

Bounds for Parametric Sequence Comparison

David Fernández-Baca* Timo Seppäläinen †
Giora Slutzki‡

Abstract

We consider the problem of computing a global alignment between two or more sequences subject to varying mismatch and indel penalties. We prove a tight $3(n/2\pi)^{2/3} + O(n^{1/3} \log n)$ bound on the worst-case number of distinct optimum alignments for two sequences of length n as the parameters are varied. This refines a $O(n^{2/3})$ upper bound by Gusfield et al., answering a question posed by Pevzner and Waterman. Our lower bound requires an unbounded alphabet. For strings over a binary alphabet, we prove a $\Omega(n^{1/2})$ lower bound. For the parametric global alignment of $k \geq 2$ sequences under sum-of-pairs scoring we prove a $3 \binom{k}{2} (n/2\pi)^{2/3} + O(k^{2/3} n^{1/3} \log n)$ upper bound on the number of distinct optimality regions and a $\Omega(n^{2/3})$ lower bound, partially answering a problem of Pevzner. Based on experimental evidence, we conjecture that for two random sequences, the number of optimality regions is approximately \sqrt{n} with high probability.

Keywords. Sequence alignment, multiple alignment, parametric analysis, computational biology, experimental analysis of algorithms.

*Department of Computer Science, Iowa State University, Ames, IA 50011. Supported in part by the National Science Foundation under grants CCR-9520946 and CCR-9988348. E-mail: fernande@cs.iastate.edu.

†Department of Mathematics, Iowa State University, Ames, IA 50011. Supported in part by the National Science Foundation under grant DMS-9801085. E-mail: seppalai@iastate.edu.

‡Department of Computer Science, Iowa State University, Ames, IA 50011. E-mail: slutzki@cs.iastate.edu.

1 Introduction

Optimal sequence alignment is one of the most widely used techniques for determining similarity (homology) between biological sequences. Rather than give a partial list of references to the vast literature on this subject, we refer the reader to Gusfield’s book [4]. A collection of earlier papers on the subject can be found in [10], while the review [11] gives relevant references.

An *alignment* between two strings S and T of lengths n and m , $n \leq m$, is a pair of equal-length strings $\mathcal{A} = (S', T')$ where S' (respectively, T') is obtained by inserting special *space* characters (denoted by “-”) into S (T) under the restriction that there can be no character position in which both S' and T' have spaces. A *match* is a position in which S' and T' have the same character. A *mismatch* is a position in which S' and T' have different characters, neither of which is a space. An *indel* is a position in which one of S' and T' has a space. A *gap* is a sequence of one or more consecutive spaces in S' or T' .

In scoring an alignment matches are rewarded, while mismatches, indels, and gaps are penalized. Various alignment scoring criteria have been proposed (see, again, [4]); this paper deals with *global alignments*. Let α , β , and γ denote the mismatch, space, and gap penalties. The *score* of a global alignment \mathcal{A} with w matches, x mismatches, y indels, and z gaps is

$$\text{score}_{\mathcal{A}}(\alpha, \beta, \gamma) = w - x\alpha - y\beta - z\gamma. \quad (1)$$

Different penalty choices yield different optimum alignments. It can be shown that, for any pair of strings, the (α, β, γ) space is decomposed into convex polyhedral *optimality regions* such that, for each region R , there exists an alignment that is optimal for all points in the interior and, furthermore, R is maximal for this property [5]. Clearly, if we fix any parameter we get a decomposition of the space for the other two into convex polygonal regions. Parametric sequence alignment [5, 6, 7, 11, 12] studies the properties of such parameter-space decompositions as well as the methods for constructing them.

Henceforth, we will consider only alignments that do not penalize gaps; i.e., $\gamma = 0$. This paper is motivated by the following result.

Theorem 1 (Gusfield et al. [5]) *For global alignment with varying mismatch and space penalties, the number of optimality regions of the associated decomposition of the (α, β) plane is $O(n^{2/3})$.*

All known algorithms to build the decomposition of the parameter space induced by two sequences (see, e.g., [6]) run in time proportional to the number of regions multiplied by the work needed to compute a single global alignment, which is $O(nm)$. Thus, Theorem 1 implies a $O(n^{5/3}m)$ bound on the time to build the decomposition.

In this paper we show that the bound of Theorem 1 is tight when the alphabet is unbounded, thereby answering a question posed by Pevzner and Waterman [9]. In fact, we prove that the exact bound is $3(n/2\pi)^{2/3} + O(n^{1/3} \log n)$. For a bounded (specifically, binary) alphabet, we are only able to give an $\Omega(n^{1/2})$ lower bound. Our lower bound proofs are constructive: We show how to generate for each n a pair of sequences of length n that achieve the claimed bounds. Building the parameter-space decompositions induced by these families of sequences requires, respectively, $\Omega(n^{8/3})$ and $\Omega(n^{5/2})$ time.

After considering parametric two-sequence comparison, we study the alignment of $k \geq 2$ sequences of length n under *sum-of-pairs* scoring, which is a direct generalization of (1). We show that the parameter-space decomposition in this case has the same structure as for two sequences (see Sections 2 and 5 for details on this structure). Thus, the decomposition can be built in time proportional to the number of regions multiplied by the work needed to compute a single multiple alignment. Additionally, we prove a $3 \binom{k}{2} (n/2\pi)^{2/3} + O(n^{1/3} \log n)$ upper bound on the number of optimality regions, while our construction for the two-sequence case implies an $\Omega(n^{2/3})$ lower bound. Our results partially answer a problem posed by Pevzner [8]. Note that the upper bound is polynomial in the number of sequences, which contrasts with the fact that all known multiple alignment algorithms are exponential in k .

In addition to the analytical bounds just described, we also conducted experimental studies of two-sequence alignment. The results strongly suggest that for randomly-generated pairs of sequences of length n , the expected number of optimality regions is approximately \sqrt{n} . Furthermore, extreme variation seems to be rare: none of the randomly-generated examples had significantly more or fewer than \sqrt{n} regions. Experimental evidence also suggests that alphabet-size dependence is a relatively minor factor. Thus, we conjecture that, with high probability, the number of regions for randomly-chosen pairs of strings is \sqrt{n} plus lower-order terms.

Our results follow mainly from combinatorial and number-theoretic ar-

guments and thus it is not clear whether they provide any algorithmic (or, for that matter, biological) insight into the structure of alignments for different regions. On the other hand, this kind of analysis technique is widely applicable, as illustrated in a companion paper [3].

