Mathematica Pannonica
19/1 (2008), 125–135

MAXIMAL RANK DISTANCE FOR BINARY SEQUENCES

Liviu P. **Dinu**

Facultatea de matematică, Bucharest, Romania

Andrea **Sgarro**

DMI, University of Trieste and CBM Area Science Park, Trieste, Italy

Dedicated to Professor Maurer on the occasion of his 80^{th} birthday

Received: April 2008

MSC 2000: 68 R 15, 05 A 99, 68 Q 15, 91 F 20, 92 D 20

Keywords: Rank distance, low-complexity distances, sequence distances, string distances, combinatorics of sequences, mathematical linguistics, DNA sequences.

Abstract: The rank distance is a low-complexity and robust distance between sequences, which has been used in computational linguistics and bioinformatics. We tackle the problem of maximising rank distances; in particular, we solve the problem of exhibiting sequences at largest rank distance from a given binary sequence.

1. Introduction

Initially, rank distances between sequences were used in computational linguistics, cf. [5], [6]; later their use was extended to such application domains as is bioinformatics, cf. [8], or authorship identification, cf. [9]. The reasons for the practical success of the rank distance are

E-mail addresses: liviu.p.dinu@gmail.com, sgarro@units.it

basically two:

i) it is quite *quick* to compute

ii) it is *robust* with respect to small modifications of the sequences

As for the first point, the computational effort is only linear in the sequence length n, as happens with the unsophisticated *Hamming distance*, but unlike what happens with the more sophisticated *edit distance*, whose computational complexity is quadratic in n, cf. [1]. Problems which are NP-complete (i.e. computationally intractable) for the edit distance, such as the median problem, cf. [11], are quite tractable for the rank distance, [7]. As for the second point, think e.g. of a sequence x and "by mistake" rotate it one position to the right to get y: if, say, x is 01 repeated $n/2$ times the Hamming distance between x and y is as high as n, and so the percentage "error" scaled to the maximal possible value for Hamming distances is as high as 100%. Instead, while the rank distance is still linear in n, the percentage error goes to zero with the sequence length n, and so is practically negligible for very long sequences as are, say, DNA strings; cf. below Prop. 1 in Sec. 2, where we show that maximal rank distances are quadratic in n. In this paper we cope precisely with the problem of finding maximal values for the rank distance, concentrating on the binary case.

We move to formal definitions. Let $\mathcal{A}$ be an *alphabet* of size $|\mathcal{A}| \geq 2$. The elements of $\mathcal{A}^n$ are the sequences of length n. If x is such a sequence, n_a denotes the number of occurrences of letter a in x; a *composition class* $C = C_x$ comprises a sequence x with all its permutations, and it is uniquely identified by its *composition vector*, i.e. by the $|\mathcal{A}|$ non-negative integers n_a, $\sum_a n_a = n$. The combinatorics of composition classes is extensively dealt with e.g. in [2], part 1, Ch. 2. We find it convenient to *index* sequences: the first[1] letter a which occurs in x will have index 1, the second letter a will have index 2, and so on for all of its n_a occurrences, and for all letters $a \in \mathcal{A}$. For example, the ternary sequence $x = ababbca$ of length 7 will be indexed to $a_1b_1a_2b_2b_3c_1a_3$. Once x has been indexed the *rank* of the indexed letter a_i is the position j which a_i occupies in x, numbering subsequent positions from 1 to n; this rank is denoted by $\mathrm{ord}(a_i|x)$. With the same sequence $x = ababbca$ one has $\mathrm{ord}(a_1|x) = 1$, $\mathrm{ord}(b_1|x) = 2$, $\mathrm{ord}(a_2|x) = 3$, $\ldots$, $\mathrm{ord}(a_3|x) = 7$.

Definition 1. Given two sequences x and y in the same composition

[1] Needless to say, we "read" sequences left to right.

class $C \subset \mathcal{A}^n$ their rank distance is defined as:

$$d(x,y) = \sum_{a \in \mathcal{A}} \sum_{1 \le i \le n_a} \bigl|\mathrm{ord}(a_i|x) - \mathrm{ord}(a_i|y)\bigr|.$$

The double sum has n terms.[2] For example

$$\begin{aligned} d(aabc, caba) &= |\mathrm{ord}(a_1|x) - \mathrm{ord}(a_1|y)| + |\mathrm{ord}(a_2|x) - \mathrm{ord}(a_2|y)| + \\ &\quad + |\mathrm{ord}(b_1|x) - \mathrm{ord}(b_1|y)| + |\mathrm{ord}(c_1|x) - \mathrm{ord}(c_1|y)| = \\ &= |1-2| + |2-4| + |3-3| + |4-1| = 6. \end{aligned}$$

We stress that the rank distance is defined *only* for sequences which have the *same* composition. The rank distance is a metric distance over $C = C_x$, in particular the triangle inequality $d(x,z) + d(z,y) \ge d(x,y)$ holds; for a proof cf. [4]. The rank distance is an *even* integer; if two sequences differ by a single *twiddle* (i.e. by a single exchange between consecutive positions) their rank distance is equal to 2.

