11. Sankoff, D.: Matching Sequences Under Deletion/Insertion Constraints. *Proc. Nat. Acad. Sci. U. S. A.* **69** (1972) 4-6.
12. Sleator, D. D., Tarjan, R. E.: Self-Adjusting Binary Search Trees. *J. ACM* **32(3)** (1985) 652-686.
13. Wagner, R. A., Fischer, M. J.: The String-to-String Correction Problem. *J. ACM* **21(1)** (1974) 168-173.
14. Waterman, M. S.: General Methods of Sequence Comparison. *Bull. Math. Bio.* **46** (1984) 473-501.

Matrix Longest Common Subsequence Problem, Duality and Hilbert Bases

Pavel A. Pevzner, Michael S. Waterman

Department of Mathematics and of Molecular Biology
University of Southern California
Los Angeles, California, 90089-1113

Abstract. Although a number of efficient algorithms for the longest common subsequence (LCS) problem have been suggested since the 1970's, there is no duality theorem for the LCS problem. In the present paper a simple duality theorem is proved for the LCS problem and for a wide class of partial orders generalizing the notion of common subsequence. An algorithm for finding generalized LCS is suggested which has the classical dynamic programming algorithm as a special case. It is shown that the generalized LCS problem is closely associated with the minimal Hilbert basis problem. The Jeroslav-Schrijver characterization of minimal Hilbert bases gives an $O(n)$ estimation for the number of elementary edit operations for generalized LCS.

1 Introduction

Biological molecules can be represented as long strings of letters from a finite alphabet, 4 letters for DNA and 20 letters for proteins. Currently a large effort is being expended in the experimental determination of these genetic sequences from various organisms. Biologists ask which known sequences are evolutionary related to a newly determined sequence. The primary events in sequence evolution are *substitution*, when one letter is replaced by another, and *insertion* or *deletion* of a letter. These are the edit operations in the *minimum edit distance* problem. In this paper, we explore duality theorems and primal-dual algorithms for minimum edit distance problem. In particular we demonstrate that some advanced algorithms for the minimum edit distance problem are different implementations of the primal-dual algorithm.

The simplest and most often studied minimum edit distance problem in computer science is the *longest common subsequence* (LCS) problem, which is to find a longest subsequence common to two sequences. The LCS problem is equivalent to the problem of finding the minimum number of inserted or deleted letters to transform one sequence into the other. We give a matrix generalization of LCS, A-LCS, for 2×2 matrices A. Several alignment problems of biological interest are included in this family of A-LCS problems.

A-LCS is a path in a comparability graph for a partial order. The classical Needleman-Wunsch ([NW70]) dynamic programming algorithm decomposes each long arc in this graph into short arcs, thereby achieving its efficiency. We study similar reductions for A-LCS and apply the Jeroslav-Schrijver characterization of Hilbert bases to this problem. We give a geometric interpretation of elementary edit

operations for A-LCS and demonstrate that the number of elementary edit operations equals the size of the minimum Hilbert basis in the corresponding cone. A paper treating these and related topics in more detail will appear in [PW92].

2 Examples and Definitions

A *partially ordered set* or briefly a *poset* is a pair $(P, \prec)$ such that P is a set and $\prec$ is a transitive and irreflexive binary relation on P, i.e. $p \prec q$ and $q \prec r$ imply $p \prec r$. A *chain* is a subset of P where any two elements are comparable, and an *antichain* is a subset where no two elements are comparable. A *sequence* in a poset is an ordered chain $p_1 \prec p_2 \ldots \prec p_t$. Partial orders $\prec$ and $\prec^*$ are called *conjugate* ([KT82]) if for any two distinct $p_1, p_2 \epsilon P$ the following condition holds:

$$p_1 \text{ and } p_2 \text{ are } \prec\text{-comparable} \iff p_1 \text{ and } p_2 \text{ are } \prec^*\text{-incomparable}$$

Let w be an arbitrary non-negative integer valued function on P:

$$w : P \longrightarrow Z^+$$

For a partial order $\prec$, a sequence $p_1 p_2 \ldots p_k$ in P, maximizing

$$\sum_{i=1}^{k} w(p_i), \tag{1}$$

is called a *longest $\prec$-sequence*.

