J. Mol. Biol. (1994) 235, 1–12

REVIEW ARTICLE

Sequence Alignment and Penalty Choice

Review of Concepts, Case Studies and Implications

Martin Vingron and Michael S. Waterman

Departments of Mathematics and Molecular Biology
University of Southern California
Los Angeles, CA 90089-1113, U.S.A.

Alignment algorithms to compare DNA or amino acid sequences are widely used tools in molecular biology. The algorithms depend on the setting of various parameters, most notably gap penalties. The effect that such parameters have on the resulting alignments is still poorly understood. This paper begins by reviewing two recent advances in algorithms and probability that enable us to take a new approach to this question. The first tool we introduce is a newly developed method to delineate efficiently all optimal alignments arising under all choices of parameters. The second tool comprises insights into the statistical behavior of optimal alignment scores. From this we gain a better understanding of the dependence of alignments on parameters in general. We propose novel criteria to detect biologically good alignments and highlight some specific features about the interaction between similarity matrices and gap penalties. To illustrate our analysis we present a detailed study of the comparison of two immunoglobulin sequences.

Keywords: sequence alignment; gap penalties; scoring matrix; phase transition; score statistics

1. Introduction

In the last decade the use of alignment programs to compare DNA or amino acid sequences has become a routine task. Computer programs to align two sequences, search a database with a given sequence or compare multiple sequences are readily available. In spite of easy access to such resources these programs often confuse the user by asking for parameters. How to choose these parameters remains guesswork to both the expert and the novice. While some of the program parameters influence how fast a program will run, the same alignment program will frequently produce significantly different alignments under different parameter settings. Here, we review local sequence alignment (see Waterman, 1984) and study the effect that parameter choice has on the resulting alignments. We will review some recent developments in computer science and statistics and apply them to develop a better understanding of the parameter dependence of sequence alignments.

This problematic side of the otherwise easy-to-use programs has sporadically attracted the attention of both biological and mathematical researchers. Fitch & Smith (1983) explicitly set out to study this question for the comparison of coding DNA. They calculated numbers of alignments until they had found all optimal alignments and could divide the parameter space into regions with identical optimal alignments within each region. In the context of sequence alignment with long gaps Gotoh (1990) also studied the results of his method under variation of the gap penalties. While of interest for its own sake, parameter dependence makes it difficult to compare alignment programs. Barton & Sternberg (1987) calculated tables of alignment scores under various gap penalties to prove the superiority of their secondary structure dependent alignment procedure. In a comparative study of alignment methods Rechid *et al.* (1989) tested several choices of parameters in order to compare the best results obtained with one method to the best results obtained with another one. In spite of the awareness of the problem it is still not understood how to choose alignment parameters rationally.

Alignment algorithms come in different colors and shades, with variations of gap penalties and similarity matrices. A common misunderstanding, however, is to mistake a program that uses different parameters (e.g. a different matrix) for a new algo-

0022-2836/94/010001-12 $08.00/0

rithm. An important distinction on the algorithmic level is whether a program calculates a global or a local alignment. The prototypic global algorithm is the classic Needleman-Wunsch (Needleman & Wunsch, 1970) algorithm. The Smith-Waterman algorithm (Smith & Waterman, 1981), on the other hand, is the best known local alignment algorithm. Both use a gap penalty function that in most implementations is a linear function. In addition to the gap penalty the algorithms may use a similarity or distance matrix on the letters of the alphabet that is used to represent the sequences. Here, we focus on the local alignment algorithm due to Smith & Waterman (1981). In cases where two sequences are similar over their entire lengths, the local algorithms should find this fact as well as the global algorithm. When two sequences share only a limited region of similarity only the local algorithm will discover this.

The current paper approaches the problem of parameter dependence of local alignment using essentially two tools. One is a newly developed algorithm to calculate the entire set of optimal alignments under all choices of one or two parameters (Fernández-Baca & Srinivasan, 1991; Gusfield *et al.*, 1992; Waterman *et al.*, 1992). More parameters can be included but the results are hard to visualize. The algorithm can, for example, be applied to the study of the influence of initial and extension gap penalties (under a linear gap penalty function). The result will be a tesselation of the plane spanned by the two parameters. Each region in this tesselation describes the set of parameters that, when used to align the given sequences, result in the same set of optimal alignments.

Another tool we use to study alignments is the statistical behavior of optimal alignment scores. These scores behave statistically quite differently for very small gap penalties as opposed to very large gap penalties. The theory of a phase transition in growth of alignment score with sequence length was developed by Waterman *et al.* (1987) and Arratia & Waterman (unpublished results). The next section reviews both the parametric alignment algorithm (section (b)) and the statistical theory (section (c)).

The first application of these methods will be to DNA. For DNA the parameters we are interested in are the values for match, mismatch and a gap penalty. The influence of these parameters will be illustrated for random sequences. The example will also illustrate the importance of the statistical behavior of local alignment score as a function of sequence length for a specific comparison. Certain key features can be observed in the DNA studies that carry over to protein sequence alignment.

The emphasis of this paper lies on the comparison of proteins. For many proteins the three-dimensional structures are known and structural superposition supplies us with a standard of truth for the alignment. We will focus on the comparison of two immunoglobulin sequences that, both on structural and on sequence levels, have been carefully studied. This enables us to analyze how the optimal alignments change under changes of gap penalties and how alignment quality is influenced by the parameter changes. Interestingly, the statistical phase transition in the parameter space has implications with respect to the choice of parameters and its understanding should make it easier to locate a good alignment.

2. Review of Concepts

(a) *Local alignment and its parameters*

As pointed out above we will concentrate on the local alignment algorithm (Smith & Waterman, 1981). A local alignment is one that matches a contiguous subsequence of the first sequence with a contiguous subsequence of the second sequence. The Smith–Waterman algorithm is motivated by scoring systems where scores for matches and mismatches have different signs, i.e. where matches increase the overall score of an alignment whereas mismatches decrease it. A good alignment then has a positive score and a poor alignment a negative score. The local algorithm finds an alignment with the highest score by considering only alignments that score positive and picking the best one from those. The algorithm is a dynamic programming algorithm. For the comparison of DNA, it requires setting a gap penalty ($\delta \geq 0$) in addition to the score for a match or identity (that we keep at 1) and the penalty for a mismatch ($\mu \geq 0$). Let the two sequences be $\mathbf{a} = a_1 a_2 \ldots a_n$ and $\mathbf{b} = b_1 b_2 \ldots b_m$. The algorithm then is:

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + 1, & \text{if } a_i = b_j \\ H_{i-1,j-1} - \mu, & \text{if } a_i \neq b_j \\ H_{i-1,j} - \delta, & \\ H_{i,j-1} - \delta, & \\ 0 & \end{cases}$$

With an initial assignment of $H_{i,0} = H_{0,j} = 0$ for $1 \leq i \leq n$, $1 \leq j \leq m$, the desired local alignment score is the maximum value of $H_{i,j}$ over the entire matrix:

$$max\{H_{ij} : 1 \leq i \leq n,\ 1 \leq j \leq m\}.$$

When applied to proteins one uses a similarity matrix that attributes a score to each possible residue pair. The score should be positive for desirable residue pairs and negative for dissimilar residue pairs in order to ensure meaningful local alignments. Gaps are usually penalized using a linear gap function that assigns an initial penalty for a gap opening and extension gap penalty for each deleted or inserted residue increasing the gap length.

