. This says that the displacement $r_\times(t)$ of the optimal path has a typical magnitude of the order of $\Delta_r(t)$. The same is true for all high-scoring paths. Paths with a larger displacement have significantly more gaps, reducing the score $S(r,t)$.

We now turn to the statistics of the score landscape $S(r,t)$ defined in the previous Section. Arratia and Waterman (1994) have shown that the average score $\overline{S(r,t)}$ is asymptotically linear in t. For a single pair of sequences, we find that both the optimal score $S_\times(t)$ and $S(r,t)$ at fixed r have the same asymptotics as the ensemble average,

$$S_\times(t) \simeq S(r,t) \simeq \overline{S(r,t)} \simeq E_0(\gamma)\, t\ , \tag{8}$$

since the score is cumulative over the path. The regime of validity is again $t \gg t_0(\gamma)$. The coefficient function $E_0(\gamma)$ is the limit score per aligned element for two random sequences. Using the normal form (4) of the scoring function, $E_0(\gamma)$ is a positive, monotonically decreasing function of γ, which tends to 0 in the gapless limit $\gamma \to \infty$. This function has been calculated in a variational scheme (Bundschuh and Hwa, 1998), which turns out to be a very good approximation for not too large values of γ. Fig. 4 shows the linear growth of the average optimal score $S_\times(t)$ and the extracted data for $E_0(\gamma)$.

Eq. (8) has an important consequence. The difference between the optimal score $S_\times(t)$ and other values $S(r,t)$ grows slower than t, which explains that score maxima for uncorrelated sequence pairs are not very pronounced. In fact, the local variations of the score landscape are described by scaling laws with fractional exponents, which are related to

those for the alignment paths. As described above, the score landscape $S(r,t)$ for rooted alignment paths at given t looks random for displacements $|r| \lesssim \Delta_r(t)$ (see Fig. 2(b)), while scores for larger values of $|r|$ are significantly lower. The typical amplitudes of the random fluctuations can, for example, be characterized by the mean square score difference $\Delta_S^2(t) \equiv \overline{(S(r=-\Delta_r(t)/2,t) - S(r=\Delta_r(t)/2,t))^2}$. This determines also the score difference between different high-scoring paths.

Conjecture 2 *The mean square score difference $\Delta_S^2(t)$ for mutually uncorrelated sequences has the asymptotic form*

$$\Delta_S^2(t) \simeq B^2(\gamma)\, t^{2/3} \tag{9}$$

valid for $t \gg t_0(\gamma)$.

Remarks to Conjecture 2:

(i) The dependence on the alignment parameters lies only in the prefactor, while the exponent 2/3 is universal. The function $B(\gamma)$ is related to $A(\gamma)$ as discussed below.

(ii) The scaling laws of Conjectures 1 and 2 have precisely the same form as for a directed polymer in a random medium, with independent random variables $s(r,t)$. Hence, the mutual correlations between the $s(r,t)$ are irrelevant for the scaling of $\Delta_r^2(t)$ and $\Delta_S^2(t)$. (Details can be found in Drasdo, Hwa, and Lässig (1999); see also the discussion by Cule and Hwa (1998) for a number of related physics problems.) Nevertheless, correlation effects between the variables $s(r,t)$ can be observed in other characteristics of the score landscape. The most important one is the *single-point* score variance, which is asymptotically linear in t as discussed in Appendix D; see also the discussion by de los Rios and Zhang (1998) for a related system. In the LCS case, the score variance is directly related to the variance of the LCS length by (5).

(iii) $\Delta_S^2(t)$ can be evaluated efficiently from single sequence pairs if boundary conditions corresponding to unrooted alignment paths are used; see Appendix D.

Fig. 5 combines our numerical evidence for Conjectures 1 and 2 and shows that displacement and score statistics are indeed closely related. Fig. 5(a) contains a log-log plot of the mean square displacement $\Delta_r^2(t)$ for different values of γ. The ensemble averages are seen to have the same asymptotic behavior as the auto-correlation function $C_r(t)$ for a single pair of long sequences. Fig. 5(b) shows the mean square score difference $\Delta_S^2(t)$ evaluated as described in Appendix D. The data in (a) and (b) are asymptotically straight lines; the asymptotic behavior sets in rather quickly for most values of γ. The respective slopes of these lines are 4/3 and 2/3, in accordance with the exponents given in (6) and (9). The intercepts of the asymptotic lines with the vertical axis then determine the coefficient functions $A(\gamma)$ and $B(\gamma)$; see Figs. 5(c,d). Finally, we show autocorrelation data for a pair of unrelated cDNA sequences in Figs. 5(e,f). The same scaling is found, justifing our modeling of individual sequences as Markov chains.

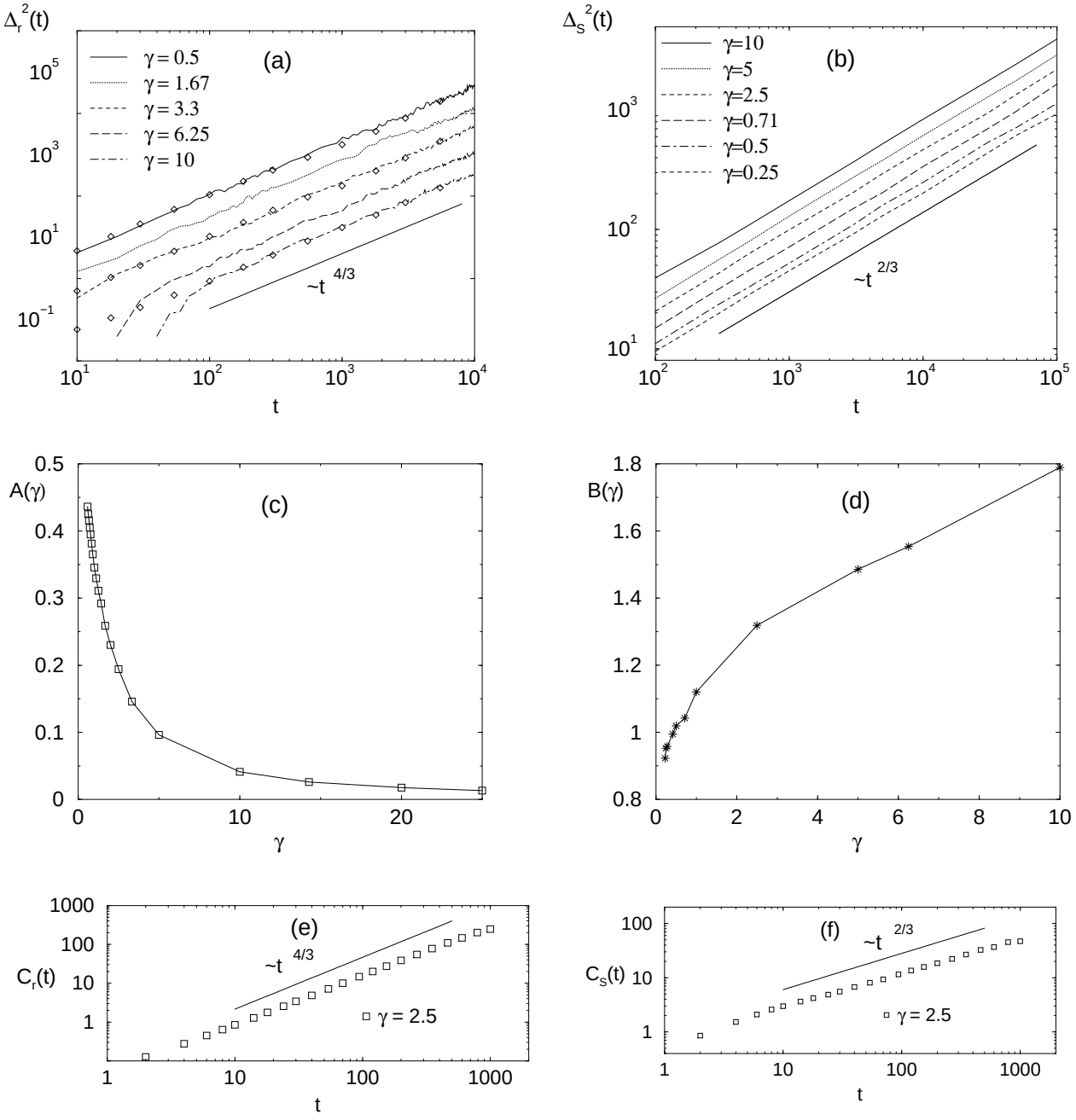

Fig. 5: (a) Mean square displacement $\Delta_r^2(t)$ (lines) and auto-correlation function $C_r(t)$ (diamonds) of the optimal alignment path for several values of γ. The averages are obtained from an ensemble of 200 mutually uncorrelated sequence pairs; the auto-correlation data are from a single sequence pair of length $N = 10^5$. (b) Mean square score difference $\Delta_S^2(t)$ for the same ensemble as in (a). (c) The coefficient $A(\gamma)$ extracted from (a). (d) The coefficient $B(\gamma)$ extracted from (b). (e,f) Auto-correlation functions $C_r(t)$ and $C_S(t)$ (defined in an analogous way) for a pair of unrelated cDNA seqences (P.lividius cDNA for COLL2alpha gene (Exposito *et al.*, 1995) and Drosophila melanogaster (cDNA1) protein 4.1 homologue (coracle) mRNA, complete cds. (Fehon *et al.*, 1994)).

Confinement and tilt scores

A related set of scaling laws governs the change in the average optimal score S_0 when the alignment paths are subject to various *constraints*. For example, the constraint $-W/2 < r(t) < W/2$ artificially confines the alignment paths to a strip of width W on the alignment grid. This constraint is easily implemented in the alignment algorithm as described in Appendix C. It becomes effective if W is smaller than typical displacements $\Delta_r(t)$ of the optimal unconstrained path, i.e., for $t \gg t_W(\gamma) \equiv W^{3/2} t_0(\gamma)$. The confinement lowers the score maximum $S_\times(t)$ since the optimal confined path $r_\times(t)$ can no longer take advantage of random agglomerations of matches outside the strip. We define the average *confinement cost* $S_c(W;t) \equiv \overline{S(W;t)} - E_0(\gamma)t < 0$.