Let $I = \{1, 2 \ldots, n\}$ and $J = \{1, 2 \ldots, m\}$. As discussed in the introduction, our interest is in the comparison of two sequences $\mathbf{s} = s_1 s_2 \ldots s_n$ and $\mathbf{t} = t_1 t_2 \ldots t_m$. For this reason we study $P \subseteq I \times J$ (often $p = (i, j) \epsilon P$ denotes $s_i = t_j$). Let $p_1 = (i_1, j_1)$ and $p_2 = (i_2, j_2)$ be two arbitrary elements in $I \times J$. Denote

$$\Delta i = \Delta i(p_1, p_2) = i_2 - i_1,$$
$$\Delta j = \Delta j(p_1, p_2) = j_2 - j_1,$$
$$\Delta = \Delta(p_1, p_2) = (\Delta i, \Delta j)$$

Consider a few examples of partial orders on $I \times J$ (Fig.1).

(i) Common subsequences(CS): $p_1 \prec_1 p_2 \Leftrightarrow \Delta i > 0, \Delta j > 0$
(ii) Common forests(CF): $p_1 \prec_2 p_2 \Leftrightarrow \Delta i \geq 0, \Delta j \geq 0$
(iii) Common inverted subsequences(CIS): $p_1 \prec_3 p_2 \Leftrightarrow \Delta i > 0, \Delta j < 0$
(iv) Common inverted forests(CIF): $p_1 \prec_4 p_2 \Leftrightarrow \Delta i \geq 0, \Delta j \leq 0$

Partial orders $\prec_1$ and $\prec_3$ are particular cases of a partial order defined by an arbitrary 2×2 matrix $A = (a_{ij})$:

$$p_1 \prec_A p_2 \Leftrightarrow A\Delta^T > 0. \tag{2}$$

For $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ we have $\prec_1$, and for $A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ we have $\prec_3$. A partial order defined by A is called an *A-order*, and a sequence in A-order is called an *A-sequence*. Similarly, define *$\overline{A}$-order* and *$\overline{A}$-sequence* by the inequality

$$p_1 \prec_{\overline{A}} p_2 \Leftrightarrow A\Delta^T \geq 0. \tag{3}$$

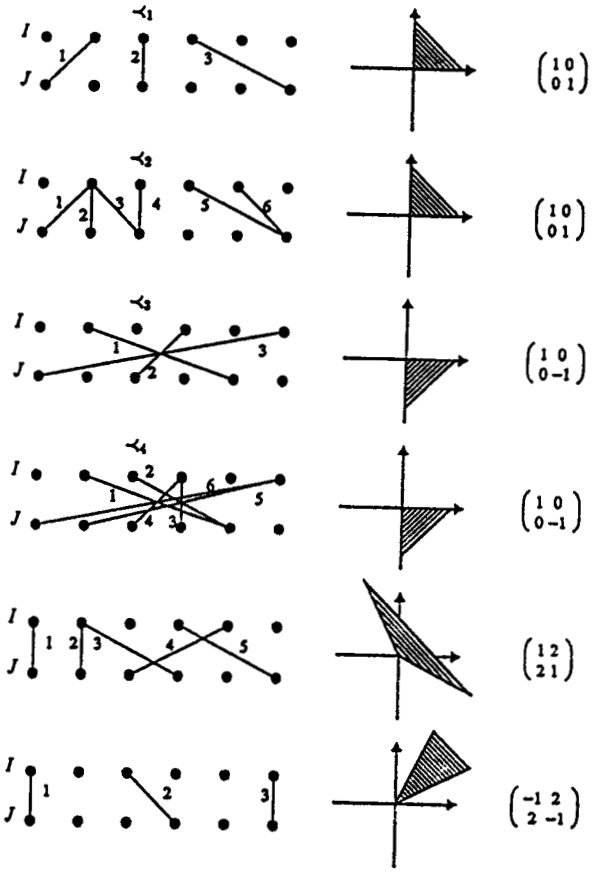

Fig. 1. Examples of sequences and corresponding cones for various partial orders.

The matrix A determines a *cone* in R^2 (Fig.1). The set of vectors Δ fulfilling (2) is designated *cone*(A), while the set of vectors Δ fulfilling (3) is denoted *cone*$(\overline{A})$. For partial order A, an A-sequence $p_1 p_2 \ldots p_k$ in P maximizing (1) is called a *longest common sequence* for A or A-LCS ($\overline{A}$-LCS is defined similarly). For $P = I \times J$, $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and w defined for sequences $\mathbf{s}$ and $\mathbf{t}$ according to the rule:

$$w(p) = w(i, j) = \begin{cases} 1, s_i = t_j \\ 0, \text{otherwise} \end{cases},$$

the problem (1) coincides with the longest common subsequence problem.

