June 1, 1999

Consistent Equivalence Relations: a Set-Theoretical Framework for Multiple Sequence Alignment

Burkhard Morgenstern[1], Jens Stoye[2,3], and Andreas Dress[2]

[1] GSF – National Research Center for Environment and Health, Institute of Biomathematics and Biometry, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany. Present address: Rhône-Poulenc Rorer, JA3-3, Rainham Road South, Dagenham, Essex RM10 7XS, U.K.
[2] Research Center for Interdisciplinary Studies on Structure Formation (FSPM), Universität Bielefeld, Postfach 100131, 33501 Bielefeld, Germany
[3] Present address: Deutsches Krebsforschungszentrum, Theoretische Bioinformatik (Abt. H0300), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

Abstract

Recently, Morgenstern *et al.* have proposed a new mathematical definition of sequence alignment (Morgenstern *et al.*,1996). In this paper, we discuss this definition in more detail. We demonstrate that it provides an appropriate conceptual framework in which problems arising in the context of sequence alignment can be treated systematically.

1 Introduction

In the standard theory of sequence alignment, an alignment of N sequences $\mathbf{s}_1, \ldots, \mathbf{s}_N$ is defined as a matrix, whose rows are so-called 'padded sequences' $\mathbf{s}_1^*, \ldots, \mathbf{s}_N^*$, which are obtained from the original sequences by insertion of additional characters designated as 'blanks', 'gap characters' or 'neutral elements' (cf. Waterman, 1995; p. 186).

This formal definition corresponds to the way alignments are constructed by standard alignment algorithms: Following a method developed in 1970 by Needleman and Wunsch (Needleman & Wunsch, 1970), alignments are constructed by inserting *gaps* into sequences (and by penalizing them by so-called *gap penalties*).

This point of view clearly is appropriate if the sequences in question are closely related and only a few gaps have to be inserted to align them properly. However, in general, related sequences may share only limited regions of similarity, separated e.g. by introns in DNA sequences or by loop regions in proteins. In such cases, the classical alignment definition has a certain drawback, at least from the theoretical point of view: the way gaps may have to be inserted into the sequences to highlight the related parts of the sequences properly, is by no means unique.

Consider e.g. the sequences $\mathbf{s}_1 = abcdddddfgh$ and $\mathbf{s}_2 = abceefgh$. A meaningful alignment of these sequences will certainly align their first three as well as their last three letters – but it seems pointless to align the d's and e's in the middle of the sequences, since these parts of the sequences share no similarity at all.

If we want to describe such an alignment in standard terms, there are several matrices all of them describing essentially the same alignment (see Figure 1, alignments A_1 – A_3).

To avoid such ambiguities, we propose another definition of what should be called an alignment. While the standard definition focuses on the gaps introduced into the sequences – i.e. on the *non-aligned residues* –, we define an alignment by specifying exclusively the *aligned residues*, and we simply ignore the non-aligned residues in our definition.

In our exposition, we will freely use standard terminology from basic set theory (including standard terminology referring to binary relations, in particular equivalence relations and partial (quasi-)orders, defined on a set X) as recalled in the Appendix of this note. In addition, the Appendix will also provide the proof for a number of specific facts which we will require in our exposition and which are not yet standard knowledge or folklore. This way, our exposition can be kept concise while those facts can be established within a simultaneously more natural and considerably more general environment.

Remark: While designing the concepts and establishing the proofs discussed in the Appendix, it became apparent that an even more general notion of what really is an alignment should be studied which encompasses the two

concepts discussed here as its two extreme cases and allows for even more transparent and straight forward proofs of all the basic facts needed to be established in this context. This notion will be presented and discussed in a forthcoming paper entitled 'What really *are* alignments?'.

2 Alignments as Equivalence Relations

In this section, we give a formal mathematical definition of what we want to call an alignment, and we draw some immediate conclusions from this definition. We assume familiarity with the basic concepts of the mathematical theory of sets and relations (see for instance Bourbaki, 1968).

Let us consider a finite set $\mathcal{A}$, called the *alphabet*. The disjoint union

$$\bigcup_{n=0}^{\infty} \mathcal{A}^n =: \mathcal{A}^*$$

is called the *sequence space* over $\mathcal{A}$. Given an alphabet $\mathcal{A}$, any map

$$\mathbf{s} = \mathbf{s}_I : I \to \mathcal{A}^*$$

from a (finite) index set I into $\mathcal{A}^*$ is called a *sequence family* (over $\mathcal{A}$). For every $i \in I$, we denote by L_i the *length* of the sequence $\mathbf{s}(i)$, i.e. the unique number $n \in I\!N_0$ $(:= \{0, 1, \ldots\})$ with $\mathbf{s}(i) \in \mathcal{A}^n$. For simplicity, we write $\mathbf{s}_i$ rather than $\mathbf{s}(i)$ to designate the i-th sequence.

The *site space* of a sequence family $\mathbf{s}_I$ is now defined to be the set

$$\mathcal{S} = \mathcal{S}(\mathbf{s}_I) := \{[i|j] : i \in I, 1 \leq j \leq L_i\}.$$

If I' is a subset of I, we denote the restriction $\mathbf{s}_I|I'$ by $\mathbf{s}_{I'}$ and we denote $\mathcal{S}(s_{I'})$ by $\mathcal{S}(I')$. In addition, for $i \in I$, we write $\mathcal{S}(i)$ instead of $\mathcal{S}(\{i\})$, so $\mathcal{S}(i)$ is the set of all sites belonging to the i-th sequence:

$$\mathcal{S}(i) := \{[i|j] : 1 \leq j \leq L_i\}.$$

To describe an alignment of a sequence family $\mathbf{s}_I$ by pointing out which residues are aligned with each other – rather than by focusing on where to introduce gaps –, we now define an alignment of $\mathbf{s}_I$ as a binary relation A defined on the site space $\mathcal{S}$ and we'll write $([i|j], [i'|j']) \in A$ if we want to express that the j-th site of the i-th sequence is aligned with the j'-th site of the i'-th sequence.

Obviously, the relation A is symmetrical as well as transitive provided we consider each site $[i|j] \in \mathcal{S}$ as being aligned with itself, that is, we assume A also to be reflexive. In other words, we assume an alignment A to be an equivalence relation defined on the site space $\mathcal{S}$, and we will also write $x \overset{A}{\sim} y$ instead of $(x, y) \in A$ for any pair $x, y \in \mathcal{S}$ of aligned sites.

On the other hand, we do not want to regard every equivalence relation defined on $\mathcal{S}$ as an alignment of the sequence family $\mathbf{s}_I$. We require an alignment to be 'consistent' in a certain sense with the linear order of the sites in the individual sequences. It is therefore convenient to define a certain partial order '$\preceq$' on the site space $\mathcal{S}$, the so-called 'direct sum' of the linear orders given on the single sequences, which is defined by

$$[i|j] \preceq [i'|j'] \quad \Leftrightarrow \quad i = i' \text{ and } j \leq j'.$$

(The relation '$\preceq$' corresponds to the 'implicit constraints' introduced by Myers *et al.*, 1996.)

