Efficient reconstruction of sequences from their subsequences or supersequences

Vladimir I. Levenshtein *
Keldysh Institute for Applied Mathematics, RAS,
125047 Moscow, Miusskaya Sq. 4, Russia

Abstract

In the paper two combinatorial problems for the set F_q^n of sequences of length n over the alphabet $F_q = \{0, 1, ..., q-1\}$ are considered. The maximum size $N_q^-(n,t)$ of the set of common subsequences of length $n-t$ and the maximum size $N_q^+(n,t)$ of the set of common supersequences of length $n+t$ of two different sequences of F_q^n are found for any nonnegative integers n and t. The number $N_q^-(n,t)+1$ (respectively, $N_q^+(n,t)+1$) is equal to the minimum number N of different subsequences of length $n-t$ (supersequences of length $n+t$) of an unknown sequence $X \in F_q^n$ which are sufficient for its reconstruction. Simple algorithms to recover $X \in F_q^n$ from $N_q^-(n,t)+1$ of its subsequences of length $n-t$ and from $N_q^+(n,t)+1$ of its supersequences of length $n+t$ are given.

1 Introduction

We denote by F_q^n the set of sequences $X = (x_1, ..., x_n)$ over the alphabet $F_q = \{0, 1, \ldots, q-1\}$, $q \geq 2$. We shall also use the multiplicative writing $X = x_1 \cdots x_n$ considering X as a *word* of length n over the alphabet F_q. Let $F_q^* = \bigcup_{n=0}^{\infty} F_q^n$ be the set of all words over F_q including the *empty word* of length 0. For any m, $m = 0, 1, ..., n$, we call $Y = (x_{i_1}, ..., x_{i_m})$, where $1 \leq i_1 < \cdots < i_m \leq n$, a *subsequence* of $X = (x_1, ..., x_n)$ of length m and call X a *supersequence* of Y of length n. One can consider that Y is obtained from X by *deletions* of $t = n - m$ letters whose place numbers differ from $i_1, ..., i_m$ and that X is obtained from Y by *insertions* of $t = n - m$ letters between (and also before and after) the letters of Y. It should be noted that the same subsequence or supersequence

*This research was supported by the Russian Foundation for Basic Research (grant 98-01-00146).

can be obtained many times. For instance, there are eight possibilities to delete three letters in $X = 0110110$ in order to obtain its subsequence $Y = 0110$. For any $X \in F_q^*$ and integer t, $t \geq 0$, denote by $D_t(X)$ the set of all different words obtained from X by deletions of t letters and by $I_t(X)$ the set of all different words obtained from X by insertions of t letters. In other words, if X has length n, then $D_t(X)$ is the set of all of its subsequences of length $n - t$ and $I_t(X)$ is the set of all of its supersequences of length $n + t$. (We shall assume that $D_t(X)$ is empty if $t > n$ or $t < 0$.)

We deal with problems of reconstruction of an unknown $X = (x_1, \ldots, x_n) \in F_q^n$ with the help of the minimum number of its different subsequences or supersequences of a given length. A typical question is whether one can recover an unknown sequence $X \in F_3^7$ if one knows 25 of its different subsequences of length four written as columns of the following matrix:

$$\begin{array}{ccccccccccccccccccccccccc}
2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 0 & 2 & 2 & 2 & 0 \\
1 & 1 & 1 & 2 & 2 & 2 & 0 & 0 & 2 & 2 & 0 & 0 & 0 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 2 & 2 \\
0 & 1 & 2 & 0 & 1 & 2 & 1 & 2 & 0 & 1 & 1 & 2 & 1 & 0 & 1 & 2 & 1 & 2 & 1 & 2 & 1 & 1 & 2 & 1 & 1
\end{array}$$

What is a procedure to recover this X if it is possible? In the general form this gives rise to two combinatorial problems to find, for any n, t, and q,

$$\max_{X,Z \in F_q^n;\ X \neq Z} |D_t(X) \cap D_t(Z)| \tag{1}$$

(i.e., the maximum number of common subsequences of two different sequences) and

$$\max_{X,Z \in F_q^n;\ X \neq Z} |I_t(X) \cap I_t(Z)| \tag{2}$$

(i.e., the maximum number of common supersequences of two different sequences). It is clear that the number (1) increased by 1 is the minimum number N such that an arbitrary $X = (x_1, \ldots, x_n) \in F_q^n$ can be reconstructed if one knows N of its different subsequences of length $n - t$, and that the number (2) increased by 1 is the minimum number N such that an arbitrary $X = (x_1, \ldots, x_n) \in F_q^n$ can be reconstructed if one knows N of its different supersequences of length $n + t$.

In the paper we introduce two functions $N_q^-(n,t)$ and $N_q^+(n,t)$ and prove that they are equal to (1) and (2), respectively, for all n, t, and q. Moreover, we describe simple algorithms to recover $X \in F_q^n$ using any $N_q^-(n,t) + 1$ different elements of $D_t(X)$ (if they exist) or any $N_q^+(n,t) + 1$ different elements of $I_t(X)$ (we verify that such a number of elements in $I_t(X)$ exist for any X).

The paper is organized as follows. In Section 2 we describe the known results on the number of subsequences and supersequences of a sequence and present main arguments which can be used to prove these results and new statements as well. In Section 3 we define $N_q^-(n,t)$, show that $N_q^-(n,t)$ is a lower bound on (1), and find some properties of this function. In Section 4 we describe and substantiate an algorithm to recover $X \in F_q^n$ if one knows $N_q^-(n,t)+1$ of its different subsequences of length $n-t$. This algorithm implies a proof of the fact that $N_q^-(n,t)$ is an upper bound on (1). We illustrate this algorithm by the foregoing example. In Section 5 we define $N_q^+(n,t)$ and show that $N_q^+(n,t)$ is equal to (2). In Section 6 we describe and ground an algorithm to recover $X \in F_q^n$ if one knows $N_q^+(n,t)+1$ of its different supersequences of length $n+t$. In Section 7 we discuss likeness and difference between the problems to reconstruct a sequence from its subsequences or supersequences and investigate their relation with some known and new combinatorial problems.

These results on reconstruction of sequences from their subsequences or supersequences are given without proofs in [12] and [13] together with solutions to other combinatorial and theoretico-information problems of effective reconstruction of sequences.

Now we present some definitions and notations which are used below. Let $U \subseteq F_q^m$ and $a \in F_q$. We denote by aU the set of words $\{aY : Y \in U\}$. For any $i = 1, 2, ..., m$, we denote by $U^{a,i}$ the subset of all words $Y = y_1 \cdots y_m \in U$ such that $y_i = a$ is the first occurence of the letter a in the word Y and denote by $\overline{U^{a,i}}$ the set of all different words of length $m-i$ which are obtained from words of $U^{a,i}$ by deleting the first i letters. For instance, if $U = \{0100, 2100, 1120, 1210\} \subset F_3^4$, then and $U^{1,2} = \{0100, 2100\}$ and $\overline{U^{1,2}} = \{00\}$. We shall omit i in $U^{a,i}$ and $\overline{U^{a,i}}$ when $i = 1$. Moreover, if $U = D_t(X)$ or $U = I_t(X)$ we shall write $D_t^a(X)$ and $I_t^a(X)$ instead of U^a.