Let $\mathcal{C} = \{C\}$ be a family of subsets of a set P. $\mathcal{C}' \subseteq \mathcal{C}$ is called a *cover* of a function w if:

$\forall p \epsilon P$ there exist at least $w(p)$ subsets in family $\mathcal{C}'$ containing an element p

For $w \equiv 1$ on P, $\mathcal{C}'$ is a cover if and only if each $p \epsilon P$ is contained in at least one of subsets $C \epsilon \mathcal{C}'$. The number of elements in $\mathcal{C}'$ is called the *size* of the cover $\mathcal{C}'$ and a cover of minimum size is called a *minimum cover* of w by $\mathcal{C}$.

3 Duality for Longest $\prec$-sequence Problems

The following lemma is proven by an easy application of Dilworth's theorem [D50].

Lemma 1. *Let $\prec$ and $\prec^*$ be conjugate partial orders on P. Then the length of a longest $\prec$-sequence in P equals the size of a minimum cover of w by $\prec^*$-sequences.*

The binary relation on P defined by $p_1 \sqsubset p_2 \iff p_1 \prec p_2$ or $p_1 \prec^* p_2$ can easily be shown to be linear

Lemma 2. *$\sqsubset$ is a linear order on P.*

Let $\mathcal{P} = p_1 p_2 \dots p_l$ be an arbitrary sequence of the members of P, and $\mathcal{P}_i = p_1 p_2 \dots p_i$. Let $\mathcal{C}_i = \{C_1, C_2, \dots, C_j\}$ be a cover of $\mathcal{P}_i$ by $\prec^*$-sequences and let $p_1^{max}, p_2^{max}, \dots, p_j^{max}$ be the $\prec^*$-maximum elements in $C_1, C_2, \dots, C_j$ correspondingly. Consider an algorithm for constructing a cover $\mathcal{C}_{i+1}$ from $\mathcal{C}_i$.

Algorithm 1.

Let k be the minimum index $(1 \le k \le j)$ fulfilling the condition (Fig.2)

$$p_k^{max} \prec^* p_{i+1}, \tag{4}$$

and if the condition (4) fails for all k assume $k = j + 1$.

If $k < j + 1$, add p_{i+1} to C_k and define

$$\mathcal{C}_{i+1} = \{C_1, C_2, \dots, C_{k-1}, C_k \bigcup \{p_{i+1}\}, C_{k+1}, \dots, C_j\}.$$

If $k = j + 1$ add $\{p_{i+1}\}$ as a new $\prec^*$-sequence to the cover $\mathcal{C}_{i+1}$:

$$\mathcal{C}_{i+1} = \{C_1, C_2, \dots, C_j, C_{j+1} = \{p_{i+1}\}\}.$$

Define also a *reference* $ref(p)$ for p_{i+1} by

$$ref(p_{i+1}) = \begin{cases} p_{k-1}^{max} & \text{if } k > 1 \\ \emptyset & \text{otherwise} \end{cases}$$

□

Assume $\mathcal{C}_1$ consists only of the set $C_1 = \{p_1\}$ and $ref(p_1) = \emptyset$. Applying algorithm 1 $|P| - 1$ times, we will construct a cover $\mathcal{C}_l$ of P and a set of references $ref(p)$ for each $p \epsilon P$. The size of the cover $\mathcal{C}_l$ depends on the choice of the ordering of P. The following proposition shows that if $\mathcal{P}$ is the ordering of P in $\sqsubset$, then algorithm 1 gives a primal-dual algorithm for simultaneous solutions of (i) the longest $\prec$-sequence problem and (ii)the minimum $\prec^*$-cover problem (we suppose for simplicity that $w \equiv 1$).