An alignment A extends the partial order '$\preceq$' to a quasi partial order '$\preceq_A$' on the set $\mathcal{S}$ which is given as the transitive closure of the union $A \cup \preceq$ (note that $\preceq$ as well as A are binary relations defined on the site space $\mathcal{S}$, that is, they are subsets of the cartesian product $\mathcal{S}^2$, so we may apply the usual set-theoretical operations to them). In other words, for any two sites $x, y \in \mathcal{S}$, we write '$x \preceq_A y$' if and only if there exists a chain $x = x_0, x_1, \ldots, x_k = y$ of sites in $\mathcal{S}$ with either $x_{i-1} \preceq x_i$ or $x_{i-1} \overset{A}{\sim} x_i$ for each $i \in \{1, \ldots, k\}$ (see Figure 2). Now we can define precisely what 'consistent with the linear order of the sites in the individual sequences' means: For every $i \in I$, the restriction of the extended quasi order $\preceq_A$ to the subset $\mathcal{S}(i)$ has to coincide with the 'original' or 'natural' linear order relation defined on $\mathcal{S}(i)$. In other words, for every $i \in I$ and $j, j' \in \{1, \ldots, L_i\}$, we must have $[i|j] \preceq_A [i|j']$ if and only if $j \leq j'$ holds.

In formal terms, we propose

Definition 1 *(a) For a binary relation R defined on a set X, we denote by R_t its* **transitive closure** *and by R_e its* **equivalence closure**, *i.e. R_t and R_e are the smallest transitive relation and the smallest equivalence relation, respectively, which are defined on X and contain R.*

(b) Let $\mathbf{s}_I$ be a sequence family. On the site space $\mathcal{S} = \mathcal{S}(\mathbf{s}_I)$, we define a partial order '$\preceq$' by

$$[i|j] \preceq [i'|j'] \quad \Leftrightarrow_{def} \quad i = i' \text{ and } j \leq j'.$$

(c) For every binary relation R defined on the site space $\mathcal{S}$ and for every $x \in \mathcal{S}$ and $i \in I$, we define

$$\begin{aligned} \preceq_R &:= (R_e \cup \preceq)_t = (R \cup R^{-1} \cup \preceq)_t, \\ \prec_R &:= \preceq_R \setminus (\preceq_R \cap \preceq_R^{-1}), \\ b^{\downarrow}(i \mid x, R) &:= \min(j \in \{1, \ldots, L_i + 1\} \mid j = L_i + 1 \text{ or } x \preceq_R [i|j]), \end{aligned}$$

and

$$b^{\uparrow}(i \mid x, R) := \max(j \in \{0, \ldots, L_i\} \mid j = 0 \text{ or } [i|j] \preceq_R x).$$

(d) An equivalence relation A defined on the site space $\mathcal{S}$ is called an **alignment** *of the sequence family $\mathbf{s}_I$ if, for every $i \in I$, the restriction of the relation $\preceq_A$ to the subset $\mathcal{S}(i)$ coincides with the linear order given on $\mathcal{S}(i)$ in which case A will be called* **($\preceq$ −)consistent**.

More generally, we'll say that a binary relation $R \subseteq \mathcal{S}^2$ **induces** *an alignment of $\mathbf{s}_I$ if its equivalence closure R_e is an alignment of $\mathbf{s}_I$ in which case we'll also say that R is* **($\preceq$ −)consistent**.

Clearly, R is consistent if and only if $b^{\uparrow}(i \mid x, R) \leq b^{\downarrow}(i \mid x, R)$ holds for all $x \in \mathcal{S}$ and $i \in I$, and in this case we have $x \overset{R_e}{\sim} [i|j]$ for some $j \in \{1, \ldots, L_i\}$ if and only if $j = b^{\uparrow}(i \mid x, R) = b^{\downarrow}(i \mid x, R)$ holds. Consequently, we will refer to these numbers as the **consistency bounds** *(of i relative to x and R).*

(e) *Finally, for any subset $I' \subseteq I$, an alignment A is defined to be an I'-***maximal** *alignment if $\mathcal{S}(I')^2$ is contained in $\preceq_A \cup \preceq_A^{-1}$.*

Note that $x, y \in \mathcal{S}, i \in I$ and $x \preceq_R y$ implies

$$b^{\downarrow}(i \mid x, R) \leq b^{\downarrow}(i \mid y, R)$$

and

$$b^{\uparrow}(i \mid x, R) \leq b^{\uparrow}(i \mid y, R).$$

The following fact is established in the Appendix:

Lemma 1 *Let A be an equivalence relation defined on the site space $\mathcal{S} = \mathcal{S}(\mathbf{s}_I)$ of some sequence family $\mathbf{s}_I$. Then each of the following four properties implies the other three:*

(a) *A is an alignment of $\mathbf{s}_I$.*

(b) *Both of the following two conditions hold:*

$$\preceq_A \cap \preceq_A^{-1} \quad \subseteq \quad A, \tag{1}$$

$$A \cap \preceq \quad \subseteq \quad D_{\mathcal{S}} \tag{2}$$

(Recall that for any set X, D_X denotes the diagonal relation

$$\{(x, x) | x \in X\}.)$$

(c) *There exist a partially ordered set $(\tilde{\mathcal{S}}, \preceq_{\tilde{\mathcal{S}}})$ and a strictly monotonously increasing mapping*

$$\phi : (\mathcal{S}, \preceq) \to (\tilde{\mathcal{S}}, \preceq_{\tilde{\mathcal{S}}})$$

such that, for all $x, y \in \mathcal{S}$, $\phi(x) = \phi(y)$ holds if and only if $x \overset{A}{\sim} y$ holds, i.e. x is aligned with y.

(c') *There exist a linearly ordered set $(\tilde{\mathcal{S}}, \preceq_{\tilde{\mathcal{S}}})$ and a strictly monotonously increasing mapping*

$$\phi : (\mathcal{S}, \preceq) \to (\tilde{\mathcal{S}}, \preceq_{\tilde{\mathcal{S}}})$$

such that for all $x, y \in \mathcal{S}$, $\phi(x) = \phi(y)$ holds if and only if $x \overset{A}{\sim} y$ holds.

Since the 'padded sequences' $\mathbf{s}_1^*, \ldots, \mathbf{s}_N^*$ used in the standard alignment definition may be regarded as strictly monotonously increasing maps $\mathcal{S}(1) \to I\!N$, ..., $\mathcal{S}(N) \to I\!N$, our definition is closely related to the standard definition (cf. Chan *et al.*, 1992); however, we avoid the above mentioned ambiguity.

More precisely: We may divide the entity of all standard, i.e. matrix-based alignments of a given sequence family $\mathbf{s}_I$ into equivalence classes such that two 'matrix alignments' A and B are equivalent if and only if the two corresponding maps $\varphi_A : \mathcal{S}(\mathbf{s}_I) \to I\!N$ and $\varphi_B : \mathcal{S}(\mathbf{s}_I) \to I\!N$ satisfy the condition $\varphi_A(x) = \varphi_A(y) \iff \varphi_B(x) = \varphi_B(y)$ for all $x, y \in \mathcal{S}(\mathbf{s}_I)$ or – equivalently – B may be obtained from A by permuting successively pairs of two consecutive columns for which each sequence has at least one gap among its entries in those two columns. For instance, the first three 'matrix alignments' shown in Figure 1 belong to one single equivalence class. Clearly (see also the Appendix), there is a one-to-one correspondence between these equivalence classes and the alignments of $\mathbf{s}_I$ as defined in Definition 1 – the I-maximal alignments corresponding exactly to those equivalence classes of standard alignments which consist of just one such alignment only.

