

Similarity Detection and Localization

Terence Hwa^{1,*} and Michael Lässig^{2,†}

¹Physics Department, State University of New York, Stony Brook, New York 11794-3800

²Max-Planck Institut für Kolloid- und Grenzflächenforschung, Kantstrasse 55, 14513 Teltow, Germany

(Received 9 November 1995)

The detection of similarities between long DNA and protein sequences is studied using concepts of statistical physics. It is shown that mutual similarities can be detected by sequence alignment methods only if their amount exceeds a threshold value. The onset of detection is a critical phase transition viewed as a localization-delocalization transition. The *fidelity* of the alignment is the order parameter of that transition; it leads to criteria to select optimal alignment parameters.

PACS numbers: 87.10.+e, 05.70.Jk, 74.60.Ge

Evaluating the similarities between long strings of the alphabet is a challenging task, which arises in many fields ranging from data processing to biology [1]. Standard applications involve the comparisons of copies of a message sequence blurred by an imperfect transmission or reproduction process. A particularly important example is the evolution in biological systems, a process that mutates gene sequences in various ways including local *substitutions*, *insertions*, and *deletions*. Molecular biologists routinely compare newly sequenced genes to known ones in databases, a first step towards unveiling the structure and function of the new findings. This so-called *sequence alignment* is the most widely used mathematical and computational tool in molecular biology [2].

A (global) *alignment* of two sequences $\{P_i\}$ and $\{Q_j\}$ defined as an ordered set of pairings (P_i, Q_j) and of unpaired elements $(P_i, -)$ and $(-, Q_j)$ called gaps, each letter P_i and Q_j belonging to exactly one pairing or gap (see Fig. 1). The optimal alignment of the two sequences is determined by minimization of a cost or “energy” function E favoring pairs of matching elements $(P_i = Q_j)$ over mismatches $(P_i \neq Q_j)$ and gaps. A simple and commonly used energy function is the sum over all matches, mismatches, and gaps of the alignment contributing the respective energies -1 , $\mu > 0$, and $\delta > 0$ each [3,4]. Many more general energy functions have been discussed [2].

Clearly, the optimal alignment depends strongly on these energy parameters, and so does its *fidelity*, i.e., the extent to which it captures mutual correlations between the sequences compared. In particular, if the evolution process involves random insertions and deletions, one has to allow for gaps in the optimal alignment so that mutually correlated regions of the sequences can actually align. For gene sequences, a more stringent criterion is the *biological relevance* of an alignment, that is, the extent to which the matched regions actually indicate functional similarities between different proteins. Finding alignment parameters that lead to high relevance is a difficult problem, which has not been solved systematically so far. For biological applications, “optimal” parameters for various types of

genes are often deduced empirically from sequence pairs whose functional alignments are already known [5].

In this Letter, we apply ideas and methods of statistical physics to sequence alignment, introducing a different conceptual approach towards the parameter selection problem. Using simple stochastic models of evolution, we mutate an ancestor sequence to obtain daughter sequences $\{P_i\}$ and $\{Q_j\}$ with well-defined mutual correlations, namely, the ensemble of all pairs $(P_i = Q_j)$ of unmutated daughters of the same ancestor element. These sequences are then aligned using a given energy function E , and the fidelity of the optimal alignment is quantified as the fraction of correctly recovered such pairs $(P_i = Q_j)$. For long sequences, we find that the mutual correlations cannot be detected if their amount is below a threshold value. A critical transition separates this *low-similarity phase* of zero fidelity from the *high-similarity phase* where the fidelity is finite. (This phase transition is distinct from the “transition” to the so-called local alignment regime discussed in the literature [6].)