The rest of this paper is organized as follows. Section 2 reviews parametric and non-parametric sequence alignment. Readers familiar with these subjects can proceed directly to Section 3, which presents a tight bound for strings over an unbounded alphabet. Section 4 gives a lower bound for strings over a binary alphabet. Multiple sequence alignment is discussed in Section 5. Experimental results are shown in Section 6.

2 Preliminaries

Alignment graphs. The dynamic programming algorithm for computing an optimum global alignment between strings $S = s_1s_2 \cdots s_n$ and $T = t_1t_2 \cdots t_m$ can be viewed as a procedure for finding a maximum-weight path in a weighted *alignment graph* G . The nodes of G are arranged in an $(n + 1) \times (m + 1)$ grid; rows (columns) are numbered consecutively from top to bottom (left to right), from 0 to n (m). We denote the nodes of G by their coordinates (i, j) . Every node has an edge directed to its right neighbor and an edge to its neighbor below it; these edges have weight $-\beta$. Additionally, for $1 \leq i \leq n$, $1 \leq j \leq m$, there is a diagonal edge $\sigma(i, j)$ directed into vertex (i, j) from vertex $(i - 1, j - 1)$. The weight of $\sigma(i, j)$ is 1 if $s_i = t_j$ and $-\alpha$ otherwise. See Figure 1. It is convenient to imagine that, as shown in Figure 1, the n horizontal and m vertical “strips” of G are labeled by successive characters of S and T , respectively.

Each path Γ from $(0, 0)$ to (n, m) corresponds to a unique alignment. Henceforth, unless otherwise stated, we will only consider paths of this sort. Horizontal moves along a path correspond to spaces inserted in S ; vertical moves correspond to spaces inserted in T ; diagonal moves correspond to either matches or mismatches. A path and its corresponding alignment are illustrated in Figure 1. Note that, if an alignment has w matches, x mismatches and y spaces, we must have

$$2w + 2x + y = n + m. \tag{2}$$

We write $score(\Gamma)$ to denote the score of the alignment associated with Γ as a function of α and β ; $score_0(\Gamma)$ will denote $score(\Gamma)$ with α fixed at 0.

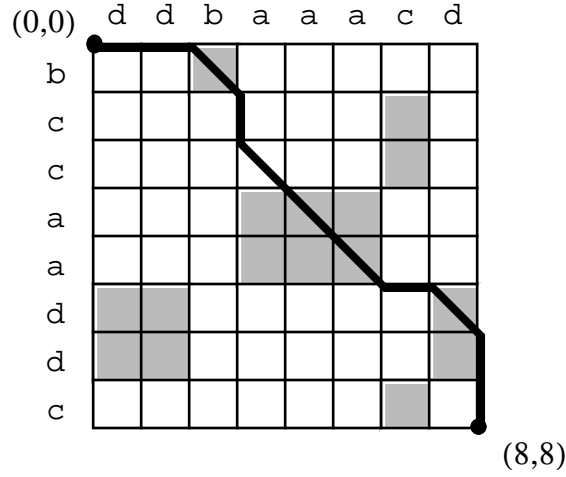


Figure 1: The alignment graph for $S = \text{bccaaddc}$ and $T = \text{ddb-aaacd--}$. Shaded areas are match blocks. The path Γ shown corresponds to the alignment $(\text{--bccaa-ddc, ddb-aaacd--})$; $\text{score}(\Gamma) = 4 - \alpha - 6\beta$.

(Recall that $\gamma = 0$ throughout the paper.)

It is helpful to imagine that G is tiled with *match blocks* in the following way. If $[r..l]$ and $[p..q]$ are maximal ranges such that $s_i = t_j$ for $i \in [r..l]$ and $j \in [p..q]$, then there is a rectangular tile whose upper left corner is $(r-1, p-1)$ and whose lower right corner is (l, q) (see Figure 1).

Parametric global alignment. The lemma below is a consequence of (2).

Lemma 2 (Gusfield et al. [5]) *Consider any two alignments with corresponding paths Γ and Γ' in the alignment graph. Then $\text{score}(\Gamma) = \text{score}(\Gamma') = (n+m)/2$ for $\alpha = -1, \beta = -1/2$.*

The previous result leads to the following characterization of the structure of the optimality regions for global alignment.

Lemma 3 (Gusfield et al. [5]) *All optimality regions on the (α, β) plane are semi-infinite cones, and are delimited by the coordinate axes or by lines of the form $\beta = c + (c + 1/2)\alpha$ for some constant c .*

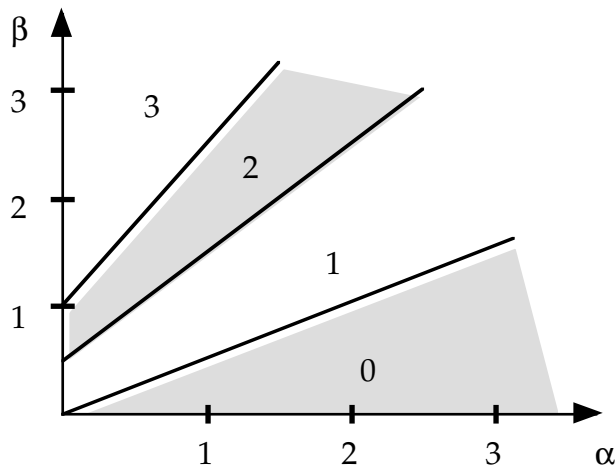


Figure 2: Decomposition of the parameter space induced by $S = 100111$ and $T = 011000$. The corresponding optimum alignments are $\mathcal{A}_0 = (100111---, --0-11000)$, $\mathcal{A}_1 = (10011--1, --011000)$, $\mathcal{A}_2 = (1001-11, -011000)$, and $\mathcal{A}_3 = (100111, 011000)$.

Thus, for any fixed $\alpha_0 \geq 0$, the line $\alpha = \alpha_0$ intersects all optimality regions except, possibly, the lowest one; see Figure 2. The lowest region R deserves special attention. Lemma 3 implies that the area in the positive quadrant delimited by $\beta = 0$ and $\beta = \alpha/2$ is contained in R , because that area cannot be further decomposed into subregions. Now, depending on the input strings, R may or may not be delimited above by the line $\beta = \alpha/2$. Suppose it is (as is the case for the sequences in Figure 2), and let \mathcal{A} , \mathcal{A}' be alignments that are optimal, respectively, for R and the region immediately above R . Since the boundary of both regions contains the origin, $score(\mathcal{A}) = score(\mathcal{A}')$ for $(\alpha, \beta) = (0, 0)$. Since at that point mismatches and spaces carry zero weight, equality can only hold if both alignments have the same number of matches. Note that this is the only case in which optimal alignments for two consecutive regions (in bottom-to-top order) may have the same number of matches; in all other cases, the number of matches must decrease. We refer to parameter-space decompositions where the lowest region is delimited above by $\beta = \alpha/2$ as decompositions containing a *special region*.