3 Alignments and Consistency

How can we decide algorithmically whether or not a given binary relation A defined on the site space $\mathcal{S} = \mathcal{S}(\mathbf{s}_I)$ of some sequence family $\mathbf{s}_I$ induces an alignment? Clearly, we might start with $A_0 := \emptyset$ and then proceed by successively enlarging the subset in question by pairs (x, y) from A, always checking for consistency. More generally, we may assume that we are given some pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k) \in \mathcal{S}^2$ of elements from $\mathcal{S}$, and that we may want to construct an alignment A of $\mathbf{s}_I$ recursively in a greedy fashion by putting $A_0 := \emptyset$ as above and then putting

$$A_\kappa := \begin{cases} (A_{\kappa-1} \cup \{(x_\kappa, y_\kappa)\})_e, & \text{if } A_{\kappa-1} \cup \{(x_\kappa, y_\kappa)\} \text{ is consistent,} \\ A_{\kappa-1}, & \text{otherwise} \end{cases}$$

(cf. Abdeddaïm, 1997 a/b; Morgenstern *et al.*, 1996.) In both cases, we need to know whether or not the union of an alignment $A \subseteq \mathcal{S}^2$ and a single pair $(\hat{x}, \hat{y}) \in \mathcal{S}^2$ is consistent. This question is answered by Proposition 1 which is also established in the Appendix:

Proposition 1 *Let $\mathbf{s}_I$ be a sequence family, let A be an alignment of $\mathbf{s}_I$, assume $\hat{x}, \hat{y} \in \mathcal{S} = \mathcal{S}(\mathbf{s}_I)$, and put $A' := A \cup \{(\hat{x}, \hat{y})\}$. Then*

$$x \preceq_{A'} y \iff \begin{cases} x \preceq_A y, \text{ or} \\ (x \preceq_A \hat{x} \text{ and } \hat{y} \preceq_A y), \text{ or} \\ (x \preceq_A \hat{y} \text{ and } \hat{x} \preceq_A y). \end{cases}$$

In particular, A' is consistent if and only if neither $\hat{x} \prec_A \hat{y}$ nor $\hat{y} \prec_A \hat{x}$ holds.

Corollary 1 *A is not properly contained in any other alignment if and only if A is I-maximal, i. e. if we have*

$$\mathcal{S}^2 = \preceq_A \cup \preceq_A^{-1}$$

in which case $\preceq_A$ is a **total** *quasi-order.*

(Recall that a relation R on a set X is called *total* if any two elements of X are "comparable" via R, i.e. if one has $R \cup R^{-1} = X^2$.)

Corollary 2 *Given an alignment A, an element $\hat{x} \in \mathcal{S}$ and an element $\hat{y} = [i_1|j_1] \in \mathcal{S}$, the relation $A' := A \cup \{(\hat{x}, \hat{y})\}$ is a proper consistent extension of A if and only if one has $b^\uparrow(i_1 \mid \hat{x}, A) < j_1 < b^\downarrow(i_1 \mid \hat{x}, A)$; moreover, whether or not A' is consistent, its consistency bounds can be computed, for any $x \in \mathcal{S}$ and $i \in I$, by the simple formulae*

$$b^\downarrow(i \mid x, A') = \begin{cases} \min(b^\downarrow(i \mid x, A), b^\downarrow(i \mid \hat{y}, A)) & \text{if } x \preceq_A \hat{x}, \\ \min(b^\downarrow(i \mid x, A), b^\downarrow(i \mid \hat{x}, A)) & \text{if } x \preceq_A \hat{y}, \\ b^\downarrow(i \mid x, A) & \text{else} \end{cases} \tag{3}$$

and

$$b^{\uparrow}(i \mid x, A') = \begin{cases} \max(b^{\uparrow}(i \mid x, A), b^{\uparrow}(i \mid \hat{y}, A)) & \text{if } \hat{x} \preceq_A x, \\ \max(b^{\uparrow}(i \mid x, A), b^{\uparrow}(i \mid \hat{x}, A)) & \text{if } \hat{y} \preceq_A x, \\ b^{\uparrow}(i \mid x, A) & \text{else,} \end{cases} \tag{4}$$

– note that this is well-defined as $x \preceq_A \hat{x}$ and $x \preceq_A \hat{y}$ *(or* $\hat{x} \preceq_A x$ and $\hat{y} \preceq_A y$*) implies*

$$\min(b^{\downarrow}(i \mid x, A), b^{\downarrow}(i \mid \hat{y}, A)) = \min(b^{\downarrow}(i \mid x, A), b^{\downarrow}(i \mid \hat{x}, A)) = b^{\downarrow}(i \mid x, A)$$

(or

$$\max(b^{\uparrow}(i \mid x, A), b^{\uparrow}(i \mid \hat{y}, A)) = \max(b^{\uparrow}(i \mid x, A), b^{\uparrow}(i \mid \hat{x}, A)) = b^{\uparrow}(i \mid x, A),$$

respectively).

Clearly, these formulae allow us to formally set up the above mentioned recursive procedure for constructing alignments greedily for any given sequence $(x_1, y_1), \ldots, (x_k, y_k)$ of pairs of elements from $\mathcal{S}$ that we wish to align.

If the sequence family under consideration consists of N sequences $\mathbf{s}_1, \ldots, \mathbf{s}_N$ of total length $L := L_1 + \ldots + L_N = \#\mathcal{S}$, equations (3) and (4) allow us to compute the new values of $b^{\downarrow}$ and $b^{\uparrow}$ in $O(NL)$ time and space whenever a new pair (x_i, y_i) is included into the growing alignment.

Remark: Note that, with $x = [i_0, j_0]$, (3) is equivalent with

$$b^{\downarrow}(i \mid x, A') = \begin{cases} b^{\downarrow}(i \mid \hat{y}, A) & \text{if } b^{\uparrow}(i_0 \mid [i, b^{\downarrow}(i, \hat{y})] < j_0 \leq b^{\uparrow}(i_0 \mid \hat{x}) \\ b^{\downarrow}(i \mid \hat{x}, A) & \text{if } b^{\uparrow}(i_0 \mid [i, b^{\downarrow}(i, \hat{x})] < j_0 < b^{\uparrow}(i_0 \mid \hat{y}) \\ b^{\downarrow}(i \mid x, A) & \text{else} \end{cases}$$

and that (4) can be rewritten in a similar way which allows for a rather fast updating procedure, essentially of the order N^2.

Based on a graph theoretical approach, Abdeddaïm has proposed an algorithm which – consistency provided – includes the pairs $(x_1, y_1), \ldots, (x_k, y_k)$ successively into a growing alignment by computing the values of $b^{\downarrow}$ and $b^{\uparrow}$ every time a new pair (x_i, y_i) is included into the alignment. This algorithm takes $O(kL + L^2)$ time and $O(NL)$ space for processing *all* pairs $(x_1, y_1), \ldots, (x_k, y_k)$. (Abdeddaïm, 1997 b).

Next, we generalize formulae (3) and (4) to the case in which an alignment A is extended by incorporating an arbitrary *pairwise alignment* P. To this end, we define:

Definition 2 *Let* $\mathbf{s}_I$ *be a sequence family. An alignment* P *of* $\mathbf{s}_I$ *is called a* **pairwise alignment** *if there exist* $i, j \in I$ *so that* P *is the equivalence closure of a consistent relation* $R = \{(x_1, y_1), \ldots, (x_k, y_k)\} \subseteq \mathcal{S}(i) \times \mathcal{S}(j)$ *(or – equivalently – if using the notation introduced in the Appendix, we have* $girth_{\preceq}(supp(P - D_{\mathcal{S}})) \leq 2$*).*

So let A be an arbitrary alignment, assume $(x_1, y_1), \ldots, (x_k, y_k) \in \mathcal{S}(i) \times \mathcal{S}(j)$ such that $P := \{(x_1, y_1), \ldots, (x_k, y_k)\}_e$ is a pairwise alignment and assume that $A' := A \cup P$ is a consistent extension of A.