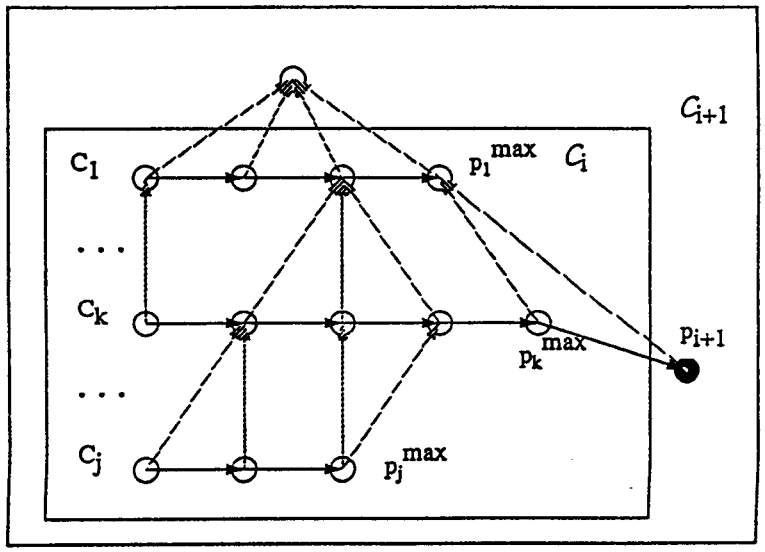

Fig. 2. Addition of p_{i+1} to the $\prec^*$-sequence C_k in the cover $\mathcal{C}_i = \{C_1, C_2, \dots, C_j\}$. The references $ref(p) = q$ correspond to the (dashed) arcs (p, q).

Proposition 3. *If $\mathcal{P} = p_1 p_2 \dots p_l$ is the ordering of P in $\sqsubset$, then algorithm 1 constructs a minimum cover $\mathcal{C}_l = \{C_1, C_2, \dots, C_t\}$ of P by $\prec^*$-sequences. A traceback of references ref(p) defines a longest $\prec$-sequence of length t for each $p \epsilon C_t$.*

Proof. We show that for each i $(1 \le i \le l)$ the cover $\mathcal{C}_i = \{C_1, C_2, \dots, C_j\}$ satisfies the condition

$$\forall k > 1, \forall p \epsilon C_k : \quad ref(p) \prec p. \tag{5}$$

Trivially this condition holds for $\mathcal{C}_1$. We suppose that it holds for $\mathcal{C}_i$ and prove it for $\mathcal{C}_{i+1}$. Consider two following cases:

1) $k < j + 1$ (see condition (4), algorithm 1).
   In this case $ref(p_{i+1}) = p_{k-1}^{max}$. Since $\mathcal{P}$ is the $\sqsubset$-ordering then $p_{k-1}^{max} \sqsubset p_{i+1}$ and therefore either $p_{k-1}^{max} \prec p_{i+1}$ or $p_{k-1}^{max} \prec^* p_{i+1}$. Since k is the minimum index fulfilling $p_k^{max} \prec^* p_{i+1}$, then $p_{k-1}^{max} \prec p_{i+1}$ and therefore condition (5) holds for $\mathcal{C}_{i+1}$.
2) $k = j + 1$.
   In this case $ref(p_{i+1}) = p_j^{max}$. Since $\mathcal{P}$ is the $\sqsubset$-ordering then either $p_j^{max} \prec p_{i+1}$ or $p_j^{max} \prec^* p_{i+1}$. Since $k = j + 1$, condition (4) fails for each $k \le j$. Therefore $p_j^{max} \prec p_{i+1}$ and condition (5) holds for $\mathcal{C}_{i+1}$.

Obviously each cover $\mathcal{C}_l = C_1 C_2 \dots C_t$ fulfilling condition (5) determines (through the traceback procedure) a $\prec$-sequence of length t for each $p \epsilon C_t$. According to lemma 1 each such sequence is a $\prec$-longest sequence, and $\mathcal{C}_l$ is a minimum cover of P by $\prec^*$-sequences. □

We remark that algorithm 1 is a modification of the maximum path algorithm for the *comparability graph* of partial order $\prec$. According to the Dushnik-Miller theorem ([DM41]), partial order $\prec$ has a conjugate partial order $\prec^*$ if and only if the *dimension* of $\prec$ is ≤ 2. The linear order $\sqsubset$ together with the linear order $\sqsubset'$ defined by the rule

$$p_1 \sqsubset' p_2 \Longleftrightarrow p_1 \prec p_2 \text{ or } p_2 \prec^* p_1$$

yield a 2-dimensional representation of the partial order $\prec$. Notice also that lemma 1 is closely related to 'minimal antichain cover-longest chain' version of Dilworth's theorem and Robacker's theorem ([Fulk71]) about 'maximum cut packing-minimum path length' in graphs.

4 Duality for A-LCS Problems

Lemma 4. *Let* $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ *be an arbitrary* 2×2 *matrix and* $A^* = \begin{pmatrix} a_{11} & a_{12} \\ -a_{21} & -a_{22} \end{pmatrix}$. *Then the partial orders* A *and* $\overline{A^*}$ *are conjugate.*

The next lemma is an immediate corollary of Lemmas 1 and 4:

Lemma 5. *Let* $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ *be a* 2×2 *matrix and* $A^* = \begin{pmatrix} a_{11} & a_{12} \\ -a_{21} & -a_{22} \end{pmatrix}$. *Then*

(i) *the length of A-LCS equals the size of a minimum cover of* w *by* $\overline{A^*}$ *-sequences, and*

(ii) *the length of* $\overline{A^*}$*-LCS equals the size of a minimum cover of* w *by A-sequences .*

Applying lemma 5 to the matrix $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ we derive the following theorem:

Theorem 6. .