Conjecture 3 *The average confinement cost has the asymptotic form*

$$S_c(W;t) \simeq E_c(W) \cdot t \tag{10}$$

for $t \gg t_W(\gamma)$, *and*

$$E_c(W) \simeq -C(\gamma)\, W^{-1} \,, \tag{11}$$

for $W \gg 1$.

In a similar way, the alignment may be constrained by restricting *both* ends of the alignment path to given values of r. Consider, for example, an optimal rooted path (starting at $(r = 0, t = 0)$) with endpoint fixed at $x = r(T)$. It is forced to have an average *tilt* $\theta \equiv x/T$, which increases its number of gaps and decreases its number of matches. This is quantified by the *tilt cost* $S_t(\theta;t) \equiv \overline{S(r = \theta t;t)} - E_0(\gamma)t < 0$.

Conjecture 4 *The average tilt cost has the asymptotic form*

$$S_t(\theta;t) \simeq E_t(\theta) \cdot t \tag{12}$$

for $t \gg t_0(\gamma)$, *with*

$$E_t(\theta) \simeq -D(\gamma)\theta^2 \,. \tag{13}$$

for small tilt angles, $|\theta| < t_0^{-1}(\gamma)$.

Conjectures 3 and 4 have also been verified numerically. Fig. 6(a) shows the confinement cost per unit of t, $E_c(W)$, as a function of $1/W$ for several values of γ. The data sets fall on straight lines, supporting the conjectured scaling form (11). The slopes of these lines then give the coefficient $C(\gamma)$ shown in Fig. 6(b). The tilt cost $E_t(\theta)$ is shown in Fig. 7(a) as a function of θ^2 for various values of γ. We find again straight lines and extract the coefficient $D(\gamma)$ from their slopes (Fig. 7(b)).

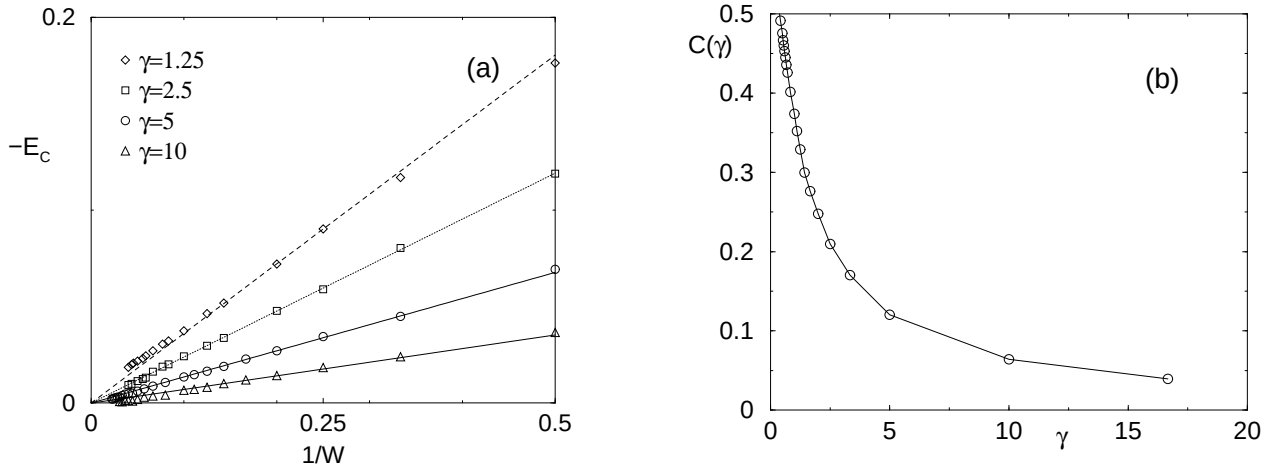

Fig. 6: (a) The confinement cost E_c as a function of $1/W$ for various values of γ. The averages are obtained from an ensemble of 200 mutually uncorrelated random sequences. (b) The coefficient $C(\gamma)$ obtained from the slope of the lines in (a).

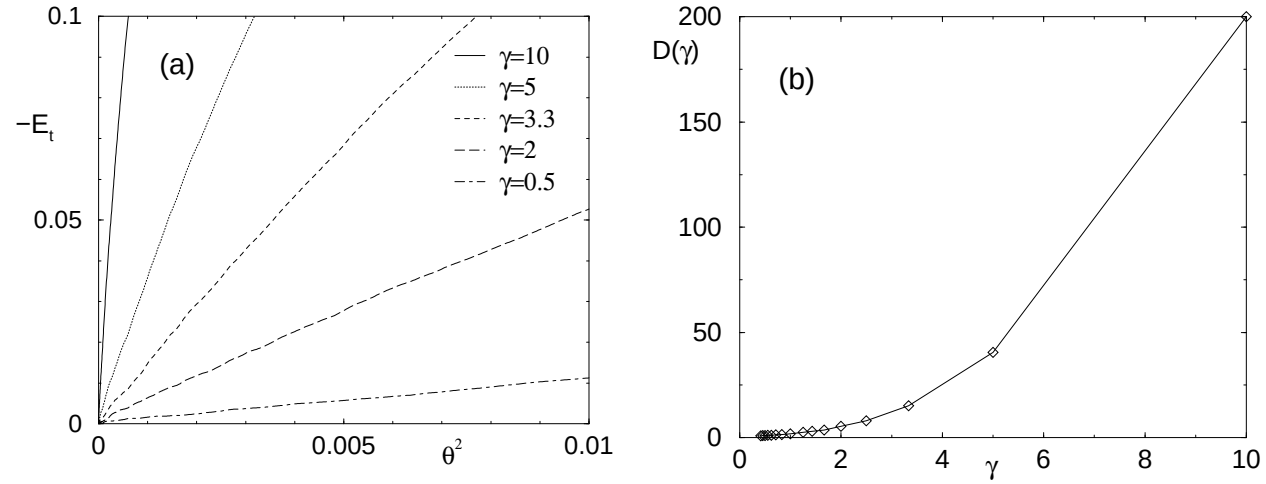

Fig. 7: (a) The tilt cost E_t as a function of θ^2 for various values of γ. (b) The coefficent $D(\gamma)$ obtained from the slope of the lines in (a).

Parameter dependence and link to gapless alignment

Conjectures 1 to 4 all have the same structure: they describe power laws with universal exponents and parameter-dependent coefficients. These relations contain variables of longitudinal distance (t), displacement (r), and score (S). Taking r as the basic variable, the amplitudes A, B, C, D are given in terms of the (a priori arbitrary) normalization factors of t and S, namely $t_0(\gamma)$ and $s_0(\gamma)$. To define the normalization factors, we rewrite Conjectures 1 and 2 as $\Delta_r^2(t) \simeq (t/t_0)^{4/3}$ and $\Delta_S^2(t) \simeq s_0^2(t/t_0)^{2/3}$, respectively. Hence, $A = t_0^{-2/3}$ and $B = s_0/t_0^{1/3}$. The scales t_0 and s_0 define the lower boundaries in longitudinal distance and score of the asymptotic scaling regime described by Conjectures 1 to 4. It is then a simple matter of dimensional analysis to express the remaining amplitudes as $C = s_0/t_0$ and $D = s_0 \cdot t_0$. Hence, we have only *two independent* amplitudes, and there are universal amplitude relations, e.g., $C = A \cdot B$.

Of course, these universal relations do not yet fix the parameter dependence of the amplitudes. To obtain this dependence, recall that for large γ, $t_0(\gamma)$ is the average distance between gaps of the optimal alignment. Conjectures 1 to 4 refer to alignments with a large number of gaps, i.e., to sequences of length $N \gg t_0(\gamma)$. In the limit $\gamma \to \infty$, however, the average distance $t_0(\gamma)$ between gaps diverges. Hence, for given sequences and sufficiently large γ, we always have $N \ll t_0(\gamma)$; the optimal alignment is gapless. Consistency between the statistics of gapped and gapless alignments then imposes a set of matching conditions at the crossover scale $t_0(\gamma)$. The r.m.s. score difference $\Delta_S(t)$ between to gapless alignments in neighboring diagonals $r = r_1$ and $r = r_1 + 1$ grows as $\Delta_S(t) \propto t^{1/2}$. The occurrence of a gap requires $\Delta_S(t)$ to exceed the gap cost γ; this happens, by definition, for $t \propto t_0(\gamma)$ and $\Delta_S \propto s_0(\gamma)$, and leads to an average score gain per unit of t, $E_0(\gamma) \propto s_0(\gamma)/t_0(\gamma)$. Thus we have

$$t_0^{1/2}(\gamma) \propto s_0(\gamma) \propto E_0^{-1}(\gamma) \propto \gamma \quad (\gamma \gg 1) \; . \tag{14}$$

We conclude that for large γ, there is only a *single independent* amplitude function (up to γ-independent factors) in Conjectures 1 to 4, which is moreover linked to the coefficient $E_0(\gamma)$ in (8),

$$A^{3/4}(\gamma) \propto B^{-3}(\gamma) \propto C(\gamma) \propto D^{-1/3}(\gamma) \propto E_0(\gamma) \; . \tag{15}$$

Numerically we find the relations (15) to hold approximately in the entire interval $\gamma > \gamma_0$. This is shown in Fig. 8. The amplitude data of Figs. 5(b), 6(c,d), 7(b), and 8(b), raised to the appropriate powers according to (15) and adjusted by γ-independent proportionality factors, all collapse approximately onto a single curve, which can be fitted as

$$E_0(\gamma) = \frac{0.722}{\gamma + 1.257} \; . \tag{16}$$

In the LCS regime ($\gamma < \gamma_0$), optimal alignment paths are independent of γ and scores are linear in γ as given by (5). Hence, we have $t_0(\gamma) = t_0(\gamma_0)$ and $s_0(\gamma) \propto E_0(\gamma) = (\sqrt{c-1}/2 + \gamma)\ell - \gamma$, where $\ell \equiv \overline{L}/N$. The numerical value of this constant, $\ell = 0.654...$ is very close to the expression $\ell = 2/(1 + \sqrt{c})$ (with $c = 4$) conjectured by Arratia (private