Then, applying Proposition 1 and Corollary 2 repeatedly, we get

Corollary 3 *For all $x, y \in \mathcal{S}$ and $i \in I$, one has*

$$x \preceq_{A'} y \iff \begin{cases} x \preceq_A y, \text{ or} \\ (x \preceq_A x_\kappa \text{ and } y_\kappa \preceq_A y \text{ for some } \kappa \in \{1, \ldots, k\}), \text{ or} \\ (x \preceq_A y_\kappa \text{ and } x_\kappa \preceq_A y \text{ for some } \kappa \in \{1, \ldots, k\}) \end{cases}$$

as well as

$$\begin{aligned} b^{\downarrow}(i \mid x, A') \quad = \quad & \min(\{b^{\downarrow}(i \mid x, A)\} \cup \{\min(b^{\downarrow}(i \mid x_\kappa, A), b^{\downarrow}(i \mid y_\kappa, A)) \mid \\ & \kappa \in \{1, \ldots, k\} \text{ and } x \preceq_A x_\kappa \text{ or } x \preceq_A y_\kappa\}) \end{aligned}$$

and

$$\begin{aligned} b^{\uparrow}(i \mid x, A') \quad = \quad & \max(\{b^{\uparrow}(i \mid x, A)\} \cup \{\max(b^{\uparrow}(i \mid x_\kappa, A), b^{\uparrow}(i \mid y_\kappa, A)) \mid \\ & \kappa \in \{1, \ldots, k\} \text{ and } x_\kappa \preceq_A x \text{ or } y_\kappa \preceq_A x\}) \end{aligned}$$

4 Consistency of Pairwise Alignments

Recently, Stoye has shown the following fact concerning consistency of pairwise standard alignments: Consider a set of N sequences and pairwise standard alignments $A_{i,j}$ for all $1 \leq i < j \leq N$. If any three of these pairwise alignments are consistent, then all of them are *simultaneously* consistent (Stoye, 1997).

It is remarkable that this result depends crucially on the underlying alignment definition: Generally, triplewise consistency of a family of pairwise alignments in the sense of Definition 2 does *not* imply *simultaneous* consistency of the family. Consider, for instance, the site space

$$\mathcal{S} := \{[i|j] : i \in \{1, \ldots, 4\}, j \in \{1, 2\}\}$$

and the pairwise alignments

$$P_{i,i'} := \begin{cases} D_{\mathcal{S}} \cup \{([i|1], [i'|2])\}_e & \text{if } i' = i+1 \pmod 4, \\ D_{\mathcal{S}} & \text{otherwise.} \end{cases}$$

Here, any three of the pairwise alignments are consistent – but the family $(P_{i,i'})_{i,i' \in \{1,\ldots,4\}}$ is *not* consistent.

The point is that standard alignments contain more information than consistent equivalence relations in the sense of Definition 1. We will discuss this difference in the last section of this paper.

However, the following theorem states that triplewise consistency of pairwise alignments *does* imply simultaneous consistency if pairwise alignments are *maximal* in a certain sense.

Theorem 1 *Let $\mathbf{s}_I$ be a sequence family and suppose that for every pair $(i,j) \in I^2$ there is a pairwise alignment $P_{i,j}$ of the sequences $\mathbf{s}_i$ and $\mathbf{s}_j$ such that the following two conditions hold:*

(a) For every pair (i,j), the alignment $P_{i,j}$ is $\{i,j\}$-maximal.

(b) For all $i,j,k \in I$, the family $(P_{i,j}, P_{i,k}, P_{j,k})$ is consistent.

Then the union $\bigcup_{i,j \in I} P_{i,j}$ is an alignment of $\mathcal{S}$.

Proof: For any $x, y \in \mathcal{S}$ with, say, $x \in \mathbf{s}_i$ and $y \in \mathbf{s}_j$, we have by definition either $x \preceq_{P_{i,j}} y$ or $y \preceq_{P_{i,j}} x$. In other words, we have

$$\mathcal{S} = \bigcup_{i,j \in I} (\preceq_{P_{i,j}} \cup \preceq_{P_{i,j}}^{-1}).$$

Now our assertion follows from Corollary 7. □

Definition 3 *(a) Let $\mathbf{s}_I$ be a sequence family. For every alignment A of $\mathbf{s}_I$ and every $i,j \in I$, we define*

$$T_{i,j}(A) := (A \cap (\mathcal{S}(i) \times \mathcal{S}(j)))_e \,. \tag{5}$$

(b) Consider a real-valued, monotonically increasing function w defined on the set of all pairwise alignments of $\mathbf{s}_i$ and $\mathbf{s}_j, i, j \in I$. A pairwise alignment P of $\mathbf{s}_i$ and $\mathbf{s}_j$ is called a **(w-)optimal alignment** *of the sequences $\mathbf{s}_i$ and $\mathbf{s}_j$, if $w(Q) \leq w(P)$ holds for all pairwise alignments Q of $\mathbf{s}_i$ and $\mathbf{s}_j$.*

(c) If the set I is finite, we call an alignment A of $\mathbf{s}_I$ a **(sum-of-pairs) (w-) optimal alignment** *of $\mathbf{s}_I$, if*

$$\sum_{i,j \in I} w(T_{i,j}(B)) \leq \sum_{i,j \in I} w(T_{i,j}(A)) \tag{6}$$

holds for all alignments B of $\mathbf{s}_I$.

Remark 1 *$T_{i,j}(A)$ is the largest pairwise alignment of the sequences $\mathbf{s}_i$ and $\mathbf{s}_j$ contained in A.*

Theorem 2 *Consider a finite set I, a sequence family $\mathbf{s}_I$, a real-valued, monotonically increasing function w defined on the set of all pairwise alignments of $\mathbf{s}_i$ and $\mathbf{s}_j, i, j \in I$, and let $P_{i,j}$ be a pairwise alignment of the sequences $\mathbf{s}_i$ and $\mathbf{s}_j$ for every $i, j \in I$.*

If for every $i, j \in I$, $P_{i,j}$ is a optimal alignment of the sequences $\mathbf{s}_i$ and $\mathbf{s}_j$ and if the family $(P_{i,j})$ is consistent, then

$$A := \left(\bigcup_{i,j \in I} P_{i,j} \right)_e$$

is a (w-)optimal alignment of $\mathbf{s}_I$.

Proof: The inclusion $P_{i,j} \subseteq T_{i,j}(A)$ implies $w(P_{i,j}) = w(T_{i,j}(A))$ for all $i, j \in I$. Therefore, we have for every alignment B of the sequence family $\mathbf{s}_I$

$$\sum_{i,j \in I} w(T_{i,j}(B)) \leq \sum_{i,j \in I} w(P_{i,j}) = \sum_{i,j \in I} w(T_{i,j}(A)). \tag{7}$$

□

Note that this last observation also holds for optimal 'standard' alignments (Stoye, 1997; Lemma 2.15).

