Chapter 51
Parametric Optimization of Sequence Alignment*

D. Gusfield[†] K. Balasubramanian[†] D. Naor[†]

Abstract

The *optimal alignment* or the *weighted minimum edit distance* between two DNA or amino acid sequences for a given set of weights is computed by classical dynamic programming techniques, and is widely used in Molecular Biology. However, in DNA and amino acid sequences there is considerable disagreement about how to weight matches, mismatches, insertions/deletions (indels) and gaps. *Parametric Sequence Alignment* is the problem of computing the optimal valued alignment between two sequences as a *function* of variable weights for matches, mismatches, spaces and gaps. The goal is to partition the parameter space into regions (which are necessarily convex) such that in each region one alignment is optimal throughout and such that the regions are maximal for this property. In this paper we are primarily concerned with the structure of this convex decomposition, and secondarily with the complexity of computing the decomposition. The most striking results are the following: For the special case where only matches, mismatches and spaces are counted, and where spaces are counted throughout the alignment, we show that the decomposition is surprisingly simple: all regions are infinite; there are at most $n^{2/3}$ regions; the lines that bound the regions are all of the form $\beta = c + (c + 0.5)\alpha$; and the entire decomposition can be found in $O(knm)$ time, where k is the actual number of regions, and $n < m$ are the lengths of the two strings. These results were found while implementing a large software package to do parametric sequence analysis, and in turn have led to faster algorithms for those tasks.

1 Introduction

Finding the minimum cost edit distance, or the best alignment, of two DNA, RNA or amino acid sequences has become almost the standard technique for sequence comparison in molecular biology. They are used to determine whether and where two sequences are similar (homologous), to determine evolutionary history between species, to find consensus sequences and other significant functions. There are literally hundreds of papers written on this topic and its applications to biology. For an introduction and small reflection of this literature see [D1] [D2] [FS] [GO] [GD] [PL] [SK] [vH] [W].

However, in all present methods for optimal sequence alignment, specific substitution, insertion/deletion (indel) and gap penalties must be specified, and the (biological) significance of the alignment depends heavily upon the "right" choice of the weights. There is considerable disagreement among molecular biologists about the correct choice, and it is probably the case that there is no unique choice for the parameters (as pointed out by [AV] with respect to gap penalties). The significance of an alignment is based either on biological grounds, or on its sensitivity to the choice of parameters. Instead of repeatedly varying the parameter weights and solving for the optimal alignment, other parametric methods ought to be employed. As an example, J. Kruskal and D. Sankoff [SK, pp. 290–293] demonstrate the difficulties in finding the relative weights for gaps/substitutions and other operations by a specific example: the comparison of human and *E. Coli* 5S RNA (two sequences of 120 characters each over a 4-letter alphabet). Their solution involves varying one parameter, the number of *indels*, until its appropriate value is found. Similarly, the paper by Fitch and Smith [FS] demonstrate how the biologically accepted alignment may easily be missed if inappropriate weights are used.

There are two ways around the problem of choosing a "correct" choice of parameters. The first is to compute, for a given set of initial guesses for parameter values, the optimal alignment $\mathcal{A}$ at that point and, in addition, a maximal region P such that $\mathcal{A}$ is optimal throughout P. The other, more global, approach is to find the entire decomposition of the parameter space into such maximal regions. During the process of developing and analyzing a computer program which is based on these approaches, we were able, in several cases, to characterize this decomposition, and this considerably simplified the algorithm for finding it.

We formally define the *Parametric Alignment* problem in Section 1.1 and summarize our results in Section 1.2. In Sections 2 and 3 we consider the 2-parameter case, where only substitution and indel weights can vary. Two different variants, the global and the local alignment, are studied. Section 4 deals with the 3-parameter

*Research partially supported by grant DE-FG03-90ER60999 from the Department of Energy, and grants CCR-8803704 and CCR-9103937 from the National Science Foundation.

[†]Computer Science, University of California, Davis

case. Finally, in section 5 we discuss a richer variant of the problem, where independent weights for each type of mismatch are specified.

1.1 DefinitionsThe *edit distance* between two sequences is the minimum weighted sequence of *edit operations* (insertion, deletion and substitution of single characters) that must be performed to transform one sequence into another. This has found widespread use as a measure of sequence similarity. It is often more important to know what the actual edit operations are rather than just the total cost or value. These operations can be represented as an *alignment*, and it is often the alignment that is searched for.

An *alignment* of two sequences $\mathcal{S}_1$ and $\mathcal{S}_2$ of lengths n and m ($\geq n$) respectively is obtained by introducing spaces into the two sequences such that the lengths of the two resulting sequences are identical, and placing these two resultant sequences one upon the other subject to the constraint that no column contains two spaces. Any column that contains two identical characters is called a *match*. Any column that contains two dissimilar characters is called a *mismatch* and any column that contains a space will be referred to as a *space* or *indel*. The correspondence between them and the edit operations is straightforward: a mismatch represents a substitution, a space is either an insertion or a deletion, depending whether it is introduced in the source string or the target string, and a match is an untouched character. A series of one or more contiguous space characters in the same sequence will be referred to as *gap*.