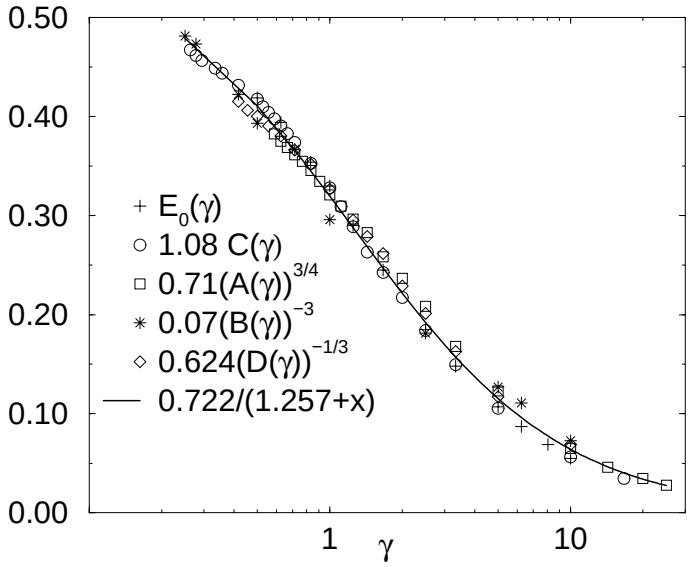

Fig. 8: Parameter dependence of the amplitudes A, B, C, D, and E_0 as given by (15), together with a fit curve of the form (16).

communication; see Steele, 1986). This conjecture has recently been proved (Bundschuh and Hwa, 1999; Boutet de Monvel, 1999) for the first passage percolation problem where the match/mismatch score $s(r,t)$ are independent random variables.

4 Alignment of Correlated Sequences

Displacement fluctuations of the evolution path

As discussed in Section 2, the mutual correlations between sequences can be represented by the evolution path $R(t)$ on the alignment grid. This path has displacement fluctuations due to the random distribution of insertions and deletions, see Figs. 2(c) and 3. However, the statistics of these fluctuations is different from that of the alignment paths discussed in the previous Section. Since the evolution is modeled as a Markov process, the mean square displacement $\Delta_R^2(t) \equiv \overline{(R(t_1+t) - R(t_1))^2}$ has the form

$$\Delta_R^2(t) = q|t| \tag{17}$$

characteristic of a Markov random walk, with q given by Eq. (2) (see Appendix A). The overbar denotes an ensemble average over realizations of the evolution process with given values of U and q. The ensemble average (17) can also be obtained from the auto-correlation function of a single (sufficiently long) evolution path $R(t)$ as in (7).

Score gain over uncorrelated sequences

For sequences with mutual correlations (i.e., $U > 0$), the morphology of the optimal alignment path $r_\times(t)$ and the score statistics are more complicated than for uncorrelated sequences,

since in addition to the random matches, there are now the native matches along the evolution path $R(t)$. Due to these competing score contributions, the problem seems to be beyond the means of even an approximate analytical approach. However, it turns out that the statistics of *weakly* correlated sequences (in a sense defined below) is described with remarkable accuracy by the scaling theory developed in the previous Section.

Consider a pair of correlated sequences of length $N \gg 1$ with an optimal alignment of finite fidelity $\mathcal{F} > 0$ at a given value of γ. Since the optimal alignment path $r_\times(t)$ and the evolution path $R(t)$ have a finite fraction of common bonds, the displacement fluctuations of $r_\times(t)$ remain confined to a "corridor" centered around the path $R(t)$ (see Fig. 2(c)). The width r_c of this corridor can be defined by the mean square *relative* displacement

$$r_c^2 \equiv \overline{(r_\times(t) - R(t))^2}\,, \tag{18}$$

averaged over an ensemble of mutually correlated sequences with evolution parameters U, q. By Eq. (6), we can associate a longitudinal scale $t_c = r_c^{3/2}\, t_0(\gamma)$ with r_c. t_c describes the characteristic interval in t between intersections of the alignment path and the evolution path. In other words, these two paths form "bubbles" of typical width r_c and length t_c; see Fig. 2.

Alignments between mutually correlated sequences produce an average score larger or equal to the average score for uncorrelated sequences at the same value of γ. This score gain is due to the native pairs contained in the alignment and is defined as $\delta S(t;\gamma,U,q) \equiv \overline{S_\times(t;\gamma,U,q)} - E_0(\gamma)\, t$, where $E_0(\gamma)$ is the coefficient function in (8).

Conjecture 5 *The score gain over uncorrelated sequences has the asymptotic form*

$$\delta S(t;\gamma,U,q) \simeq \delta E(\gamma,U,q) \cdot t \tag{19}$$

for $t \gg t_c$, with $\delta E(\gamma,U,q) > 0$.

Remarks to Conjecture 5:

(i) This conjecture says that the scale t_c is a *correlation length*; i.e., points t_1 and t_2 on the alignment path are essentially uncorrelated if $|t_2 - t_1| \gg t_c$. (This property can be shown for closely related physics problems.) In the regime $t \gg t_c$, the fidelity and the width r_c thus become asymptotically independent of t. The score gain $\delta S(t)$ accumulates contributions from uncorrelated regions along the alignment path, leading to linear behavior.

(ii) The ensemble average can be generated from a single pair of sequences with $N, N' \gg t_c$.

We have verified the asymptotic linearity of $\delta S(t)$; see Fig. 9(a). The γ-dependence of δE at fixed evolution parameters is shown in Fig. 9(b) (plotted as a function of $C(\gamma)$ rather than γ). It is seen to be closely related to that of the fidelity, also shown in Fig. 9(b). *This makes the score gain*, and not the total score, *the most important alignment observable.* The common parameter dependence of δE and $\mathcal{F}$ can be understood rather systematically in the framework of scaling theory, to which we now turn.

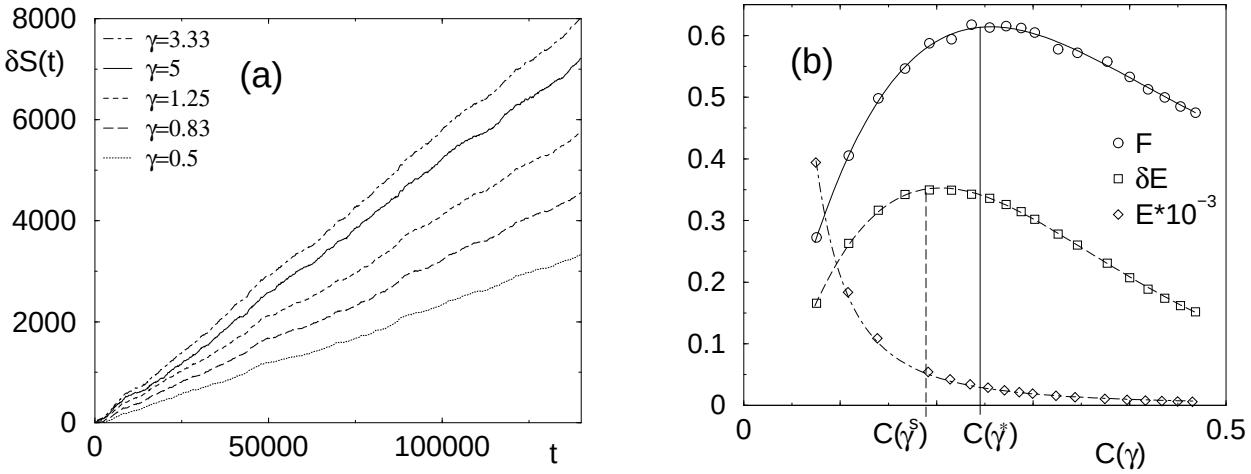

Fig. 9: (a) The score gain over uncorrelated sequences, $\delta S(t;\gamma,U,q)$ as a function of t for several γ, obtained from a single pair of sequences with mutual correlations ($U = 0.33$, $q = 0.11$). The slopes clearly depend non-monotonically on γ. (b) The fidelity $\mathcal{F}$, the score gain per element δE, and the total score per element $E \equiv S_\times(t)/t$ as functions of $C(\gamma)$. $\mathcal{F}$ and δE have maxima at close by parameter values γ^* and γ^s, respectively. The fidelity at the point of maximal score gain, $\mathcal{F}(\gamma^s)$, is very close to the fidelity maximum $\mathcal{F}(\gamma^*)$. These optimal parameters cannot be inferred from the parameter dependence of the total score E.

Scaling theory for correlated sequences

There is a considerable amount of alignment data even for the simple scoring function and evolution model considered in this paper. The fidelity $\mathcal{F}(\gamma,U,q)$ and the score gain $\delta E(\gamma,U,q)$ can be shown as functions of $C(\gamma)$ like in Fig. 9(b) for each U and q. However, for *weakly* correlated sequences (e.g., $U \ll 1$, such that $r_c \gg 1$), the alignment data can in fact be presented in a simpler way. The simplification is due to a relationship between these data at different values of the alignment and evolution parameters. This relationship can be exhibited by using a scaled gap strength $x \equiv C(\gamma)/U$ and a scaled indel frequency $y \equiv q/U^2$:

Conjecture 6 *For long and weakly correlated sequences ($t \gg t_c \gg t_0(\gamma)$), the fidelity and the score gain take the form*

$$\delta E(\gamma,U,q)/U \simeq \varepsilon(x,y)\ , \quad \mathcal{F}(\gamma,U,q) \simeq f(x,y)\ . \tag{20}$$

Remarks to Conjecture 6:

(i) This conjecture is valid if $t \gg t_c$ (so that the score gain becomes linear in t according to Conjecture 5) *and* $t_c \gg t_0(\gamma)$. The latter condition says that there are many gaps in a correlation interval, i.e., $r_c \gg 1$. The numerics shows that scaling sets in already for r_c of order 1. For $t \gg t_0(\gamma) \gg t_c$, the score gain is still linear according to Conjecture 5, and $r_c \ll 1$. This case can be treated by the statistics of gapless alignments but is never realized for weakly correlated sequences.