5 Discussion

Listing pairs of – presumably – homologue residues or sites is a natural way of describing sequence alignments (see e.g. Smith and Waterman, 1981; Kruskal, 1983; Vingron and Argos, 1991; Miller, 1993; Miller *et al.*, 1994). If two sequences are to be aligned, some authors call such a list of pairs of residues a 'trace' rather than an 'alignment' and use the term 'alignment' only for standard 'matrix alignments' (cf. Kruskal, 1983; Waterman, 1995). Kececioglu has used the language of graph theory to generalize the definition of a 'trace' to multiple sequence comparison (Kececioglu, 1993).

Kruskal remarks that 'alignments are richer than traces' since they contain some information about the order of adjacent gaps which is not contained in a mere listing of aligned pairs of residues (see Figure 1). However, it seems doubtful if this information can be inferred from sequence comparison alone – so if sequences are compared solely on the basis of their primary structure, it seems to be adequate if we try to avoid or, at least, make no assertions about the order of adjacent gaps and confine us to pointing out which residues of the sequences can be expected to be homologous.

Mathematically spoken, a set of aligned pairs of residues or sites is a binary *relation.* In the case of pairwise alignments, this relation can be regarded as a relation *between* the set of sites of the first sequence and that of the second one – i.e. as a subset of the cartesian product $\mathcal{S}(1) \times \mathcal{S}(2)$ (Miller *et al.*, 1994). However, since a direct generalization of this definition to multiple alignments is somewhat complicated, we prefer to define an alignment as a binary relation, defined on the set $\mathcal{S}(\mathbf{s}_I)$ of sites of *all* sequences.

Since the linear order relations on the single sequences can be described by one single partial order $\preceq$ on $\mathcal{S}(\mathbf{s}_I)$ as well, we have a natural and uniform setting to treat questions arising in the context of sequence alignment – especially questions concerning the *consistency* of alignments. The concepts of the theory of sets and relations provide a convenient mathematical framework which allows a simple and transparent treatment of these questions.

Taylor (1996) deplores the absence of a method to assess the quality of remotely related sequences that does not depend on the unreliable exact location of gaps. By successively adding consistent gapfree pairs of segments of the sequences to an initially empty set, algorithms based on our alternative alignment definition can be developed which help improving this situation.

A Some more Definitions, Results, and Proofs

In the following, we consider binary relations $T, R, A, B, \ldots$ defined on a set X. For any such relations $R, R_1, R_2 \subseteq X^2$, we put (as usual)

$$R^{-1} := \{(y,x) \in X^2 \mid (x,y) \in R\}$$

and

$$\begin{aligned} R_1 \circ R_2 \quad := \quad & \{(x,y) \in X^2 \mid \text{ there exists some } z \in X \\ & \text{with } (x,z) \in R_1 \text{ and } (z,y) \in R_2\}. \end{aligned}$$

Recall that R is called reflexive (or symmetric or transitive), if $D_X \subseteq R$ (or $R = R^{-1}$, or $R \circ R \subseteq R$, respectively) holds.

Now, assume $T \subseteq X^2$ to be a transitive binary relation, and $A, B \subseteq X^2$ to be equivalence relations defined on X. We define A to be ***T*-consistent** if the assertions

(A1) $\quad T \cap T^{-1} \subseteq A$

and

(A2) $\quad T_A \cap T^{-1} \subseteq T$

hold true, where T_A denotes the smallest transitive relation defined on X which simultaneously contains T and A. More generally, we define a binary relation $R \subseteq X^2$ to be T-consistent if the equivalence relation

$$R(T) := (R \cup (T \cap T^{-1}))_e$$

is T-consistent, and we define $T_R := T_{R(T)}$.

Clearly, if $T \cap T^{-1} \subseteq A \subseteq B$ holds, then A is necessarily T-consistent whenever B is T-consistent in view of

$$T_A \cap T^{-1} \subseteq T_B \cap T^{-1} \subseteq T.$$

More precisely, we have

Theorem 3 *For X, T, A and B as above with $A \subseteq B$, the following three assertions are equivalent:*

(i) *B is T-consistent and $T \cap T^{-1}$ is contained in A,*

(ii) *$T_B \cap T_A^{-1}$ is contained in A and $A \cap T^{-1}$ is contained in T,*

(iii) *B is T_A-consistent and A is T-consistent.*

Proof:

(i) $\Rightarrow$ (ii): Clearly, $T_B \cap T^{-1} \subseteq T$ implies $A \cap T^{-1} \subseteq T$. Now, assume $(x,y) \in T_B$ and $(y,x) \in T_A$ for some $x, y \in X$. Then there exist elements $x_0 := x, x_1, \ldots, x_n := y, x_{n+1}, \ldots, x_m := x$ with $(x_{i-1}, x_i) \in B \cup T$ for $i = 1, \ldots, n$ and $(x_{i-1}, x_i) \in A \cap T$ for $i = n+1, \ldots, m$. Clearly, this implies $(x_{i-1}, x_i) \in B \cup T$ for all $i = 1, \ldots, m$ and, hence, also $(x_i, x_{i-1}) \in T_B$. So, for all $i = n+1, \ldots, m$, we have

$$(x_i, x_{i-1}) \in T_B \cap (A^{-1} \cup T^{-1}) = (T_B \cap A) \cup (T_B \cap T^{-1}) \subseteq A \cup (T \cap T^{-1}) \subseteq A$$

which implies $(y,x) \in A$ as claimed.

(ii) $\Rightarrow$ (iii): Clearly, the first assumption implies

$$T \cap T^{-1} \subseteq T_A \cap T_A^{-1} \subseteq A \subseteq B$$

as well as

$$T_B \cap T_A^{-1} \subseteq A \subseteq T_A,$$

so it implies in particular that B is T_A-consistent. Similarly, the first and the second assumption together imply

$$T_A \cap T^{-1} \subseteq T_B \cap T_A^{-1} \cap T^{-1} \subseteq A \cap T^{-1} \subseteq T,$$

so A is also T-consistent.

(iii) $\Rightarrow$ (i): Clearly, $T \cap T^{-1} \subseteq A$ and $A \subseteq B$ implies

$$T \cap T^{-1} \subseteq B,$$

while $T_B \cap T_A^{-1} \subseteq T_A$ and $T_A \cap T^{-1} \subseteq T$ together imply

$$T_B \cap T^{-1} = T_B \cap T_A^{-1} \cap T^{-1} \subseteq T_A \cap T^{-1} \subseteq T.$$

□

Corollary 4 *Given X, T, and A as above, then A is T-consistent if and only if $T_A \cap T_A^{-1}$ is contained in A and $A \cap T^{-1}$ is contained in T.*

Proof: Put $B := A$ in Theorem 3 and use that, in this case, Assertion **(i)** is equivalent with the assertion that $A = B$ is T-consistent, so this assertion is equivalent with Assertion **(ii)** which is exactly what is claimed in Corollary 4.□

Clearly, putting $X := \mathcal{S} := \mathcal{S}(\mathbf{s}_I)$ for some family of sequences $\mathbf{s}_I$ and $T := \preceq$, the above corollary implies the equivalence of the assertions (a) and (b) in Lemma 1, while the remaining implications "$(b) \Leftrightarrow (c) \Leftrightarrow (c')$" are simple consequences of the wellknown facts that any reflexive and transitive relation T defined on a set X induces a partial order on the set $\overline{X}$ of equivalence classes

relative to the equivalence relation $T \cap T^{-1}$, and that any partial order can be extended to a linear order.