(b) *Calculating all optimal alignments*

The score of a given DNA alignment can also be described by counting the number of matches, the number of mismatches, the number of insertions or deletions. Then each of these is multiplied by its corresponding parameter values:

$$\text{No.}\{\mathit{matches}\} - \mu \times \text{No.}\{\mathit{mismatches}\} - \delta \times \text{No.}\{\mathit{insertions\ deletions}\}. \quad (1)$$

An optimal alignment is one that maximizes this expression. For proteins, an analogous expression is given by Smith *et al.* (1981).

Equation (1) for the score of an alignment shows that the score is a linear function of the parameters mismatch penalty μ and gap penalty δ. We restrict ourselves to such parameters that influence the score linearly. Linearity then implies that every alignment defines a plane in the space spanned by μ, δ and the alignment score. An optimal alignment has score described by the plane with score equal to the maximal value over all alignments at the parameter-choice (μ, δ). Note that there may be several optimal alignments associated to a plane.

Unless a point (the parameter vector) is exactly on the border between two alignments it will define a neighborhood where the best alignment-plane and with it the alignments remain unchanged with respect to possible changes of the parameters. With larger changes though a new set of optimal alignments (a plane) takes over. Figure 1(a) shows an example of the tesselation produced by drawing all the borderlines where another alignment plane takes over. Each of the resulting regions contains those parameters where one alignment-plane is optimal. All alignments in one region therefore have the same number of matches, mismatches and deletions. Every region is a convex polygon (Waterman *et al.*, 1992). This picture is produced from the comparison of two 400 bp long random DNA sequences under variation of the mismatch penalty (horizontal axis) and the gap penalty (vertical axis).

Recent advances (Fernández-Baca & Srinivasan, 1991; Gusfield *et al.*, 1992; Waterman *et al.*, 1992) have made it possible to calculate this tesselation of the plane in an efficient way, i.e. avoiding the calculation of large numbers of alignments. Linking this program to a graphical display, M. Eggert at USC built an environment that allows one to click with the cursor to a region and see a representative alignment, the score of which is given by the hyperplane in that region of the tesselation. In addition, the alignment can be compared with a reference alignment and the quality of every alignment in the tesslation can be assessed. Gusfield and collaborators are planning to release a user-friendly package.

(c) *Statistics and phase transition*

Consider comparing two random DNA sequences. Let a match score +1, a mismatch $-x$ and gaps have no penalty at all. The alignment algorithm is then free to pick the maximal number of matches that can be fit into an alignment without regard to gaps. The result is what in the computer science literature is known as the longest common subsequence (Apostolico & Guerra, 1987). It is intuitively clear that extending such an alignment as the sequences get longer is relatively easy. In fact, probabilistic considerations show that the length of such an alignment grows linearly as the sequence length increases (Chvátal & Sankoff, 1975). We expect such an alignment to contain many gaps that make it a rather unlikely alignment from the biological point of view.

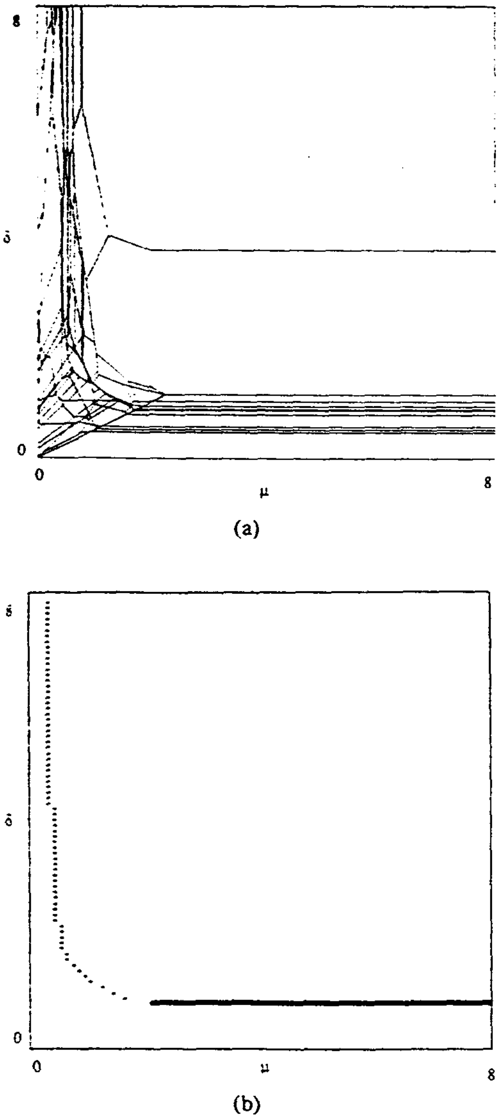

Figure 1. (a) Tesselation for the comparison of two 400 bp long random DNA sequences. The horizontal-axis is the penalty for a mismatch (μ) and the vertical axis the gap penalty (δ). (b) The phase transition curve that separates the linear and the logarithmic region.

When the gap penalty is so high that no matching region could justify a gap, the local alignment will contain only contiguous matching regions. This match will be chosen to optimize the sum over both matches and mismatches. Naturally, for two random sequences the length of this region will be small when mismatches are expensive. If, for example, we make the mismatch penalty prohibitively expensive as above, the resulting alignment contains only the longest region of exact matches and avoids both mismatches and gap altogether. For these cases it has been shown that the score of the alignment is proportional to the logarithm of the length of the sequences being compared (Waterman *et al.* 1987).

In this last case it is the local algorithm that allows the alignment to "shrink" when a longer alignment would involve paying high penalties. If, however, one would deliberately misuse the local algorithm by not penalizing mismatches (both the scores for matches and mismatches are positive), this shrinkage would not take place. Even under high gap penalties optimal alignments would match essentially the entire sequences. In such a case the score would again grow linearly in spite of a high gap penalty.

This example shows that it is not just the gap penalty that is responsible for linear or logarithmic growth of optimal alignment score. As determined by R. Arratia & M. S. Waterman (unpublished results), the statistical behavior associated with a parameter choice is determined by the sign of the expected score of a global alignment of two random sequences of equal length. Their result says that if the expected score of the global alignment is positive, then the local alignment score using the same parameters grows linearly with the length of the sequences. If the expected score of the global alignment is negative, then the local alignment score using the same parameters grows with the logarithm of the length of the sequences. When the expected score of the global alignment is 0, there is a transition between these two statistical regimes. Practically, one can approximate the expected score by calculating a large number of alignments of random sequences and average their scores. Figure 1(b) shows the transition curve for a DNA comparison where mismatch and gap penalty are varied.