An alignment $\mathcal{A}$ may therefore be characterized by the number of matches, mismatches, spaces and gaps. We denote these quantities by $w_{\mathcal{A}}, x_{\mathcal{A}}, y_{\mathcal{A}}, z_{\mathcal{A}}$ respectively (or w, x, y, z when referring to an unspecified alignment). Note however that this representation is many-to-one: different alignments could have the same 4-tuple (w, x, y, z). If α_0 is the mismatch penalty, β_0 is the space (indel) penalty and γ_0 is the gap penalty, then the *value* of an alignment is defined to be

$$v = w - \alpha_0 x - \beta_0 y - \gamma_0 z$$

The region of interest is the region where α, β, γ are all positive. We ignore the case where the weight of the matches is also a parameter since we can divide all the parameters by this value and reduce it to the above case without changing the relative order of the value of the alignments. For fixed weights, the value of the optimal (maximum value) alignment of two strings can be found by dynamic programming in $O(nm)$ time [SK], a fact that was discovered many times independently.

A given choice of parameter values α_0, β_0 and γ_0 defines an optimal alignment (not necessarily unique). Since an alignment is essentially a discreet object any alignment that is optimal for some fixed values $(\alpha_0, \beta_0, \gamma_0)$ is optimal in a certain *region* in the (α, β, γ) space. Hence, the 3-parameter space is decomposed into regions which we call *optimal regions* such that in every region one alignment is optimal throughout and the regions are maximal for this property. This decomposition is completely defined by the two sequences.

The value of an alignment is always a linear function of the parameters; hence it can be easily observed that the regions, which are bounded by the intersection of hyperplanes, are all convex polygons.

We consider two variants on the alignment problem: the *global* and the *local* alignment. By *global*, we refer to the case when all spaces are counted towards the total number of spaces regardless of where they occur. The *local* case is when we ignore any (contiguous) spaces that overhang at the *extreme ends* of the alignment, *i.e.* we are permitted to delete one suffix and one prefix (perhaps of the same sequence) without incurring any cost or losing any value. *Global* alignments are used when searching for a global homology between two sequences, that is, when the whole sequences are expected to be homologous. This is true more often of proteins than in the study of DNA. However global alignment problems also arise as subtasks of more complex alignment problems. *Local* alignments are used when searching for local homology, that is, when the sequences are expected to consist of a homologous region which should be aligned, and non-homologous suffixes or prefixes which need not participate in an alignment.

1.2 Summary of ResultsIn this paper we are concerned with the characteristics of the parametric space decomposition, and its algorithmic implications. We are also interested in questions of the type: given an arbitrary line in the parametric space, how many different regions can it go through. The latter was motivated by our computer program *PARAL* which is based on a primitive operation that, when given a line in the space, finds all the regions it crosses in time $O(knm)$, where k is the actual number of regions it goes through (using the algorithm of Eisner-Severance [ES]).

We first consider a case of particular interest, the 2-dimensional case where gaps are ignored (that is $\gamma = 0$). Here, the optimal alignment is the one that maximizes $w - \alpha x - \beta y$. We show that when *global* alignments are considered, the decomposition of the α, β space is surprisingly structured. We prove that the number of (convex) polygons in the decomposition is bounded by $n^{2/3}$. We show further that every polygon is infinite and is bounded by two rays, each of which runs along a line of the form $\beta = c + (c + 0.5)\alpha$ for some constant c. As a consequence, we show that the *entire* decomposition can be found in $O(knm)$ time, where k is the number of actual polygons in the decomposition. Hence, the amortized cost for finding a single region is $O(nm)$ time, which is also the time to find a *single* optimal alignment within that region. For *local* alignments, the decomposition becomes more complex; some of its regions can be bounded (finite), but their number is bounded by n^2. Any arbitrary line in the 2-dimensional space can therefore go through at most n^2 regions. As a consequence, the algorithm implemented in PARAL finds all the regions in time (n^3m) per region, and an alternative method of Gusfield [G], which is an adaptation of Meggido's method [M], can find all regions in time $O(nm^2 \log n)$ per region. These algorithms are not described in this abstract.

If gaps are allowed, then we obtain similar results for the decomposition of the 3-dimensional space if *global* alignments are considered. All regions, which are convex polyhedra, are unbounded cones, bordered by rays of the form $\beta = c + (c + 0.5)\alpha$, $\gamma = d + d\alpha$. An arbitrary line in the 3-dimensional space can go through at most n^2 regions if *global* alignments are computed, and through at most n^3 regions for *local* alignments.