The following lemma gives a condition that guarantees that each path in a collection $\{\Gamma_i\}_{1 \leq i \leq q}$ of paths in the alignment graph is the highest scoring

path over a non-empty range of β -values, with α fixed at 0. The proof uses a simple inductive argument and is thus omitted.

Lemma 4 *Let $\Gamma_1, \Gamma_2, \dots, \Gamma_q$ be paths in the alignment graph. Assume that $score_0(\Gamma_i) = w_i - y_i\beta$, where $y_1 > y_2 > \dots > y_q$. Let $\beta_0 = 0$, $\beta_q = \infty$, and, for $r = 1, \dots, q - 1$, $\beta_r = (w_r - w_{r+1}) / (y_r - y_{r+1})$. Suppose $\beta_0 < \beta_1 < \dots < \beta_q$. Then, for $\beta \in (\beta_{r-1}, \beta_r)$, $score_0(\Gamma_r) > score_0(\Gamma_s)$, for all $s \neq r$.*

Note that in the above lemma β_r is the β -value such that $score_0(\Gamma_r) = score_0(\Gamma_{r+1})$.

3 Exact bounds with an unbounded alphabet

We now derive a tight bound on the number of optimality regions for parametric global alignment. We note that the constant in the $O(n^{2/3})$ upper bound in Theorem 1 was not derived in [5], although that reference attributes a 0.88 bound to Robert Irving. Our analysis provides the constant (which matches Irving's) plus a lower-order term. Furthermore, we give a matching lower bound, which was not supplied in [5].

Let us number the optimality regions from bottom to top, with 0 being the index of the lowest region and k the index of the highest. Let w_i, x_i, y_i denote the number of matches, mismatches and spaces in the optimum alignment for the i th region, $0 \leq i \leq k$. For $i = 1, 2, \dots, k$, let $\Delta w_i = w_{i-1} - w_i$, $\Delta x_i = x_i - x_{i-1}$, and $\Delta y_i = y_{i-1} - y_i$. Since the number of matches and indels is a non-increasing function of the indel penalty, while the number of mismatches is nondecreasing, we have $\Delta w_i, \Delta y_i, \Delta x_i \geq 0$. These inequalities are, in fact, strict for all quantities, except, possibly, for Δw_1 , which is zero only if the decomposition has a special region.

The equation of the intersection line between the $(i - 1)$ st and the i th optimality regions is given by

$$\beta = \frac{\Delta w_i}{\Delta y_i} + \frac{\Delta x_i}{\Delta y_i} \alpha.$$

Therefore,

$$\frac{\Delta w_i}{\Delta y_i} < \frac{\Delta w_{i+1}}{\Delta y_{i+1}} \quad \text{and} \quad \frac{\Delta x_i}{\Delta y_i} < \frac{\Delta x_{i+1}}{\Delta y_{i+1}}. \quad (3)$$

We can now state our main result. It shows an upper bound similar to that presented in [5], with a constant $3/(2\pi)^{2/3} \approx 0.88105$ (compare this with Irving's aforementioned value of 0.88). The theorem also provides a constructive lower bound argument that matches the upper bound exactly.

Theorem 5 *The maximum number of optimality regions on the (α, β) plane induced by the parametric global alignment of a pair of strings of length n is $3(n/2\pi)^{2/3} + O(n^{1/3} \log n)$. For every positive integer n , there exist an alphabet and a pair of strings of length n over that alphabet whose parametric optimal global alignment induces $3(n/2\pi)^{2/3} + O(n^{1/3} \log n)$ optimality regions.*

Proof. As above, let $k + 1$ be the number of optimality regions and let w_i, x_i, y_i be the number of matches, mismatches, and spaces in the optimal alignment for region $i, i = 0, 1, \dots, k$. Note that

$$\sum_{i=1}^k \Delta w_i \leq w_0 \quad \text{and} \quad \sum_{i=1}^k \Delta y_i \leq y_0.$$

By (2),

$$2w_0 + y_0 \leq 2n,$$

therefore,

$$\sum_{i=1}^k (2\Delta w_i + \Delta y_i) \leq 2n. \tag{4}$$

We first establish a tight bound on the number of pairs $(\Delta w_i, \Delta y_i)$ satisfying (4) such that the fractions $\Delta w_i/\Delta y_i$ are irreducible and distinct. We then show that, for any such sequence of pairs, there exist two strings whose parametric optimum alignment has $k + 1$ optimality regions.

Observe that Δy_i is always even, since any space inserted into sequence S must be compensated by a space inserted into T . Let $a_i = \Delta w_i$ and $b_i = \Delta y_i/2$. We shall obtain an upper bound on the maximum number of distinct irreducible fractions $a_i/b_i, a_i, b_i > 0$ such that

$$\sum_{i=1}^k (a_i + b_i) \leq n.$$

This provides a bound on the number of regions, except for the cases where the decomposition has a special lower region. For this situation, the actual number of regions differs from the number of fractions by at most 1, which is covered by the lower-order term.

The maximum number of fractions is attained by considering each successive integer r and taking all irreducible a/b where $a + b = r$ until we get a set of fractions whose numerators and denominators add up to at most n .

Let $\phi(m)$ be the *Euler totient function* [1], giving the number of positive integers less than or equal to m that are relatively prime to m . The number of irreducible fractions whose numerators and denominators add up to r is $\phi(r)$. Thus, the largest number of distinct r 's is given by the value of s satisfying

$$\sum_{r=1}^{s+1} r\phi(r) > n \geq \sum_{r=1}^s r\phi(r).$$

To obtain s , we use the following result from analytic number theory¹[1].

$$\sum_{n \leq x} \phi(n) = \frac{3}{\pi^2}x^2 + O(x \log x) \quad (5)$$

By (5),

$$\begin{aligned} \sum_{r=1}^s r\phi(r) &= s \sum_{r=1}^s \phi(r) - \sum_{i=1}^{s-1} \sum_{j=1}^i \phi(j) \\ &= \frac{3}{\pi^2}s^3 + O(s^2 \log s) - \sum_{i=1}^{s-1} \left(\frac{3}{\pi^2}i^2 + O(i \log i) \right) \\ &= \frac{2}{\pi^2}s^3 + O(s^2 \log s). \end{aligned} \quad (6)$$

By a similar argument,

$$\sum_{r=1}^{s+1} r\phi(r) = \frac{2}{\pi^2}s^3 + O(s^2 \log s).$$

¹In all asymptotic estimates here we use a definition of big- O slightly different from the one commonly used in computer science: $f(n)$ is $O(g(n))$ if there exist constants c and n_0 such that $|f(n)| \leq cg(n)$ for all $n \geq n_0$, where $g(n) \geq 0$ for all $n \geq n_0$ (see [1]).