For any $Z \in F_q^m$, we call the vector

$$k(Z) = (k_0(Z), k_1(Z), ..., k_{q-1}(Z)),$$

where $k_i(Z)$ is the number of letters $i \in F_q$ in the word Z, the *composition* of Z. There exists a unique permutation $\theta = (\theta_0, \theta_1, ..., \theta_{q-1})$ of $F_q = \{0, 1, ..., q-1\}$ such that

$$k_{\theta_0}(Z) \geq k_{\theta_1}(Z) \geq \cdots \geq k_{\theta_{q-1}}(Z) \tag{3}$$

and $\theta_i < \theta_j$ when $k_{\theta_i}(Z) = k_{\theta_j}(Z)$. This permutation is called the *ordering permutation* for Z and vector

$$l(Z) = (k_{\theta_0}(Z), k_{\theta_1}(Z), ..., k_{\theta_{q-1}}(Z)), \tag{4}$$

is called the *ordered composition* of Z. In particular, for the word $Z = 0211012 \in F_3^7$, $k(Z) = (2, 3, 2)$ is the composition, $\theta = (1, 0, 2)$ is the ordering permutation for Z, and $l(Z) = (3, 2, 2)$ is the ordered composition.

For description of algorithms for reconstruction of sequences we introduce some functions of the threshold type. Let R^h be the set of all real vectors of length h. (In fact, we consider only vectors of R^h whose coordinates are nonnegative integers, and use $h = q$ for reconstruction from subsequences and $h = t + 1$ for reconstruction from supersequences.) For any vector $\tau = (\tau_0, \tau_1, ..., \tau_{h-1}) \in R^h$ consider the *generalized threshold function*

$$f_{\tau_0,\tau_1,...,\tau_{h-1}} : R^h \to F_h,$$

which is defined on all vectors $w = (w_0, w_1, .., w_{h-1}) \in R^h$ with

$$\sum_{i=0}^{h-1} w_i > \sum_{i=0}^{h-1} \tau_i \tag{5}$$

and is equal to the minimum number $i \in F_h$ such that $w_i > \tau_i$. The numbers $\tau_0, \tau_1, ..., \tau_{h-1}$ are called *thresholds.* This definition is correct, since the condition (5) guarantees that such a number exists.

We tacitly assume that all numbers under consideration are integers and

$$\sum_{i=l}^{j} a_i = 0, \text{ if } j < l, \text{ and } \binom{n}{i} = 0, \text{ if } i < 0 \text{ or } i > n. \tag{6}$$

2 The number of subsequences and supersequences of a sequence

In this section we describe some known results on the sets $D_t(X)$ and $I_t(X)$. We first mention the main properties of these sets for any $X = x_1 \cdots x_n \in F_q^n$:

$$D_t(X) \subseteq D_{t+1}(Y) \text{ for any } Y \in I_1(X), \tag{7}$$

$$D_t^a(X) = aD_{t-i+1}(x_{i+1} \cdots x_n) \tag{8}$$

if $x_i = a$ is the first occurrence of a letter $a \in F_q$ in the word X (i.e., $X \in (F_q^n)^{a,i}$ in our notations), and

$$I_t(X) = x_1 I_t(x_2 \cdots x_n) \bigcup \bigcup_{b \in F_q \setminus \{x_1\}} bI_{t-1}(x_1 \cdots x_n). \tag{9}$$

If $\theta = (\theta_0, \theta_1, ..., \theta_{q-1})$ is a permutation of $F_q = \{0, 1, ..., q - 1\}$ and $\theta(X)$ is the word obtained from X replacing any letter $i \in F_q$ by $\theta_i \in F_q$, $i = 0, 1, ..., q - 1$, then

$$|D_t(\theta(X))| = |D_t(X)| \text{ and } |I_t(\theta(X))| = |I_t(X)|. \tag{10}$$

It is turned out that the size of $D_t(X)$ essentially depends on the number $r(X)$ of *runs*, which are maximal intervals of the word X consisting of the same letters. For instance, $r(X) = 4$ for $X = 0220111$. It is known [9] that for any $X \in F_q^n$,

$$\binom{r(X)-t+1}{t} \le |D_t(X)| \le \binom{r(X)+t-1}{t} \tag{11}$$

and, in particular, $|D_1(X)| = r(X)$. The upper bound follows from the fact that any word of $D_t(X)$ is uniquely defined by the numbers of letters deleted in every run. The lower bound is the number of words of $D_t(X)$ which are obtained by deleting at most one letter in every two successive runs. This proof is valid for any $q \ge 2$. Since $|D_t(X)|$ depends on X, a natural problem is to find

$$D_q(n,t) = \max_{X \in F_q^n} |D_t(X)|. \tag{12}$$

It is useful to assume that $D_q(n,t) = 0$ when $n < t$ or $t < 0$. Note that from (7) and (8) it follows that

$$D_q(n,t) \le D_q(n+1,t+1), \tag{13}$$

and if $y_i = a$ is the first occurence of a letter $a \in F_q$ in the word $Y = y_1 \cdots y_m \in F_q^m$, then

$$|D_t^a(Y)| \le D_q(m-i, t-i+1). \tag{14}$$

The problem to find (12) was solved by Calabi [2] (see also [3]) with the help of the following words of length n with n runs:

$$A_{n,q} = a_1 \cdots a_n \quad \text{where } a_i = b \in F_q \text{ if } i-1 \equiv b \pmod{q}.$$

For instance, $A_{5,3} = 01201$. Calabi proved that

$$D_q(n,t) = |D_t(A_{n,q})| \tag{15}$$

and found that for any integer m, $m = 0, 1, \dots$,

$$\left(\sum_{j=0}^{q-1} z^j\right)^m \sum_{i=0}^{\infty} z^i = \sum_{h=0}^{\infty} D_q(m+h,h) z^h. \tag{16}$$

Since $D_q(t,t) = 1$, (16) is equivalent to the equality

$$D_q(n,t) = \sum_{i=0}^{q-1} D_q(n-i-1, t-i) \text{ when } n > t \ge 0. \tag{17}$$

(This also holds for negative t by our assumption.) Note that from (8) and (12)-(14) it follows that for any $X \in F_q^n$ and t, $n > t \geq 0$,

$$|D_t(X)| \leq \sum_{i=0}^{q-1} D_q(n-i-1, t-i).$$

with equality for $X = A_{n,q}$ (the latter is proved by induction on $n+t$ using the fact that any suffix Y of length m of $A_{n,q}$ coincides with $\theta(A_{m,q})$ for a (cyclic) permutation θ of F_q and hence $|D_t(Y)| = |D_t(\theta(A_{m,q}))| = |D_t(A_{m,q})|$). This gives a simple proof of (15) and (17). As a special case we have equality in (14) for any $a \in F_q$ and $Y = \theta(A_{m,q})$, where θ is a permutation of F_q. We shall also use below that for $n \geq t$,

$$D_q(n+1, t) = D_q(n, t) + D_q(n, t-1) - D_q(n-q, t-q), \tag{18}$$

which follows from (17).

Recently Hirschberg [6] reopened the Calabi results and found the following recurence on q, $q \geq 2$, for computing $D_q(n,t)$:

$$D_q(n, t) = \sum_{i=0}^{t} \binom{n-t}{i} D_{q-1}(t, t-i) \tag{19}$$

where $D_1(n,t) = 1$, if $n \geq t \geq 0$, and $D_1(n,t) = 0$, otherwise. In particular, this gives the known result (see [3])

$$D_2(n, t) = \sum_{i=0}^{t} \binom{n-t}{i} \tag{20}$$

and shows that

$$D_3(n, t) = \sum_{i=0}^{t} \binom{n-t}{i} \sum_{j=0}^{t-i} \binom{i}{j}. \tag{21}$$

Note that (19) may be obtained using in (16) the expansion

$$\left(\sum_{j=0}^{q-1} z^j\right)^m = \sum_{i=0}^{m} \binom{m}{i} z^i \left(\sum_{h=0}^{q-2} z^j\right)^i.$$

With the help of (20) Hirschberg [6] also proved that for any $X \in F_q^n$,

$$|D_t(X)| \geq \sum_{i=0}^{t} \binom{r(X)-t}{i}. \tag{22}$$

This improves upon the lower bound (11) for $t \geq 2$.