The special cases sketched at the beginning of this section of course fit into this framework. When gaps and mismatches are free, a global alignment will have positive score. Thus the local score will grow linearly as described above. Similarly, when mismatches are not penalized, the optimal global alignment (for 2 sequences of equal length) will not introduce a gap and always score positive. The local alignment for such a parameter setting will consequently grow linearly too. When both the penalties for mismatches and gaps are high then the global alignment will score less than 0 and the local alignment will be in the true sense local. Its score will grow logarithmically with the length of the sequences.

(d) *Scoring systems for proteins*

When comparing amino acid sequences, additional parameters determine the alignment. Instead of scores for matches and mismatches, a matrix is used which scores every pair of amino acid residues. We will assume that alignments are calculated under Dayhoff's PAM250 matrix (Dayhoff *et al.*, 1983). This matrix gives positive values to identities and conservative substitutions. Some identical pairs receive very high scores, e.g. matching two tryptophan residues yields +17 or two cyteine residues yield +12. Note that not all identical pairs are weighted equally. Rare substitutions are given a negative score and are therefore usually avoided by the alignment algorithm. The lowest scoring substitutions have score −8. When comparing proteins we use a linear gap penalty function, and the initiation and the extension of a gap are assigned separate values. With an initial gap penalty of g_1 and an extension gap penalty of g_2 a gap of length k costs $g_1 + g_2 \times k$.

When both of the gap penalty parameters are 0, we have essentially the same situation as in the case of DNA-alignment under 0 gap penalty. An alignment will then be long and, since it does not have to pay for gaps, will jump freely from one good match to the next one. Its score grows linearly with the length of the sequences. Again, the entire region around the origin is under the linear regime and contains alignments that are long and tend to contain many gaps. When either or both gap parameters are high, any gap will be expensive and thus avoided. Therefore, far away from the origin of this parameter space, the algorithm will seek out a possibly short well-matching (as described by the score matrix) alignment. The score of such an alignment grows logarithmically with the length of random sequences. As before, once the gap penalty is high enough, the optimal alignment is the best-scoring matching region without gaps.

The Smith-Waterman algorithm explicitly uses the 0-level of the similarity matrix. A high overall level of the matrix leads to longer alignments because fewer residue pairs score less than 0. But raising the overall level of a matrix in this way also introduces an asymmetry between deletions in the shorter and the longer sequence (Dayhoff *et al.*, 1983). When a large value is added to a matrix, the mere number of residue pairs in an alignment is more important than the placement of gaps. The number of residue pairs, however, is maximized when there is no deletion in the shorter sequence. Thus, deletions in the shorter sequence will be avoided when the matrix average is very high.

The matrix level influences the statistical behavior too. When all matrix entries are positive the score of the resulting alignments will grow linearly with the length of the sequences. To be more precise, we will introduce the expected score for a random residue pair scored by a matrix. In addition to the matrix, this requires taking into account the distribution of the amino acids in the sequences. The expected score of a residue pair is defined as the sum over all residue pairs of the score of a pair as given by the matrix weighted by their respective frequencies. Calculating this for the PAM250 matrix and the distribution of amino acids given by McCaldon & Argos (1988) results in an expected score of −0·81. If this expected score is positive for some matrix then the growth of the score of the local alignments is linear even under arbitrarily strong gap penalties. To see this, consider that the global alignment will then pick out only the main diagonal of the comparison and the expected score along that one diagonal is positive. When gap penalties are lower, the score of the

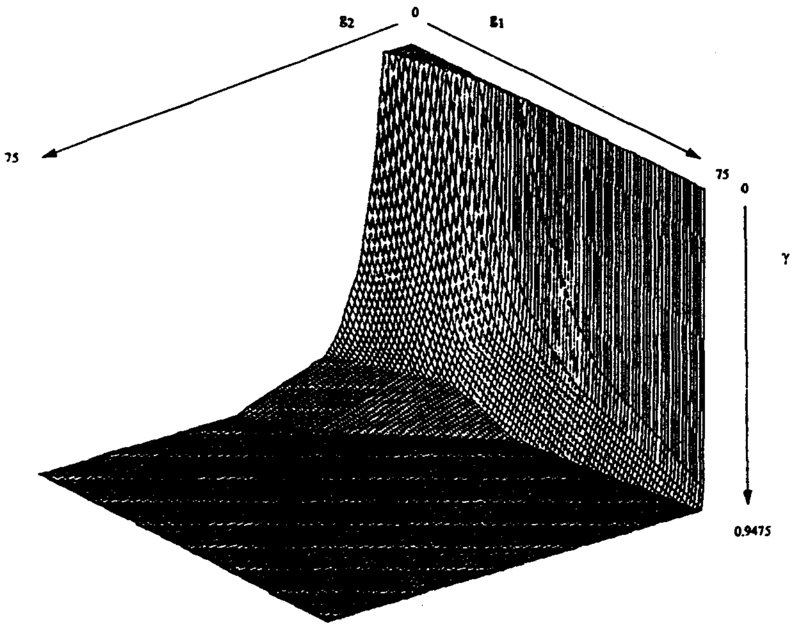

Figure 2. The phase transition in 3 dimensions. g_1 and g_2 are the penalties for initiating and extending a gap in a linear gap penalty function and γ is a constant that is added to every entry of the Dayhoff matrix.

global alignment can only increase. Therefore, there is no region of logarithmic growth nor a phase transition in the space of gap penalties when a matrix is used that has a positive expected score.

When a matrix contains only a few positive entries, then the local alignment score will grow linearly only for very small gap penalties. For matrices with most entries negative, growth will be logarithmic because a global alignment will score less than 0 in expectation. The expected value of such a matrix will be negative for a biologically reasonable distribution of amino acids. In the plane spanned by the initial and extension gap penalties one can therefore find a transition curve between the two regimes. Upon raising the overall level of the matrix the linear region grows in area and slowly pushes the transition curve out. Figure 2 shows the surface that is described by the transition curve as an increasing constant is added to every element of the matrix. The (g_1, g_2)-plane of the Figure is spanned by the two gap penalties and 0 added to the Dayhoff matrix. The linear region in this plane is tiny. Moving downwards the level of the matrix rises and the linear region grows. Towards the bottom of the graph the linear region takes over the entire plane and the logarithmic region has disappeared. Note that the bottom plane has 0·9475 added to the Dayhoff matrix, just slightly more than necessary to make its average positive.

3. Comparison of Two Random DNA Sequences

Studying the tesselation we see a reflection of the statistical behavior even in the comparison of one specific pair of sequences. Figure 1(a) shows a tesselation of the plane for the comparison of two random DNA sequences of length 400. The score for a match is kept at +1 and the penalties for mismatches and gaps are varied along the μ and δ-axis respectively. At the origin both μ and δ are 0. Close to the origin we see a messy region. This part of the parameter space is the linear region giving essentially global alignments. Alignments change very easily due to only small changes in the parameters because the alignment is long and for every parameter choice there is a subtle equilibrium between matches, mismatches and gaps. In contrast, when mismatch and gap penalty are higher, the alignments tend to be shorter and, depending on the choice of parameters, have only few to no gaps. This region is clearly under the regime of logarithmic growth of score. The tesselation there is much coarser than in the linear region. In the outer region corresponding to large gap and mismatch penalty the optimal alignment is the longest region of consecutive exact matches. No further increase of penalties can change the score.