(ii) The scaling form (20) can be understood as an asymptotic invariance property of alignment ensemble averages. Consider the scale transformations[2] $t \to b^{-1}t$, which change the length of the alignment path by a factor $b \ll T$. We require the fidelity $\mathcal{F}$ and the score gain δS to remain invariant. By Conjecture 5, this requires $\delta E \to b \cdot \delta E$. The discussion in Appendix E further indicates specific transformation rules for the parameters U, q, and γ: $U \to b \cdot U$, $C \to b \cdot C$, and $q \to b^2 \cdot q$. Parametrizing γ in (20) by the coefficient $C(\gamma)$, we have

$$\mathcal{F}(C^{-1}, U, q) = \mathcal{F}(b^{-1} \cdot C^{-1}; b \cdot U, b^2 \cdot q) \ , \quad \delta E(C^{-1}, U, q) = b^{-1} \cdot \delta E(b^{-1} \cdot C^{-1}; b \cdot U, b^2 \cdot q) \ . \tag{21}$$

By choosing $b = 1/U$, we recover the scaling form (20), with $f(x, y) = \mathcal{F}(x^{-1}, 1, y)$ and $\varepsilon(x, y) = \delta E(x^{-1}, 1, y)/U$.

According to Conjecture 6, the scaled score gain ε and the fidelity f can be represented as *one-parameter* families of functions of the variable x, parametrized by the variable y. That this is indeed the case can be seen from Fig. 10(a,b) for numerical data obtained from single sequence pairs with various values of U, q and γ. As expected from (20), the data for different parameter sets (γ, U, q) corresponding to the same (x, y) collapse approximately. This data collapse will be useful for similarity detection.

Alignment parameter optimization

The numerical fidelity and score patterns of Fig. 10(a,b) have clear maxima $f^*(y) \equiv f(x^*(y), y)$ and $\varepsilon^s(y) \equiv \varepsilon(x^s(y), y)$, attained at closeby points $x^*(y)$ and $x^s(y)$ respectively. Most importantly, the fidelity evaluated at the point of maximal score gain, $f(x^s(y), y)$, is very close to the maximum $f^*(y)$; see Fig. 11 and the example of of Fig. 9(b). For a given sequence pair, the corresponding alignments are typically very similar. We conclude that *the fidelity can be optimized efficiently by maximalization of the score gain* δE.

This optimization rule can be understood as a geometric criterion in accordance with the qualitative picture of Section 2. To see this, we compare the fluctuations $\overline{R^2(t)}$ of the evolution path for *correlated* sequences with the fluctuations $\overline{r_\times^2(t)}$ of the optimal alignment path for *uncorrelated* sequences. Equating the mean square displacements, we obtain a pair of characteristic scales $\widetilde{r}$ and $\widetilde{t}$, i.e., $\overline{R^2\left(\widetilde{t}\right)} = \overline{r_\times^2\left(\widetilde{t}\right)} \equiv \widetilde{r}^2$. From Eqs. (6) and (17), we obtain

$$\widetilde{t}(\gamma, q) = q^3/A^6(\gamma) \ , \qquad \widetilde{r}(\gamma, q) = q^2/A^3(\gamma) \ . \tag{22}$$

We call these scales the *roughness matching* scales. For $|t| < \widetilde{t}(\gamma, q)$, the displacement of the evolution path exceeds that of the optimal alignment path, while for $|t| > \widetilde{t}(\gamma, q)$, the displacement of the alignment path becomes dominant.

The definition of the roughness matching scales $\widetilde{t}$ and $\widetilde{r}$ does not involve the confinement scales r_c and t_c. However, the two sets of scales are related at the optimal parameter values

[2] Such transformations make sense only in the regime $t \gg t_0(\gamma)$, where r and t can be treated as continuum variables. This is precisely where Conjecture 6 is valid.

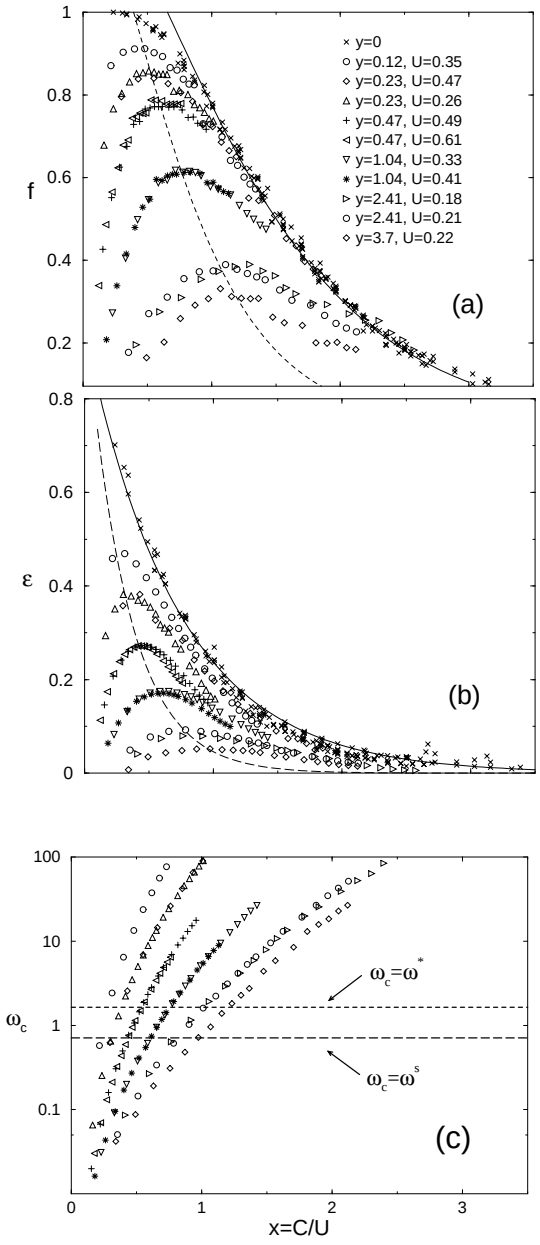

Fig. 10: (a) Fidelity $f(x,y)$ and (b) scaled score gain $\varepsilon(x,y)$ obtained from single sequence pairs with various evolution parameters U, q and alignment parameters γ. The data for different (U,q,γ) corresponding to the same $x = C(\gamma)/U$, $y = q/U^2$ collapse approximately, as predicted by the scaling theory. The lines are the theoretical loci of the maxima $(x^*(y), f^*(y))$ (short-dashed), $(x^s(y), \varepsilon^s(y))$ (long-dashed) and the theoretical limit curves $f(x,0)$, $\varepsilon(x,0)$ (solid) for the case $q = 0$; see Appendix E. (c) The geometrical ratio $\omega_c(x,y)$ given by Eq. (23) vs. x. The optimal values $x^*(y)$ and $x^s(y)$ defining the maxima of the curves in (a) and (b) are given approximately by the intersections of the curves $\omega_c(x,y)$ with the lines $\omega_c \approx \omega^*$ and $\omega_c \approx \omega^s$, respectively.

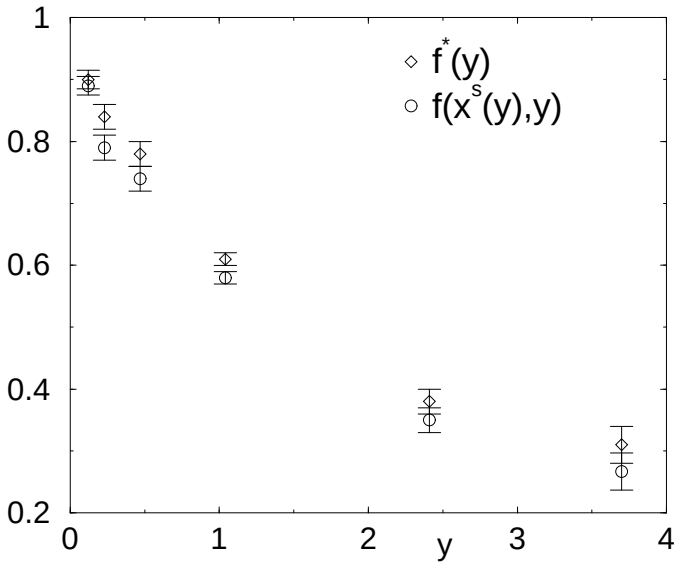

Fig. 11: Fidelity optimization by maximalization of the score gain. The fidelity maximum $f^*(y)$ is very close to the fidelity at the point of maximal score gain, $f(x^s(y), y)$.

as we now show. Noting that $\tilde{r}$ can be written in scaling form, $\tilde{r} = y^2/x^4$, we can define the geometric ratio

$$\omega_c(x, y) \equiv \frac{r_c(x, y)}{\tilde{r}(x, y)} = \frac{r_c(x, y)\, x^4}{y^2} \,. \tag{23}$$

Over the relevant parameter regime, ω_c is a monotonically increasing function of x; see the numerical data of Fig. 10(c). Comparison with Figs. 10(a,b) shows that the optimal values $x^*(y)$ and $x^s(y)$ are given by the closely related conditions $\omega_c \approx \omega^*$ and $\omega_c \approx \omega^s$, respectively, where $\omega^* \approx \omega^s$ are two parameter-independent constants of order 1. The first condition is readily interpreted in terms of the path morphology discussed in Section 2: The confinement length is proportional to the roughness matching scale (22) at the optimal alignment parameter. In other words, at $x = x^*(y)$ (i.e., $\omega_c = \omega^*$), the fluctuations of the optimal alignment path $r_\times(t)$ just match those of the evolution path $R(t)$ (see Fig. 3(b)). The shortcut regime (Fig. 3(c)) corresponds to the ascending branch ($x < x^*(y)$, i.e., $\omega_c < \omega^*$) of the fidelity curves in Fig. 11(a), while the random fluctuation regime (Fig. 3(a)) corresponds to the descending branch ($x > x^*(y)$, i.e., $\omega_c > \omega^*$).