It is interesting that $|I_t(X)|$ does not depend on $X \in F_q^n$ and

$$|I_t(X)| = I_q(n,t) \quad \text{for any } X \in F_q^n \tag{23}$$

where

$$I_q(n,t) = \sum_{i=0}^{t} \binom{n+t}{i} (q-1)^i. \tag{24}$$

(We assume that $I_q(n,t) = 0$ for $n \geq 0$ and $t < 0$.) It was published in [10] for $q = 2$; the extension to the general case is immediate. The fact that $|I_t(X)|$ does not depend on $X \in F_q^n$ follows from (9) by induction on $n + t$, and (24) can be easily found by calculation of $|I_t(0^n)|$. Note that from (9) and (23) it follows that for $n \geq 1$,

$$I_q(n,t) = I_q(n-1,t) + (q-1)I_q(n,t-1) \tag{25}$$

and hence

$$I_q(n,t) = \sum_{i=0}^{t} I_q(n-1,t-i)(q-1)^i \tag{26}$$

since $I_q(n,0) = 1$ for any $n \geq 0$.

3 The maximum number of common subsequences

For any n, t, and q, $q \geq 2$, consider the function

$$N_q^-(n,t) = \sum_{i=1}^{q-1} D_q(n-i-1,t-i) + D_q(n-2,t-1) \tag{27}$$

and note that $N_q^-(n,t) = 0$ if $n \leq t$ or $t \leq 0$. From (17) it follows that for $n \geq t+1$,

$$N_q^-(n,t) = D_q(n,t) - D_q(n-1,t) + D_q(n-2,t-1). \tag{28}$$

Our aim is to prove that for any n and t, $n > t \geq 0$, the maximum number of common subsequences of length $n-t$ of two different sequences of F_q^n equals $N_q^-(n,t)$. In this section we prove that this number is not smaller than $N_q^-(n,t)$. Moreover, we find some properties of the function $N_q^-(n,t)$ which are needed to ground an algorithm to recover any $X \in F_q^n$ if one knows $N_q^-(n,t) + 1$ of its different subsequences of length $n-t$. This algorithm will be described in the next section. In particular, its existence allows us to complete the proof of the statement on the maximum number of common subsequences.

Lemma 1 *For any n, t, and q such that $n \geq t+1 \geq 2$, $q \geq 2$,*

$$\max_{X,Z \in F_q^n;\ X \neq Z} |D_t(X) \cap D_t(Z)| \geq N_q^-(n,t).$$

Proof. Let $X = \theta(A_{n,q})$ for a permutation $\theta = (\theta_0, ..., \theta_{q-1})$ of F_q. Then $X = \theta_0\theta_1 Y$ where Y is a word of length $n-2 \geq 0$. All elements of disjoint sets $D_{t-1}^{\theta_1}(\theta_1 Y) = \theta_1 D_{t-1}(Y)$, $D_{t-2}^{\theta_2}(Y)$,...,$D_{t-2}^{\theta_{q-1}}(Y)$, and $\theta_0 D_{t-1}(Y)$ are common subsequences of length $n-t$ of X and $Z = \theta_1\theta_0 Y$. The word Y can be obtained from $A_{n-2,q}$ using a permutation of F_q. Therefore, for $m = n-2$, $a = \theta_j$, $j = 2, ..., q-1$, $i = j-1$, there are equalities in (14), with t replaced by $t-2$, and hence

$$|D_t(X) \cap D_t(Z)| \geq \sum_{i=1}^{q-2} D_q(n-2-i, t-1-i) + 2D_q(n-2, t-1) = N_q^-(n,t),$$

and the lemma follows. □

It is worth pointing out that in case $n = t+1 \geq 2$,

$$N_q^-(t+1,t) = \min(t+1,q) = D_q(t+1,t) \tag{29}$$

and

$$\max_{X,Z \in F_q^{t+1};\ X \neq Z} |D_t(X) \cap D_t(Z)| = N_q^-(t+1,t). \tag{30}$$

This follows from (27), (28), Lemma 1, and the fact that for any $X \in F_q^{t+1}$, $D_t(X) \subseteq F_q$ and $|D_t(X)| \leq t+1$. As a consequence of (29) and (30) for any $X \in F_q^n$ we have

$$n \geq t+2 \text{ if } |D_t(X)| \geq N_q^-(n,t)+1 \text{ and } n \geq t+1 \geq 2. \tag{31}$$

Lemma 2 *If $n \geq t+2$, then*

$$N_q^-(n,t) = \sum_{i=0}^{q-1} N_q^-(n-i-1, t-i) \tag{32}$$

and for any h, $0 \leq h \leq q-1$,

$$N_q^-(n,t) - \sum_{i=0}^{h-1} N_q^-(n-i-1,t-i)$$

$$\geq \sum_{i=h+1}^{q-1} D_q(n-i-1, t-i) + D_q(n-h-2, t-h-1). \tag{33}$$

Proof. For $h = 0$ equality in (33) holds due to (27). For $1 \le h \le q$ using (27), (28), and (18) we have

$$N_q^-(n,t) - \sum_{i=0}^{h-1} N_q^-(n-i-1,t-i) = \sum_{i=1}^{q-1} D_q(n-i-1,t-i) + D_q(n-2,t-1)$$

$$- \sum_{i=0}^{h-1} D_q(n-i-1,t-i) + \sum_{i=0}^{h-1} D_q(n-i-2,t-i) - \sum_{i=0}^{h-1} D_q(n-i-3,t-i-1)$$

$$= \sum_{i=h}^{q-1} D_q(n-i-1,t-i) - D_q(n-1,t) + D_q(n-2,t-1)$$

$$+ D_q(n-2,t) - D_q(n-h-2,t-h)$$

$$= \sum_{i=h}^{q-1} D_q(n-i-1,t-i) - D_q(n-h-2,t-h) + D_q(n-q-2,t-q). \quad (34)$$

For $h = q$ this proves (32). For $1 \le h \le q-1$, applying once more (18) for $D_q(n-h-1,t-h)$ one can verify that (34) is equal to

$$\sum_{i=h+1}^{q-1} D_q(n-i-1,t-i) + D_q(n-h-2,t-h-1)$$

$$+ D_q(n-q-2,t-q) - D_q(n-h-q-2,t-h-q).$$

According to (13), this completes the proof. □

Corollary 1 *For any n, t, and q, $n \ge t+1 \ge 2$, $q \ge 2$,*

$$N_q^-(n,t) = \sum_{i=0}^{t-1} \binom{n-t-1}{i} \left(D_{q-1}(t,t-i-1) + D_{q-1}(t-1,t-i-1) \right). \quad (35)$$

Proof. Using (28) and (19) we get

$$N_q^-(n,t) = D_q(n,t) - D_q(n-1,t) + D_q(n-2,t-1)$$

$$= \sum_{i=1}^{t} \binom{n-t-1}{i-1} D_{q-1}(t,t-i) + \sum_{i=0}^{t-1} \binom{n-t-1}{i} D_{q-1}(t-1,t-i-1).$$

□

Note that for $t \geq i \geq 0$,

$$D_1(t,t-i)=1,\ D_2(t,t-i)=\sum_{j=0}^{t-i}\binom{i}{j},\ D_{q-1}(t,t-i)\leq (q-1)^i,$$

Therefore, Corollary 1 implies

$$N_2^-(n,t)=2\sum_{i=0}^{t-1}\binom{n-t-1}{i}, \tag{36}$$

$$N_3^-(n,t)=\sum_{i=0}^{t-1}\binom{n-t-1}{i}\sum_{j=0}^{t-i-1}\left(\binom{i+1}{j}+\binom{i}{j}\right), \tag{37}$$

and

$$N_q^-(n,t)\leq q\sum_{i=0}^{t-1}\binom{n-t-1}{i}(q-1)^i$$

with equality for $q=2$. For small t direct calculations with the help of Corollary 1 show that

$$N_q^-(n,0)=0,\quad N_q^-(n,1)=2,\quad N_q^-(n,2)=2n-3-\delta_{q,2}, \tag{38}$$

$$N_q^-(n,3)=(n-2)^2-(3n-10)\,\delta_{q,2}-\delta_{q,3}, \tag{39}$$

where $\delta_{i,j}$ is the Kronecker symbol.