Thus,

$$n = \frac{2}{\pi^2} s^3 + O(s^2 \log s)$$

and, therefore,

$$s = \left(\frac{\pi^2}{2}\right)^{1/3} n^{1/3} + O(\log n).$$

The total number of pairs is

$$\sum_{r=1}^s \phi(r) = 3 \left(\frac{n}{2\pi}\right)^{2/3} + O(n^{1/3} \log n). \quad (7)$$

This concludes the proof of the upper bound on the number of regions. To prove the lower bound, we show that for all n of the form

$$n = \sum_{r=1}^s r \phi(r), \quad (8)$$

there exist two strings of length n whose parametric optimal alignment induces $3(n/2\pi)^{2/3} + O(n^{1/3} \log n)$ optimality regions. For this purpose, let F_n be the set of fractions implied by the preceding argument; i.e.,

$$F_n = \bigcup_{r=1}^s \{a/b : a/b \text{ irreducible and } a + b = r\}.$$

Then,

$$k = |F_n| = \sum_{i=1}^s \phi(i).$$

By a reasoning similar to the one leading to equation (7), we have that $k = 3(n/2\pi)^{2/3} + O(n^{1/3} \log n)$. Let $a_1/b_1 < a_2/b_2 < \dots < a_k/b_k$ be the sorted sequence of elements of F_n and note that

$$\sum_{i=1}^k a_i = \sum_{i=1}^k b_i = n/2.$$

In Lemma 6 below, we show that for any such sequence of fractions, there exists a pair of strings of length n whose global alignment induces $k + 1$ regions.

For n not of the form (8), we choose s such that

$$\sum_{r=1}^{s+1} r\phi(r) > n > \sum_{r=1}^s r\phi(r).$$

By earlier arguments, $s = (\pi^2/2)^{1/3}n^{1/3} + O(\log n)$. Now, use Lemma 6 to build a pair of strings S, T of length $n' = \sum_{r=1}^s r\phi(r)$ inducing $|F_{n'}| = \sum_{r=1}^s \phi(r)$ optimality regions. Substituting for s , we see that the number of regions is $3(n/2\pi)^{2/3} + O(n^{1/3} \log n)$. Let ν_1 and ν_2 be two characters not present in either S or T . Append $n - n'$ copies of ν_1 to S and $n - n'$ copies of ν_2 to T . The number of optimality regions induced by the parametric alignment of the new strings remains $3(n/2\pi)^{2/3} + O(n^{1/3} \log n)$. \square

To complete the proof of Theorem 5, we show the following.

Lemma 6 *Suppose we are given m fractions*

$$\frac{a_1}{b_1} < \frac{a_2}{b_2} < \dots < \frac{a_m}{b_m} \tag{9}$$

such that

$$N_1 = \sum_{k=1}^m a_k \leq n/2 \quad \text{and} \quad N_2 = \sum_{k=1}^m b_k \leq n/2. \tag{10}$$

Then with an alphabet of $N_1 + 2$ letters we can construct two strings of length n such that the (α, β) plane is decomposed into $m + 1$ regions.

Proof. We first define the two strings and then show that there are $m + 1$ different paths in their corresponding alignment graph that become optimal in turn as β grows and α is held at zero.

Let the alphabet be

$$\{0, 1, \omega_1, \dots, \omega_{N_1}\}.$$

Set

$$i(r) = \sum_{k=r}^m b_k \quad \text{and} \quad j(r) = \sum_{k=r}^m a_k$$

for $r = 1, \dots, m$, and

$$i(m+1) = j(m+1) = 0.$$

Denote the strings to be constructed by

$$S = s_1 s_2 \dots s_n$$

and

$$T = t_1 t_2 \dots t_n.$$

String S is defined by

$$s_{i(r)+j(r+1)+k} = \omega_{j(r+1)+k} \quad \text{for } k = 1, \dots, a_r,$$

for each $r = m, m-1, m-2, \dots, 1$, and

$$s_i = 0 \quad \text{for all other values of } i.$$

By (10), $i(1) + j(1) \leq n$ so this construction does not need more than n symbols. String T is defined by

$$t_j = \omega_j \quad \text{for } j = 1, \dots, N_1,$$

and

$$t_j = 1 \quad \text{for } j = N_1 + 1, \dots, n.$$

The only matches between S and T are the matches of the unique occurrences of ω_i in both S and T , for each $i = 1, \dots, N_1$. These matches are arranged in m blocks whose lengths are a_m, a_{m-1}, \dots, a_1 . In S the ω -block of length a_r is preceded by a block of 0's of length b_r , while in T the ω -blocks are all adjacent.

In the alignment graph we have

$$\text{weight}(\sigma(i, j)) = 1 \text{ for } (i, j) = (i(r) + j(r+1) + k, j(r+1) + k),$$

for $k = 1, \dots, a_r$, for each $r = m, m-1, \dots, 1$. For all other values of (i, j) , $\text{weight}(\sigma(i, j)) = 0$. Corresponding to the ω -blocks of matches between S and T are diagonal runs of edges of weight 1 in the alignment graph. The first run of length a_m starts at $(i, j) = (b_m, 0)$.

Now we describe $m + 1$ paths $\Gamma_{m+1}, \Gamma_m, \dots, \Gamma_1$ through the alignment graph, each starting at $(0, 0)$ and ending at (n, n) .

Γ_{m+1} is simply the main diagonal. Along this path there are no matches and no indels, only n mismatches. Since $\alpha = 0$, $score_0(\Gamma_{m+1}) = 0$.

Γ_m takes first b_m steps down, and then runs along the diagonal $(b_m + j, j)$ as $j = 1, \dots, n - b_m$, and lastly takes b_m horizontal steps from $(n, n - b_m)$ to (n, n) . This path picks up one a_m -block of matches and $2b_m$ indels, so $score_0(\Gamma_m) = a_m - 2\beta b_m = j(m) - 2\beta i(m)$.