- *The length of a longest CS equals the size of a minimum cover by CIF.*
- *The length of a longest CF equals the size of a minimum cover by CIS.*
- *The length of a longest CIS equals the size of a minimum cover by CF.*
- *The length of a longest CIF equals the size of a minimum cover by CS.*

For the LCS problem and a fixed length alphabet, algorithm 1 can be implemented in $O(nL)$ time where L is the length of longest common sequence or in $O((r+n)\log n)$ time where r is the total number of matches between the two input sequences. These improvements of the Needleman-Wunsch algorithm have been suggested by Hirschberg ([H77]) and Hunt and Szymanski ([HS77]). $\prec^*$-chains in algorithm 1 correspond to the *k-candidates* in Hirschberg's algorithm. Maximal elements of $\prec^*$-chains in algorithm 1 correspond to the *dominant matches* in Apostolico's improvement ([A86]) of Hunt-Szymanski's algorithm. Further improvements of algorithm 1 for the LCS problem can be found in ([A86], [AG87], [EGGI90]). The relationships between algorithm 1, advanced LCS algorithms and Robinson-Schensted-Knuth algorithm for Young tableaux ([Saga91]) will be considered in detail elsewhere.

5 Maximum Paths in Graphs and A-LCS Problem

A longest $\prec$-sequence problem can be reformulated as a 'maximum path' problem in the weighted directed graph $G(P, E, w)$, where E and w are defined by the rule

$$(p_1, p_2) \epsilon E \Longleftrightarrow p_1 \prec p_2$$

$$w(p_1, p_2) = w(p_2).$$

Sankoff [S72] first proposed an $O(n^2)$ algorithm for the LCS-problem in computer science, but a few authors developed closely related algorithms even earlier in molecular biology and speech processing. As a matter of fact the contribution of these authors is concerned with transformations of $G(P, E)$ to reduce computational complexity. They increased $|P|$ with a simultaneous significant decrease of $|E|$ by 'decomposition' of each 'long' arc into short arcs.

The classical Needleman-Wunsch algorithm for the LCS problem has running time $O(n^2)$ due to the special *arrangement* (systolic schedule, [CPHW91]) of vertices of G. Arranging vertices allows implementation of the maximum path algorithm for acyclic graphs in $O(|E|)$ time and gives $O(n^2)$ running time for the LCS-problem. Unfortunately the Needleman-Wunsch transformation of $G(P, E)$ into the graph with $O(n^2)$ arcs is not valid for an arbitrary A-order. Below we describe a transformation of $G(P, E)$ that decreases the number of arcs significantly, where Needleman-Wunsch transformation is a special case.

6 Algorithms for A-LCS Problems

Consider the A-LCS problem and let L_1 and L_2 be the lines $a_{11}x + a_{12}y = 0$ and $a_{21}x + a_{22}y = 0$, respectively. Without loss of generality below we suppose that $r = (i_1, j_1)$ and $s = (i_2, j_2)$ are the first integer points on the lines L_1 and L_2 fulfilling the conditions

$$|x| \leq n, |y| \leq m, \tag{6}$$

and $A = \begin{pmatrix} j_1 & -i_1 \\ -j_2 & i_2 \end{pmatrix}$.

Let $v_1, v_2 \ldots, v_k$ be the set V of all non-zero integer vectors (or vertices) of the parallelogram Π, defined by points $0, r, s, r+s$. The number of elements of V equals $\|V\| = |i_1 j_2 - i_2 j_1| + 2 = |det(A)| + 2$

Consider a graph $G^*(I \times J, E^*)$ with vertex set $I \times J$ and arc set E^* determined by V

$$(p_1, p_2) \epsilon E^* \Longleftrightarrow (p_2 - p_1) \epsilon V \tag{7}$$

Define weighting functions w and $\overline{w}$ on E^* according to the rule:

$$w(p_1, p_2) = \begin{cases} w(p_2), & \text{if } p_2 \epsilon P, \text{ and } (p_2 - p_1) \neq r, s \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$\overline{w}(p_1, p_2) = \begin{cases} w(p_2), & \text{if } p_2 \epsilon P \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

Theorems 7 and 8 below reduce A-LCS and $\overline{A}$-LCS problems to longest path problems.