4 An algorithm for reconstruction of sequences from their subsequences

In this section we describe and substantiate an algorithm to recover a sequence $X=x_1x_2...x_n \in F_q^n$ if one knows the length n of X and $N=N_q^-(n,t)+1$ different subsequences $Y_i \in D_t(X)$, $i=1,2,...,N$, where $n \geq t+2$ and $t \geq 1$. (The number t can be defined as $n-m$ where m is the length of Y_i.) This algorithm consists of some successive steps. At the first step we have the set $U=\{Y_1,...,Y_N\}$ which can be considered as a matrix of size $(n-t)\times N$ with columns Y_i. Applying a definite generalized threshold function to the ordered composition of the first row of U we find x_1 (as a letter which occurs more often than others among the first letters of subsequences Y_i) and a certain number j, $0 \leq j \leq t$, of the next successive letters of X which were deleted. Herewith, we indicate a submatrix U' of the size $(n'-t')\times N'$ whose columns are different subsequences $Y_i' \in D_{t'}(X')$

where $X' = x_{j+2}...x_n$, $n' = n - j - 1$, $t' = t - j$, and $N' = N_q^-(n', t') + 1$. If $t' = t - j = 0$, then $N_q^-(n', t') = 0$, the matrix U' consists of the only column X', and recovering $X = x_1...x_{j+1}X'$ is completed. In case $t' = t - j \geq 1$ we have $n' \geq t' + 2$ according to (31), and recovering $X \in F_q^n$ is reduced to recovering $X' \in F_q^{n'}$ with the help of $N' = N_q^-(n', t') + 1$ of its different subsequences of length $n' - t'$. Since $n' < n$ and $n' \geq t' + 2$ when $t' \geq 1$, we shall have $t' = 0$ in some steps.

Theorem 1 *Given n, t, and q such that $n \geq t + 2$, $t \geq 1$, $q \geq 2$, and $N = N_q^-(n, t) + 1$, let words $Y_i = y_{1,i}y_{2,i} \cdots y_{n-t,i}$, $i = 1, ..., N$, be different and $U = \{Y_1, ..., Y_N\} \subseteq D_t(X)$ for some $X = x_1 \cdots x_n \in F_q^n$. Let also*

$$\theta = (\theta_0, ..., \theta_{q-1}) \tag{40}$$

be the ordering permutation for the word $Z = y_{1,1}y_{1,2} \cdots y_{1,N} \in F_q^N$ and $l(Z)$ be its ordered composition (see (3) and (4)). Then the generalized threshold function $f_{\tau_0,\tau_1,...,\tau_{q-1}}$ with thresholds

$$\tau_i = N_q^-(n - i - 1, t - i), \quad i = 0, 1, ..., q - 1, \tag{41}$$

is defined on the vector $l(Z)$ and if

$$f_{\tau_0,\tau_1,...,\tau_{q-1}}(l(Z)) = j, \tag{42}$$

then $0 \leq j \leq t$ and

$$x_i = \theta_{i-1} \text{ for } i = 1, ..., j + 1. \tag{43}$$

Moreover,

$$\overline{U^{\theta_j}} \subseteq D_{t-j}(x_{j+2} \cdots x_n) \tag{44}$$

and

$$\left|\overline{U^{\theta_j}}\right| \geq N_q^-(n - j - 1, t - j) + 1. \tag{45}$$

In case $j = t$, $\overline{U^{\theta_j}}$ consists of the unique word $x_{j+2} \cdots x_n$.

Proof. Let $k(Z) = (k_0, ..., k_{q-1})$ and $l(Z) = (l_0, ..., l_{q-1})$ be the composition and the ordered composition of $Z = y_{1,1}y_{1,2}...y_{1,N} \in F_q^N$. By the definition of the ordering permutation (40), we have $l_i = k_{\theta_i} = \left|U^{\theta_i}\right|$, $i = 0, ..., q - 1$, and $l_0 \geq \cdots \geq l_{q-1}$. The function $f_{\tau_0,\tau_1,...,\tau_{q-1}}$ is defined on $l(Z)$, since by (32) and (41),

$$\sum_{i=0}^{q-1} l_i = N_q^-(n, t) + 1 = \sum_{i=0}^{q-1} \tau_i + 1, \tag{46}$$

and hence the condition (5) holds. If $j \in F_q$ is defined by (42), then

$$l_0 \geq \cdots \geq l_j \geq 1. \tag{47}$$

This implies that $D_t(X)$ contains at least $j+1$ words starting from different letters and hence $j \leq t$. Now we prove (43) by induction on i, $1 \leq i \leq j+1$. Suppose that $x_i = \theta_{i-1}$ for all i, $1 \leq i \leq h$, where $0 \leq h \leq j$, and prove that $x_{h+1} = \theta_h$. (The case $h = 0$ corresponds to a proof of the base of induction.) According to (47) $l_h \geq 1$ and hence X contains the letter θ_h. Let $x_m = \theta_h$ is the first occurence of a letter $\theta_h \in F_q$ in the word X and hence $m \geq h+1$. Our goal is to prove that

$$l_h = \left|U^{\theta_h}\right| \geq D_q(n-h-2, t-h-1)+1. \tag{48}$$

If it will be proved, then for $m \geq h+2$ we obtain a contradiction, since by (14) and (13)

$$l_h \leq \left|D_t^{\theta_h}(X)\right| \leq D_q(n-m, t-m+1) \leq D_q(n-h-2, t-h-1).$$

To prove (48), we denote by r the number of different letters of the set $F_q \backslash \{\theta_0, ..., \theta_{h-1}\}$ in the word $x_{h+1} \cdots x_{m-1}$ ($r \geq 0$) and note that $s = m-h-1-r \geq 0$ is the number of remaining letters of this word. Taking into account that $l_0 \geq \cdots \geq l_{q-1}$ and using again (14) and (13) we have

$$\sum_{i=h+1}^{q-1} l_i \leq r l_h + \sum_{p=1}^{q-h-r-1} D_q(n-m-p, t-m-p+1),$$

$$r l_h \leq \sum_{p=0}^{r-1} D_q(n-m+p, t-m+p+1),$$

and hence

$$\sum_{i=h+1}^{q-1} l_i \leq \sum_{i=h+1+s}^{q-1+s} D_q(n-i-1, t-i) \leq \sum_{i=h+1}^{q-1} D_q(n-i-1, t-i). \tag{49}$$