Next, observe that, for any transitive and reflexive relation $T \subseteq X^2$ and any pair $(y_1, y_2) \in X^2$, we have

$$(T \cup \{(y_1, y_2), (y_2, y_1)\})_t = T \cup T_{12} \cup T_{21}$$

with

$$T_{ij} := \{(u, v) \in X^2 \mid (u, y_i), (y_j, v) \in T\}$$

for $(i, j) \in \{(1, 2), (2, 1)\}$: Indeed, as $T \cup T_{12} \cup T_{21}$ contains $T \cup \{(y_1, y_2), (y_2, y_1)\}$ and is itself necessarily contained in $(T \cup \{(y_1, y_2), (y_2, y_1)\})_t$, it is enough to observe that $T \cup T_{12} \cup T_{21}$ is transitive which follows immediately from the relations

$$\begin{aligned} T \circ T &\subseteq T, \\ T \circ T_{ij} &\subseteq T_{ij}, \\ T_{ij} \circ T &\subseteq T_{ij}, \\ T_{ij} \circ T_{ij} &\subseteq T_{ij}, \end{aligned}$$

and

$$T_{ij} \circ T_{ji} \subseteq T$$

which are easily established (with $(i, j) \in \{(1, 2), (2, 1)\}$ of course, just as above).

It follows in particular that, in case $(y_1, y_2), (y_2, y_1) \notin T$, we have

$$\begin{aligned} &(T \cup \{(y_1, y_2), (y_2, y_1)\})_t \cap (T \cup \{(y_1, y_2), (y_2, y_1)\})_t^{-1} \\ &\quad = (T \cup T_{12} \cup T_{21}) \cap (T \cup T_{12} \cup T_{21})^{-1} \\ &\quad = (T \cap T^{-1}) \cup \{(u, v), (v, u) \mid (u, y_1), (y_1, u), (v, y_2), (y_2, v) \in T\} \end{aligned}$$

in view of

$$T \cap T_{12}^{-1} = T \cap T_{21}^{-1} = T_{12} \cap T^{-1} = T_{21} \cap T^{-1} = T_{12} \cap T_{12}^{-1} = T_{21} \cap T_{21}^{-1} = \emptyset$$

and

$$T_{12} \cap T_{21}^{-1} = \{(u, v) \mid (u, y_1), (y_1, u) \in T \text{ and } (v, y_2), (y_2, v) \in T\} = (T_{21} \cap T_{12}^{-1})^{-1}.$$

We can now conclude

Corollary 5 *Given X, T and A as above as well as two distinct equivalence classes $Y_1, Y_2 \subseteq X$ with respect to A, then – assuming $T \cap T^{-1} \subseteq A$ – the enlarged equivalence relation*

$$B \;:=\; A \;\cup\; Y_1 \times Y_2 \;\cup\; Y_2 \times Y_1$$

is T-consistent if and only if A is T-consistent and we have neither $(y_1, y_2) \in T_A$ nor $(y_2, y_1) \in T_A$ for some/all $y_1 \in Y_1$ and $y_2 \in Y_2$.

Proof: Clearly, $(u,v) \in T_A$ implies $(u',v') \in T_A$ for all $u', v' \in X$ with $u' \overset{A}{\sim} u$ and $v' \overset{A}{\sim} v$. So, of course, assuming $(y_1,y_2),(y_2,y_1) \notin T_A$ for some $y_1 \in Y_1$ and $y_2 \in Y_2$ is equivalent to assuming that for all $y_1, \in Y_1$ and $y_2 \in Y_2$. Moreover, the above theorem implies that B is T-consistent if and only if A is T-consistent and B is T_A-consistent which – in view of Corollary 4 – clearly implies $B \cap T_A^{-1} \subseteq T_A$ and, hence,

$$B \cap T_A = B^{-1} \cap T_A = (B \cap T_A^{-1})^{-1} \cap T_A \subseteq T_A^{-1} \cap T_A \subseteq A,$$

so we cannot have $(y_1,y_2) \in T_A$ or $(y_2,y_1) \in T_A$ for any $y_1 \in Y_1$ and $y_2 \in Y_2$ in view of $(y_1,y_2),(y_2,y_1) \in B - A$.

Vice versa, if A is T-consistent and we have $(y_1,y_2),(y_2,y_1) \notin T_A$ for some (and, hence, for all) $y_1 \in Y_1$ and $y_2 \in Y_2$, then B is T_A-consistent in view of $B \cap T_A^{-1} = (B^{-1} \cap T_A)^{-1} = (B \cap T_A)^{-1} = A \subseteq T_A$ and $T_B = (T_A \cup \{(y_1,y_2),(y_2,y_1)\})_t$ and, hence,

$$\begin{aligned} T_B \cap T_B^{-1} &= (T_A \cap T_A^{-1}) \cup \\ &\quad \{(y_1',y_2'),(y_2',y_1') \mid (y_1',y_1),(y_1,y_1'),(y_2,y_2'),(y_2',y_2) \in T_A\} \\ &= A \cup \{(y_1',y_2'),(y_2',y_1') \mid (y_1',y_1),(y_2,y_2') \in A\} \\ &= A \cup Y_1 \times Y_2 \cup Y_2 \times Y_1 = B. \end{aligned}$$

□ Other useful consequences of the above results are

Corollary 6 *Given X, T and A as above, A is a maximal T-consistent equivalence relation if and only if A is T-consistent and we have $T_A \cup T_A^{-1} = X^2$.*

Corollary 7 *If A_k $(k \in K)$ is a family of T-consistent equivalence relations so that $T_{A_{k_1}} \circ T_{A_{k_2}} \subseteq \bigcup_{k \in K} (T_{A_k} \cup T_{A_k}^{-1})$, then the following assertions are equivalent:*

(i) *$A_{k_1} \cup A_{k_2} \cup A_{k_3}$ is T-consistent for all $k_1, k_2, k_3 \in K$,*

(ii) *$A := \bigcup_{k \in K} A_k$ is a T-consistent equivalence relation,*

(iii) *$Q := \bigcup_{k \in K} T_{A_k}$ is transitive and one has $Q \cap Q^{-1} = A$.*

In particular, these three assertions are equivalent in case $X^2 = \bigcup_{k \in K} (T_{A_k} \cup T_{A_k}^{-1})$, in which case A is a maximal T-consistent equivalence relation.

Proof: (i) $\Rightarrow$ (iii): First, assume $(u,w),(w,v) \in Q$ and choose $k_1, k_2, k_3 \in K$ so that $(u,w) \in T_{A_{k_1}}$, $(w,v) \in T_{A_{k_2}}$ and $(u,v) \in T_{A_{k_3}} \cup T_{A_{k_3}}^{-1}$ according to our general assumption. Then Theorem 3, applied with respect to $B := (A_{k_1} \cup A_{k_2} \cup A_{k_3})_e$ and $A := A_{k_3}$ implies

$$(u,v) \in T_{A_{k_3}} \cup (T_{A_{k_1} \cup A_{k_2} \cup A_{k_3}} \cap T_{A_{k_3}}^{-1}) \subseteq T_{A_{k_3}} \cup A_{k_3} = T_{A_{k_3}},$$

so $\bigcup_{k \in K} T_{A_k}$ is indeed transitive.