There are several regular features in this tesselation that demand to be explained.

(i) The bottom right part contains only horizontal lines.

(ii) Straight lines through many regions near the origin.

(iii) Pencil of lines crossing the δ-axis.

The horizontal lines above and parallel to the μ-axis are easily explained. Any mismatch (costing μ) can be avoided by a deletion followed by an insertion (costing 2δ). Therefore, when a deletion costs less than half of a mismatch, it is better to avoid mismatches altogether. This holds for the part of the μ, δ-plane under the line $\delta = \mu/2$. It is in this area that the mismatch penalty μ becomes irrelevant and all dividing lines between regions are horizontal, i.e. independent of μ.

The fact that some dividing lines pass through several regions is more interesting. Let the alignments in two adjacent regions, say A and B, have a certain part of the alignment in common. Then upon changing parameters in a certain way region A will become A' and region B will become B'. If both the change between A and A' and between B and B' occurs in their common region in the same manner. then the dividing line between A and A' on the one hand and B and B' on the other hand will be a straight line. This observation does not yield a necessary explanation for the straight lines but supplies us with a plausible sufficient answer.

It is very easy to prove that the situation described is compatible with straight lines between regions. An analytic description of the dividing line between two adjacent regions can be obtained from equating the formulae for their scores (eqn (1)). This will result in a relationship between δ and μ. When the two alignments from the adjacent regions undergo the change described above the numbers of matches, mismatches and insertions/deletions change in the same way for them and the equation of the dividing line remains unchanged.

Another, perhaps even more striking feature of Figure 1, is that many of the straight lines through the linear region seem to emanate from one point in the third quadrant of the plane. For global alignments this feature has already been observed and explained by Gusfield *et al.* (1992). The lines do in fact meet in the point $(-1, -1/2)$ and form a pencil of lines. Values of -1 and $-1/2$ for mismatch and gap penalties favor insertions, deletions and mismatches. For a match-score of 1 (as is used in this Figure) this choice of parameters implies that every step through the comparison matrix receives a score of 1. Therefore any alignment that uses the entire sequences will be optimal, all having identical scores. The global alignments and their close relatives populate the linear region of the tesselation. The lines emanating from $(-1, -1/2)$ therefore appear to be the borderlines between global alignments that are optimal at $(-1, -1/2)$, and moving away from $(-1, -1/2)$ have undergone the same changes in adjacent regions that leads to the straight lines.

A close inspection of the overlay between the tesselation in Figure 1(a) and the transition curve (Fig. 1(b)) reveals that these pencil-lines extend almost exactly until the transition curve. Note that the transition curve is a generic attribute of the scoring scheme and not of the two sequences compared in the tesselation. The above analysis provides us with an explanation of this phenomenon. The pencil-lines extend as long as the alignments involved are still essentially global. This is the case in the linear region that ends at the transition curve. Beyond that curve there is a dramatic change of character of the alignments to genuine local alignments. Even with this explanation the quality of the coincidence between the combinatorial structure of the tesselation and the statistical phase transition remains re narkable.

The above analysis illustrates our approach to understanding the tesselations. We try to explain certain general features of the alignments. We cannot always give necessary reasons, which means we cannot claim that whenever a certain pattern occurs it absolutely must be due to the reason we give. Rather we present plausible circumstances that are the sufficient and usual reason for the pattern to occur and that we believe to be the predominant explanation. This will also be the basis of the following analysis of protein comparisons. We will however not always reflect this limitation in the wording of each explanation.

4. Comparison of Proteins

(a) *Test case: immunoglobulins*

We want to illustrate the above ideas with the comparison of a well-studied pair of proteins. There is of course a wide range of pairs of related protein sequences going from very similar to very distant. For two sequences that share a very high degree of similarity the resulting alignment is almost independent of the gap penalties under which it is calculated and studying it is therefore not of interest here. On the other hand most sequence-pairs for which relatedness was recognized only after the structures were solved, resemble random comparisons too closely to teach us anything new. We therefore picked as our central example one that we believe to be a good compromise between the impossible and the obvious, and one that has been well studied on the structural level. We will study the comparison of two immunoglobulin sequences, namely of heavy and light chain of the variable domain of the Fab antibody. Amzel & Poljak (1979) published a structural superposition relating these two sequences. Each of them has two cysteine residues in strands b and f, which form a disulfide bridge. Another conserved residue that could guide an alignment is a tryptophan residue in strand c of both sequences. The difficulty of the alignment stems mainly from a second tryptophan residue in the heavy chain 11 residues C-terminal from the first one, which easily leads to misalignment. The problems in comparing these two sequences and possible remedies using secondary structure information have been studied by Barton & Sternberg (1987).

Figure 3 shows the optimal and one suboptimal local alignment of the two sequences in the form of a dot-plot. The dot-plot was calculated using an algorithm to compute non-intersecting suboptimal local alignments (Waterman & Eggert, 1987). The optimal alignment connects diagonals labeled C1, Wf, C2 and d. Diagonal Wc is found as the top suboptimal alignment. C1 and C2 are the diagonals containing the correct cysteine matches and d is the diagonal matching beta strands g from the two sequences. Wc and Wf are the diagonals containing

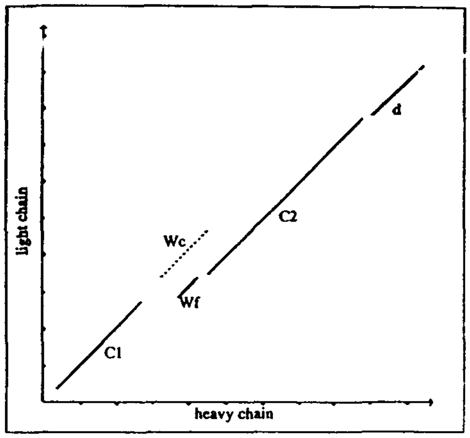

Figure 3. The optimal (strong lines) and the top sub-optimal (dotted lines) local alignments of the light and heavy chain of an immunoglobulin variable domain (Fab antibody).