By the definition of the function $f_{\tau_0,\tau_1,\ldots,\tau_{q-1}}$ with the thresholds (42) and the inductive assumption, we have

$$\sum_{i=0}^{h-1} l_i \leq \sum_{i=0}^{h-1} N_q^-(n-i-1, t-i). \tag{50}$$

Using (46), (49), (50), and Lemma 2 we get

$$l_h - 1 = N_q^-(n,t) - \sum_{i=0}^{h-1} l_i - \sum_{i=h+1}^{q-1} l_h \geq D_q(n-h-2, t-h-1)$$

and complete the proof of (43). From the definition of $\overline{U^{\theta_j}}$, (8), and (43) it follows (44). By the definition of the generalized threshold function, (42) implies

$$l_j = \left|U^{\theta_j}\right| = \left|\overline{U^{\theta_j}}\right| \geq \tau_j + 1,$$

and hence (45) holds. In case $j = t$, (44) and (45) imply that $\overline{U^{\theta_j}}$ consists of the unique word $x_{j+2} \cdots x_n$ since $|D_0(Y)| = 1$ for any $Y \in F_q^*$. $\square$

Theorem 1 substantiates the following algorithm to recover an unknown $X = x_1 \cdots x_n \in F_q^n$ with the help of a set $U = \{Y_1, ..., Y_N\} \subseteq D_t(X)$ (i. e., a set of N of its different subsequences of length $n-t$), where $n \geq t+2$, $t \geq 1$, and $N = N_q^-(n,t) + 1$.

Algorithm for recovering a sequence from its subsequences:

- Find the ordering permutation $\theta = (\theta_0, ..., \theta_{q-1})$ for the word $Z \in F_q^N$, which is formed by the first letters of Y_i, $i = 1, ..., N$, and its ordered composition $l(Z)$.
- Calculate $f_{\tau_0,\tau_1,...,\tau_{q-1}}(l(Z))$ with thresholds $\tau_i = N_q^-(n-i-1, t-i)$, $i = 0, ..., q-1$.
- If $f_{\tau_0,\tau_1,...,\tau_{q-1}}(l(Z)) = j$, recover the first $j+1$ letters of X by the formula $x_i = \theta_{i-1}$, $i = 1, ..., j+1$.
- If $t' = j - t = 0$, recover the remaining letters of X as a unique word $x_{j+2} \cdots x_n$ of the set $\overline{U^{\theta_j}}$ and stop.
- If $t' = t - j \geq 1$, repeat this procedure, with n and t replaced by n' and t', to recover the word $X' = x_{j+2} \cdots x_n$ of length $n' = n-j-1$ using $N' = N_q^-(n', t') + 1$ of its subsequences of length $n' - t'$ of the set $\overline{U^{\theta_j}}$.

Example 1 Let $q = 3$, $n = 7$, $t = 3$. By (39), $N_3^-(7,3) = 24$ and hence 25 subsequences formed by columns of the matrix in Introduction are sufficient to recover an unknown $X = x_1 \cdots x_7 \in F_3^7$. For the first row $Z \in F_3^{25}$ of this matrix, $k_0(Z) = 8$, $k_1(Z) = 4$, $k_2(Z) = 13$, and hence $\theta = (2, 0, 1)$ is the ordering permutation and $l(Z) = (13, 8, 4)$ is the ordered composition. Since $\tau_0 = N_3^-(6,3) = 15$, $\tau_1 = N_3^-(5,2) = 7$, $\tau_2 = N_3^-(4,1) = 2$ (see (38)), we have $f_{\tau_0,\tau_1,\tau_2}(l(Z)) = 1$. Hence $x_1 = 2$, $x_2 = 0$, $t' = 2$, $n' = 5$, and the problem is reduced to reconstruction of $x_3 \cdots x_7 \in F_3^5$ by $N_3^-(5,2)+1 = 8$ of its subsequences of the set $\overline{U^{\theta_1}} = \overline{U^0}$ which are the columns of the following matrix:

$$\begin{matrix} 1 & 1 & 1 & 1 & 1 & 2 & 2 & 0 \\ 2 & 2 & 2 & 0 & 0 & 0 & 0 & 2 \\ 0 & 1 & 2 & 1 & 2 & 1 & 2 & 1 \end{matrix}$$

Now for the first row $Z \in F_3^8$ of this matrix, $k_0(Z) = 1$, $k_1(Z) = 5$, $k_2(Z) = 2$, and hence $\theta = (1, 2, 0)$ and $l(Z) = (5, 2, 1)$. Since $\tau_0 = N_3^-(4,2) = 5$, $\tau_1 = N_3^-(3,1) = 2$, $\tau_2 = N_3^-(2,0) = 0$, we have $f_{\tau_0,\tau_1,\tau_2}(l(Z)) = 2$. Therefore, $x_3 = 1$, $x_4 = 2$, $x_5 = 0$. In this case $j = t = 2$ and $\overline{U^{\theta_2}} = \overline{U^0}$ defines the remaining letters $x_6 = 2$, $x_7 = 1$. Thus, we finally have $X = 2012021$.

Theorem 2 *For any n, t, and q, $n \geq t+1$, $t \geq 0$, $q \geq 2$,*

$$\max_{X,Z \in F_q^n;\ X \neq Z} |D_t(X) \cap D_t(Z)| = N_q^-(n,t).$$

Proof. This statement is trivial when $t = 0$. For $n = t + 1 \geq 2$ this also holds according to (29) and (30). For $n \geq t+2$, $t \geq 1$ this follows from Lemma 1 and Theorem 1, which grounds the algorithm above and shows a possibility to recover any $X \in F_q^n$ if one knows $N_q^-(n,t) + 1$ of its different subsequences of length $n - t$. $\square$

5 The maximum number of common supersequences

For any n, t, and q, $n \geq 0$, $q \geq 2$, consider the function

$$N_q^+(n,t) = \sum_{i=0}^{t-1} \binom{n+t}{i} (q-1)^i (1 - (-1)^{t-i}). \tag{51}$$

($N_q^+(n,t) = 0$ if $t \leq 0$ by our assumption (6)). From this definition it follows that for $n \geq 1$,

$$N_q^+(n,t-1) + N_q^+(n-1,t) = 2I_q(n,t-1), \tag{52}$$

where $I_q(n,t)$ is defined by (24). Using that $\binom{n+t}{i} = \binom{n+t-1}{i} + \binom{n+t-1}{i-1}$ we also see that for $n \geq 1$,

$$N_q^+(n,t) = N_q^+(n-1,t) + (q-1)N_q^+(n,t-1) \tag{53}$$

and hence

$$N_q^+(n,t) = 2I_q(n,t-1) + (q-2)N_q^+(n,t-1). \tag{54}$$

Since $N_q^+(n,0) = 0$, it follows that for any $n \geq 1$,

$$N_q^+(n,t) = \sum_{i=0}^{t} N_q^+(n-1,t-i)(q-1)^i, \tag{55}$$

$$N_q^+(n,t) = 2\sum_{i=1}^{t} I_q(n,t-i)(q-2)^{i-1}. \tag{56}$$

We shall use below that (56) is also valid for $n = 0$.

Theorem 3 *For any $n,\ t$, and q, $n \geq 1,\ t \geq 0,\ q \geq 2$,*

$$\max_{X,Z \in F_q^n;\ X \neq Z} |I_t(X) \cap I_t(Z)| = N_q^+(n,t)$$

where $N_q^+(n,t)$ is defined by (51).