Theorem 7. *The length of a $\overline{A}$-LCS coincides with the length of a $\overline{w}$- longest path in G^*.*

Proof. Obviously, each path $p_1, p_2, ..., p_t$ in $G^*(I \times J, E^*, \overline{w})$ corresponds to an $\overline{A}$-sequence of the same length, since the vertices of this path with $w(p_k) > 0$ correspond to elements of an $\overline{A}$-sequence. To prove the theorem, it is sufficient to prove that each $\overline{A}$-sequence $p_1, p_2, ..., p_t$ has a corresponding path in the G^*-graph of at least the same length.

Let p_{k-1}, p_k be two arbitrary sequential elements of an $\overline{A}$-sequence. Since $\Delta = \Delta_k(p_{k-1}, p_k)$ belongs to $cone(\overline{A})$, then

$$\Delta = xr + ys = \lfloor x \rfloor r + \lfloor y \rfloor s + \langle x \rangle r + \langle y \rangle s$$

($\lfloor x \rfloor$ is the integer part of x and $\langle x \rangle$ is the fractional part of x). The vector $p = \langle x \rangle r + \langle y \rangle s$ belongs to the parallelogram Π and is integer ($p = \Delta - \lfloor x \rfloor r + \lfloor y \rfloor s$). Therefore Δ is decomposed as a sum of $\lfloor x \rfloor + \lfloor y \rfloor + 1$ (or $\lfloor x \rfloor + \lfloor y \rfloor$ if $p = 0$) vectors defined by vertices from V. The decomposition for each pair p_{k-1}, p_k determines a path in G^* that visits vertices $p_1, p_2, ..., p_t$ and therefore has at least the same length as the $\overline{A}$-sequence $p_1 p_2 \dots p_t$. □

Theorem 8. *The length of an A-LCS coincides with the length of a w-longest path in G^*.*

Proof. Let $\mathcal{P} = p_1 p_2 \dots p_t$ be an arbitrary path in G^* . Consider a subsequence of $\mathcal{P}$ defined by vertices: $\mathcal{P}' = \{p_k : p_k \epsilon P, (p_k - p_{k-1}) \neq r, s\}$. Observe that $\mathcal{P}'$ is an A-sequence and according to (8) the length of this A-sequence coincides with the w-length of P:

$$\sum_{k=2}^{t} w(p_{k-1}, p_k) = \sum_{p_k \epsilon \mathcal{P}'} w(p_{k-1}, p_k) = \sum_{p_k \epsilon \mathcal{P}'} w(p_k).$$

To prove the theorem, it is sufficient to prove that each A-sequence $p_1, p_2, ..., p_t$ has a corresponding path in graph G^* with the same w-length. Let p_{k-1}, p_k be two arbitrary sequential elements of the A-sequence. As was proved in Theorem 7, $\Delta = \Delta_k(p_{k-1}, p_k) = \lfloor x \rfloor r + \lfloor y \rfloor s + p$. If $p \neq 0$, define $\mathcal{P}_k$ to be the path consisting of $\lfloor x \rfloor$ arcs r, $\lfloor y \rfloor$ arcs s and ending with p. According to (8) only the last arc of this path has positive weight, equal to $w(p_k)$. If $p = 0$, then $\lfloor x \rfloor > 0$, $\lfloor y \rfloor > 0$ since otherwise p_{k-1} and p_k would be incomparable. Let $\mathcal{P}_k$ be the path consisting of $\lfloor x \rfloor - 1$ arcs r, $\lfloor y \rfloor - 1$ arcs s and the arc $r + s$ at the end. According to (8) only the last arc of this path has positive weight, equal to $w(p_k)$. Thus each pair p_{k-1}, p_k determines a path $\mathcal{P}_k$ in $G^*(I \times J, E^*, w)$, and the only last arc of this path has positive weight $w(p_k)$. Therefore the length of the path $\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_t$ equals the length of the A-sequence $p_1, p_2, \dots, p_t$. □

7 A-LCS Problems and Hilbert Bases

According to theorems 7 and 8, finding longest A and $\overline{A}$ sequences requires about kn^2 operations, where k is the maximum vertex degree in G^* ($k = |det(A)| + 2$). For

the classical LCS problem, and for the longest CF, CIS, CIF problems, det(A)=1 and $k = 3$ (Fig.3a,b) as in the usual Needleman-Wunsch algorithm. For $A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ and $A = \begin{pmatrix} -1 & 2 \\ 2 & -1 \end{pmatrix}$, $det(A) = -3$ and $k = 5$ (Fig.3c,d), but we can further decrease k as some points in Π are non-negative integer combinations of others.