Our analysis of the phase transition is based on the known representation [3] of an alignment of two sequences $\{P_i\}, \{Q_j\}$ on the two-dimensional lattice of Fig. 2. The cells of this lattice are labeled by the index pairs (i, j) or alternatively by the rotated coordinates $t = i + j$ and $r = i - j$. The bonds encode the adjacency of letters: The diagonal bond in cell (i, j) represents the pairing (P_i, Q_j) ; horizontal and vertical bonds correspond to gaps $(P_i, -)$ and $(-, Q_j)$, respectively. Thus any alignment maps onto a lattice path that is *directed* along the t coordinate, i.e., given by a unique function $r(t)$. For daughter sequences generated from our stochastic evolution model, mutual similarities can be represented on the lattice by a path $R(t)$ joining all correlated pairs

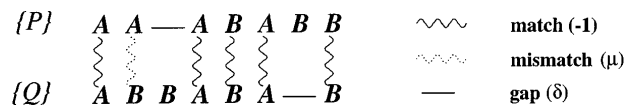


FIG. 1. One possible alignment of two binary sequences, $\{P_i\} = AAABABB$ and $\{Q_j\} = ABBABAB$, with six pairings (five matches, one mismatch) and two gaps.

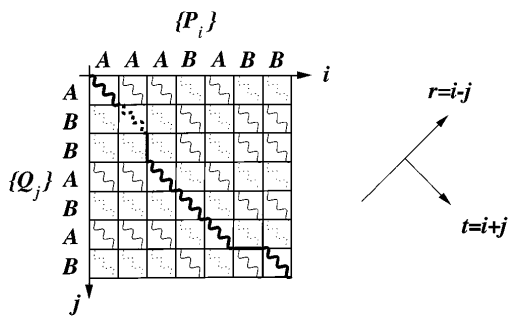


FIG. 2. The *directed* path (thick line) corresponding to the alignment in Fig. 1. Horizontal and vertical lattice bonds represent gaps, diagonal bonds represent pairings with bond energies $v_{r,t} = -(J \pm \Delta)$ for matches (full-wiggly lines) and mismatches (dashed-wiggly lines), respectively.

($P_i = Q_j$). We call such a path the “target path.” For long sequences, it induces a *morphological transition* on the optimal alignment path $r_0(t)$: In the low-similarity phase, this path is *superdiffusive* with typical fluctuations $\delta r(t) \equiv r_0(t) - R(t)$ scaling as $|\delta r(t)| \sim t^{2/3}$, while, in the high-similarity phase, it is *localized* to the target path with finite fluctuations $|\delta r(t)| \sim \xi_{\perp}$; see Fig. 3. Hence only in the high-similarity phase is the path $r_0(t)$ a faithful approximation of the target $R(t)$. The fidelity of the alignment is then simply given by the *overlap* of the two paths, i.e., by their number of intersections per unit of t . For long sequences, the overlap is proportional to the inverse localization length ξ_{\perp}^{-1} ; maximization of this “order parameter” gives a numerically and analytically accessible criterion for the choice of alignment parameters.

To establish these results, we restrict ourselves here to binary sequences (with elements $P_i, Q_j \in \{+1, -1\}$), and to the simple model introduced above with just two parameters μ and δ (generalizations are briefly discussed at the end of this Letter). With a convenient shift in the energy function, we write the total energy E of any path $r(t)$ as the sum over all its diagonal bonds with bond energies $v_{r,t} \equiv -J - \Delta P_{i=(r+t)/2} Q_{j=(r-t)/2}$; horizontal and vertical bonds take the energy 0. The parameters $J \equiv 2\delta - (\mu - 1)/2$ and $\Delta = (\mu + 1)/2$ are the effective gap and mismatch costs; they characterize the stiffness

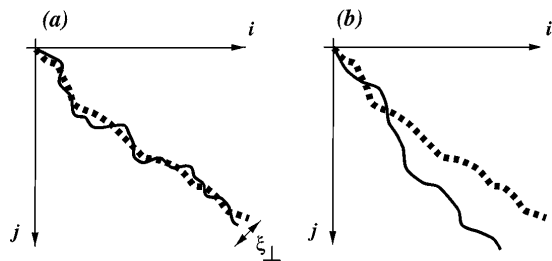


FIG. 3. Typical large-scale configurations of the optimal alignment path $r_0(t)$ for long sequences. (a) In the high-similarity phase, this path (full line) is *localized* to the target path $R(t)$ (dashed line) with typical transverse displacement of size ξ_{\perp} . (b) In the low-similarity phase, it fluctuates independently of the target path.

of the alignment paths and the site-to-site variations of the pairing potential $v_{r,t}$, respectively. The lowest energy path $r_0(t)$ depends on μ and δ only via the ratio $g \equiv \Delta/J$ in the biologically relevant limit $g \leq 1$ (or $2\delta \geq \mu$) [7].