Once $\mathcal{A}^n$ is chosen, **three problems** arise: given $x \in C$ maximise $d(x,y)$ over $\mathcal{C}$; given $\mathcal{C}$ maximise $d(x,y)$ and so find the *diameter* $\delta(\mathcal{C})$ of $\mathcal{C}$; maximise the diameters over all the composition classes $\mathcal{C}$, and so find the absolute maximum for the rank distance over $\mathcal{A}^n$.

In Sec. 2, after pointing out that we are dealing with a "constrained" version of a situation already dealt with in the literature, we solve straightaway the third problem. In Sec. 3 we solve the remaining problems in the case when $\mathcal{A}$ is binary. In the concluding Sec. 4 we tackle the general case; as for the first of the three problems, we shall have to be contented with a conjecture.

2. Preliminaries

It will turn out that a special role is played by *compact sequences*, as now defined:

Definition 2. A compact sequence x is one where patterns like $ab\ldots a$ are prohibited ($a \neq b$), i.e. all the occurrences of each letter a must occupy consecutive positions. A run of a compact sequence is a substring (i.e. an infix) made up by all the occurrences of a given letter.

Compact sequence will be denoted in boldface like **bca**, meaning that one has a run of n_b letters b, followed by a run of n_c letters c, and

[2] We agree that a sum with zero terms is equal to zero: in practice, letters a with $n_a = 0$ are omitted from the expression at the right.

by a run of n_a letters a.

In the special case of sequences without repeated letters, the rank distance is tightly related to a measure of disarray between permutations of the integers from 1 to n called *Spearman footrule*; cf. [3], which is a standard reference on "ordinal" distances. More precisely, the rank distance and Spearman footrule have the same value when the sequence x is made up of those integers written in their natural order, and y is a permutation thereof. This overlapping is enough to quickly "re-cycle" results of [3] to the case of sequences *without repeated letters*, i.e. $|\mathcal{A}| = n$, $n_a = 1$ for all the n letters $a \in \mathcal{A}$; below, in Th. 1, we mention a few facts which we need.

In the sequel the prefix of length $\lfloor \frac{n}{2} \rfloor$ of a sequence x and the suffix of the same length will be called the *first half* and the *second half* of x, respectively; if n is odd, to "cover" the whole of x one still needs an infix of length 1 corresponding to the central position. In Th. 1 below, the logical value $\langle P \rangle$ is 1 when proposition P is true, else is 0; by saying that a substring x' of x is positioned on the left or on the right of another substring y' of y, we mean that either the last position of the letters in x' is smaller or equal to the first position of the letters in y', or the first position of the letters in x' is greater or equal to the last position of letters in y', respectively.

Theorem 1 ([3], 1977)**.** *Given an n-length sequence x without repeated letters, the maximum value for the rank distance is achieved e.g. by its mirror image x^*:*

$$(1) \qquad \max_y d(x,y) = d(x,x^*) = 2\left\lfloor \frac{n}{2} \right\rfloor \left\lceil \frac{n}{2} \right\rceil = \frac{n^2 - \langle n \quad odd \rangle}{2}.$$

A sequence y which achieves maximum rank distance $d(x,y)$ is necessarily obtained from x^* in the following way: permute letters inside the first half of x^* and permute letters inside its second half; if n is odd, one may further permute the letter which occupies the central position with any letter in any other position. More generally, for any two sequences x and y, if a substring of x, x' say, has all its correspondents in a substring y' of y which is positioned either to the left or to the right of x', then one can freely permute the letters in x' and in y' without changing the value of $d(x,y)$.

As an example, take $n = 4$, $x = abcd$, $x^* = dcba$; one has $d(x,y) \leq$ $\leq d(x,x^*) = 8 = d(x, cdab)$. With respect to sequences where letters appear only once, our problem is *constrained*: certain permutations are not allowed because they involve equal letters. This implies that the new distances can never exceed the value given in (1). To better see how the constraints work, observe that indexing as in Def. 1 does take the "new" rank distance back to the "old" case of non-repeated letters, just think of a larger letter alphabet whose elements are the indexed letters: the problem, however, is that not all sequences over the larger alphabet are allowed, because an indexed letter is constrained to have a position in the sequence which precedes any occurrence of the same letter with a higher index. In other words: in a way we are still dealing with sequences without repeated letters, but now our workspace does *not* include all of them.