For the simple evolution model and scoring function studied in this paper, the fidelity and score patterns of Fig. 10(a,b) can even be predicted theoretically (see Appendix E). This will certainly become more difficult for models with more parameters. However, the geometrical link between the maxima of the fidelity and of the score gain is expected to be preserved. This has indeed been found for local alignments (Olsen, Hwa, and Lässig, 1999a) and for probabilistic alignments used for maximum likelihood inference (Kschischo and Lässig, 1998).

Similarity detection

The evolution process used in this paper is closely related to a more realistic process for the divergent evolution of *two* daughter sequences $Q^{(1)}$ and $Q^{(2)}$ from a closest common ancestor sequence Q. Modeling the two evolution paths as independent Markov processes with respective parameters U_1, q_1 and U_2, q_2, one can show that the evolution path linking $Q^{(1)}$ and $Q^{(2)}$ is again a Markov process with parameters $U = U_1 U_2$ and $q = q_1 + q_2 + O(q^2)$.

For practical alignments, however, the evolutionary parameters U and q are unknown. Since they enter the definition of the basic variables x and y, knowledge of the optimal parameters $x^*(y)$ and $x^s(y)$ seems to be of little use for applications. However, these parameters can be reconstructed from alignment data, as we will now show for a specific example.

Consider three sequences $Q^{(1)}$, $Q^{(2)}$ and $Q^{(3)}$ related by the evolution tree of Fig. 12(a). The evolutionary distances τ_i are defined in terms of the mutual similarity coefficients U_{ij} by

$$-\log U_{ij} = \tau_i + \tau_j \qquad (i, j = 1, 2, 3)\ . \tag{24}$$

We whish to determine τ_1, τ_2 and τ_3 from pairwise alignments of the sequences[3]. Fig. 12(b) shows the alignment data δE_{ij} as defined in Eq. (19) for each of these pairs, plotted as a function of $C(\gamma)$. To fit the data curve $\delta E_{ij}(C)$ to the corresponding scaled score gain curve $\varepsilon_{ij}(x)$ of Fig. 10(b), we have to divide both axes of the diagram by U_{ij}. In this way, we can determine the *a priori* unknown factors U_{ij}, and hence the evolutionary distances τ_i, see Fig. 12(b). For this example, we obtain $U_{12} \approx 0.54$, $U_{13} \approx 0.43$, $U_{23} \approx 0.415$, and $\tau_1 \approx 0.22$, $\tau_2 \approx 0.33$, $\tau_3 \approx 0.55$, which is to be compared with the actual values $\tau_1 = 0.27$, $\tau_2 = 0.38$, and $\tau_3 = 0.61$ used to produce the sequences.

Finally, high-fidelity pairwise alignments of these sequences are found for parameters $\gamma^*_{ij} \approx \gamma^s_{ij}$ as expected from the above (see Appendix E).

5 Discussion

We have presented a statistical scaling theory for global gapped alignments. Alignments of mutually uncorrelated sequences are found to be governed by a number of *universal scaling laws*: ensemble averages such as the mean square displacement of the alignment path or the variance of the optimal score follow power laws whose exponents do not depend on the scoring parameters. The parameter dependence is contained entirely in the prefactors. This universality is comparable to the diffusion law describing a large variety of random walk processes on large scales, the only parameter dependence being the value of the diffusion constant. In contrast to diffusive random walks, however, we find optimal alignment paths to be strongly non-Markovian on all length scales due to random agglomerations of matches and mismatches. Hence, the exponents take nontrivial values. The scaling laws also govern the displacement statistics of the optimal path $r_\times(t)$ of a single pairwise alignment, and the

[3]In this example, we use effective indel rates $-\log(1 - q_{ij}) = \Gamma(\tau_i + \tau_j)$ with $\Gamma = 0.2$, but this choice is not crucial.

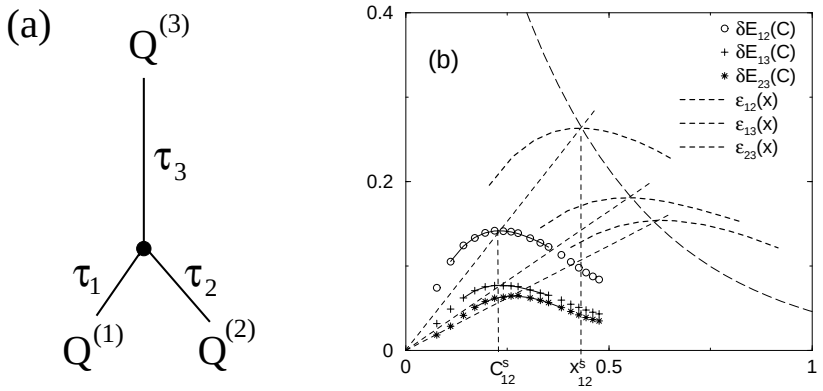

Fig. 12: (a) Evolution tree linking three sequences $Q^{(1)}$, $Q^{(2)}$, and $Q^{(3)}$. The sequences have evolutionary distances τ_1, τ_2, and τ_3 to the branching point of the tree, as defined by Eq. (24), and have lengths $N_1 \approx N_2 \approx N_3 \approx 5000$. (b) Alignment data δE_{12}, δE_{13} and δE_{23} for pairwise alignments of the sequences at different values of γ, shown as a function of $C(\gamma)$. ε_{12}, ε_{13}, and ε_{23} obtained by rescaling the raw alignment data by respective factors U_{12}, U_{13}, and U_{23} such that the maxima of the rescaled curves fall on the theoretical locus $(x^s(y), \varepsilon^s(y))$ (long-dashed curve, cf. Fig. 10(b)). This determines the *a priori* unknown similarity coefficients U_{ij}, and hence the evolutionary distances τ_i.

associated statistics of scores $S(r,t)$. These properties makes the concepts discussed here applicable to individual alignment problems.

The scaling theory is also relevant for the statistics of mutually correlated sequence pairs. Two important quantities are the *score gain* over uncorrelated sequences and the alignment *fidelity*. Both quantities strongly depend on the evolutionary parameters linking the two sequences and on the alignment parameters. For a simple Markovian evolution model and for linear scoring functions, we have obtained a quantitative description of this parameter dependence. In particular, the alignment parameter of maximal fidelity turns out to be closely related to the parameter of maximal score gain, which makes it possible to construct the alignment of maximal fidelity from a systematic analysis of score data. Moreover, the underlying evolutionary parameters (the mutual similarity U and the effective indel rate q) can also be inferred from this analysis.

It is important to understand inhowfar the results of this paper carry over to more refined algorithms for the alignment of realistic sequences. The universal scaling laws for uncorrelated sequences should prove to be very robust under changes of the scoring function (such as scoring matrices distinguishing between transitions and transversions) as well as changes in the sequences (the number of different letters and their frequencies). As corroborated by preliminary numerical results, such changes reduce to a different parameter dependence of the amplitude functions A, B, C, and D. In particular, we find the universal scaling laws to be preserved for the alignment of bona fide uncorrelated cDNA sequences, which also validates the Markov model for single sequences. While not affecting the asymptotic universality, some scoring functions (for example, systems with affine gap cost distinguishing between gap initiation and gap extension) may introduce intermediate regimes where the

score and fidelity curves are modified. Nevertheless, the fidelity and the score gain remain key quantities of an alignment, and their optimal values are closely related. This makes it possible to construct optimal alignments on the basis of a statistical analysis of score data. This link and the underlying scaling theory are also crucial to the analysis of local alignment algorithms, as we have shown recently (Hwa and Lässig, 1997; Drasdo *et al.*, 1998).

Acknowledgments. The authors are grateful to Stephen Altschul, Steven Benner, Ralf Bundschuh, Richard Durbin, Martin Vingron, and Michael Waterman for conversations and suggestions. TH acknowledges the financial support of an A. P. Sloan Research Fellowship, an Arnold and Mabel Beckman Foundation Young Investigator Award, and the hospitality of the Max-Planck Institute at Teltow where much of the work was carried out. DD acknowledges the financial support of the grant no. 342/4-3 from the Deutsche Forschungsgemeinschaft.

Appendix A: Evolution model

The Markov process governing the evolution of a daughter sequence Q' from an ancestor sequence Q is specified by the flux diagram of Fig. 13.

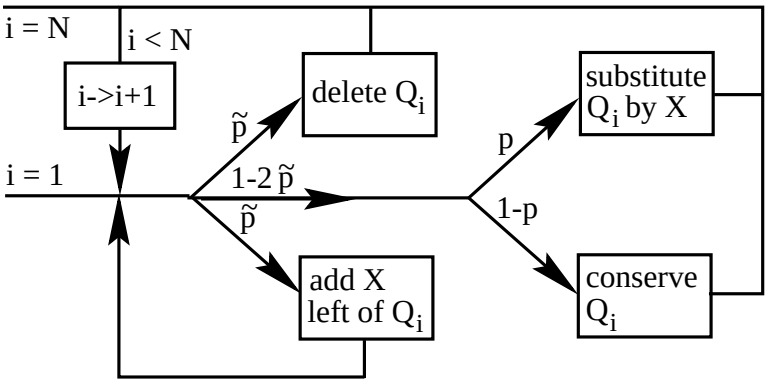

Fig. 13: Flux diagram of the Markov evolution process. A realization generates a daughter sequence $Q' = \{Q'_j\}$ from an ancestor sequence $Q = \{Q_i\}$. The process is characterized by the insertion/deletion probability $\tilde{p}$ and the substitution probability p. X denotes a random letter.