To model the behavior of typical alignment paths in the *low-similarity phase*, we take the sequences $\{P_i\}$ and $\{Q_j\}$ from an ensemble of unbiased *random sequences* (i.e., $\overline{P_i} = \overline{Q_j} = 0$, $\overline{P_i P_{i'}} = \delta_{i,i'}$ and $\overline{Q_j Q_{j'}} = \delta_{j,j'}$) with no mutual correlations, $\overline{P_i Q_j} = 0$. (Averages over this ensemble are denoted by overbars.) The pairing potential $v_{r,t}$ then becomes a *random potential* with average $\overline{v_{r,t}} = -J$ and variance

$$\begin{aligned} \overline{v_{r,t} v_{r',t'}} - J^2 &= \Delta^2 \overline{P_{(r+t)/2} P_{(r'+t')/2} Q_{(r-t)/2} Q_{(r'-t')/2}} \\ &= \Delta^2 \delta_{r,r'} \delta_{t,t'}. \end{aligned} \quad (1)$$

It induces random fluctuations on the alignment paths.

The large-scale statistics of these fluctuations can be derived from the partition function of the alignment paths [8] in a path integral representation [9]. We show [10] that the continuum “action” of this path integral takes the form [11]

$$S = \int d\tau \left[\frac{\gamma}{2} \dot{\rho}^2 + \eta(\rho(\tau), \tau) + O(\dot{\rho}^4, \eta \dot{\rho}^2, \dots) \right]. \quad (2)$$

It describes a “coarse-grained” alignment path $\rho(\tau)$ with a finite line tension γ (and $\dot{\rho} \equiv d\rho/d\tau$) in a coarse-grained random potential $\eta(\rho, \tau)$ characterized by its second moment $\overline{\eta(\rho, \tau) \eta(\rho', \tau')} \sim \Delta^2 \delta(\rho - \rho') \delta(\tau - \tau')$. All other short-ranged moments [12], as well as the terms omitted in (2), are irrelevant variables. We conclude that the large-distance behavior of alignments is governed by the well-known universality class of a directed path in a two-dimensional *Gaussian* random potential [9]. Many properties of this universality class are known exactly. Typical fluctuations of the *optimal path* $r_0(t)$ are (asymptotically) superdiffusive, $\overline{[r_0(t) - r_0(t')]^2} \simeq A|t - t'|^{4/3}$, reflecting the tendency of the optimal alignment to use gaps to gain an excess number of matches over mismatches. This goes along with the variance $\overline{E_0^2(N)} - \overline{E_0(N)}^2 \simeq BN^{2/3}$ of the optimal energy E_0/N for paths of length $N \gg 1$. We have verified these properties numerically and have obtained the nonuniversal amplitudes A and B ; an example is shown in Fig. 4. Scaling arguments, which are supported by our numerics, yield $A \simeq g^{4/3}$ and $B \simeq g^{4/3} J^2$ for small values of g (which are appropriate for the alignment of sequences with not too many insertions and deletions; see below).