Proof. We prove by induction on $n+t$ that for any $a, b \in F_q$ and $X, Z \in F_q^{n-1}$ such that $aX \neq bZ$,

$$|I_t(aX) \cap I_t(bZ)| \leq N_q^+(n,t) \tag{57}$$

with equality when $a \neq b$ and $Z = X$. This is true when $n+t = 1$ and, hence, $n = 1$, $t = 0$, $N_q^+(1,0) = 0$. Assuming that this statement holds when $n+t < m$ we will prove it for $n+t = m$. Consider first the case when $a \neq b$ and X, Z may be identical. By (9) we have

$$\begin{aligned} |I_t(aX) \cap I_t(bZ)| &= (q-2)\, |I_{t-1}(aX) \cap I_{t-1}(bZ)| \\ &+ |I_t(X) \cap I_{t-1}(bZ)| + |I_t(Z) \cap I_{t-1}(aX)|. \end{aligned}$$

Note that

$$|I_t(X) \cap I_{t-1}(bZ)| \leq I_q(n,t-1) \quad \text{and} \quad |I_t(Z) \cap I_{t-1}(aX)| \leq I_q(n,t-1).$$

Since $I_{t-1}(aX) \subseteq I_t(X)$ for any $a \in F_q$, we have equalities in the both cases when $Z = X$. Using (54) we obtain by induction

$$|I_t(aX) \cap I_t(bZ)| \leq (q-2)N_q^+(n,t-1) + 2I_q(n,t-1) = N_q^+(n,t)$$

with equality when $Z = X$. Let now $a = b$ and $X \neq Z$. Then, by (9) and induction,

$$|I_t(aX) \cap I_t(aZ)| = |I_t(X) \cap I_t(Z)| + (q-1)\, |I_{t-1}(aX) \cap I_{t-1}(aZ)|$$

$$\leq N_q^+(n-1,t)+(q-1)N_q^+(n,t-1)=N_q^+(n,t),$$

and the proof is complete. $\square$

Note that for $q=2$, by (54),

$$N_2^+(n,t)=2I_2(n,t-1)=2\sum_{i=0}^{t-1}\binom{n+t-1}{i}. \tag{58}$$

Given $X=x_1\cdots x_n\in F_2^n$ and $b\in F_2$, $b\neq x_1$, we have $|I_t^b(X)|\leq I_2(n,t-1)$. Therefore, with the help of any $U\subseteq I_t(X)$, $|U|=N_2^+(n,t)+1$, we can recover x_1 as a letter which occurs more often among the first letters of all words of U (similarly to the case of recovering a sequence from its subsequences). However, we will see that for $q\geq 3$ this is in general not true. Nevertheless, now we shall show how it is possible to recover the first letter x_1 of an arbitrary $X=x_1\cdots x_n\in F_q^n$ using the first $t+1$ letters of any $N_q^+(n,t)+1$ words of $I_t(X)$. For any $a,b\in F_q$, $a\neq b$, and any set U of words, denote by $U_{a,b}$ the subset of U consisting of all words which contain the both a and b and have the following property: the first occurence of a precedes to that of b.

Lemma 3 *Given $X=x_1\cdots x_n\in F_q^n$, $n\geq 1, U\subseteq I_t(X)$, $|U|=N_q^+(n,t)+1$, we have*

$$\sum_{i=1}^{t+1}\left|U^{x_1,i}\right|=|U|=N_q^+(n,t)+1. \tag{59}$$

Moreover, if $\sum_{i=1}^{t+1}\left|U^{b,i}\right|=N_q^+(n,t)+1$ for some $b\in F_q$, $b\neq x_1$, then

$$|U_{x_1,b}|>|U_{b,x_1}|. \tag{60}$$

Proof. Each word $Y\in I_t(X)$ is obtained from X in a result of t or less insertions before the first letter x_1. This implies $\sum_{i=1}^{t+1}\left|U^{x_1,i}\right|=|U|$ and gives (59). Note that for any $b\in F_q$, $b\neq x_1$, and any $i=2,...,t+1$, there exist $(q-1)^{i-1}-(q-2)^{i-1}$ words of length $i-1$, which contain b and do not contain x_1. Using (26), (56), (52) and (53) we have

$$|U_{b,x_1}|\leq\sum_{i=2}^{t+1}\left((q-1)^{i-1}-(q-2)^{i-1}\right)I_q(n-1,t-i+1)$$

$$=(q-1)\sum_{i=0}^{t-1}I_q(n-1,t-i-1)(q-1)^i-(q-2)\sum_{i=1}^{t}I_q(n-1,t-i)(q-2)^{i-1}$$

$$= (q-1)I_q(n,t-1) - \frac{1}{2}(q-2)N_q^+(n-1,t) = \frac{1}{2}N_q^+(n,t).$$

This completes the proof, because, by the condition of the lemma, $U_{x_1,b}$ and U_{b,x_1} form a partition of the set U, $|U| = N_q^+(n,t)+1$, into two parts. $\square$

It is worth pointing out that in the case of insertions,

$$|I_t(X)| \geq N_q^+(n,t)+1 \text{ for any } X \in F_q^n,\ n \geq 1,\ t \geq 0. \tag{61}$$

In particular, this follows from the fact that

$$I_q(n,t) - N_q^+(n,t) = \sum_{i=0}^{t} \binom{n+i-1}{i} (q-1)^i (q-2)^{t-i} \tag{62}$$

which can be proved by induction on $n+t$ with the help of (25) and (53). Note also that (51) for small t implies

$$N_q^+(n,0) = 0,\ N_q^+(n,1) = 2,\ N_q^+(n,2) = 2(q-1)(n+2), \tag{63}$$

$$N_q^+(n,3) = 2 + (q-1)^2(n+2)(n+3).$$

6 An algorithm for reconstruction of sequences from their supersequences

Now we shall substantiate a simple algorithm to recover an arbitrary $X = x_1 \cdots x_n \in F_q^n$, $n \geq 1$, if one knows its length n and $N = N_q^+(n,t)+1$ different supersequences $Y_i \in I_t(X)$, $i = 1,2,...,N$, $t \geq 1$. This algorithm consists of some successive steps. At the first step we have the set $U = \{Y_1, ..., Y_N\}$ which can be considered as a matrix of size $(n+t) \times N$ with columns Y_i. For any $a \in F_q$ we define a real vector $m(a)$ of length $t+1$. We recover x_1 with the help of the vectors $m(a)$ and the first $t+1$ rows of this matrix. This completes recovering $X = x_1 \cdots x_n$ if $n = 1$. In case $n \geq 2$ we apply a definite generalized threshold function in $t+1$ variables to the vector $m(a)$, where $a = x_1$, and find a certain number j, $0 \leq j \leq t$, of letters which were inserted before x_1 or between x_1 and x_2. Herewith we indicate a submatrix U' of the matrix U which has the size $(n'+t') \times N'$ and consists of different columns $Y_i' \in I_{t'}(X')$ where $X' = x_2 \cdots x_n$, $n' = n-1$, $t' = t-j$, and $N' = N_q^+(n',t')+1$. If $t' = t-j = 0$, then $N_q^+(n',t') = 0$, the matrix U' consists of the only column X', and recovering $X = x_1X'$ is completed. In case $t' = t-j \geq 1$ recovering $X \in F_q^n$ is reduced to recovering $X' \in F_q^{n'}$ with the help of $N' = N_q^+(n',t')+1$ of its different supersequences of length $n'+t'$. Since $n' = n-1$, $t' \leq t$, we shall have $n' = 1$ or $t' = 0$ in some steps.