For $r = m - 1, m - 2, \dots, 1$, path Γ_r picks up the blocks of matches of length a_m, a_{m-1}, \dots, a_r , so altogether $j(r)$ matches. To do so requires $2i(r)$ indels, for the path takes b_k steps down before the match block of length a_k , for $k = m, m - 1, \dots, r$, then after the a_r -block the path stays on the same diagonal until it reaches the bottom row at point $(n, n - i(r))$, and then takes $i(r)$ horizontal steps. Thus $score_0(\Gamma_r) = j(r) - 2\beta i(r)$. The sequence of paths is illustrated in Figure 3.

Next we show that among the Γ_r 's each one is optimal on a particular β -interval. Define $\beta_0 < \beta_1 < \dots < \beta_{m+1}$ by $\beta_0 = 0$, $\beta_r = (1/2)a_r/b_r$ for $r = 1, \dots, m$, and $\beta_{m+1} = \infty$. Thus, for $r = 1, \dots, m$, β_r is the β -value of the intersection of $score_0(\Gamma_r)$ and $score_0(\Gamma_{r+1})$. By Lemma 4, for $r = 1, \dots, m + 1$: if $\beta \in (\beta_{r-1}, \beta_r)$, then $score_0(\Gamma_r) > score_0(\Gamma_k)$ for $k \neq r$.

To conclude the proof, we show that for $\alpha = 0$ and any $\beta \geq 0$, one of the Γ_r 's is an optimizing path. Let Γ be an arbitrary path starting at $(0, 0)$ and ending at (n, n) . Let r be the minimal index among $\{1, 2, \dots, m + 1\}$ such that Γ intersects the diagonal $\{(i(r) + j(r + 1) + k, j(r + 1) + k) : 0 \leq k \leq n - i(r) - j(r + 1)\}$. Then the path must contain at least $2i(r)$ indels. The number of matches on Γ cannot be larger than the total number of matches on this diagonal and above, that is, no larger than $j(r)$. Thus the value attained by Γ is bounded by

$$score_0(\Gamma) \leq j(r) - 2\beta i(r) = score_0(\Gamma_r).$$

In other words, some Γ_r is always among the maximizers. \square

To our knowledge, it is open whether the bound we have just proved can be extended to fixed-size alphabets; indeed, we suspect that this might be impossible. The best we have been able to achieve is $\Omega(\sqrt{n})$, which is shown in the next section.

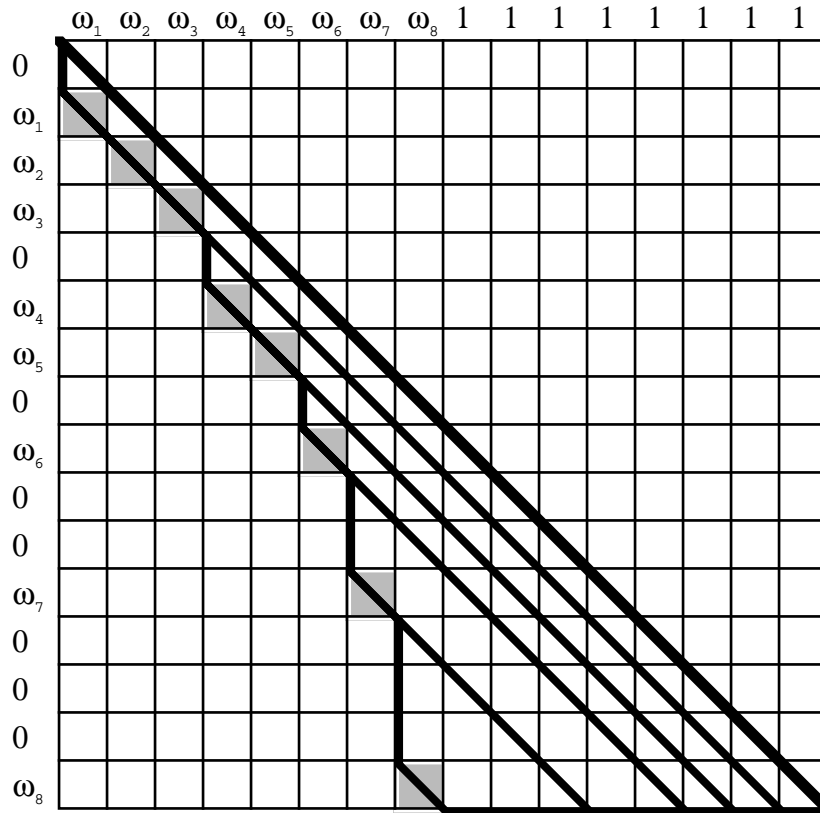


Figure 3: Alignment graph for the two strings corresponding to the fractions $1/3, 1/2, 1/1, 2/1, 3/1$.

4 A lower bound with a finite alphabet

We now prove a lower bound when the alphabet is binary. Observe that, in this case, the match blocks form a checkerboard pattern (see Figure 4).

For each integer $m \geq 1$, define the binary string B_m as

$$B_m = \begin{cases} 0 & \text{if } m = 1 \\ B_{m-1} \cdot 1^m & \text{if } m \text{ is even} \\ B_{m-1} \cdot 0^m & \text{if } m > 1 \text{ and } m \text{ is odd.} \end{cases}$$

Note that B_m is of the form $0^1 1^2 0^3 \dots b^m$, where $b = 0$ for m odd and $b = 1$ for m even, and that its length is $m(m+1)/2$. Let B_m^c denote the bitwise complement of B_m .

Lemma 7 *The global alignment of B_m and B_m^c induces a decomposition of the (α, β) plane into m optimality regions.*

Proof. Note that the alignment graph is symmetric along the main diagonal. Thus, we assume without loss of generality that each optimum solution corresponds to a path whose edges lie either on or below this diagonal.

We define a sequence $\Gamma_1, \Gamma_2, \dots, \Gamma_m$ of paths through the alignment graph for B_m and B_m^c . Let

$$s(r) = \sum_{i=1}^r i.$$

Γ_m is just the main diagonal. Thus, $score_0(\Gamma_m) = 0$. Γ_{m-1} takes one step down, then runs along the diagonal $(1+j, j)$ for $j = 1, 2, \dots, n-1$, and, finally, takes one horizontal step from $(n, n-1)$ to (n, n) . Generally Γ_{m-i} , $1 \leq i \leq m-1$, has three parts:

- (i) First, pick as many matches as possible from i match blocks by repeating the following step for $j = 1, \dots, i$:

Take one step down, then go along the diagonal $(j+k, k)$ for j steps, from $(s(j), s(j-1))$ to $(s(j)+j, s(j))$.

The total number of matches collected along this segment of the path is $s(i)$. The total number of indels is i .