A much more complex case arises when different weights are assigned to every possible mismatch (for example, in Amino Acid sequences, where the alphabet size is 20, there are 190 possible mismatch weights). In this most general setting, we show a sub-exponential bound on the number of regions that any straight line can intersect.

2 Two Parameter Global Alignment

In this section we consider the global alignment problem where only two parameters α and β are given. We ignore the number of gaps, so $\gamma = 0$. The objective function, therefore, is to maximize $w - \alpha x - \beta y$, where w, x and y are the numbers of matches, mismatches and indels respectively. We are interested in bounding the number of regions in the (first quadrant of) the α, β plane. We establish the following lemmas and observations for this purpose.

LEMMA 2.1. *For any alignment $\mathcal{A}$ with corresponding tuple (w, x, y): $2w + 2x + y = N$, where $N = n + m$ is the sum of the sequence lengths.*

Proof : Any character can be part of exactly one match, mismatch or indel. A match or a mismatch involves two characters. Thus the total number of characters that form part of a match is $2w$. Similarly the total number of characters involved in mismatches is $2x$. A indel involves only one character from the input sequences and since we do not ignore any indels in counting their total number it follows that the number characters involved in indels is y. The lemma follows.

Recall that $m > n$.

LEMMA 2.2. *For any alignment $\mathcal{A}$, $w + x \leq n$*

Proof: A match or mismatch involves one character from *each* sequence. Hence their total cannot exceed the number of characters in the shorter sequence.

COROLLARY 2.1. *For any alignment, $\mathcal{A}$, $m - n \leq y \leq m + n$*

COROLLARY 2.2. *In all alignments of two sequences, y is always odd or always even depending on whether $m + n$ is odd or even. There are therefore only $n + 1$ distinct values for y.*

THEOREM 2.1. *Any line forming a boundary between two regions is of the form $\beta = c \ + (c + \frac{1}{2})\alpha$, for some $c > -1/2$.*

Proof : At a given point (α, β) an alignment has the value $v = w - \alpha x - \beta y$. The left hand side of lemma 2.1, which can be rewritten as $w + x + y/2 = (n + m)/2$, is also a linear combination of w, x and y, with $\alpha = -1$ and $\beta = -1/2$. This suggests that this point is of some significance. Therefore consider the point $(-1, -1/2)$ on the α, β plane. At this point all alignments have the same value: $v = w + x + y/2 = (n + m)/2$ (from lemma 2.1). In other words, *all the value planes must meet at $\alpha = -1, \beta = -1/2, v = (n + m)/2$.* It follows then that any intersection between two such planes must also pass through that point. Now, let $\beta = c + c_1\alpha$ (for some c_1 and c) be a boundary (intersection) line. Since $(-1, -1/2)$ is a point on that line, $-1/2 = c - c_1$. Hence $c_1 = c + 1/2$. In other words, $\beta = c \ + (c + \frac{1}{2})\alpha$, for some c. Since we are only interested in the quadrant where β and α are positive, it follows that any line passing through that quadrant and the point $(-1, -1/2)$ must have a positive slope. Hence $c > -1/2$.

COROLLARY 2.3. *All the regions of optimality are semi-infinite regions bounded by lines of the form $\beta = c \ + (c + \frac{1}{2})\alpha$ or by the co-ordinate axes.*

An example (using two made up sequences) of the decomposition of the parameter space into regions of optimality, displaying the above property, is shown in Figure 1.

2.1 The number of regions We now examine the number of regions possible in any decomposition. A *breakpoint* along any given line is the point where the line moves between two adjacent regions.

LEMMA 2.3. *Along any horizontal line we never encounter break points in the region $\alpha > 2\beta$.*

Proof : Consider an alignment which contains at least one mismatch. A single mismatch may always be replaced by two indels, one in each sequence, so alignment containing mismatches can be changed into one with no mismatches without affecting the number of matches. Thus if the cost of a mismatch, α, is greater than twice that of an indel, β, it follows that any optimal alignment for those parameters will have no mismatches.

Consider now a horizontal line in the region $\alpha > 2\beta$. Any alignment that is optimal at a point in this region can have no mismatches. Thus the value of such an optimal alignment remains constant through this region on any horizontal line (where β is constant). Hence there are no breakpoints on a horizontal line in this region.

LEMMA 2.4. *There are at most $n+1$ regions.*

Proof: Lemma 2.3, with $\beta = 0$, shows that we will encounter no break points along the α axis and that therefore all the region boundaries must intersect the positive β axis. In other words the β axis intersects all the regions. Let the alignments encountered, in order of increasing β, be $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_{k+1}$. Since $\alpha = 0$ the value of an alignment (along this line) is simply $v_i(\beta) = w_i - \beta y_i$. Since our objective function aims to maximize the alignments value, it can be seen that $y_{i+1} < y_i$ for all $\mathcal{A}_i$ $(i < k)$. By corollary 2.2, the y_i can attain only $n+1$ distinct values and the lemma follows.