Theorem 4 *Given n, t, and q such that $n \geq 1$, $t \geq 1$, $q \geq 2$, and $N = N_q^+(n,t) + 1$, let words Y_i, $i = 1, ..., N$, be different and $U = \{Y_1, ..., Y_N\} \subseteq I_t(X)$ for some $X = x_1 \cdots x_n \in F_q^n$. Let also for any $a \in F_q$,*

$$m(a) = (m_0(a), ..., m_t(a)), \text{ where } m_i(a) = \left|U^{a,i+1}\right|. \tag{64}$$

Then

$$\sum_{i=0}^{t} m_i(a) = \sum_{i=1}^{t+1} \left|U^{a,i}\right| = N_q^+(n,t) + 1 \tag{65}$$

holds for $a = x_1$ and validity of (65) for some $a \in F_q$, $a \neq x_1$, implies $|U_{x_1,a}| > |U_{a,x_1}|$. Moreover, in case $n \geq 2$, the generalized threshold function $f_{\tau_0,\tau_1,...,\tau_t}$ with thresholds

$$\tau_i = N_q^+(n-1, t-i)(q-1)^i, \; i = 0, 1, ..., t, \tag{66}$$

is defined on the vector $m(x_1)$ and if

$$f_{\tau_0,\tau_1,...,\tau_t}(m(x_1)) = j \;, \tag{67}$$

then $0 \leq j \leq t$ and

$$\overline{U^{x_1,j+1}} \subseteq I_{t-j}(x_2 \cdots x_n),$$

$$\left|\overline{U^{x_1,j+1}}\right| \geq N_q^+(n-1, t-j) + 1.$$

In case $j = t$, $\overline{U^{x_1,j+1}}$ consists of the unique word $x_2 \cdots x_n$.

Proof. The facts that

$$\sum_{i=0}^{t} m_i(x_1) = \sum_{i=1}^{t+1} \left|U^{x_1,i}\right| = N_q^+(n,t) + 1 \tag{68}$$

and that validity of (65) for some $a \in F_q$, $a \neq x_1$, implies $|U_{x_1,a}| > |U_{a,x_1}|$ were proved in Lemma 3. They allow one to recover x_1 with the help of the first $t+1$ rows of the matrix U. In case $n \geq 2$ from (68) and (55) it follows that

$$\sum_{i=0}^{t} m_i(x_1) > N_q^+(n,t) = \sum_{i=0}^{t} N_q^+(n-1, t-i)(q-1)^i = \sum_{i=0}^{t} \tau_i$$

and hence $f_{\tau_0,\tau_1,...,\tau_t}$ with thresholds (66) is defined on the vector $m(x_1)$. Let (67) hold and hence $0 \leq j \leq t$. Since $\left|U^{x_1,j+1}\right| > \tau_j$, the set $U^{x_1,j+1}$ contains $N_q^+(n-1, t-j)(q-1)^j + 1$ or more different words and hence $\overline{U^{x_1,j+1}}$ contains $N_q^+(n-1, t-j) + 1$ or more different words. By the definition of the set $\overline{U^{x_1,j+1}}$

all its elements belong to $I_{t-j}(x_2 \cdots x_n)$. If $j = t$, $\overline{U^{x_1,j+1}}$ consists of the only word $x_2 \cdots x_n$ of the set $I_0(x_2 \cdots x_n)$. $\square$

Theorem 4 grounds the following algorithm to recover an unknown $X = x_1 \cdots x_n \in F_q^n$ with the help of a set $U = \{Y_1, ..., Y_N\} \subseteq I_t(X)$ (i. e., a set of N of its different supersequences of length $n+t$) where $n \geq 1$, $t \geq 1$, and $N = N_q^+(n,t) + 1$.

Algorithm for recovering a sequence from its supersequences:

- Find for any $a \in F_q$ the vector $m(a) = (m_0(a), ..., m_t(a))$, where $m_i(a) = \left|U^{a,i+1}\right|$, and the non-empty set F of (permissible as the first one) letters $a \in F_q$, for which $\sum_{i=0}^{t} m_i(a) = N_q^+(n,t) + 1$.
- Recover x_1 as a unique letter $a \in F$ such that $|U_{a,b}| > |U_{b,a}|$ for any $b \in F \backslash \{a\}$ when $|F| \geq 2$, and stop if $n = 1$.
- If $n \geq 2$, calculate the function $f_{\tau_0,\tau_1,...,\tau_t}(m(x_1))$ with thresholds $\tau_i = N_q^+(n-1, t-i)(q-1)^i$, $i = 0, 1, ..., t$.
- If $f_{\tau_0,\tau_1,...,\tau_t}(m(x_1)) = j$ and $t' = t - j = 0$, recover $x_2 \cdots x_n$ as a unique word of the set $\overline{U^{x_1,t+1}}$ and stop.
- If $t' = t - j \geq 1$, repeat this procedure, with n and t replaced by n' and t', to recover the word $X' = x_2 \cdots x_n$ of length $n' = n-1$ with the help of $N' = N_q^+(n', t') + 1$ of its supersequences of length $n' + t'$ of the set $\overline{U^{x_1,j+1}}$.

Example 2 Let $q = 3$, $n = 3$, $t = 2$, and hence $N_3^+(3,2) = 20$ (see (63)). Recover an unknown $X = x_1x_2x_3 \in F_3^3$ with the help of the set U from 21 of its supersequences written by columns of the following matrix:

$$\begin{array}{ccccccccccccccccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 \\
0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 0 & 1 & 2 \\
2 & 2 & 0 & 0 & 1 & 0 & 0 & 0 & 2 & 2 & 2 & 0 & 0 & 1 & 0 & 0 & 0 & 2 & 2 & 2 & 2 \\
0 & 0 & 1 & 1 & 0 & 2 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 2 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 2 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1
\end{array} \tag{69}$$

We have $m(0) = (9,2,5)$, $m(1) = (9,2,1)$, $m(2) = (3,14,4)$, and hence $F = \{2\}$ and $x_1 = 2$, because $\sum_{i=0}^{2} m_i(a) = 21$ holds only for $a = 2$. Using that $n > 1$, $\tau_0 = 16$, $\tau_1 = 4$, $\tau_2 = 0$, we get $f_{16,4,0}(3,14,4) = 1$. Since $j = 1$ and $t' = t - j = 1$, we go to reconstruction of $X' = x_2x_3$ of length $n' = 2$ with the help of $N_3^+(2,1) + 1 = 3$ words of length $n' + t' = 3$ belonging to the set $\overline{U^{x_1,j+1}} = \overline{U^{2,2}}$. This set contains

7 different words but we can use any three from them, for instance, written by columns of the following matrix:

$$\begin{array}{ccc} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{array}$$

Now we have $m(0) = (2,1)$, $m(1) = (1,2)$, $m(2) = (0,0)$, and $F = \{0,1\}$, because $m_0(a) + m_1(a) = 3$ holds for $a = 0$ and $a = 1$. Since $|U_{0,1}| = 2 > |U_{1,0}| = 1$, $x_2 = 0$. Using that $n \geq 2$, $\tau_0 = 2$, $\tau_1 = 0$, we get $f_{2,0}(2,1) = 1$. In this case $t' = 0$ and the set $\overline{U^{0,2}}$ consists of the letter $x_3 = 1$. Thus, $x = 201$.

This example shows (see (69)) that it is in general not possible to define x_1 applying the majority function to the first row of the initial matrix as in the case of deletions.