To study the *high-similarity phase*, we construct sequences $\{P_i\}$ and $\{Q_j\}$ with mutual correlations as they would arise if these sequences developed from a common ancestor sequence (again taken to be an unbiased random sequence) through independent mutations over a time θ . Consider first the evolution through point substitutions only: On each sequence, elements are randomly

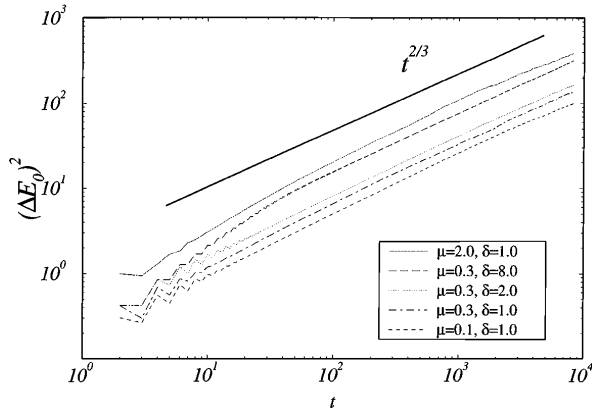


FIG. 4. Energy variations $(\Delta E_0)^2 \equiv \overline{E_0^2(t)} - \overline{E_0(t)}^2 \sim t^{2/3}$ of the optimal alignments. The results are obtained from a sample of 100 pairs of uncorrelated random sequences for each set of the alignment parameters listed.

chosen with rate Γ and replaced by an unbiased random number ± 1 . Then $\{P_i\}$ and $\{Q_j\}$ remain unbiased random sequences and have mutual correlations $\overline{P_i Q_j} = p^2 \delta_{i,j}$, where $p = \exp(-\Gamma\theta)$ is the fraction of elements on each sequence that are still unaffected by the mutations after a time θ . It follows that the pairing potential still has the variance (1) (up to an irrelevant [10] term $\sim \delta_{t,t'} \delta_{r,-r'}$), but the average is now $\overline{v_{r,t}} = -J - U \delta_{r,0}$: mutual correlations generate a target path $R(t) = 0$ (along the diagonal of the alignment lattice) that attracts the alignment paths with strength $U = gp^2J$. This interaction reflects the excess number of matches ($P_i = Q_i$) of unmutated daughter elements of the ancestor sequence. It turns out to change the optimal alignment drastically [13]: The path $r_0(t)$ becomes *localized*, i.e., its mean square displacement from the target path $r_0^2(t)$ grows no longer superdiffusively, but reaches a finite value $\xi_\perp^2 \equiv \lim_{t \rightarrow \infty} r_0^2(t)$ for long sequences. In recent years, this localization has been studied carefully in the context of flux pinning in type-II superconductors [13,14]. It is governed [14] by the competition of the potential energy $\sim U/\xi_\perp$ per unit of t gained from the overlap with the target path, and the random energy cost V/ξ_\perp (with a constant $V \simeq g^2J$ for $g^2 \ll 1$ [10]) due to the confinement of the alignment path. (This is very similar to the competition of potential and kinetic energy determining the localization of a quantum particle in a potential well.) In the limit $p = 1$ of identical sequences, the optimal alignment is obviously $r_0(t) = 0$; the path is tightly bound to the target. As p decreases, the fluctuations around the target increase, and the overlap decreases. It is found [13] that the path $r_0(t)$ remains localized to the target even for an arbitrarily weak attraction $U > 0$, although the localization length becomes very large for $U \ll V$,

$$\xi_\perp \approx \exp(bV/U) \quad (3)$$

(with a universal constant $b \approx 2.8$). Thus the fidelity of the optimal alignment, ξ_\perp^{-1} , tends to zero with vanishing U ; this signals a continuous transition to the superdiffusive behavior at $U = 0$.

We now turn to a more realistic model of evolution that includes local insertions and deletions: We generate the sequences $\{P_i\}$ and $\{Q_j\}$ from a common ancestor through point substitutions as before, then we randomly choose a fraction $\tilde{p} \ll 1$ of the sites on each sequence, and insert or delete a random element ± 1 at the chosen sites. This modifies the mutual correlations,

$$\overline{P_i Q_j} = p^2 \delta_{i-j, R(i+j)}, \quad (4)$$

and hence the pairing potential $\overline{v_{r,t}} = -J - U \delta_{r, R(t)}$. The target path $R(t)$ is no longer along the diagonal of the alignment lattice, but along the trajectory of a Gaussian random walk with mean square displacement $[\overline{R(t) - R(t')}]^2 = \tilde{p}|t - t'|$ (see Fig. 3). Since a fraction $2\tilde{p}$ of the target path involves gaps, the attractive strength of the target is shifted to