Let $w_i - \alpha x_i - \beta y_i$ and $w_j - \alpha x_j - \beta y_j$ be the planes corresponding to the values of two alignments $\mathcal{A}_i$ and $\mathcal{A}_j$ respectively. The equation of the line (or line segment) forming the boundary between the two regions with alignments $\mathcal{A}_i$ and $\mathcal{A}_j$ is:

$$\beta = \frac{w_i - w_j}{y_i - y_j} + \frac{x_j - x_i}{y_i - y_j}\alpha$$

We will call this the *ratio form* of the boundary line[1]. The ratio form of the boundary line suggests an added constraint on the number of mismatches x, and this constraint can be used to further refine lemma 2.4.

THEOREM 2.2. *The number of regions is bounded by $O(n^{\frac{2}{3}})$* .

Proof: We have seen that the boundaries between regions are of the form $\beta = c + (c + \frac{1}{2})\alpha$. Thus we may specify a boundary simply be specifying the slope $m = c + 1/2$. Consider the β axis which, as we have seen, intersects all the regions. Let the alignments encountered, in order of increasing β, be $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_{k+1}$. Let them be separated by boundary lines with slopes $m_1, m_2, \ldots, m_k$ respectively. We have seen that $y_{i+1} < y_i$ for all $\mathcal{A}_i$ $(i < k)$. The slopes m_i are positive since we are interested only in the quadrant where α and β are positive, and any line from the point $(-1, -1/2)$ that intersects that quadrant must have a non-zero (and non-infinite) positive slope. Let $\Delta x_i = x_{i+1} - x_i$ and $\Delta y_i = y_i - y_{i+1}$. Since $y_{i+1} < y_i$, Δy_i is positive. From the ratio form of the boundary line, $m = \Delta x_i / \Delta y_i$ therefore, Δx_i is positive. Thus a boundary slope can be identified by a pair of values $(\Delta x_i, \Delta y_i)$, and there can be no more boundaries than there are distinct $(\Delta x_i, \Delta y_i)$ pairs.

Now, $\sum_i \Delta x_i \leq n$ and $\sum_i \Delta y_i \leq (m+n) - (m-n) = 2n$. Therefore

$$3n \geq \sum_{i=1}^{k} (\Delta x_i + \Delta y_i)$$

For all t consider the number of boundaries such that $(\Delta x_i + \Delta y_i) = t$. Since the pairs have to be distinct there can be only $t-1$ such pairs. Intuitively the way to maximize the number of $(\Delta x_i, \Delta y_i)$ pairs is to generate as many as possible which sum to two, then three and so on as long as the constraint on the total is maintained. Let s be the largest value such that

$$\begin{aligned} 3n &\geq \sum_{t=2}^{s} t(t-1) \\ &= \frac{1}{3}(s-1)s(s+1) \end{aligned}$$

Therefore $s = O(n^{\frac{1}{3}})$. The number of regions is therefore

$$\begin{aligned} k \quad &\leq \textstyle\sum_{t=1}^{s+1}(t-1) \quad &= \frac{s(s+1)}{2} \\ &= O(s^2) &= O(n^{\frac{2}{3}}) \end{aligned}$$

A closer analysis based on the same observation and using Euler's function $\Phi(i)$, the number of integers less than and relatively prime to i, can be used to show that $k \leq 0.88n^{2/3}$.

THEOREM 2.3. *All regions in the decomposition can be found in $O(nm)$ time per region, i.e. the time per region is no more than the time to compute a single alignment at a fixed α, β point.*

Proof : The Eisner-Severance method [ES] finds all the breakpoints (intersections between regions) along any single line or direction in $O(nm)$ time per breakpoint We have seen that a single line, the β axis, intersects all the regions. Hence all the regions can be found in $O(knm)$ time by the Eisner-Severance method

[1] Writing the intersection of the two planes in this way also gives an interesting interpretation of the boundary between two regions. The slope of the boundary is the rate at which mismatches are exchanged for indels in the two adjacent alignments, and the intercept of the boundary line is the rate at which matches are exchanged for indels.

where k is the number of regions. This actually gives us the intercepts along the β axis but this information is enough to determine the boundary lines since each line must pass through $(-1, -1/2)$.

COROLLARY 2.4. *The entire decomposition can be found in $O(n^{5/3}m)$ time.*

3 Two Parameter Local Alignment

In this section we consider the regions of optimality generated by looking at local instead of global alignments. In this case certain spaces occurring at the very end of the alignment may be disregarded. This renders lemma 2.1 invalid. However we note that lemma 2.2 remains valid. We may also make the following weaker observation.