The statistical properties of this Markov process are straightforward to compute. Using the notation $t \equiv i + j$ and $R \equiv j - i$, we find $R(t)$ is asymptotically a Gaussian random variable with

$$\overline{R(t)} = 0 \,, \qquad \overline{R^2(t)} = qt \,, \tag{25}$$

where q is given by Eq. (2). This implies in particular that the length N' of the daughter sequence is also a Gaussian random variable with

$$\overline{N'} = N \,, \qquad \overline{(N' - N)^2} = 2qN \,. \tag{26}$$

To show Eqs. (25), we start from the recursion relation

$$w(R, t+1) = \tilde{p}\,(w(R-1,t) + w(R+1,t)) + (1 - 2\tilde{p})w(R, t-1), \tag{27}$$

where $w(R,t)$ is, up to a normalization factor, the probability to find the evolution path at position R for a given t. Asymptotically this can be replaced by a differential relation in a Kramers-Moyal-expansion in R and t (see, e.g., Risken, 1989),

$$\begin{aligned} w(R,t) + \frac{\partial w(R,t)}{\partial t} \;\approx\; & \tilde{p}\left(w(R,t) - \frac{\partial w(R,t)}{\partial R} + \frac{1}{2}\frac{\partial^2 w(R,t)}{\partial R^2}\right. \\ & \left. + w(R,t) + \frac{\partial w(R,t)}{\partial t} + \frac{1}{2}\frac{\partial^2 w(R,t)}{\partial R^2}\right) \\ & + (1 - 2\tilde{p})(w(R,t) - \frac{\partial w(R,t)}{\partial t}), \end{aligned} \tag{28}$$

which reduces to

$$\frac{\partial w(R,t)}{\partial t} = q\,\frac{\partial^2 w(R,t)}{\partial R^2}. \tag{29}$$

with q given by Eq. (2). small this corresponds to the For the initial condition describing *rooted* evolution paths, i.e., $R(t=0)=0$, the solution of (29) is indeed a Gaussian with the moments (25).

Appendix B: Scoring function

Given a three-parameter scoring function S of the form (3), the *optimal* global alignment of two sequences Q and Q' remains invariant under the linear transformations

$$S \to aS + b \quad (a > 0)\ . \tag{30}$$

This shows that the optimal global alignment depends only on a single effective parameter. Written in terms of the scoring parameters, the transformations (30) read

$$\mu_\pm \to a\mu_\pm + 2b'\ , \quad \mu_g \to a\mu_g + b' \tag{31}$$

with $b = Nb'$. To arrive at the normal form (4) of the scoring function used in this paper, we compute the score average m and variance v^2 of a pairing of random elements,

$$m = \frac{1}{c}\mu_+ + \frac{c-1}{c}\mu_-\ , \tag{32}$$

$$v^2 = \frac{1}{c}\mu_+^2 + \frac{c-1}{c}\mu_-^2 - m^2\ , \tag{33}$$

and choose $a = 1/v$ and $2b' = -m/v$. Hence, (4) is normalized in such a way that a pairing of two random elements has average score 0 and score variance 1. Expressed in terms of the original scoring parameters, the effective gap cost is

$$\gamma = \frac{1}{v}\mu_g - \frac{m}{2v}\ . \tag{34}$$

Appendix C: Alignment Algorithm

The dynamic programming algorithm generates the score landscape $S(r,t)$ for all grid points by the recursion relation

$$S(r,t) = \max\left\{ \begin{array}{l} S(r-1,t-1) - \gamma \\ S(r+1,t-1) - \gamma \\ S(r,t-2) + s(r,t) \end{array} \right\} \tag{35}$$

with

$$s(r,t) = \left\{ \begin{array}{ll} \sqrt{c-1} & \text{if} \quad Q'_{(r+t)/2} = Q_{(r-t)/2} \\ -\frac{1}{\sqrt{c-1}} & \text{if} \quad Q'_{(r+t)/2} \neq Q_{(r-t)/2} \end{array} \right. . \tag{36}$$

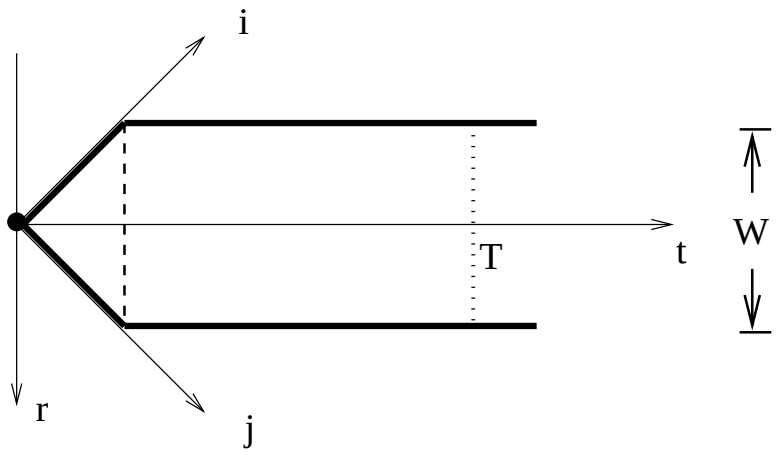

Fig. 14: Restricted alignment grid (bounded by thick lines) used for the evaluation of the recursion relation (35). With initial condition (i), the alignment paths are pinned at their initial point (dot) defined to be at $t = 0$. With initial condition (ii), the score is prescribed along the dashed line defined to be at $t = 0$, namely $S(r, t{=}0) = 0$.

This recursion relation is evaluated on a restricted alignment grid shown in Fig. 14, which limits the computing time to a value $\sim T \times W$. The width of the strip is chosen according to the specific tasks (see below). Across the strip, we use periodic boundary conditions, i.e., $S(r - W/2, t) = S(r + W/2, t)$. (Similar results are obtained for open boundary condition.)

Two types of initial conditions are used in the text:

(i) $r(t = 0) = 0$ (with $t \equiv i + j$), corresponding to alignment paths *rooted* at the point $(r = 0, t = 0)$.

(ii) $S(r, t{=}0) = 0$ for $-W/2 < r < W/2$ (with $t \equiv i + j - W/2$), corresponding to *unrooted* alignment paths starting at an arbitrary point $(r, t = 0)$.

Evaluation of the recursion relation stops at $t = T$. Hence, the optimal alignment path $r_\times(t)$ ends at the point $x \equiv r_\times(T)$ given by $S(x, T) = S_\times(t) \equiv \max_r S(r, T)$. If this maximum occurs for different values of x, one of them is chosen at random. The entire path $r_\times(t)$ is then found by backtracking it from its endpoint x. Degeneracies are again resolved by a random choices. This is justified since degenerate optimal paths have a typical distance of order 1 only.

To compute the unconstrained fluctuations of optimal alignments for uncorrelated sequences, W has to be sufficiently large so that the result becomes independent of it: The necessary condition is $W \gg \Delta_r(t)$. The mean square displacement $\Delta_r^2(t)$ and the tilt cost $E_t(\theta)$ are evaluated with initial condition (i); in the latter case, also the endpoint $x = \theta \cdot T$ is pinned. The mean square score differences $\Delta_S^2(t)$ and $C_S(\rho, t)$ are computed with initial condition (ii). On the other hand, the confinement cost $E_c(W)$ is determined by choosing $W \ll \Delta_r(t)$ so that the result becomes independent of T and of the initial condition.

For correlated sequences, we again choose W large enough, i.e., $W^2 \gg \overline{(\Delta R(T))^2} + r_c^2$, so that the result becomes independent of it. For $T \gg t_c$, quantities defined per unit of t such as $\mathcal{F}$ and δE will also become independent of the initial condition.

Appendix D: Statistics of the score landscape

For simplicity, we discuss the score landscape not for the full alignment grid but for a strip region $-W/2 < r < W/2$ with $W \gg \Delta_r(t)$ and initial conditions corresponding to unrooted alignment paths (see Appendix C). The ensemble averages then become invariant under translations of r and can, hence, be evaluated efficiently as averages over r. For example, the mean square score difference $C_S(\rho,t) \equiv \overline{(S(r+\rho,t) - S(r,t))^2}$ for arbitrary r and $\rho > 0$ is given by

$$C_S(\rho,t) \simeq W^{-1} \sum_{r=-W/2}^{r=W/2} (S(r+\rho,t) - S(r,t))^2 \tag{37}$$

for $W \to \infty$.

The following conjecture is an extension of Conjecture 2, which describes the score landscape in more detail.

Conjecture 7 *The mean square score difference $C_S(\rho,t)$ for mutually uncorrelated sequences has the asymptotic form*

$$C_S(\rho,t) \simeq \hat{B}^2(\gamma) t^{2/3}\, g\left[\rho/\Delta_r(t)\right] + \alpha(\gamma) \cdot \min(\rho,t) \tag{38}$$

valid for $t \gg t_0(\gamma)$. The scaling function $g[x]$ is normalized such that $g[1] = 1$; it has the asymptotics $g[x] = g_1 x$ for $x \ll 1$ and $g[x] \simeq g_\infty$ for $x \gg 1$, with g_1 and g_∞ being constants of order 1. The correction term has a coefficient $\alpha(\gamma) \ll 1$.

Remarks to Conjecture 7:

(i) For small distances $\rho \ll \Delta_r(t)$, Conjecture 7 gives

$$C_S(\rho,t) \simeq \left[g_1 B^2(\gamma)/A(\gamma) + \alpha(\gamma)\right] \rho\,. \tag{39}$$

Since the first term turns out to be larger than 1 and $\alpha(\gamma) \ll 1$ for all γ, the α term is always negligible. The asymptotic linearity $C_S(\rho,t) \sim \rho$ has been proved recently for a version of the LCS problem corresponding to $\gamma = \gamma_0$ (Bundschuh and Hwa, 1999).

(ii) For $\rho = \Delta_r(t)$, Conjecture 7 reduces to Conjecture 2,

$$\Delta_S^2(t) \equiv C_S(\Delta_r(t),t) \simeq B^2(\gamma)\, t^{2/3}\,. \tag{40}$$

with $B^2(\gamma) = \hat{B}^2(\gamma) + \alpha(\gamma) A(\gamma) \approx \hat{B}^2(\gamma)$; the α term is again negligible against the scaling term.