- (ii) Next, go along the diagonal $(i+k, k)$ for $n - (s(i) + i)$ steps, from $(s(i) + i, s(i))$ to $(n, n - i)$. This picks up $(m - i - 1)i$ matches and zero indels.

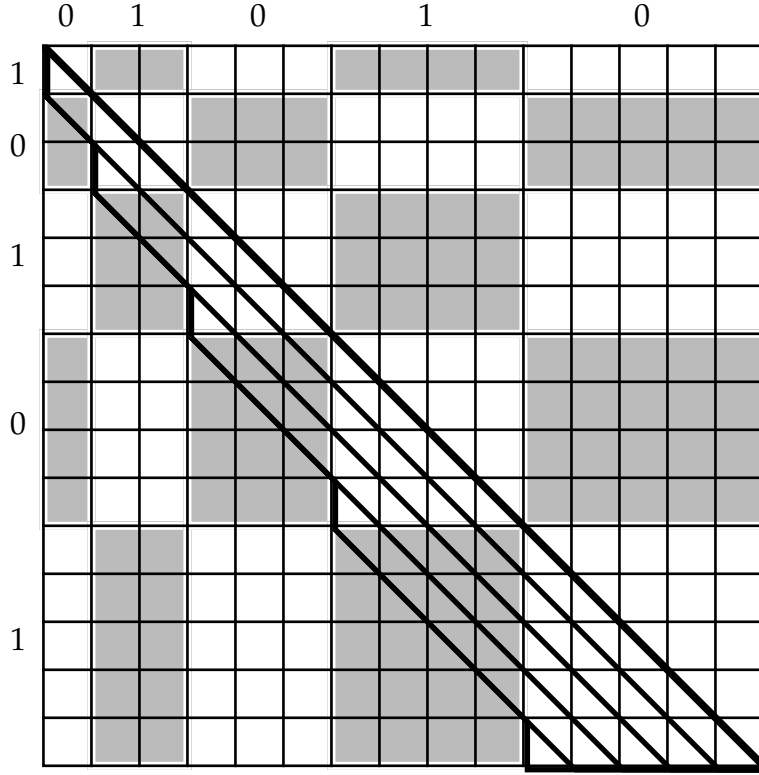


Figure 4: Alignment graph for B_s and B_s^ϵ .

- (iii) Finally, go i steps horizontally from $(n, n - i)$ to (n, n) . This picks up i indels.

The sequence of paths is illustrated in Figure 4.

Observe that

$$\text{score}_0(\Gamma_{m-i}) = s(i) + (m - 1 - i)i - 2i\beta,$$

i.e.,

$$\text{score}_0(\Gamma_i) = s(m - i) + (i - 1)(m - i) - 2(m - i)\beta.$$

Thus, for $i = 1, \dots, m - 1$, the β -value such that $\text{score}_0(\Gamma_i) = \text{score}_0(\Gamma_{i+1})$ is $\beta_i = i/2$. Let $\beta_0 = 0$ and $\beta_m = \infty$. We claim that for $r = 1, \dots, m$, Γ_r is an optimum path for $\beta \in (\beta_{r-1}, \beta_r)$.

Since the conditions of Lemma 4 apply to the set of paths defined above, we have the following for $r = 1, \dots, m$: if $\beta \in (\beta_{r-1}, \beta_r)$, then $score_0(\Gamma_r) > score_0(\Gamma_k)$ for $k \neq r$. It remains to show that, for every path Γ starting at $(0, 0)$ and ending at (n, n) and every $\beta \geq 0$, $score_0(\Gamma) \leq \max_{1 \leq i \leq m} score_0(\Gamma_i)$. For this purpose, let r be the largest index among $\{0, 1, \dots, m-1\}$ such that Γ intersects diagonal $(r+k, k)$. This path must contain at least $2r$ indels and can have no more matches than Γ_{m-r} . Thus, $score_0(\Gamma) \leq score_0(\Gamma_{m-r})$ for $\beta \geq 0$ and the proof is complete. \square

Theorem 8 *For every n , there exist pairs of binary strings of length n whose parametric optimal global alignments induce a decomposition of the (α, β) plane into $\sqrt{2n} + O(1)$ optimality regions.*

Proof. For n of the form $n = i(i+1)/2$, we take as our strings B_i and B_i^c . By Lemma 7, these induce i optimality regions. Since $i = \sqrt{2n} + O(1)$, the bound follows.

For all other n , choose i such that

$$\frac{(i+1)(i+2)}{2} \geq n \geq \frac{i(i+1)}{2}.$$

Let D_n be the string obtained by appending $n - i(i+1)/2$ 1's if i is even and the same number of 0's if i is odd. Then, strings D_n and D_n^c induce i optimality regions. Since, again, $i = \sqrt{2n} + O(1)$, the proof is complete. \square

5 Multiple sequence alignment

A *multiple alignment* \mathcal{A} of strings S_1, \dots, S_k , where S_i has length n_i , is obtained by inserting spaces in each string so that the resulting strings have the same length l . The result is a matrix with k rows and l columns, such that each character and space of each string appears in exactly one column. Note that any such \mathcal{A} induces a pairwise alignment of S_i and S_j in a natural way: remove all rows of \mathcal{A} except those corresponding to S_i and S_j and strike out any columns containing two spaces. This will be called the *induced pairwise alignment of S_i and S_j* .

Multiple sequence alignment is at least as important in biology as pairwise alignment, because it allows one to identify preserved patterns among a variety of species, which can give clues about similarities in molecular structure or function. Different ways have been proposed for scoring a multiple alignment \mathcal{A} . A common scheme is the *sum-of-pairs (SP) score* [2], which is just the sum of the scores of all pairwise alignments induced by \mathcal{A} . That is, if $score_{ij}(\mathcal{A})$ denotes the score of the pairwise alignment between S_i and S_j induced by \mathcal{A} , the SP score of \mathcal{A} is

$$score(\mathcal{A}) = \sum_{i < j} score_{ij}(\mathcal{A}). \quad (11)$$

Here

$$score_{ij}(\mathcal{A}) = w_{ij} - \alpha x_{ij} - \beta y_{ij}, \quad (12)$$

where w_{ij} , x_{ij} , and y_{ij} denote the number of matches, mismatches, and spaces in the pairwise alignment between S_i and S_j induced by \mathcal{A} .

By (2),

$$2w_{ij} + 2x_{ij} + y_{ij} = n_i + n_j, \quad (13)$$

which leads to the following extension of Lemma 2.