LEMMA 3.1. *For any local alignment,*

$$2w + 2x + y \leq N$$

where $N = n + m$ is the sum of the sequence lengths.

However, we note that any "extreme" spaces *will not* be counted (since any additional space will always decrement the value). Thus if space is "counted", it must be the case that there is at least one match/mismatch on either side of it, else it would have therefore, if $y > 0$ then we must have at least two matches/mismatches consuming 4 characters. Hence,

LEMMA 3.2. *For any alignment,*

$$y \leq m + n - 4$$

The weakening of these conditions leads one to suspect that the picture of all the optimal regions may be more complicated and not show the structure that we observed in the global alignment case. This has indeed been observed to be the case. The example in Figure 2, using the same sequences as in Figure 1 illustrates this increased complexity.

We further note that lemmas 2.3 remains valid in the local alignment case since it did not require any of the abovementioned conditions for its proof.

LEMMA 3.3. *There are at most $O(n^2)$ optimal regions.*

Proof: Consider two alignments $\mathcal{A}_1$ and $\mathcal{A}_2$ which have tuples (w, x, y_1) and (w, x, y_2), *i.e.*, they differ only in the number of indels. Let us assume that they are both optimal in some region. Without loss of generality, assume that $y_1 < y_2$. since we are interested in the region where the mismatch and indel penalties are always positive, it follows that $\mathcal{A}_1$ will always have a larger value than $\mathcal{A}_2$, which contradicts our assumption that there was some region where $\mathcal{A}_2$ was optimal.

In other words, it is not possible for two different alignments to have the same value of w and x. Thus there can be only as many regions as there are distinct pairs w, x. This proves the lemma.

4 Three Parameter Alignment

We now consider the case where gap penalties can also vary as a parameter γ, hence the parameter space is now 3-dimensional. The consideration of gap penalties is very important in the context of DNA or Protein sequences since in many cases the alignments accepted as "standard" or best by biologists cannot be obtained by the dynamic programming approach unless a specific non-zero penalty is added for each gap. For an example of this see [FS]. As in the 2-dimensional case, we can prove a simpler decomposition when global alignments are considered. Recall that z is the number of gaps in the alignment. The following observation holds in general:

LEMMA 4.1. $z \leq 2n - 1$

Proof: Any space that is introduced in the long sequence must be opposite to a character in the short sequence, hence the number of spaces, and therefore the number of gaps, in the long sequence is bounded by n. Likewise, any gap in the short string must be bracketed by characters at both ends, hence the number of gaps in the short string is bounded by $n - 1$. The claim follows.

4.1 Global Alignments with GapsWe first note

that lemmas 2.1 and 2.2 and corollaries 2.2 and 2.1 from section 2 are still valid in this three parameter global alignment casee, for exactly the same reasons, since we are not changing the way we count indels, mismatches or spaces.

Let us describe a line in 3-dimensions α, β and γ as a function of one parameter, α, by two equations $\beta = c_0 + c_1\alpha$ and $\gamma = c_2 + c_3\alpha$.

THEOREM 4.1. *Any line forming a boundary between three or more regions is of the form $\beta = c + (c + \frac{1}{2})\alpha$, $\gamma = d + d\alpha$.*

Proof: Consider the point $(-1, -1/2, 0)$ on the α, β, γ plane. At this point all the alignments have the same value: $v = w + x + y/2 = (n + m)/2$ (from lemma 2.1). In other words all the value hyperplanes must meet at $\alpha = -1, \beta = -1/2, \gamma = 0, v = (n + m)/2$. By the same reasoning as in theorem 2.1 it follows that the intersection between two such hyperplanes is a plane containing this point and the intersection between three or more hyperplanes must be a line passing through this point. Again using the same reasoning it can be seen that the family of these boundary lines passing through the point $\alpha = -1, \beta = -1/2, \gamma = 0$ is described by the conditions $\beta = c + (c + \frac{1}{2})\alpha$, $\gamma = d + d\alpha$.

Corollary 4.1. *All optimal regions are semi-infinite "conic" regions bounded by lines of the form* $\beta = c + (c + \frac{1}{2})\alpha$, $\gamma = d + d\alpha$.

This implies that just as the two parameter global alignment essentially had only one degree of freedom (in the sense that alignments that are optimal at some point must also be optimal at some point on the α or β axis) so also the three parameter global alignment essentially has only two degrees of freedom since any alignment that is optimal at some interior point must also be optimal on one of the three co-ordinate planes where one of the three parameters is zero. In other words, if we compute the decomposition for the three co-ordinate planes we have all the information required to describe the entire decomposition.

Theorem 4.2. *There are at most $O(n^2)$ different alignments that are optimal somewhere in the α, β, γ parameter space.*

Proof : We know, from Theorem 2.2 that there can be only $O(n^{2/3})$ regions on the $\gamma = 0$ plane. Using the same arguments as in lemma 3.3, we can see that on the $\beta = 0$ plane there can be atmost as many regions as there are distinct (w, x) pairs and on the $\alpha = 0$ plane there can be atmost as many regions as there are distinct (w, y) pairs. Both of these are n^2.