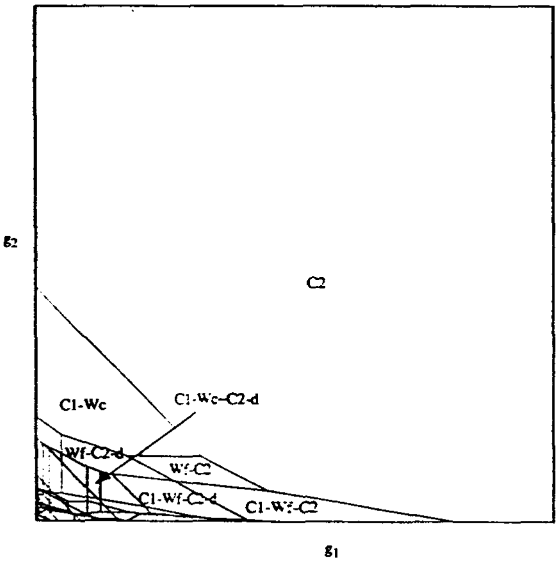

Figure 4. Tesselation for the comparison of the light and heavy chain of an immunoglobulin variable domain (Fab antibody). The horizontal axis is the initiating gap penalty (g_1) and the vertical axis the extension gap-penalty (g_2). Values along the axes range from 0 to 40. The unmodified PAM250 matrix is used.

the correct and false tryptophan matches, respectively. The labeled regions are parts of certain diagonals that will be encountered again analyzing tesselations. Table 1 lists the corresponding matching regions. We will use these diagonals to denote alignments made up of different combinations of diagonals.

(b) *Varying gap penalties when matrix average is negative*

Figure 4 shows the tesselation for the two immunoglobulins described above. The scoring matrix used is Dayhoff's PAM250 matrix. As pointed out above (section 2(d)) its average is around −0·81. We want to discuss these alignments starting in the upper right corner. The outer region, where any gap is highly penalized, contains the alignment made up only of diagonal C2 from the dot-plot. As one moves towards the origin, other diagonals come into play. For example, the C1-Wc alignment region to the left of the C2 region has a high extension gap penalty and starts next to the y-axis where the initial gap penalty is zero. Accordingly, this alignment has one gap of only length 1 (extending a gap is expensive). The next adjacent region links the C2 diagonal to Wf. This alignment has a gap of length 2, which is made possible by the decreased extension gap penalty. The other alignment that borders the C2 region contains C1-Wf-C2. Here the extension gap penalty is so low that, in spite of a higher initial penalty, two gaps of length 10 and 2 are allowed.

Going from this region (C1-Wf-C2) to the left decreases the initial gap penalty. The next alignment attaches diagonal d, which results in the alignment C1-Wf-C2-d. A biologist looking for an alignment matching up the entire sequences and containing only a moderate number of gaps would easily believe this one alignment to be the correct one. In addition it is found under the kind of gap penalty choice that is usually recommended (high initial, low extension penalty) and a rather large region in the space of gap penalties gives this alignment. Nevertheless, this is not the alignment closest to the structural correspondence between the sequences. It is the alignment that is also easily found using the (global) Needleman–Wunsch algorithm and that was criticized by Barton & Sternberg (1987).

An alignment that matches the tryptophan residues correctly is present in our tesselation. The shaded regions match this part of the sequences correctly. The difference between the shaded cells is that as one moves from left to right in the para-

Table 1

Important matching regions (diagonals) from immunoglobin variable domain light (top sequence) and heavy (bottom sequence) chain

```
C1:
        QSVLTQPPSVSGAPGQRVTISCTGSSSNI
        |  | |      | |     |  | .
        QVQLEQSGPGLVRPSQTLSLTCTVSGSTF

Wc:
        AGNHVKWYQQLPGTAPKLLIF
               |   | ||
        SNDYYTWVRQPPGRGLEWIGY

Wf:
        AGNHVKW
         |
        PGRGLEW

C2:
        YQQLPGTAPKLLIFHNNARFSVSKSGSSATLAITGLQAEDEA
        |                          |   |                  |
        YVFYHGTSLKDTIPLRSRVTMLVDTSKNQFSLRLSSVTAADTA

d:
        YDRSLRVFGGGTKLTV
                     ||
        IAGCIDVWGQGSLVTV
```

meter plane the initial gap becomes more expensive and it is more "cost effective" to link the crucial diagonals without any further stops. When the initial gap penalty is low the diagonals are connected and between diagonals the alignment will match additional residues. The vertical division lines between the shaded regions are a reflection of the fact that the alignments in these regions bridge the same total number of gapped letters but distributed into different numbers of gaps. We have denoted all the alignments in the shaded region with the abbreviation C1-Wc ~ C2-d where the ~ means that the gap does not link Wc and C2 directly but has additional matches in between. This structure of adjacent regions with vertical division lines is encountered frequently, and we will call it a band of regions. Closer to the origin from the shaded band there is another band. This one however has too many gaps and contains alignments not as good as those in the shaded band. The innermost alignment regions show no recognizable order and contain the alignments with un-biologically low gap penalties. In the statistical sense this clearly is in the linear region.

(c) *Positive and varying matrix average*

As mentioned in the introduction, we change the Dayhoff matrix by addition of a constant to every matrix entry. Interestingly, after adding a positive constant to the matrix, our view of the good alignments in the tesselation improves. Figure 5 shows the alignment regions for the same pair of immunoglobulins but with 1·5 added to the entire matrix.

Now the alignments are generally longer because the local alignment algorithm extends them further. For example. the alignment in the outer region is still based on the diagonal C2 but has increased from 46 residue pairs to 85 residue pairs. Certain alignments are essentially unchanged compared with Figure 4. These comprise Wf-C2-d and C1-Wf-C2-d. Both of them only grew at the ends but are unchanged internally. Specifically, the positions of the gaps (the breakpoints where an alignment skips from one diagonal to a subsequent one) are the same as in the corresponding alignments from Figure 4†. The band of good alignments C1-Wc ~ C2-d (shaded) is much clearer in Figure 5 than in Figure 4. It now comprises more regions and has grown in area too. Alignments like C1-C2 and C1-C2-d are new to this plot. We conclude that raising the matrix level allows C1 to be longer. thus avoiding the additional gap-initiation used to include Wf in between C1 and C2. Toward the origin the tesselation is considerably less messy. Only immediately adjacent to the origin can the alignments with high numbers of gaps survive.

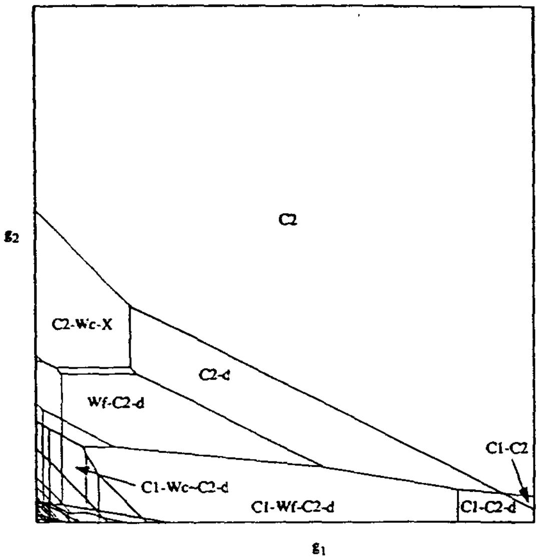

Figure 5. Tesselation for the comparison of the light and heavy chain of an immunoglobulin variable domain (Fab antibody). The horizontal axis is the initiating gap penalty (g_1) and the vertical axis the extension gap-penalty (g_2). Values along the axes range from 0 to 40. A constant of 1·5 has been added to every entry in the PAM250 matrix.