7 Concluding remarks

The results above show that there is a significant difference between the problems of finding the maximum numbers of common subsequences and common supersequences of a sequence, which extends to the algorithms of its reconstruction using the minimum number of elements. This difference is eventually caused by the fact that the number $|D_t(X)|$ of subsequences of length $n - t$ of a sequence $X \in F_q^n$ depends on X (see (11) and (22)) while the number $|I_t(X)|$ of supersequences of length $n + t$ does not (see (24)). In particular, for the maximum number $N_q^+(n,t)$ of common supersequences we have the explicit formula (51), but for the maximum number $N_q^-(n,t)$ of common subsequences only recurrent formulas (28) and (35) are known which give simple expressions (36), (37) for small q (similar to the case of $D_q(n,t)$, see (19)-(21)). As far as the algorithms is concerned, the condition $|I_t(X)| \geq N_q^+(n,t) + 1$ is satisfied for all $X \in F_q^n$ (see (61)) while the condition $|D_t(X)| \geq N_q^-(n,t) + 1$ holds if X has a sufficiently large number of runs. On the other hand, note that for $N_q^-(n,t)$ and $N_q^+(n,t)$ the same recurrences are valid as for $D_q(n,t)$ and $I_q(n,t)$ (cf. (32) and (55) with (17) and (26)). Moreover, the members of these recurrences are thresholds of the generalized threshold functions which are used in the both algorithms.

Some statements of the paper can be useful for other combinatorial problems. In particular, the minimum number n such that $D_q(n, n-m) = q^m$ is the minimum length of a (universal) sequence $X \in F_q^n$ for which $D_{n-m}(X) = F_q^m$, and for any m and n, $n \geq m > 0$, there does not exist a (universal) sequence $X \in F_q^m$ for which $I_{n-m}(X) = F_q^n$. Note also that the inequality $I_q(m, n-m) \geq N$ or $\sum_{i=0}^{n-m} \binom{n}{i} (q-1)^i \geq N$ is a necessary condition for a set $U \subseteq F_q^n$, $|U| = N$, to have a common subsequence of length m. This gives an upper bound on m and can be used in algorithms to find a longest common subsequence of a set of sequences (see [5]).

It should be mentioned that the considered problem differs from the known problem of reconstruction of a sequence $X = (x_1, ..., x_n) \in F_q^n$ by the *multiset* $M_m(X)$ of all $\binom{n}{m}$ of subsequences $Y = (x_{i_1}, ..., x_{i_m})$, where $1 \leq i_1 < \cdots < i_m \leq n$ (reconstruction by fragments). This problem consists of finding the minimum number $m = m_q(n)$ such that $M_m(X) \neq M_m(Z)$ for any different $X, Z \in F_q^n$ and is far from being conclusive (see [15], [8], [14], [7]).

The problems considered belong to the class of metric problems introduced in [12]. Let G be a finite or countable graph with the set V of vertices and $B_t(X)$ be the metrical ball (in the path metric) of radius t centered at $X \in V$. (It is assumed that $B_t(X)$ is finite if V is countable.) Fix a set $A \subseteq V$ (a set of objects) and a set $B \subseteq V$ (a set of observations). A problem is to find the minimum number N such that an arbitrary $X \in A$ can be reconstructed with the help of any set $U \subseteq B_t(X) \bigcap B$ if $|U| \geq N$. One can see that this minimum number $N = N(A, B, t)$ is equal to

$$\max_{X,Z \in A;\ X \neq Z} \left| B_t(X) \bigcap B_t(Z) \bigcap B \right|$$

increased by 1. In particular, we can consider the graph with the set $V = F_q^*$ of vertices and the path metric $d(X, Z)$, which is equal to the minimum number of deletions and insertions translating $X \in F_q^*$ into $Z \in F_q^*$. This distance $d(X, Z)$, introduced in [9] (see also [11]), can be also defined as the difference of lengths of a shortest common supersequence and a longest common subsequence of X and Z. If we set $A = F_q^n$ and $B = F_q^{n-t}$ or $B = F_q^{n+t}$, we get the problems considered. Similar problems were solved in [12], [13] for some other cases, including the Hamming and Johnson graphs with $V = A = B = F_q^n$. However, this approach gives rise to many new combinatorial problems. In particular, let S_q be the set of all $q!$ permutations of F_q and let H be a subset of S_q such that $h^{-1} \in H$ if $h \in H$. Denote by G_H the graph with the set $V = S_q$ of vertices, for which $X \in S_q$ and $Z \in S_q$ are adjacent if and only if there exists $h \in H$ translating X into Z. The problems of finding the diameter of G_H and constructing effective algorithms for calculation of the path distance $d(X, Y)$ have been investigated as combinatorial problems and problems of computational molecular biology (see survey papers [4], [1]). In this connection it is worth pointing out that for some graphs G_H the problems to find $N(S_q, S_q, t)$ and a simple algorithm to recover an arbitrary $X \in S_q$ using any $N(S_q, S_q, t) + 1$ of elements of $B_t(X)$ are also problems of essential interest.

References

[1] A. Apostolico and R. Giancarlo, "Sequence alignments and molecular biology", J. of Comp. Biology, vol. 5, no. 2, pp. 173–196, 1998.

[2] L. Calabi, "On the computation of Levenshtein's distances", TN-9-0030, Parke Math. Labs., Inc., Carlisle, Mass., 1967.

[3] L. Calabi and W.E. Hartnett, "Some general results of abstract coding theory with applications to the study of codes for the correction of synchronization errors", Inform. and Control, vol. 15, pp. 235–249, 1969.

[4] M.M. Glukhov, A.Ju. Zubov, "On length of symmetric and alternating permutation groups for different systems of generators (survey)", Matematicheckie Problemy Kiberneticki, vol. 8, M.: Nauka. Fizmatlit, 1999 (in Russian).

[5] D.S. Hirschberg, "Algorithms for the longest subsequence problem", J. of ACM, vol. 24, no. 4, pp. 664–675, 1977.

[6] D.S. Hirschberg, "Bounds on the number of string subsequences", in Proc. 10th Symp. on Combinatorial Pattern Matching, Warwick UK, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 1999.

[7] I. Krasikov, Y. Roditty, "On a reconstruction problem for sequences", J. Combin. Theory, Ser. A, vol. 77, no. 2, pp. 344–348, 1997.

[8] V.K. Leont'ev, Ju.G. Smetanin, "Reconstruction of vectors from a set of their fragments", DAN, vol. 302, no. 6, pp. 1319-1322, 1988 (in Russian). English translation in Soviet Math. - Dokl., vol. 32, no. 2, pp. 438-441, 1989

[9] V.I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals", DAN, vol. 163, no. 4, pp. 845-848, 1965 (in Russian). English translation in Soviet Phys. - Dokl., vol. 10, no. 8, pp. 707-710, 1966.

[10] V.I. Levenshtein, "Elements of coding theory", in Discrete Mathematics and Math. Problems of Cybernetics, Nauka, Moscow, 1974, pp. 207–305 (in Russian). German translation in Diskrete Mathematik und mathematische Fragen der Kybernetik, Academie-Verlag, Berlin, 1980, pp. 198-279.

[11] V.I. Levenshtein, "On perfect codes in deletion/insertion metric", Discretnaya Mathematika, vol. 3, no. 1, pp. 3-20, 1991 (in Russian). English translation in Discrete Math. Appl., vol. 2, no. 3, pp. 241–258, 1992.

[12] V.I. Levenshtein, "Reconstruction of objects from a minimum number of distorted patterns", DAN, vol. 354, no. 5, pp. 593-596, 1997 (in Russian). English translation in Doklady Mathematics, vol. 55, no. 3, pp. 417–420, 1997.

[13] V.I. Levenshtein, "Efficient reconstruction of sequences", Submitted to IEEE Trans. on Information Theory.

[14] B. Manvel, A. Meyerowitz, A. Schwenk, K. Smith, P. Stockmeyer, "Reconstruction of sequences", Disc. Math., vol. 94, pp. 209-219, 1991.

[15] A.I. Zenkin and V.K. Leont'ev, "On a nonclassical problem of recognition", Zh. Vychisl. Mat. Mat. Fiz., vol. 24, no. 6, pp. 925-931, 1984 (in Russian).