Next, assume $(u,v),(v,u) \in Q$ and choose $k_1, k_2 \in K$ with $(u,v) \in T_{A_{k_1}}$ and $(v,u) \in T_{A_{k_2}}$. Then, as above, we have $(u,v) \in T_{A_{k_1} \cup A_{k_2}} \cap T_{A_{k_2}}^{-1} \subseteq T_{A_{k_2}}$ and, hence, $(u,v) \in A_{k_2} \subseteq A$.

So, we have indeed $Q \cap Q^{-1} \subseteq A$, while the opposite inclusion is trivial.

(iii) $\Rightarrow$ (ii): Clearly, the assumption $Q \cap Q^{-1} = A$ implies in particular that A is an equivalence relation, while the assumption that Q is transitive implies that the transitive relation T_A generated by $T \cup A$ coincides with Q. So, finally, we have

$$T_A \cap T_A^{-1} = A \text{ and } A \cap T^{-1} = \bigcup_{k \in K} (A_k \cap T^{-1}) \subseteq T,$$

so A is indeed a T-consistent equivalence relation.

(ii) $\Rightarrow$ (i): This is trivial. □

Next, consider as above a transitive and reflexive relation $T \subseteq X^2$ and an arbitrary binary relation $R \subseteq X^2$. Define the **support** $supp(R)$ of R by

$$supp(R) := \{x \in X \mid \text{ there exists } y \in X \text{ with } (x,y) \in R \cup R^{-1}\}$$

and, for any subset $Y \subseteq X$, define its **girth** relative to T by

$$girth_T(Y) := \max(\#Y' \mid Y' \subseteq Y \text{ and } (Y'^2) \cap T \subseteq D_X).$$

We claim

Theorem 4 *A relation R is T-consistent if and only if any relation $R' \subseteq R$ with $\#R' \leq girth_T(supp(R'))$ is T-consistent; in particular, R is T-consistent if and only if R' is T-consistent for all $R' \subseteq R$ with $\#R' \leq girth_T(supp(R))$.*

Proof: Clearly, if R is T-consistent, then so is R' for every subset $R' \subseteq R$.

Next, observe that R is T-consistent if and only if there is no pair (u,v) in $T' := (R \cup R^{-1} \cup T)_t$ with $(v,u) \in T \setminus T^{-1}$, that is, if and only if for every (cyclic) sequence $x_0, x_1, \ldots, x_k := x_0 \in X$ of elements from X with $(x_{i-1}, x_i) \in R \cup R^{-1} \cup T$ for all $i = 1, \ldots, k$, we have $(x_{i-1}, x_i) \in T^{-1}$ for all $i \in \{1, \ldots, k\}$ with $(x_{i-1}, x_i) \in T$. So, assuming that R is not T-consistent, we can find a shortest such sequence $x_0, x_1, \ldots, x_k \in X$ with $(x_{i-1}, x_i) \in R \cup R^{-1} \cup T$ for all $i = 1, \ldots, k$ which is *bad*, that is, for which, say, $(x_0, x_1) \in T \setminus T^{-1}$ holds. We claim that the union $R' \subseteq R$ of

$$\{(x_{i-1}, x_i) \mid i \in \{1, \ldots, k\} \text{ and } (x_{i-1}, x_i) \in R \setminus T\}$$

and

$$\{(x_i, x_{i-1}) \mid i \in \{1, \ldots, k\} \text{ and } (x_{i-1}, x_i) \in R^{-1} \setminus (R \cup T)\}$$

is a subrelation of R with

$$\#R' \leq girth_T(supp(R'))$$

which will surely establish Theorem 4.

This is evident in case $k = 2$ as this implies $\#R' = 1$. So, the minimality of k together with $(x_0, x_1) \in T \setminus T^{-1}$ implies in particular $(x_1, x_2) \notin T$ and $(x_{k-1}, x_k) = (x_{k-1}, x_0) \notin T$.

Next, put

$$J := \{i \in \{1, \ldots, k\} \mid (x_{i-1}, x_i) \notin T\} \subseteq \{2, \ldots, k\}$$

and

$$Y := \{x_i \mid i \in J\}.$$

Clearly, we have $\#R' \leq \#J$ and $Y \subseteq supp(R')$. So, it remains to show that we have

$$girth_T(Y) = \#J,$$

that is

$$(x_i, x_j) \notin T$$

for all $i, j \in J$ with $i \neq j$. So, we assume the opposite and derive a contradiction: Indeed, if $(x_i, x_j) \in T$ and $i < j$, we necessarily would have $i < j - 1$ (in view of $(x_{j-1}, x_j) \notin T$ by definition of J), so the sequence $x_0, x_1, \ldots, x_i, x_j, \ldots, x_k = x_0$ would be a shorter bad sequence. Next, if $(x_i, x_j) \in T$ and $j < i$ and, in addition, $(x_i, x_j) \notin T^{-1}$, then $y_0 := x_j, y_1 := x_{j+1}, \ldots, y_{i-j} := x_i, y_{i-j+1} := x_j$ would be a shorter bad sequence in view of $2 \leq j$ and, hence, $i - j + 1 \leq k - 2 + 1 = k - 1$.

And finally, if $(x_i, x_j) \in T, j < i$ and $(x_i, x_j) \in T^{-1}$, then we have $(x_{i'}, x_{j'}) \in T$ for $i' := j$ and $j' := i$ from J with $i' < j'$, so we can argue as in the first case.

Clearly, this establishes Theorem 4. □

A simple consequence is

Corollary 8 *Given a sequence family $\mathbf{s}_I$ and an equivalence relation A defined on $\mathcal{S} = \mathcal{S}(I)$. Then A is an alignment if and only if every subset A' of A of cardinality at most $\#I$ induces an alignment.*

Finally, assume as above that X is a set, that $T \subseteq X^2$ is a reflexive and transitive relation defined on X, that $A \subseteq X^2$ is a T-consistent equivalence relation defined on X and that $B \subseteq X^2$ is an arbitrary binary relation defined on X so that B is T-consistent.

Then we have:

Theorem 5 *The equivalence relation $A \cup B$ is T-consistent if and only if $A \cup B'$ is T-consistent for all subsets B' of B with $girth_T(B') \geq \#B' + 1$.*

Proof: As above, it is sufficient to show that in case $A \cup B$ is *not* T-consistent there exists a subset $B' \subseteq B$ with $girth_T(B') \geq \#B' + 1$ so that $A \cup B'$ is *not* T-consistent.

And also as above, we proceed – assuming that $A \cup B$ is not T-consistent – by choosing some shortest bad sequence, that is, a sequence $x_0, x_1, \ldots, x_k = x_0 \in X$ with

$$(x_{i-1}, x_i) \in A \cup B \cup B^{-1} \cup T$$

for all $i = 1, \ldots, k$ and $(x_{i-1}, x_i) \in T \setminus T^{-1}$ for at least one $i \in \{1, \ldots, k\}$. As we have assumed B to be T-consistent, it is clear that there must exist some $i \in \{1, \ldots, k\}$ with

$$(x_{i-1}, x_i) \in A \setminus (T \cup B \cup B^{-1})$$

and, without loss of generality, we may now assume that this holds for $i := 1$. As above, we consider the set

$$J := \{i \in \{1, \ldots, k\} \mid (x_{i-1}, x_i) \notin A \cup T\} \subseteq \{2, \ldots, k\},$$

we put

$$\begin{array}{lcl} j_0 & := & \min(J), \\ B' & := & \{(x_{i-1}, x_i) \mid i \in J \text{ and } (x_{i-1}, x_i) \in B\} \cup \{(x_i, x_{i-1}) \mid i \in J \text{ and } (x_{i-1}, x_i) \notin B\}, \end{array}$$