Note that due to the addition of a constant the average value of the matrix is of course increased by this constant and is now positive. This implies that the average score of the global alignment between two random sequences is positive even for high gap penalties. Thus. the local alignment score grows linearly. no matter how strongly gaps are penalized. Consequently. in this case there is no phase transition.

Figure 6 illustrates the result of adding a constant to every matrix entry. The two gap penalties are kept equal to each other and increase along the vertical axis. For technical reasons in our plot the constant added to the matrix (γ) decreases along the horizontal axis. The dark shaded region next to the horizontal axis contains the fine part of the tesselation. Along the horizontal axis we highlighted where 0 is added to the PAM250 matrix (the situation of Fig. 4), where 1·5 is added (the situation of Fig. 5) and where 0·81 is added. such that the average of the resulting matrix is approximately 0. For example. the alignments along a vertical line through the point where nothing is added to the PAM250 matrix. correspond to the alignments along the main diagonal of Figure 4.

Vertical division lines in Figure 6 indicate growth at the ends of an alignment. The top labeled regions are all essentially the diagonal C2. which we have already seen to be the best local alignments without gaps. Upon adding a constant to the matrix this diagonal grows at the ends. On the top right the alignment contains only the best-matching core of diagonal C2. This then grows towards the left (with increasing matrix level) until in the top left region

† The reason is that no matter what the gap penalty. the point where the gap is introduced is optimized so as to achieve the highest sum of scores along the 2 diagonals.

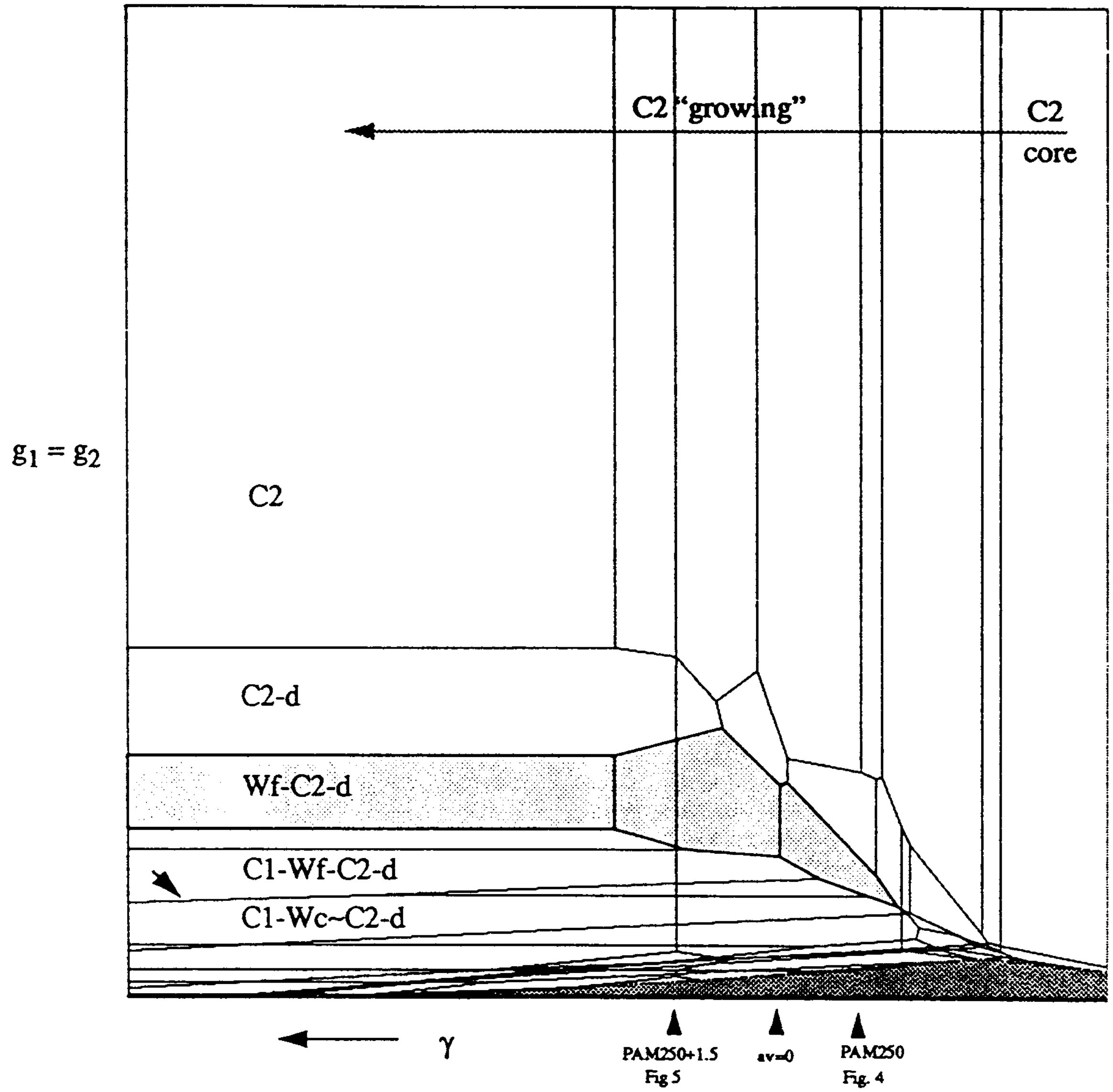

Figure 6. Tesselation for the comparison of the light and heavy chain of an immunoglobulin variable domain (Fab antibody). The constant γ added to the Dayhoff matrix increases along the horizontal axis towards the left. The vertical axis represents simultaneous growth of initial and extension gap penalty.

the shorter sequence is entirely fitted into the longer one.

Similarly, the regions below, which are divided by vertical lines, are alignments growing at the end in the same manner. This allows linking some alignments from Figures 4 and 5. For example, the alignment Wf-C2-d was contained in both Figures 4 and 5, differing only in how far the alignment extends at the ends. In Figure 6 the alignment labeled Wf-C2-d is part of a set of four shaded cells that show the growth of this alignment. The rightmost of these shaded cells contains the Wf-C2-d alignment exactly as it occurs in Figure 4. The vertical division lines between the shaded cells represent additions of a few residues at a time at the ends of the alignment. The Wf-C2-d alignment from Figure 5 was calculated with 1.5 added to the matrix, a value that in Figure 6 is exactly the division line between two regions.