(iii) The correction term becomes visible only for large distances $\rho \gg \Delta_r(t)$, i.e., for paths with no element pairs in common. Consider, in particular, the mean square score difference $C_S(\rho,t)$ for distances $\rho > t$. Since the corresponding optimal paths are statistically independent, this reduces to twice the single-point score variance,

$$C_S(\rho > t, t) = 2\mathrm{Var}[S(r,t)] \equiv 2\overline{S^2(r,t)} - 2\overline{S(r,t)}^2\,. \tag{41}$$

According to Conjecture 7, we have

$$2\mathrm{Var}[S(r,t)] \simeq \alpha(\gamma)t + g_\infty \hat{B}^2(\gamma) t^{2/3} \; . \tag{42}$$

Hence, the α term will eventually dominate the scaling term for sufficiently large t. Eq. (42) also describes the variance of the optimal score, $\mathrm{Var}[S_\times(t)]$. For $\gamma < \gamma_0$, in particular, $S_\times(t)$ is linear in the length $L(t)$ of the LCS. Hence,

$$\mathrm{Var}[L(t)] \simeq \alpha_0 t + O(t^{2/3}) \tag{43}$$

with $\alpha_0 \equiv \alpha(\gamma)/(\gamma + \sqrt{c-1}/2)^2$. The linear asymptotics is in agreement with the rigorous bound by Steele (1982). Chvátal and Sankoff (1975) had conjectured $\mathrm{Var}[L(t)]$ to be of order $O(t^{2/3})$. Indeed, the $O(t^{2/3})$ term in (43) turns out to remain dominant even for $t \approx 10^4$ since $\alpha_0 \ll 1$.

(iv) We emphasize again that the α term in Conjecture 7 is *spurious*, i.e, it does not affect the scaling of the optimal alignment path. The reason is that finding the optimal rooted path $r_\times(t)$ amounts to evaluating *score differences* of paths within a distance $\rho \lesssim \Delta_r(t)$, where the α term is negligible according to (39) and (40). This does not contradict the asymptotic dominance of this term for the *single-point* variance (38). Indeed, the existence of spurious contributions to the single-point score is easy to understand. Consider, for example, changing the potential by a (fictitious) amount depending only on the sequence Q but not on Q', i.e., $s(r,t) \to s(r,t) + \tilde{s}((r+t)/2)$. This changes the score of any alignment containing all elements of Q, $S \to S + \sum_{i=1}^{N} \tilde{s}(i)$. However, since this shift is the same for all such alignments, all score differences remain invariant, and so does the optimal path $r_\times(t)$. The α terms above turn out to be generated by a similar mechanism which can be traced back to correlations between the random variables $s(r,t)$; see Drasdo, Hwa, and Lässig (1999).

Conjecture 7 has also been verified numerically. The log-log plot of $C_S(\rho,t)$ for several t and $\gamma = \gamma_0$ is shown in Fig. 15(a). The rescaled data $C_S(\rho,t)/B^2 t^{2/3}$ plotted as functions of the rescaled variable $x \equiv \rho/\Delta_r(t)$ collapse for $x \lesssim 1$ to a single function $g(x)$ (see Fig. 15(b)), as predicted by Conjecture 7 with the α term neglected. This term is visible only for larger values of ρ. Plotting C_S/t vs. ρ/t (Fig. 15(c)) exhibits its functional form $\alpha \min(r,t)$ and determines $\alpha(\gamma_0) \approx 0.012$ from the slope of the ascending straight lines. The saturation value reached for $\rho/t > 1$ gives the single point score variance $\Delta_S(t)$, which is seen to follow Eq. (42). That $\Delta_S(t)$ grows faster than $t^{2/3}$ has recently been noted by Boutet de Monvel (1999), who erroneously attributed it to a new asymptotic regime $\Delta_S(t) \sim t^{0.836}$. The data for larger values of γ look similar to those in Fig. 15. However, $\alpha(\gamma)$ is found to be a rapidly decreasing function of γ, rendering the α term unobservable for $\gamma \gtrsim 2$. For details, see Drasdo, Hwa, and Lässig (1999).

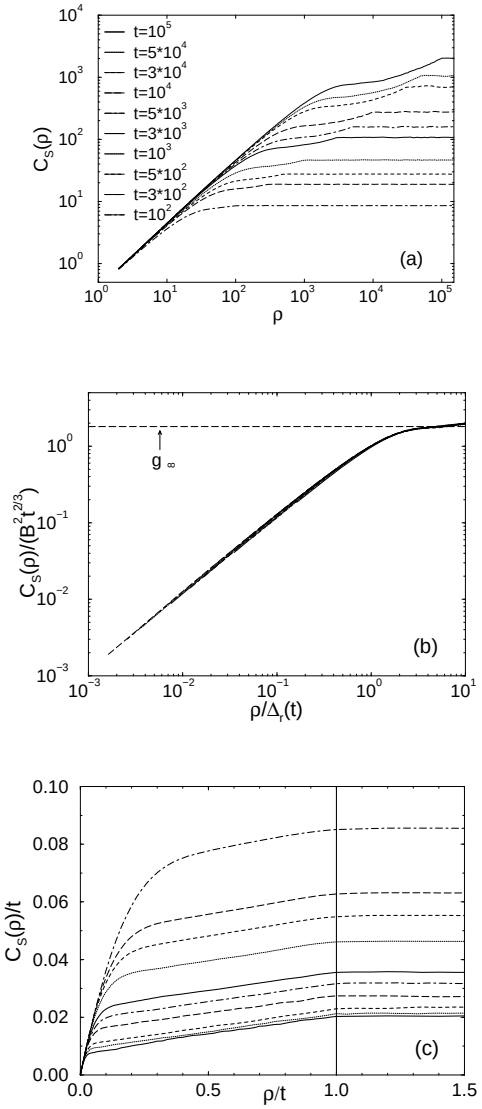

Fig. 15: (a) The mean square score difference $C_S(\rho,t)$ as a function of ρ for several t and $\gamma = 0.25 < \gamma_0$. (b) The rescaled data $C_S(\rho,t)/B^2t^{2/3}$ plotted as functions of the rescaled variable $x \equiv \rho/\Delta_r(t)$ collapse for $x \lesssim 1$ to a single function $g(x)$. (c) The rescaled data $C_S(\rho,t)/t$ plotted as functions of the rescaled variable ρ/t show the asymptotic form $C_S(\rho,t) \simeq \alpha(\gamma)\min(\rho,t) + O(t^{2/3})$ for $\rho \gg \Delta_r(t)$; we obtain $\alpha(0.25) \approx 0.012$.

Appendix E: Variation Theory and Alignment Parameter Optimization

Given the evolution parameters U, q and the alignment parameter γ, the confinement length r_c and the score gain δE can be calculated approximately in a "variational approach", treating r_c as an *independent* continuum variable to be determined a posteriori from an extremal condition. We assume that mutual correlations act as a constraint on the displacement fluctuations of the alignment path, producing a tilt cost E_t and a confinement cost E_c as discussed in Section 3. These costs must be outweighed by the score gain due to native matches in order to produce a net gain $\delta E > 0$. The different score contributions take the following forms:

(i) If the optimal alignment path $r_\times(t)$ is confined to a corridor of width r_c around the fluctuating path $R(t)$, then at the scale t_c, $r_\times(t)$ has a typical tilt of $\theta \sim \overline{R^2(t_c)}^{1/2}/t_c = q/r_c$ with respect to the main diagonal of the alignment grid, implying a tilt cost

$$E_t(r_c; q, \gamma) \sim -D(\gamma) \left(\frac{q}{r_c}\right)^2 \tag{44}$$

according to Conjecture 4.

(ii) The confinement cost to an untilted corridor of width r_c is $E_c = C(\gamma)/r_c$. The tilt reduces the effective width of the corridor so that the confinement cost takes the form

$$E_c(r_c; q, \gamma) \sim -C(\gamma) \frac{1 + q/[C^2(\gamma) r_c]}{r_c} \,. \tag{45}$$

On the other hand, the gain in score per unit of t due to the native matches is simply $E_n = U \cdot \mathcal{F}$, as it is clear from the definition of the fidelity $\mathcal{F}$. We need to express $\mathcal{F}$ in terms of r_c. Naively one would expect $\mathcal{F} \sim 1/r_c$. A detailed analysis shows that this is correct up to a logarithmic correction (Hwa and Nattermann, 1995, Kinzelbach and Lässig, 1995, Hwa and Lässig, 1996) leading to

$$E_n(r_c; U) \sim U \, \frac{1 + \log r_c}{r_c} \,. \tag{46}$$

The net score gain is the sum of these contributions, $\delta E = E_c + E_t + E_n > 0$. The alignment parameter enter the expressions (44), (45), and (46) only via the coefficients $C(\gamma)$ and $D(\gamma) \sim C^{-3}(\gamma)$. The scale transformations $E_c \to b \cdot E_c$, $E_t \to b \cdot E_t$, $E_n \to b \cdot E_n$ amount to the transformations $C \to b \cdot C$, $U \to b \cdot U$, $q \to b^2 \cdot q$, leading to the scaling form (20). Absorbing all unknown proportionality factors into the definition of the variables x, y, and ε, we obtain the scaled energy gain

$$\delta\mathcal{E}(r_c; x, y) \equiv \delta E/U = -\frac{x}{r_c} - \frac{y}{x}\left(1 + \frac{y}{x^2}\right)\frac{1}{r_c^2} + \frac{1 + \log r_c}{r_c} \,. \tag{47}$$

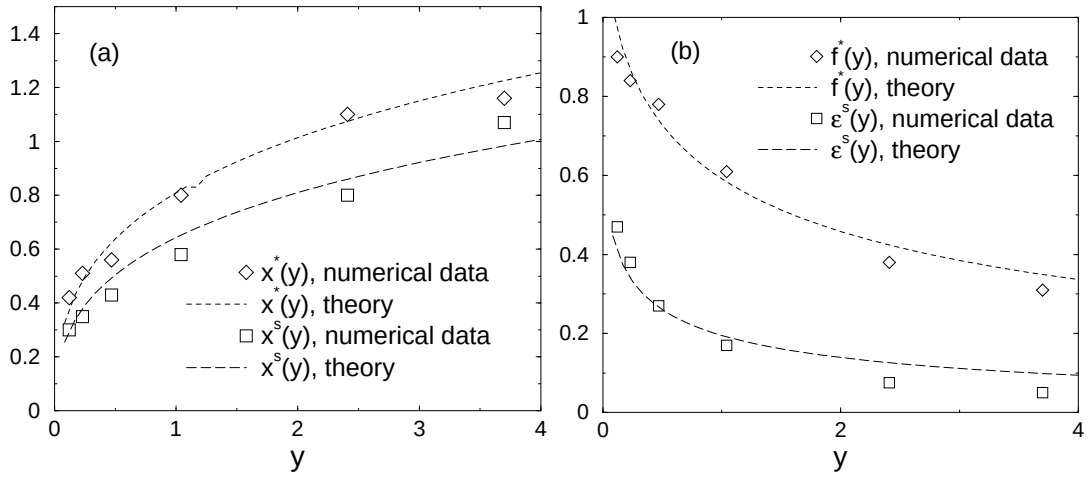

Fig. 16: Alignments of maximal fidelity and of maximal score gain. Theoretical predictions for the curves (a) $x^*(y)$, $x^s(y)$ and (b) $f^*(y)$, $\varepsilon^s(y)$, compared to numerical data obtained from fits to the curves of Fig. 7.