By similar arguments, and using the fact that $w_i + x_i + (c_1 - c_0)y_i + (c_3 - c_2)z_i > w_{i+1} + x_{i+1} + (c_1 - c_0)y_{i+1} + (c_3 - c_2)z_{i+1}$, it is easy to show that in the case of **Local alignments with gaps** (1) an arbitrary line can go through at most n^3 regions, and that (2) the number of regions is at most n^3.

5 The Case of Richer Weights and Penalties

In the previous results, the total penalty for mismatches was just the product of the mismatch penalty α and the *number* of mismatches. While this is sufficient in many biological applications, many other applications use a richer set of weights and penalties. In detail, for each pair (a, b) of unequal characters in the alphabet, there is a number $w(a, b)$ which is the base penalty for aligning these mismatching characters. One may also specify a character-dependent penalty for aligning a particular character with a space, and also a positive weight for aligning two matching characters which depends on the particular pair of characters. There are several commonly used pair-dependent weight and penalty schemes in the biological literature. The most widely referred to is called the Dayhoff matrix [SD].

With such pair-dependent weights and penalties, the value of an alignment A is computed as $M(A) - MS(A) - S(A)$ where $M(A)$ is the sum of all the (pair-dependent) weights contributed by matching pairs of characters in A, $MS(A)$ is the sum of all the (pair-dependent) penalties contributed by mismatching pairs of characters, and $S(A)$ is the sum of all the (pair-dependent) penalties contributed by characters opposite spaces.

One might want to parametricly study the effect of changing these pair-dependent weights, but this seems too unwieldly. A simpler question that is still of importance is how to balance the influence of the term contributed by matches verses the terms contributed by mismatches and spaces. So for a given alignment A, its parametric value is $M(A) - \alpha MS(A) - \beta S(A)$. As before, the α, β space decomposes into maximal convex regions where a particular alignment is optimal throughout. The results in the previous sections depend on $M(A), MS(A)$ and $S(A)$ being the *number* of matches, mismatches and spaces in A respectively, and breakdown under this richer weight/penalty structure. We don't know non-trivial bounds on the number of regions in the parametric decomposition, but we can prove that along any line, the number is sub-exponential.

Theorem 5.1. *With pair-dependent weights and penalties, the number of breakpoints encountered along any line L in the parametric decomposition is at most $(2m)^{\log_2 n}$.*

Proof: Along a line L in α, β space, the value of β is linearly dependent on α, so by adjusting the base penalties for spaces, we have a one parameter problem. In that problem, the parametric value of an alignment A can be assumed to be $M(A) - \alpha[MS(A) + S(A)]$. Consider the dynamic programming table used to find the optimal alignment, once a fixed value of α is specified. An optimal alignment is specified by a path in that table from cell $(1, 1)$ to cell (n, m). We associate a single optimal alignment in each region of the decomposition, and hence a single optimal path in each region. Thus as we move along L (with changing α) through changing regions, the corresponding path changes. Let S be the set of paths which correspond to the regions encountered along L. Let $T(n, m)$ denote the maximum possible size of S in any n by m table.

Each path in S goes through row $n/2$. Consider a fixed cell $(n/2, k)$. The number of paths in S which go through cell $(n/2, k)$ is bounded by $T(n/2, k) + T(n/2, m - k) \leq 2T(n/2, m)$. The reason for the plus (rather than a product) is that changes in the optimal path before row $n/2$ occur as α is changing and are totally independent with changes in the optimal path that occur after row $n/2$. Hence $T(n, m) \leq 2mT(n/2, m)$ which implies that $T(n, m) \leq (2m)^{\log_2 n}$.

6 Program Description

We have developed a program, PARAL, which allows the user to specify two sequences, a range for α and β, and the desired type of alignment (global or local) with or without gaps. Then, when a specific choice for α and β is given, the program computes an optimal alignment $\mathcal{A}$ for that choice and then determines and displays the region P in the α, β space that $\mathcal{A}$ is optimal for. The user may explore the interesting part of the space by repeatedly specifying values for α and β which have not been placed yet in a region. It can also generate all the regions in the entire decomposition systematically without having to choose any specific points.

The implementation is based on the following primitive: given a point p and a direction l, find the first point p' along l in which l crosses to a different region, and also find the new alignment at p'. This primitive can be implemented in $O(knm)$ time, where k is the actual number of regions that l goes through, by using the method of [ES] that finds all breakpoints along a line. It can also be implemented in $O(nm^2 \log n)$ time, independent of k, were each successive breakpoint is found by Gusfield's [G] adaptation of Megiddo's method [M]. We have adopted the first approach in PARAL.