A number of regions on the upper left side of the Figure are divided by horizontal lines and are unbounded towards the left. These alignments are not influenced by a further increase in the constant that is added to the matrix because they already extend as far as possible. They start and end with the N and C termini of one of the sequences. Further below, the division lines between the region are tilted downwards (highlighted by an arrow) as the matrix level increases. This means that these alignments will slowly disappear as one moves to the left. The distinguishing feature of the upper, persistent, alignments is that their length (number of aligned pairs) is maximal because they have no deletions in the shorter sequence. The number of aligned residue pairs is therefore given by the length of the shorter sequence. The lower ones have a deletion in the shorter sequence (a gap of length 1 to get from C1 to Wc) that decreases the number of aligned residue

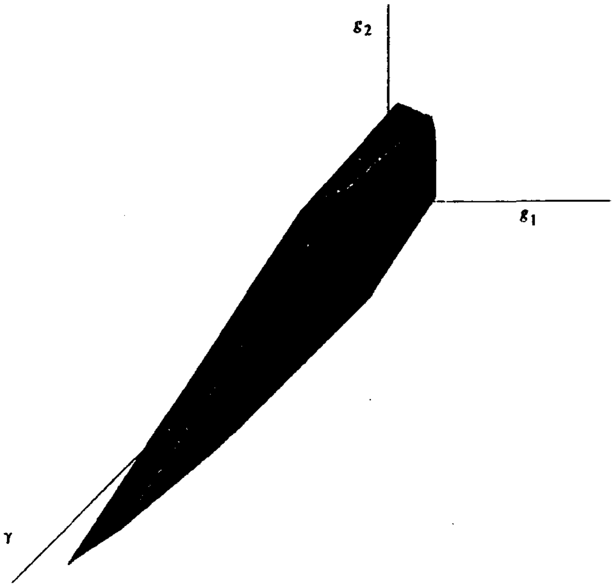

Figure 7. One region of the tesselation from Fig. 5 extended into 3 dimensions by raising the constant γ (towards the front) that is added to the Dayhoff matrix. The plane in the back corresponds to that shown in Fig. 5. The region is the one shaded cell where the arrow in Fig. 5 ends.

pairs in the alignment. As pointed out in section 2(d), when the matrix level is raised a longer alignment will sooner or later score better than a shorter one. This explains why the alignment regions on the bottom left are slowly disappearing.

Figure 7 illustrates this bias against deletions in the shorter sequence by following an alignment region in three dimensions. The alignment in this region is one of the C1-Wc ~ C2-d family. The three-dimensional polytope has as its backside one of the shaded regions from Figure 5 (with 1·5 added to the Dayhoff matrix). As one moves to the front the constant added to the matrix increases. Due to the one deletion in the shorter sequence (when the alignment links C1 to Wc) raising the matrix makes this alignment more and more disadvantageous. It is thus slowly pushed out of sight.

(d) *Other examples, specifically local alignments*

The results of the above analysis of the comparison of two immunoglobulins is by no means singular. Among the other cases that we looked at were the comparison of human hemoglobin alpha-chain and lupin leghemoglobin, and *Lactobacillus casei* and *Escherichia coli* dihydrofolate reductases. The first pair shares very little homology when measuring the percentage identical residues in the correct alignment (only 16%) and has frequently been used to study and test alignment methods. All the features described for the immunoglobulins can also be seen in this comparison, including the fine and coarse tesselation for weak and strong gap penalties and the band of regions containing good alignments. The other pair, the dihydrofolate reductases, are considerably more similar. Correspondingly, the tesselation is rather coarse and most of the regions contain reasonable alignments. As expected the choice of gap penalty is much less crucial for such a simple example.

We also calculated the tesselation for a pair of structurally related proteins that show very little sequence similarity, namely azurin and plastocyanin. In such a difficult case the resulting tesselation is almost indistinguishable from the tesselation produced by random sequences. Features that are also inherent to random sequences like the fine *versus* coarse tesselation remain. The band structure that seems to be a distinguishing feature of similar sequences is hardly discernible any more. Neither could we draw any conclusions about the best choice of gap penalties from such a case.

The examples discussed above all share similarity over their entire length and therefore constitute test cases that do not specifically address a local alignment algorithm. We therefore also studied examples that are in the true sense local alignments. One such case can be found in the proteins containing a helix-turn-helix motif. We chose to compare phage lambda repressor with the lambda Cro protein. These two share a well-studied region of about 20 amino acid residues without gaps (Sauer *et al.*, 1982). Surprisingly, using the Dayhoff matrix the correct alignment was not found at all. Only when we tried other matrices (e.g. that proposed by Gribskov & Burgess, 1986) were there gap penalties under which the correct alignment was found. The band structure is lost with this rather short alignment that comprises only one diagonal.

A second truly local example we studied was the comparison of the *E. coli* CAP protein (which binds cAMP) and a bovine cGMP-gated ion channel. Both of these proteins contain the common binding site for cAMP and cGMP. For computational reasons we had to restrict the 690 amino acid residue ion channel sequence to about 240 residues, which contained the region of similarity to CAP. According to Kaupp *et al.* (1992), alignment of the approximately 100 residue site requires two gaps. In this case using the Dayhoff matrix picks out the first of the three regions of similarity under high gap penalties. Upon lowering the penalties the second diagonal is added. The third diagonal, however, is so subtle that it is not found at all. As with the helix-turn-helix comparison the good alignments are found in the logarithmic region and in the linear region longer, biologically irrelevant alignments take over.

5. Discussion of Results

(a) *Tesselation features*

Depending on which parameters are studied we find different systematic features of the tesselation.

In the plots varying mismatch and insertion deletion-penalty, we found the pencil of lines characterizing the global alignments.

In the plots varying the initiation and extension gap penalties, bands of regions separated by vertical lines indicate alignments that contain the same number of insertions and deletion distributed over a different number of gaps. This we interpret as an indicator of well-matching regions defining the outlines of an alignment.

In the plots varying the gap penalty and a constant added to the score matrix, vertical lines indicate growth of an alignment at the ends without internal change.

With every set of parameters one expects to find different features. Such systematic features reflect the limitations in the change that alignments undergo when these parameters vary. This clearly contradicts our intuition of a confusing multitude of optimal alignments.

(b) *Role of the matrix*

We also found it surprising that most of the alignments corresponding to larger regions in Figure 4 can be described in terms of only a few diagonals. The fact that we compare biologically related sequences seems to be responsible for this. Their comparison shows diagonals most of which are clearly above noise-level and are following each other without huge gaps in between. Any reasonable alignment will use the good diagonals and string them together in a way that is determined by the gap penalty. Where there are decisions to be made as in our example between Wf and Wc, the gap penalty might decide whether the alignment includes a good diagonal that requires a larger gap or a less attractive one that might be easier to reach. In this sense, the matrix defines the "players". It defines which diagonals will be considered by an alignment.

Our analysis is not aimed at judging how well a certain matrix defines these diagonals. Other attributes of a matrix seem to be more generic. Consider Figure 6 where the average of the matrix decreases along the horizontal axis. At the very right the matrix average is around -2 and the C2 alignment has taken over the entire logarithmic region. The linear region (shaded region along the horizontal axis) on the other hand is populated only by biologically meaningless alignments containing lots of gaps. Following the good alignments coming from the left, one sees that they are literally squeezed in between the linear and the logarithmic regions. At the end they either disappear entirely or are reduced to a tiny area around the phase transition. We conclude that the matrix average should be negative but not too negative.

(i) At the same time the matrix average should be kept negative for the following reasons.