$$U = [gp^2(1 - \tilde{p}) - 2\tilde{p}]J, \quad (5)$$

and now changes sign at the threshold $p^2 = p_c^2 = 2\tilde{p}/g + O(\tilde{p}^2)$. Above the threshold ($p^2 > p_c^2$), the interaction remains attractive ($U > 0$). In this phase, the optimal alignment path turns out to be localized to the target in much the same way as before [13], with a finite mean square displacement $\xi_\perp^2 \equiv \lim_{t \rightarrow \infty} [r_0(t) - R(t)]^2$ given by (3) and (5) and a finite fidelity $\sim \xi_\perp^{-1}$. As p^2 approaches p_c^2 from above, the fidelity again tends to zero continuously. For $p^2 < p_c^2$, the interaction becomes repulsive, and the path $r_0(t)$ is again superdiffusive. This is the low-similarity phase, where the optimal alignment does not reflect the mutual correlations: Its fidelity is zero. The singular behavior (3) of the fidelity applies to sequences whose length exceeds the correlation length $\xi_\parallel \sim \xi_\perp^{3/2}$; for shorter sequences, the fidelity is reduced by finite-size effects.

The concept of an order parameter is useful for the selection of optimal alignment parameters μ and δ . As follows from Eqs. (3), (5), and $V \simeq g^2J$, the correlations (4) with given parameters p and \tilde{p} can be detected only for $g > g_c(p, \tilde{p}) = 2\tilde{p}/p^2 + O(\tilde{p}^2)$. They are recovered best if g takes the value $g^*(p, \tilde{p}) = 4\tilde{p}/p^2 + O(\tilde{p}^2)$ obtained by maximizing ξ_\perp^{-1} for fixed p and $\tilde{p} \ll p^2$. This is a linear condition on μ and δ .

There is a number of related alignment issues relevant to applications in biology, where our results apply in a similar way; a detailed account will be published elsewhere [10]. (i) For alphabets with k different letters ($k = 4$ for DNA and $k = 20$ for proteins), the effective attraction strength increases with k ; this leads to a much higher fidelity and a reduced correlation length. (ii) A higher biological relevance of the alignment is often achieved if gap initiations are penalized by a higher energy than gap extensions. Such refinements lead to alignment paths that have, in addition to their line tension γ , a finite ‘‘bending rigidity.’’ (iii) The simultaneous alignment of n sequences is described by a directed alignment path $(r_1, \dots, r_{n-1})(t)$ in $n - 1$ transversal dimensions. For $n > 2$, the detection threshold is increased to

finite $U > 0$, making similarity detection more difficult. The divergence of the localization length close to the transition is then given by a *power law* instead of Eq. (3). (iv) The same is true even for the alignment of two sequences if there are long-ranged intrasequence correlations that fall off sufficiently slowly (as may be the case for the noncoding regions of the genome). (v) Unlike the *global* algorithms discussed so far, *local* alignment algorithms match only a contiguous piece of sequence $\{P_i\}$ with a different piece of sequence $\{Q_j\}$; they are appropriate for finding mutual similarities that exist only within these two pieces. It has been noted [6] that the regimes of local and global alignment are separated by a transition line in the space of parameters (μ, δ) . In the global alignment regime, the sequences are aligned up to small boundary regions. As the transition line is approached, these regions grow up to a length $O(N)$ in the local alignment regime. This boundary-induced criticality is analogous to wetting transitions and thus quite different from the transition described in this Letter.

In summary, we have described a unique approach to similarity detection, identifying sequence alignment algorithms with physics problems defined on a lattice with quenched disorder. We show that the successful detection of correlations between sequences depends on the kinds of mutations they undergo, as well as on the specific choice of the alignment parameters. This is demonstrated for simple stochastic mutation processes modeling biological evolution. In such systems, correlations can only be detected if their amount exceeds a threshold value; the onset of detection is shown to be a critical phase transition with universal characteristics. Most importantly, it is the *fluctuation statistics* at this transition that determines the fidelity of the optimal alignment. Using that order parameter we derive criteria for the optimal choice of alignment parameters given a limited knowledge of the mutation process—in contrast to the current approach of finding these parameters empirically. Conversely, the knowledge of optimal alignment parameters for a given class of proteins can be used to infer the nature of mutations suffered by those proteins.