Given a point p, its optimal region P is ideally found as follows: first, an arbitrary direction l is chosen and the next point p' is computed. Given the alignment at p', one boundary of P can now be determined by intersecting both alignments. The procedure is repeated, say, clockwise, along the new boundary, until all boundaries of P are found. The idealized method needs more detail to handle degeneracies that can occur if more than three regions meet at a point. We omit the details here.

Alignments with parametrized gap penalties (the 3-dimensional case) are also handled, however without a 3-D display; instead, a display of the projected decomposition on any plane is given. A more complete paper on the program is in progress [GBBMN].

7 Acknowledgement

Thanks to Rob Irving for working out the constant in the $O(n^{2/3})$ bound on the number of regions in the case of 2-D global alignment (Theorem 2.2).

References

[AV] P. Argos and M. Vingron, *Sensitivity Comparison of Protein Amino Acid Sequences,* in Methods in Enzymology, Volume 183: *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences,* Edited by R. Doolittle, 352-365.

[D1] R.F. Doolittle, *Of Urfs and Orfs: A Primer on How to Analyze Derived Amino Acid Sequences,* University Science Books, 1986.

[D2] R.F. Doolittle, edt. *Methods in Enzymology, Volume 183: Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences.*

[ES] M. Eisner and D. Severance, *Mathematical Techniques for Efficient Record Segmentation in Large Shared Databases,* Journal of the ACM, 23:619-635, 1976.

[FS] W.M. Fitch and T.F. Smith, *Optimal Sequence Alignments,* Proc. Natl. Acad. Sci. USA 80(1983) 1382–1386.

[GO] O. Gotoh, *Optimal Sequence Alignment Allowing for long Gaps,* Bull. Math. Biol., 52 (3):359-373, 1990.

[GD] M. Gribskov, J. Devereux, *Sequence Analysis Primer,* Stokton Press, 1991.

[G] D. Gusfield, *Parametric Combinatorial Computing and a Problem of Program Module Distribution,* Journal of the ACM, 30:551-563 (1983).

[GBBMN] D. Gusfield, K. Balasubramanian, J. Bronder, D. Mayfield, D. Naor, *PARAL: A Method and Computer Package for Optimal String Alignment using Variable Weights.*

[M] N. Meggido, *Combinatorial Optimization with Rational Objective Functions,* Math. Oper. Res., 4:414-424, (1979)

[PL] Pearson, W.R. and D.J. Lipman, *Improved Tools for Biological Sequence Comparison,* Proc. Natl. Acad. Sci. USA, 85 (1988), pp. 2444–2448.

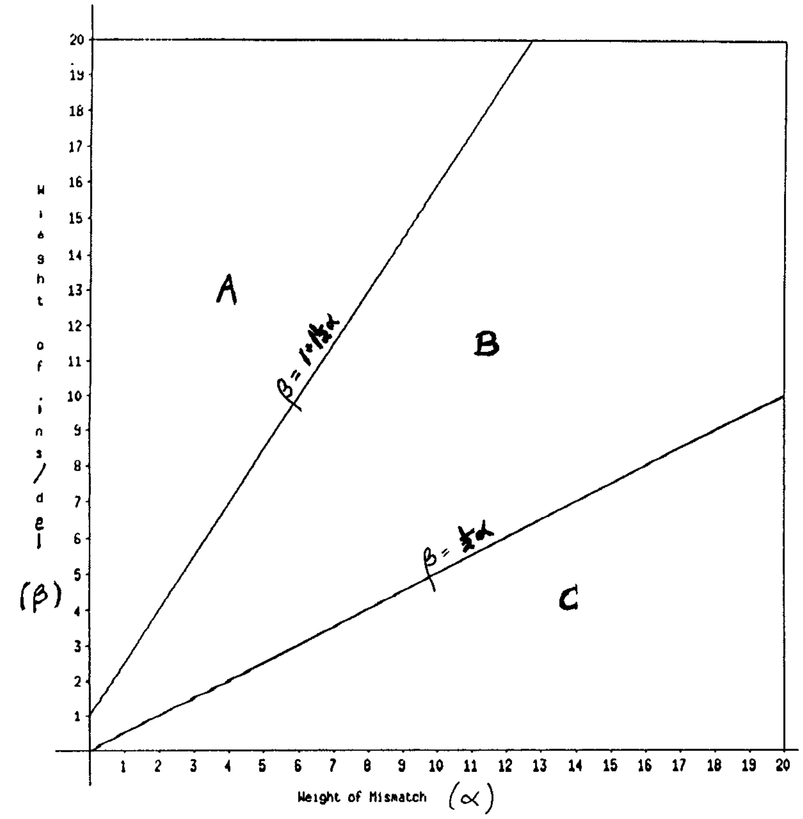
[SAL] G.D. Schuler, S.F. Altschul, D.J. Lipman, *A Workbench for Multiple Alignment Construction and Analysis,* in press in "Proteins: Structure Function and Genetics".