(ii) A strongly positive matrix average introduces a bias against deletions in the shorter sequence (see section 2(d)).

(iii) Statistics are better understood in the logarithmic region than in the linear region (Arratia *et al.*. 1990; Karlin & Altschul, 1990; C. Neuhauser, unpublished results).

(iv) A negative matrix average ensures the existence of a phase transition curve within the parameter space spanned by the gap penalties, and this in turn is helpful in locating reasonable gap penalties (see below).

(c) *Gap penalty choice*

Once a matrix is specified and fulfils the above criterion of having its average slightly under 0, gap penalties can be calibrated specifically for it. This choice depends on the purpose one has in mind. The main distrinction is between searching a database and comparing two sequences. In the first case the objective is the best contrast between sequences related to the query and all others. In the latter case the alignment should be as good and as complete as possible. We speculate that good contrast will be achieved rather with a few well-matching diagonals than an overall comparison. This implies a gap penalty choice well in the logarithmic region for the purpose of database searching.

For the comparison of two sequences, however, the strong "core"-alignment must be extended to encompass further matching regions. In the tesselation this will typically happen as one moves from an outer region towards the transition curve. This suggests choosing gap penalties near the transition curve when the objective is a good alignment of two sequences. Figure 8 illustrates this point by overlaying the transition curve for the PAM250 matrix with some choices of gap penalties. The o denotes the Wf-C2 alignment. * is the C1-Wf-C2-d alignment and + is C1-Wc ~ C2-d.

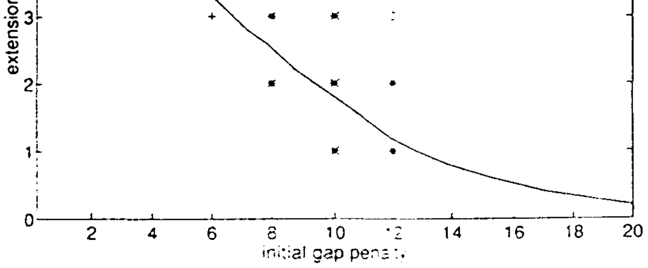

Figure 8. The phase transition in the parameters initial and extension gap penalty for the PAM250 matrix together with the location of certain alignments. o denotes the Wf-C2 alignment. * the C1-Wf-C2-d alignment and + is C1-Wc ~ C2-d alignment.

As expected we are not able to say anything systematic about the exact location of the qualitatively best alignments in the tesselation. The traditional choice of "large initial, small extension" certainly does not suffice to find these regions. The study of the entire tesselation and the identification of the above-mentioned bands can aid in identifying good alignments.

(d) *Conclusion*

The above analysis has helped us to understand how parameters influence alignments and how we can calibrate certain parameters. In addition, there is the insight that there are many more systematic (combinatorial and statistical) features about optimal alignments than previously known that can aid in searching for a biologically valid alignment. On the other hand, one needs to avoid overinterpreting the result of an alignment program. It may well be that the feature that strikes the researcher's eye can be explained totally without recourse to biology from the observations made above only. A thorough knowledge of the systematic changes in alignments should therefore help us not to be led astray by essentially uninteresting consequences of the mathematical model used.

We thank Mark Eggert for programming assistance and preparation of several Figures. This work was supported by the National Science Foundation (DMS 90-05833, DMS 87-20208) and the National Institutes of Health (GM36230).

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Amzel, L. M. & Poljak, R. J. (1979). Three-dimensional structure of immunoglobulins. *Annu. Rev. Biochem.* **48**, 961–997.

Apostolico, A. & Guerra, C. (1987). The longest common subsequence problem revisited. *Algorithmica*, **2**, 315–336.

Arratia, R., Gordon, L. & Waterman, M. S. (1990). The Erdös-Rényi law in distribution for coin tossing and sequence matching. *Ann. Statist.* **18**, 539–570.

Barton, G. J. & Sternberg, M. J. E. (1987). Evaluation and improvements in the automatic alignment of protein sequences. *Protein Eng.* **1**, 89–94.

Chvátal, V. & Sankoff, D. (1975). Longest common subsequences of two random sequences. *J. Appl. Probab.* **12**, 306–315.

Dayhoff, M. O., Barker, W. C. & Hunt, L. T. (1983). Establishing homologies in protein sequences. *Methods Enzymol.* **91**, 524–545.

Fernández-Baca, D. & Srinivasan, S. (1991). Constructing the minimization diagram of a two-parameter problem. *Operat. Res. Letters*, **10**, 87–93.

Fitch, W. M. & Smith, T. F. (1983). Optimal sequence alignments. *Proc. Nat. Acad. Sci., U.S.A.* **80**, 1382–1386.

Gotoh, O. (1990). Optimal sequence alignment allowing for long gaps. *Bull. Math. Biol.* **52**, 359–373.

Gribskov, M. & Burgess, R. R. (1986). Sigma factors from *E. coli*, *B. subtilis*, phage SPO1, and phage T4 are homologous proteins. *Nucl. Acids Res.* **14**, 6745–6763.

Gusfield, D., Balasubramanian, K. & Naor, D. (1992). Parametric optimization of sequence alignment. *Proceedings of the Third Annual ACM-SIAM Symposium on Discrete Algorithms, January 1992*, 432–439.

Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 2264–2268.

Kaupp, U. B., Vingron, M., Altenhofen, W., Bönigk, W., Eismann, E. & Ludwig, J. (1992). Cyclic nucleotide-gated channels—a family of proteins involved in vertebrate photoreception and olfaction. In *Signal Transduction in Protoreceptor Cells* (Hargrave, P. A., Hofmann, K. P. & Kaupp, U. B., eds), pp. 195–213. Springer-Verlag.

McCaldon, P. & Argos, P. (1988). Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences. *Proteins: Struct. Funct. Genet.* **4**, 99–122.

Needleman, S. B. & Wunsch, Ch. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.

Rechid, R., Vingron, M. & Argos, P. (1989). A new interactive protein sequence alignment program and comparison of its results with widely used algorithms. *Comp. Appl. Biosci.* **5**, 107–113.

Sauer, R. T., Yocum, R. R., Doolittle, R. F., Lewis, M. & Pabo, C. O. (1982). Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. *Nature (London)*, **298**, 447–451.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.

Smith, T. F., Waterman, M. S. & Fitch, W. M. (1981). Comparative biosequence metrics. *J. Mol. Evol.* **18**, 38–46.

Waterman, M. S. (1984). General methods of sequence comparison. *Bull. Math. Biol.* **46**, 473–500.

Waterman, M. S. & Eggert, M. (1987). A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. Mol. Biol.* **197**, 723–725.

Waterman, M. S., Gordon, L. & Arratia, R. (1987). Phase transitions in sequence matches and nucleic acid structure. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 1239–1243.

Waterman, M. S., Eggert, M. & Lander, E. (1992). Parametric sequence comparisons. *Proc. Nat. Acad. Sci. U.S.A.* **89**, 6090–6093.

Edited by F. Cohen

(Received 21 April 1993; accepted 26 August 1993)