We thank Michael Zhang for a review of current issues in sequence alignment. T.H. is supported by an A.P. Sloan Research Fellowship and an ONR Young Investigator Award through ONR N00014-95-1-1002. M.L. is grateful for the hospitality of SUNY at Stony Brook, where much of this work was done.

*Current address: Physics Department, University of California at San Diego, La Jolla, CA 92093-0319.

Electronic address: hwa@ucsd.edu

†Electronic address: lassig@mpikg-teltow.mpg.de

- [1] See, e.g., *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, edited by D. Sankoff and J. B. Kruskal (Addison-Wesley, Reading, MA, 1983).
- [2] See reviews in *Mathematical Methods in DNA Sequences*, edited by M. S. Waterman (CRC Press, Boca Raton, FL, 1989); S. F. Altschul *et al.*, *Nature Genetics* **6**, 119 (1994).
- [3] S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.* **48**, 444 (1970).
- [4] T. F. Smith and M. S. Waterman, *Adv. Appl. Math.* **2**, 482 (1981).
- [5] W. S. Fitch and T. Smith, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1382 (1983); M. Schöniger and M. S. Waterman, *Bull. Math. Biol.* **54**, 521 (1992).
- [6] M. S. Waterman, L. Gordon, and R. Arratia, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 1239 (1987).
- [7] For $g > 1$ (or $\mu > 2\delta$), every mismatch in the optimal alignment is replaced by a pair of gaps, and the optimal alignment path and energy no longer depend on μ . However, this parameter regime is clearly “unbiological” and will not be discussed here.
- [8] This partition function is defined as the sum over *all* alignment paths $r(t)$, each weighed by a Boltzmann factor given by its energy E and the “temperature” β^{-1} , $Z(\beta^{-1}) \equiv \sum_{\text{paths}} \exp(-\beta E)$. [See, M. S. Waterman, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3123 (1983); M. Q. Zhang and T. G. Marr, *J. Theor. Biol.* **174**, 119 (1995).]
- [9] See review article by M. Kardar, in *Fluctuating Geometries in Statistical Mechanics and Field Theory*, Proceedings of the Les Houches Summer School, Session LXII, edited by F. David, P. Ginsparg, and J. Zinn-Justin (Elsevier, Amsterdam, 1996), and references therein.
- [10] T. Hwa and M. Lässig (to be published).
- [11] The path integral is understood to be discretization-invariant under reflections $\tau \rightarrow -\tau$ so that only even powers of $\dot{\rho}$ appear in the action. We set $\eta(\rho, \tau) = 0$, suppressing a constant term that affects only the normalization of the partition function.
- [12] The random potential has long-ranged correlations as well; e.g., $\overline{v_{r_1, t_1} \cdots v_{r_4, t_4}}$ is nonzero if $(r_1, t_1), \dots, (r_4, t_4)$ are the corners of a rectangle on the lattice of Fig. 2. Such correlations are relevant at the Gaussian fixed point ($\Delta = 0$) but irrelevant at the zero-temperature fixed point, essentially because they can only be felt by rare fluctuations around the minimum energy path [10].
- [13] L.-H. Tang and I. F. Lyuksyutov, *Phys. Rev. Lett.* **71**, 2745 (1993); L. Balents and M. Kardar, *Phys. Rev. B* **49**, 13 030 (1994); T. Hwa and T. Nattermann, *Phys. Rev. B* **51**, 455 (1995); H. Kinzelbach and M. Lässig, *J. Phys. A* **28**, 6535 (1995).
- [14] T. Nattermann and R. Lipowsky, *Phys. Rev. Lett.* **61**, 2508 (1988).