[SK] D. Sankoff and J. Kruskal, Editors, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison,* Addison-Wesley, 1983.

[SD] R. Schwarz and M. Dayhoff, *Matrices for detecting distant relationships,* In Atlas of protein sequences, National Biomedical Research Foundation, 1979, pages 353-358.

[vH] G. von Heijne, *Sequence Analysis in Molecular Biology,* Academic Press 1987.

[W] Waterman, M.S., *Sequence Alignments,* in M.S. Waterman, ed., *Mathematical Methods for DNA Sequences,* CRC Press (1989), pp. 53–92.



STRINGS ARE:

```
dcaabaccaabaacaa
baabbcbcaabccaac
```

ALIGNMENT OF REGION A:

```
dcaabaccaabaacaa
 | |  ||||   |
baabbcbcaabccaac
```

$v(\alpha,\beta) = 7 - 9\alpha - 0\beta.$

ALIGNMENT OF REGION B:

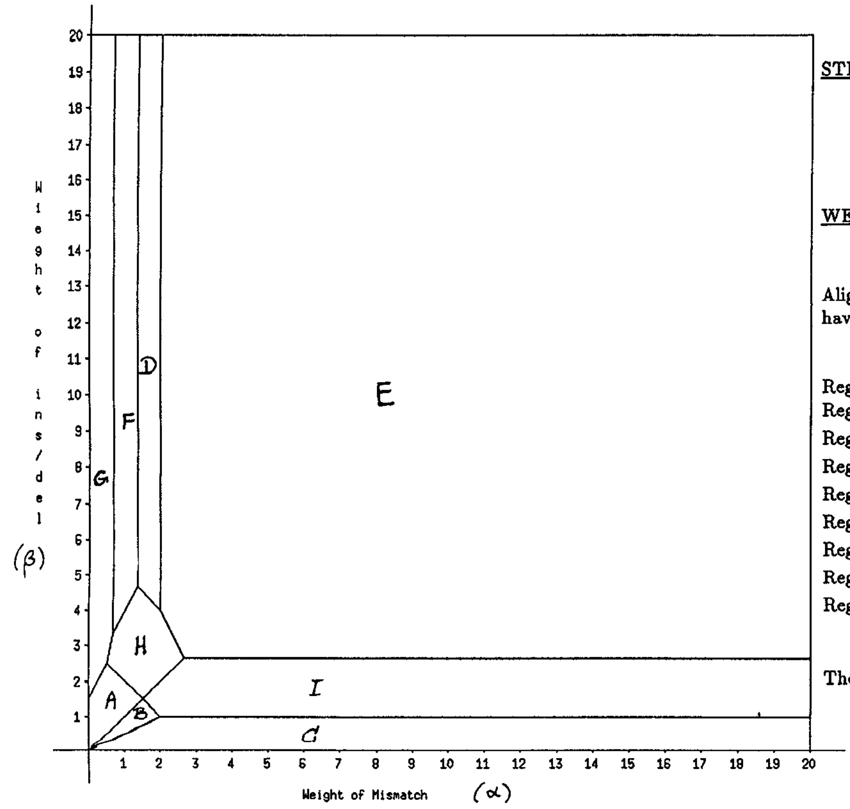
```
dcaabac_caabaacaa_
 ||| | ||||  |||
b_aabbcbcaab_ccaac
```

$v(\alpha,\beta) = 11 - 3\alpha - 4\beta.$

ALIGNMENT OF REGION C:

```
_dcaab_ac_caaba_acaa_
  |||  | ||||   |||
b__aabb_cbcaab_c_caac
```

$v(\alpha,\beta) = 11 - 0\alpha - 10\beta$



STRINGS ARE:

```
dcaabaccaabaacaa
baabbcbcaabccaac
```

WEIGHT FUNCTIONS:

Alignments optimal in regions A,B,...,I have the following weight functions:

Region A: $v(\alpha,\beta) = 11 - 3\alpha - 2\beta$
Region B: $v(\alpha,\beta) = 11 - 1\alpha - 4\beta$
Region C: $v(\alpha,\beta) = 11 - 0\alpha - 6\beta$
Region D: $v(\alpha,\beta) = 2 - 1\alpha - 0\beta$
Region E: $v(\alpha,\beta) = 0 - 0\alpha - 0\beta$
Region F: $v(\alpha,\beta) = 6 - 4\alpha - 0\beta$
Region G: $v(\alpha,\beta) = 8 - 7\alpha - 0\beta$
Region H: $v(\alpha,\beta) = 8 - 2\alpha - 1\beta$
Region I: $v(\alpha,\beta) = 8 - 0\alpha - 3\beta$

The actual alignments are not shown