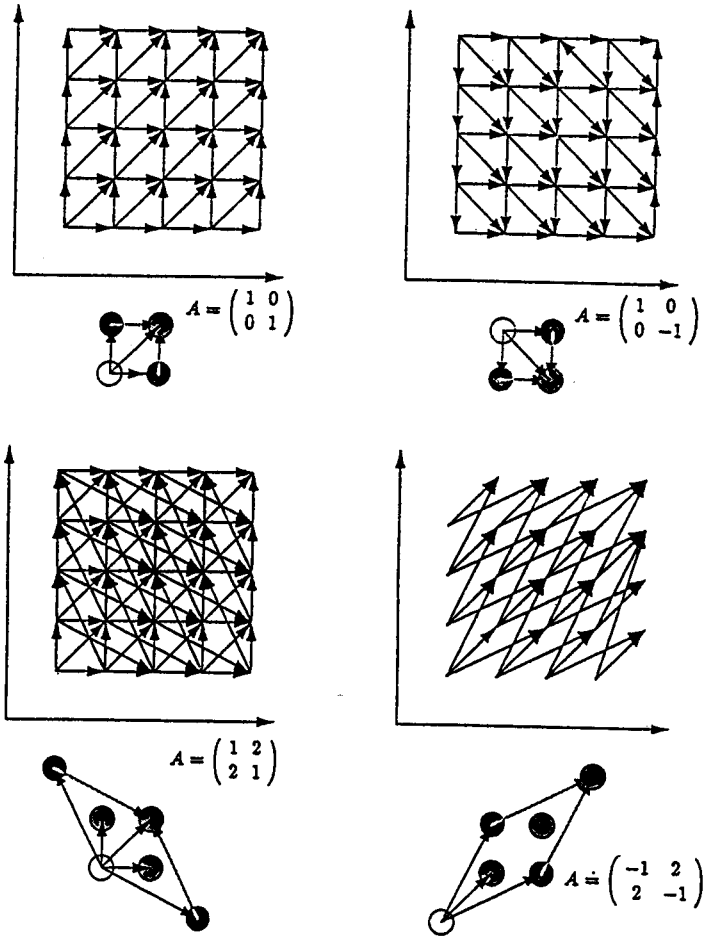

Fig. 3. Examples of parallelograms Π for various matrices A and the corresponding graphs G^*..The dark nodes in Π correspond to r and s.

We describe a procedure to eliminate arcs in the graphs G^* for A- and $\overline{A}$-LCS problems. A set H of integer vectors in the cone $K = \{x : Ax \geq 0\}$ is called a *Hilbert*

basis of K if each integer vector in K is a non-negative integer linear combination of vectors in H. Let H be a Hilbert basis of the cone $\{x : Ax \geq 0\}$, and let $H(A)$ be the intersection of H and Π, where Π is the parallelogram defined by A.

Observe that theorem 7 still holds even if we define E^* by

$$(p_1, p_2) \epsilon E^* \Longleftrightarrow (p_2 - p_1) \epsilon H(A)$$

instead of by (7). Similarly theorem 8 still holds even if we define E^* by

$$(p_1, p_2) \epsilon E^* \Longleftrightarrow (p_2 - p_1) \epsilon H(A) \bigcup \{r + s\}$$

instead of by (7). These observations allow further transformations of G^* excluding arcs which do not belong to the Hilbert basis.

For example when $A = \begin{pmatrix} -1 & 2 \\ 2 & -1 \end{pmatrix}$, $H(A) = \{(1,2), (1,1), (2,1)\}$. Therefore we can reduce the maximum degree of G^* for this matrix to 3. Unfortunately we can't guarantee that $H(A)$ contains $O(1)$ integer points for an arbitrary A-matrix. Some interesting questions arise:

1. Given A, find $H(A)$ of minimal size.
2. Find the maximum size of a minimal Hilbert basis $H(A)$ for matrix $A = (a_{ij})$ with $|a_{ij}| \leq n$.

Fortunately using the following characterization of Hilbert bases ([S81]), we can find $H(A)$ in $O(n)$ time and prove that the maximum size of minimal Hilbert basis of $H(A)$ is $O(n)$.