Lemma 9 *Any two multiple alignments \mathcal{A} and \mathcal{A}' have the same SP score for $\alpha = -1$, $\beta = -1/2$.*

Proof. By (12) and (13) if we fix $\alpha = -1$, $\beta = -1/2$, then for any i, j , $score_{ij}(\mathcal{A}) = score_{ij}(\mathcal{A}') = w_{ij} + x_{ij} + y_{ij}/2 = (n_i + n_j)/2$. Hence, $score(\mathcal{A}) = score(\mathcal{A}') = \sum_{i < j} (n_i + n_j)/2$ at $\alpha = -1$, $\beta = -1/2$. \square

Thus, we have the following generalization of Lemma 3.

Lemma 10 *All optimality regions for parametric multiple sequence global alignment are semi-infinite cones, and are delimited by the coordinate axes or by lines of the form $\beta = c + (c + 1/2)\alpha$ for some constant c .*

A direct consequence of the structure implied by the preceding result is that the entire parameter space decomposition can be constructed in time proportional to that required to find a single multiple alignment (which is

$O(n^k)$, assuming $n_i = n$ for all n) multiplied by the number of regions. The reasoning is identical to that of Theorem 2.3 of [5]. All that remains is to bound the number of regions.

Theorem 11 *The number of optimality regions in the (α, β) plane induced by the parametric multiple global alignment of k strings of length n over an unbounded alphabet is at most $3 \binom{k}{2} n / 2\pi^{2/3} + O(n^{1/3} k^{2/3} (\log n + \log k))$. For k fixed, this bound is tight within a constant factor.*

Proof. The proof is similar to that of Theorem 5. Number the optimality regions from bottom to top and let $w^{(l)}$, $x^{(l)}$, $y^{(l)}$ be the total number of matches, mismatches, and spaces in all pairwise alignments induced by the optimal multiple alignment \mathcal{A}_l for region l . That is,

$$w^{(l)} = \sum_{i < j} w_{ij}^{(l)},$$

and analogously for $x^{(l)}$, and $y^{(l)}$. Thus,

$$\text{score}(\mathcal{A}_l) = w^{(l)} - \alpha x^{(l)} - \beta y^{(l)}.$$

The equation of the intersection line between the $(l-1)$ th and the l th region is given by

$$\beta = \frac{\Delta w^{(l)}}{\Delta y^{(l)}} + \frac{\Delta x^{(l)}}{\Delta y^{(l)}} \alpha,$$

where $\Delta w^{(l)} = w^{(l-1)} - w^{(l)}$, $\Delta x^{(l)} = x^{(l)} - x^{(l-1)}$, and $\Delta y^{(l)} = y^{(l-1)} - y^{(l)}$. Therefore,

$$\frac{\Delta w^{(l)}}{\Delta y^{(l)}} < \frac{\Delta w^{(l+1)}}{\Delta y^{(l+1)}} \quad \text{and} \quad \frac{\Delta x^{(l)}}{\Delta y^{(l)}} < \frac{\Delta x^{(l+1)}}{\Delta y^{(l+1)}}. \quad (14)$$

By (13),

$$2w^{(0)} + 2x^{(0)} + y^{(0)} = \sum_{i < j} \left(2w_{ij}^{(0)} + 2x_{ij}^{(0)} + y_{ij}^{(0)} \right) = (k-1) \sum_{i=1}^k n_i.$$

Hence, if all n_i 's equal n , we must have

$$w^{(0)} + \frac{y^{(0)}}{2} \leq \binom{k}{2} n.$$

Now, if there are $r + 1$ optimality regions, we must have

$$\sum_{l=1}^r \left(\Delta w^{(l)} + \frac{\Delta y^{(l)}}{2} \right) \leq w^{(0)} + \frac{y^{(0)}}{2},$$

and, therefore,

$$\sum_{l=1}^r \left(\Delta w^{(l)} + \frac{\Delta y^{(l)}}{2} \right) \leq \binom{k}{2} n. \quad (15)$$

Thus, our goal is to count the number of distinct fractions $\Delta w^{(l)}/\Delta y^{(l)}$ subject to (15). Note that $\Delta y^{(l)}$ must be even since, for every i, j , the number of spaces in the pairwise alignments between S_i and S_j induced by \mathcal{A}_i and \mathcal{A}_{i-1} differ by an even amount. Thus, our problem is equivalent to bounding the number of distinct irreducible fractions a_i/b_i such that

$$\sum_{i=1}^r (a_i + b_i) \leq \binom{k}{2} n.$$

The upper bound now follows from a reasoning similar to that for Theorem 5; we omit the details.

We can obtain k sequences whose parametric global alignment induces $\Omega(n^{2/3})$ regions by letting S_1 and S_2 be the strings S and T from the lower bound construction of Theorem 5 and letting S_3, \dots, S_k be copies of S_2 . \square

This theorem and the discussion that precedes it imply the following.

Corollary 12 *The entire parameter space decomposition for k -sequence global alignment under SP scoring can be computed in $O(k^{4/3}n^{k+2/3})$ time.*

6 Experimental Results

Experience in attempting to construct pairs of sequences that exhibit the worst-case behavior of Theorem 5 suggests that sequences with such behavior are rare. A natural but challenging question is to determine the expected number of regions. While we have no analytical results, we do have experimental data from randomly-generated pairs of strings. Figure 5 is representative of what we encountered. The plot was obtained by considering

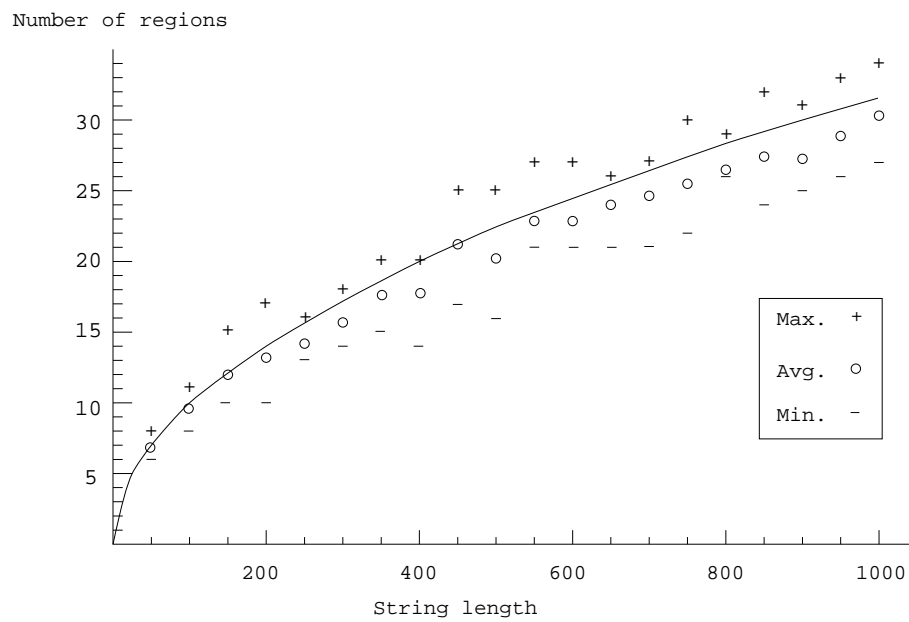


Figure 5: The number of regions as a function of string length n . The solid curve is \sqrt{n} .