Maximizing (47) then determines the actual value of $r_c(x,y) = r_c(U,q,\gamma)$ by a variational principle:

$$\varepsilon(x,y) = \max_{r_c} \delta\mathcal{E}(r_c;x,y) \; . \tag{48}$$

The numerical solution of Eqs. (47), (48) produces loci of the fidelity and score gain maxima, $(x^*(y), f^*(y))$ and $(x^s(y), \varepsilon^s(y))$, as shown in Figs. 9(a,b) and 16. The theory is seen to predict the functional form of the sequence data in a reasonable way, except in the region $f \sim 1$ (i.e., $r_c \sim 1$) where the continuum approximation valid in the regime of *weak* similarity breaks down. (The unknown γ-independent proportionality factors for the scaling variables x, y, ε and for $\mathcal{F}$ have been determined by fits to the data.)

The functional dependences of Fig. 16 can be used to construct high-fidelity alignments and to estimate the fidelity maximum. For the evolution tree discussed in Section 4, we read off $C^s_{12} \approx 0.23$, $C^s_{13} \approx 0.225$, and $C^s_{23} \approx 0.254$ from Fig. 12(b) and use the approximate relations $C^*_{ij}/C^s_{ij} = x^*_{ij}/x^s_{ij} \approx 1.2$ for $0.1 < y < 4$ as well as the function $C(\gamma)$ discussed in Section 3 to obtain the optimal alignment parameters $\gamma^*_{12} \approx 1.52$, $\gamma^*_{13} \approx 1.59$, $\gamma^*_{23} \approx 1.25$. The scaled score maxima $\varepsilon^s_{12} \approx 0.26$, $\varepsilon^s_{13} \approx 0.18$, $\varepsilon^s_{23} \approx 0.15$ determine the expected fidelities $\mathcal{F}^*_{12} \approx 0.75$, $\mathcal{F}^*_{13} \approx 0.58$, $\mathcal{F}^*_{23} \approx 0.52$ as seen from Fig. 16(b). They are in good agreement with the actual maxima $\mathcal{F}^*_{12} = 0.8$, $\mathcal{F}^*_{13} = 0.65$, $\mathcal{F}^*_{23} = 0.55$ computed by comparing directly to the evolutionary paths.

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 (3)*, 403 – 10.

Altschul, S.F. 1993. A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol. 36 (3)*, 290 - 300.

Arratia, R., Morris, P. and Waterman, M.S. 1988. Stochastic scrabbles: a law of large numbers for sequence matching with scores. *J. Appl. Probab. 25* 106 – 19.

Arratia, R. and Waterman, M.S., 1994. A phase transition for the score in matching random sequences allowing deletions. *Ann of Appl. Prob. 4*, 200 – 25.

Benner, S.A., Cohen, M.A. and Gonnet, G.H. 1993. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol. 229 (4)*, 1065 – 82.

Bishop, M.J. and Thompson, E.A. 1986. Maximum likelihood alignment of DNA sequences. *J. Mol. Biol. 190 (2)*, 159 - 65.

Boutet de Monvel, J. 1999. Extensive simulations for longest common subsequences. Eur. Phys. J. B 7, 293 – 308.

Bundschuh, R. and Hwa, T. 1999. An Analytical Study of the Phase Transition Line in Local Sequence Alignment with Gaps. to appear in Proceedings of the 3rd Annual International Symposium on Computational Molecular Biology (RECOMB99).

Chvátal V. and Sankoff D. 1975. Longest common subsequences of two random sequences. J. Appl. Prob. 12, 306 – 315.

Cule, D. and Hwa, T. 1998. Static and Dynamic Properties of Inhomogeneous Elastic Media on Disordered Substrate. Phys. Rev. B. 57, 8235 – 8253.

De los Rios P. and Zhang Y.C. 1998. Directed polymers on a factorized disorder landscape. Phys. Rev. Lett. 81, 1083-6.

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. A statistical theory of sequence alignment with gaps. *Proceeding of The Sixth International Conference on Intelligent Systems for Molecular Biology*, J. Glasgow *et al.* eds, 52 – 58 (AAAI Press, Menlo Park).

Drasdo, D., Hwa, T. and Lässig, M. 1999. In preparation.

From MEDLINE; 96096722, cf 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.

From MEDLINE; 94215495, cf. 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.

Fisher, D.S. and Huse, D.A. 1991. Directed paths in a random potential. *Phys. Rev. B 43 (13)*, 10728 - 10742.

Hardy, P. and Waterman, M.S. 1997. *The sequence alignment software library at USC.* ¿From `http://www-hto.usc.edu/software/.`
Hwa, T. and Fisher, D.S. 1994. Anomalous fluctuations of directed polymers in random media. *Phys. Rev. B 49*, 3136 – 54.

Hwa, T. and Lässig, M. 1996. Similarity detection and localization, *Phys. Rev. Lett. 76*, 2591 - 2595.

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*, S. Istrail, P. Pevzner, and M.S. Waterman eds, 109-116 (ACM Press, 1998).

Hwa, T. and Nattermann, T. 1995. Disordered induced depinning transition, *Phys. Rev. B 51*, 455 - 469.

Hwa, T. and Lässig, M. Optimal detection of sequence similarity by local alignment. Proc. of the Second Annual Conference on Computational Molecular Biology (RECOMB98), in press. E-print cond-mat/9712081.

Kardar, M. 1987. Replica Bethe ansatz studies of two-dimensional interfaces with quenched random impurities. *Nucl. Phys. B 290*, 582 - 602.

Karlin, S. and Altschul, S.F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natn. Acad. Scie. U.S.A. 87 (6)*, 2264 - 8.

Karlin, S. and Altschul, S.F. 1993. Applications and statistics for multiple high-scoreing segments in molecular sequences. *Proc. Natn. Acad. Scie. U.S.A. 90 (12)*, 5873 - 7.

Kinzelbach, H. and Lässig, M. 1995. Depinning in a random medium. *J. Phys. A: Math. Gen. 28*, 6535 - 6541.

Koretke, K.K., Lutheyschulten, Z., Wolynes, P.G. 1996. Self-consistently optimized statistical mechanical energy functions for sequence structure alignment *Prot. Sci. 5*, 1043-1059.

Kschischo M. and Lässig M 1999. Preprint.

Lässig, M. 1998. On growth, disorder, and field theory. *J. Phys. C 10*, 9905-9950.

Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol. 48 (3)*, 443 - 53.

Olsen R., Hwa. T., and Lässig, M. 1999a. Optimizing Smith-Waterman alignments. Pacific Symp. on Biocomputing 99, 302 – 13. World Scientific.

Olsen, R., Bundschuh, R., and Hwa., T. 1999b. Rapid assessment of extremal statistics for gapped local alignment. to appear in Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology. AAAI Press.

Onuchic, J.N., LutheySchulten, Z., Wolynes, P.G. 1997. Protein folding funnels: the nature of the transition state ensemble. *Ann. Rev. Phys. Chem. 48*, 545-600, and references therein.

Pearson, W.R. 1991. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics 11 (3)*, 635 - 650.

Risken, H. 1989. *The Fokker-Planck Equation.* Springer Verlag.

Sankoff D. and Kruskal J. 1983. *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison.* Addison Wesley, Reading (Mass.)

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. MOl. Biol. 147*, 195 – 7.

Steele, J.M. 1982. Long common subsequences and the proximity of two random strings. SIAM J. Appl. Math. 42, 731 – 7.

Steele, J.M. 1986. An Efron-Stein inequality for nonsymmetric statistics. Ann. Stat. 14, 753 – 758.

Thorne, J.L., Kishino, H. and Felsenstein, J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequence *J. Mol. Evol. 33 (2)*, 114 - 24

Thorne, J.L., Kishino, H., and Felsenstein, J. 1992. Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol. 34 (1)*, 3 - 16.

Vingron, M. and Waterman, M.S. 1994. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Biol 235 (1)*, 1 - 12.

Wang, J., Onuchic, J., Wolynes, P.G. 1996. Statistics of kinetic pathways on biased rough energy landscapes with application to protein folding. *Phys. Rev. Lett. 76*, 4861-4864.

Waterman, M.S., Gordon, L. and Arratia, R. 1987. Phase transitions in sequence matches and nucleic acid structure. *Proc. Natl. Acad. Sci. U.S.A. 84 (5)*, 1239 - 43.

Waterman, M.S. 1989. *In* Waterman, M.S, ed., *Mathematical Methods for DNA Sequences.* CRC Press.

Waterman, M.S. Eggert, M. and Lander, E. 1992. Parametric sequence comparisons. *Proc. Natn. Acad. Scie. U.S.A. 89 (13)*, 6090 - 3.

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

Watermann, M.S. and Vingron, M. Sequence comparison significance and Poisson approximation. *Stat. Sci.* 9, 367 – 381.

Zhang, M.Q. and Marr, T.G. 1995. Alignment of molecular sequences seen as random path analysis. *J. Theo. Biol. 174 (2)*, 119 - 29.