Theorem 9. *A set of all integer vectors in K which are not non-negative integer linear combinations of other integer vectors in K is the minimal Hilbert basis in K.*

The theorem implies a sufficient condition for a point (i,j) to lie outside the minimal Hilbert basis of the plane lattice Π : if all points in 1-neighbourhood of (i, j) (the set $\{(i', j') : |i' - i| \leq 1, |j' - j| \leq 1\}$) belong to Π, then (i, j) does not belong to the minimal Hilbert basis of Π. Therefore (i, j) can belong to the minimal Hilbert basis only if its 1-neighbourhood intersects the boundary of Π. This implies that cardinality of the minimal Hilbert basis of Π is at most $O(n + m)$ and that it is easy to find the minimal Hilbert basis in $O(n)$ time. Therefore the number of arcs in G can be reduced to $O(n^3)$, and this yields $O(n^3)$ A- and $\overline{A}$-LCS algorithms for an arbitrary A-matrix.

The LCS problem is often discussed in the terms of two elementary edit operations insertions and deletions. Generalized LCS problems require at least $|H(A)|$ elementary edit operations; each operation corresponds to a vector from the Hilbert basis. It is worth noting that, although the number of elementary edit operations can be as large as $O(n)$, an integer analog of the Caratheodory theorem ([CFS86]) implies that each arc between comparable elements can be decomposed as a sum of only 4 elementary edit operations.

For arbitrary 2×2 non-singular matrices A and B each A-LCS problem can be reduced to a B-LCS by the transformation $\begin{pmatrix} i' \\ j' \end{pmatrix} = B^{-1} \cdot A \begin{pmatrix} i \\ j \end{pmatrix}$. For arbitrary

A and $B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ this transformation reduces the A-LCS problem for n-letter words to the classical LCS problem for $O(n^2)$ words and yields a $O(n^4)$ algorithm for the A-LCS problem. Nevertheless, the Hunt-Szymanski algorithm applied after such transformation yields A-LCS algorithm with running time at most $O(n^2 \log n)$.

8 Acknowledgements

The research was supported in part by the National Science Foundation (DMS 90-05833) and the National Institute of Health (GM-36230). We are grateful to Gian-Carlo Rota, Anatoly Rubinov and Martin Vingron for helpful discussions as well as Andras Sebo and Alexander Vainshtein for useful comments on Hilbert bases.

References

[A86] Apostolico A.: Improving the worst-case perfomance of the Hunt-Szymanski strategy for the longest common subsequence of two strings. Inform. Process. Lett. **23** (1986) 63-69

[AG87] Apostolico A., Guerra C.: The longest common subsequence problem revisited. Algoritmica **2**(1987) 315-336

[CPHW91] Chow E.T., Hunkapiller T., Peterson J.C., Zimmerman B.A., Waterman M.S.: A systolic array processor for biological information signal processing. Proc. of International Conference on Supercomputing (ICS-91) June 17-21,1991 (to appear)

[CFS86] Cook W.,Fonlupt J.,Schrijver A.: An integer analogue of Caratheodory's theorem. J. of Combinatorial Theory (B) **40** (1986) 63-70

[D50] Dilworth R.P.: A decomposition theorem for partially ordered sets Ann. Math. **51** (1950) 161-165

[DM41] Dushnik B., Miller E.W.: Partially ordered sets. Am. J. Math. **63** (1941) 600-610

[EGGI90] Eppstein D., Galil Z., Giancarlo R., Italiano G. F. Sparse dynamic programming; Extended Abstract *Proc. first ACM-SIAM SODA* (1990) 513-522

[Fulk71] Fulkerson D.R.: Blocking and antiblocking polyhedra. Mathematical programming. **1** (1971) 168-194

[H77] Hirscberg D.S. Algorithms for the longest common subsequence problem. J. ACM **24** (1977) 664-675

[HS77] Hunt J.W., Szymanski T.G.: A fast algorithm for computing longest common subsequences. Comm. ACM **20** (1977) 350-353

[KT82] Kelly D., Trotter W.T.: Dimension theory for ordered sets. In I.Rival (ed.) Ordered sets Reidel, Dordrecht/Boston (1982)

[NW70] Needleman S.B., Wunsch C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J.Mol.Biol. **48** (1970) 443-453

[PW92] Pevzner P., Waterman M.: Generalized sequence alignment and duality. Adv. in Appl. Math. (1992) (in press)

[Saga91] Sagan B.E.: The symmetric group. Representations, combinatorial algorithms and symmetric functions. Wadsworth and Brooks/Cole (1991)

[S72] Sankoff D.: Matching sequences under deletion-insertion constraints. Proc. Nat. Acad. Sci. USA **69** (1972) 4-6

[S81] Schrijver A.: On total dual integrality. Linear algebra and its applications. **38** (1981) 27-32