pairs of strings of lengths between 50 and 1000. For each length, 10 pairs of strings over an alphabet of size 10 were generated uniformly at random. The figure plots the maximum, minimum, and average for each length. The striking feature of this diagram is how closely the average number of regions approximates the square root of the string length. In fact, even the extreme values are rather tightly clustered around \sqrt{n} , leading one to suspect that the probability distribution has a sharp peak close to \sqrt{n} . Thus, we make the following conjecture.

Conjecture 1 *The expected number of optimality regions in the (α, β) plane induced by the parametric global alignment of a pair of strings of length n is $\Theta(\sqrt{n})$ and the probability distribution is sharply concentrated around its peak.*

We suspect that the constant implicit in the Θ -bound above is close to 1.

We chose the alphabet size in the preceding experiment by trying various sizes to find one that tended to maximize the number of regions over the entire range of string lengths considered. Alphabet size does have some effect on the number of regions. To examine this more closely, we did the following experiment. Choose a string length n and then consider alphabet size i for i ranging from 2 to some upper bound U . For each successive value of i (by increments of one), generate some fixed number p of random pairs of strings over an alphabet of size i and compute the average number of regions. Figure 6 shows the results obtained when $(n, U, p) = (100, 100, 10)$, and are representative of what we encountered for other values of n . As might be expected, binary strings induce the smallest average number of regions, but this average peaks rather early (in the case of $n = 100$, the peak occurred at alphabet size around 10), after which it trails slowly. Similar plots for other string lengths show that the peak depends on n . While it is not clear what the dependence might be, we conjecture that the alphabet size that maximizes the expected number of regions is in the neighborhood of \sqrt{n} . In any event, dependence on alphabet size seems relatively minor compared to the importance of string length. To further confirm this, we repeated the experiments reported in Figure 5 with alphabet size varying as \sqrt{n} and $\sqrt{2n}$. The results were not substantially different.

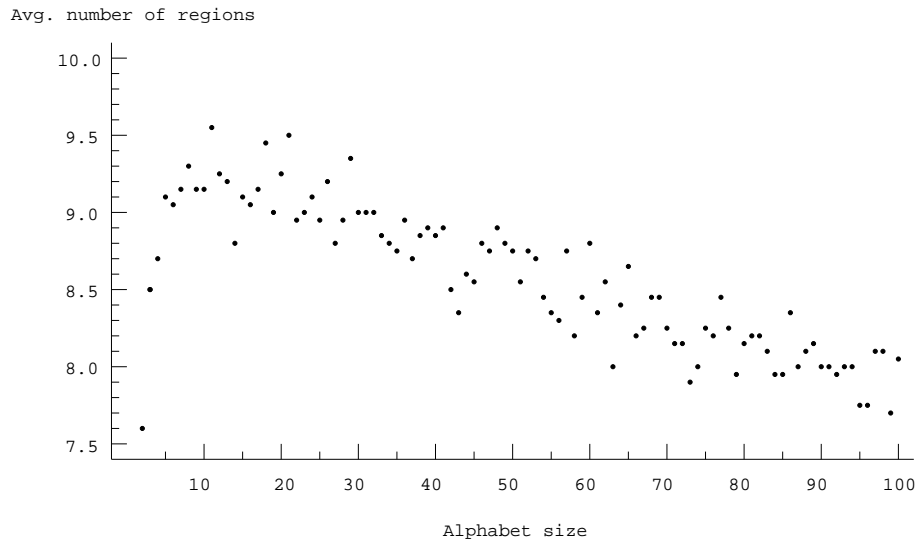


Figure 6: Effect of alphabet size on the number of regions.

Acknowledgments

We thank Steven LaValle for letting us use the Robotics Lab computers to run the computational experiments reported here and Cliff Bergman for help with displaying the results graphically.

References

- [1] T. M. Apostol. *Introduction to Analytic Number Theory*. Springer-Verlag, New York, 1976.
- [2] H. Carrillo and D. Lipman. The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.*, 48:1073–1082, 1988.
- [3] D. Fernández-Baca, T. Seppäläinen, and G. Slutzki. Parametric multiple sequence alignment and phylogeny construction. In R. Giancarlo and D. Sankoff, editors, *Combinatorial Pattern Matching*, volume 1848 of *Lecture Notes in Computer Science*, pages 69–83. Springer-Verlag, 2000.

- [4] D. Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge–New York–Melbourne, 1997.
- [5] D. Gusfield, K. Balasubramanian, and D. Naor. Parametric optimization of sequence alignment. *Algorithmica*, 12:312–326, 1994.
- [6] D. Gusfield and P. Stelling. Parametric and inverse-parametric sequence alignment with XPARAL. In Russell F. Doolittle, editor, *Computer methods for macromolecular sequence analysis*, volume 266 of *Methods in Enzymology*, pages 481–494. Academic Press, 1996.
- [7] X. Huang, P. A. Pevzner, and W. Miller. Parametric recomputing in alignment graphs. In M. Crochemore and D. Gusfield, editors, *Combinatorial Pattern Matching*, volume 807 of *Lecture Notes in Computer Science*, pages 87–101. Springer-Verlag, 1994.
- [8] P. A. Pevzner. *Computational Molecular Biology*. MIT Press, Cambridge, MA, 2000.
- [9] P. A. Pevzner and M. S. Waterman. Open combinatorial problems in computational molecular biology. In *Proc. Third Israeli Symposium on Theory of Computing and Systems*, pages 158–173. IEEE Computer Society Press, 1995.
- [10] D. Sankoff and J. B. Kruskal, editors. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, MA, 1983.
- [11] M. Vingron and M. S. Waterman. Sequence alignment and penalty choice: Review of concepts, case studies, and implications. *J. Mol. Biol.*, 235:1–12, 1994.
- [12] M. S. Waterman, M. Eggert, and E. Lander. Parametric sequence comparisons. *Proc. Natl. Acad. Sci. USA*, 89:6090–6093, 1992.