we note that $j_0 \geq 2$ and $B' \subseteq B$ must hold, we put

$$Y := \{x_i \mid i \in J\} \cup \{x_{j_0 - 1}\}$$

and we claim that, for $i, j \in J \cup \{j_0 - 1\}$, we have $(x_i, x_j) \in T$ if and only if $i = j$ holds which surely implies our Theorem as it implies

$$\#B' \leq \#J = girth_T(Y) - 1 \leq girth_T(supp(B')) - 1.$$

So, assume $i, j \in J \cup \{j_0 - 1\}, i \neq j$ and $(x_i, x_j) \in T$. We have to derive a contradiction. If $i < j$, we must have $i < j - 1$, so either the sequence $x_0, x_1, \ldots, x_i, x_j, \ldots, x_k = x_0$ would be a shorter bad sequence, or we have $(x_i, x_j) \in T^{-1}$ and $x_i, x_{i+1}, \ldots, x_j, x_i$ would be a shorter bad sequence as both are shorter and at least one of them must contain a pair of consecutive elements in $T \setminus T^{-1}$. If $j < i$, then we can't have $j := 1$ and $i := n$ in view of $(x_n, x_1) = (x_0, x_1) \in A \setminus T$, so either the sequence $x_j, x_{j+1}, \ldots, x_i, x_j$ would be a shorter bad sequence or we would have $(x_i, x_j) \in T \cap T^{-1}$, hence $j < i-1$ and, therefore, $x_0, x_1, \ldots, x_j, x_i, x_{i+1}, \ldots, x_k$ would be a shorter bad sequence. $\square$

Acknowledgement

The work of J. S. was supported by the German Research Council (DFG) graduate program (GK Strukturbildungsprozesse) and by the German Academic Exchange Service (DAAD).

References

Abdeddaïm, S. 1997 a, Incremental computaton of transitive closure and greedy alignment. *Proc. of 8-th Annual Symposium on Combinatorial Pattern Matching.* Lecture Notes in Computer Science 1264, 167 - 179.

Abdeddaïm, S. 1997 b, Improvement in incremental computaton of transitive closure and greedy alignment, In preparation.

Bourbaki, N. 1968. *Elements of mathematics, theory of sets.* Addison Wesley, Reading, MA, USA.

Chan, S.C., Wong, A.K.C., and Chiu, D.K.Y. 1992. A survey of multiple sequence comparison methods. *Bull. Math. Biol.* 54, 563-598.

Kececioglu, J. 1993. The maximum weight trace problem in multiple sequence alignment. In Apostolico, A., Crochemore, M., Galil, Z., and Manber, U. (editors), *Proc. of 4-th Annual Symposium on Combinatorial Pattern Matching.* Lecture Notes in Computer Science 684, 106-119.

Kruskal, J.B. 1983. An overview of sequence comparison. In Sankoff, D. and Kruskal, J.B. (editors), *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison*, pages 1-44. Addison-Wesley, Reading, MA, USA.

Miller, W. 1993. Building multiple alignments from pairwise alignments. *CABIOS* 9, 169-176.

Miller, W., Boguski, M., Raghavachari, B., Zhang, Z., and Hardison, R. 1994. Constructing aligned sequence blocks. *J. Comp. Biol.* 1, 51-64.

Morgenstern, B., Dress, A., and Werner, T. 1996. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. USA* 93, 12098-12103.

Myers, G., Selznick, S., Zhang, Z., and Miller, W., 1996. Progressive multiple alignment with constraints. *J. Comp. Biol.* 3, 563-572.

Needleman, S., and Wunsch, C. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443-453.

Smith, T., and Waterman, M. 1981. Comparison of biosequences. *Adv. Appl. Math.* 2, 482-489.

Stoye, J. 1997. Divide-and-conquer multiple sequence alignment. *Dissertation Thesis*, Technische Fakultät, Universität Bielefeld, Germany. Report 97-02.

Taylor, W.R., 1996. Multiple protein sequence alignment: Algorithms and gap insertion. In Doolittle, R.F. (editor), *Computer methods for macromolecular sequence analysis.* Methods in Enzymology 266, 343-367.

Vingron, M., and Argos, P. 1991. Motif recognition and alignment for many sequences by comparison of dot-matrices. *J. Mol. Biol.* 218, 33-43.

Vingron, M., and Pevzner, P. 1993. Multiple sequence comparison and n-dimensional image reconstruction. In Apostolico, A., Crochemore, M., Galil, Z., and Manber, U. (editors), *Proc. of 4th Annual Symposium on Combinatorial Pattern Matching.* Lecture Notes in Computer Science 684, 243-253.

Waterman, M. 1995. *Introduction to computational biology.* Chapman & Hall, London.

$$A_1 = \begin{pmatrix} a & b & c & d & d & d & d & d & - & - & f & g & h \\ a & b & c & - & - & - & - & - & e & e & f & g & h \end{pmatrix}$$

$$A_2 = \begin{pmatrix} a & b & c & - & - & d & d & d & d & d & f & g & h \\ a & b & c & e & e & - & - & - & - & - & f & g & h \end{pmatrix}$$

$$A_3 = \begin{pmatrix} a & b & c & d & d & - & - & d & d & d & f & g & h \\ a & b & c & - & - & e & e & - & - & - & f & g & h \end{pmatrix}$$

$$A_4 = \begin{pmatrix} a & b & c & d & d & d & d & d & f & g & h \\ a & b & c & - & - & e & e & - & f & g & h \end{pmatrix}$$

$$\vdots$$

Figure 1: Ambiguity of standard alignments. The standard alignment definition is ambiguous: Several matrices can express essentially one single way to align two sequences (A_1 – A_3; cf. Kruskal, 1983). In contrast, A_4 shows one of many further possible alignments. Clearly, all these alignments will have to be taken into account when searching for the best standard alignment.

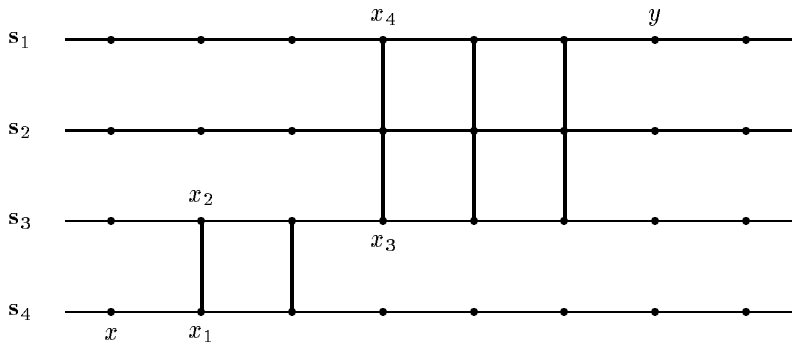

Figure 2: Quasi partial ordering induced by an alignment. An alignment A (indicated by vertical lines) extends the partial order '$\preceq$' given on the set $\mathcal{S}$ of all sites of a given sequence family to a quasi partial order '$\preceq_A$' which is the 'transitive closure' of the union $A \cup \preceq$: We have $x \preceq x_1$, $x_1 \overset{A}{\sim} x_2, x_2 \preceq x_3, x_3 \overset{A}{\sim} x_4, x_4 \preceq y$ and therefore $x \preceq_A y$.