Actually, in spite of the constraints which we have imposed, the value in (1) is also the absolute maximum of $d(x, y)$ over $\mathcal{A}^n$, as we now show. Take a composition class which is *balanced*, in the sense of the following definition:

Definition 3. A composition class $\mathcal{C}$ is balanced when it contains compact sequences x with the following property: letters a which appear in any of the two halves of x do not appear in the other half, $a \in \mathcal{A}$.

In practice, the composition $\{n_1, n_2, \ldots, n_{|\mathcal{A}|}\}$ of a balanced composition class[3] corresponds to a YES-instance of PARTITION; this is a famous NP-complete problem, but in our case the alphabet size $|\mathcal{A}|$ can be thought of as fixed, and so its computational complexity evaporates; cf. e.g. [10]. Now, if one takes a compact sequence x as in the definition of balanced classes, one soon checks that $d(x, x^*)$ is equal to the value given in (1). Clearly, balanced classes are always to be found, whatever n and whatever the size of $\mathcal{A}$, and so one can always find in $\mathcal{A}^n$ two sequences x and y with $d(x, y)$ equal to the unconstrained maximum as in (1):

Proposition 1.

$$\max_{\mathcal{C} \in \mathcal{A}^n} \max_{x, y \in C} d(x, y) = \frac{n^2 - \langle n \text{ odd} \rangle}{2}.$$

[3] With an obvious trick if n is odd: just add a dummy 1 to the composition vector.

3. Binary sequences

In this section $\mathcal{A} = \{0, 1\}$. By C_i we denote the composition class where 1 occurs i times, $0 \leq i \leq n$. In C_i there are two compact sequences[4] denoted by **01** and **10**, respectively. We begin by three lemmas. The first shows that the contribution of 1's to the rank distance is exactly equal to the contribution of 0's.

Lemma 1. $d(x, y) = 2 \sum\limits_{1 \leq j \leq i} \left|\text{ord}(1_j|x) - \text{ord}(1_j|y)\right|$.

Proof. Let y and z differ by a single twiddle, i.e. by a single exchange between consecutive positions.[5] Then $|d(x, y) - d(x, z)| = 2$: to prove this, we shall use the last statement in Th. 1. Say the twiddle involves a couple of distinct bits at positions j and $j+1$ in y, and say the matching for the corresponding indexed bits is with two bits in x at positions $u <$ $< v$. We shall check that neither $v \leq j$ nor $j + 1 \leq u$ can hold true, which is precisely what we need to use the theorem. We rule out the first case by an inverse argument. Assume that the bits 0 and 1 occur in this order both in x and in y; let their ranks be r and s, which means that the given 0 is preceded by $r - 1$ bits 0 both in x and in y, while the given 1 is preceded by $s - 1$ bits 1. This gives $v \geq r + s$ thinking of x (we did not write an equality sign, because there might be more 0's in x between positions $u + 1$ and position $v - 1$) and $j = r + s - 1$ thinking of y; however, this is incompatible with $v \leq j$. All the other cases are dealt with in the same way. Now, $d(x, x) = 0$ verifies the lemma and y can be always obtained from x by successive twiddles. ♢

Since one can compute rank distances taking care only of 1's, the following Lemma 2 needs no proof:

Lemma 2. *Let $x, y \in \mathcal{C}_i$ be two sequences. If every* 1 *in x is positioned on the left of the corresponding* 1 *in y, then $d(x, y) \leq d(x, \mathbf{01})$. Symmet-*

[4] To avoid fastidious specifications, below we tacitly rule out $C_0 = \{\mathbf{0}\}$ and $C_n =$ $= \{\mathbf{1}\}$; clearly, Th. 2 is void on these one-element classes, while its Cor. 1 is trivially true.

[5] Implicitly, we prove that the binary rank distance is *exactly* twice the (minimal) number of twiddles needed to bring x to y. Already with ternary sequences this is not the case, e.g. $d(abc, cba) = 4 = 2 \times 2$, while *three* twiddles are needed to bring abc to cba. As proven in [3] (the proof soon extends to the case of repeated letters), the rank distance is *at most* twice the number of twiddles needed to transform x into y; actually, along the transformation, some twiddles contribute 2 to the rank distance, while some others contribute 0: to see the reason why the latter fact happens, just think of the last statement in Th. 1. In the binary case *all* twiddles contribute 2.

rically, if every 1 *in* x *is positioned on the right of the corresponding* 1 *in* y*, then* $d(x,y) \leq d(x,\mathbf{10})$.

Lemma 3. *For any real numbers* $a_0 \leq a_1 \leq a_2 \leq \ldots \leq a_k \leq a_{k+1}$*, the function* $f(u) : [a_0, a_{k+1}] \to R$ *given by* $f(u) = |a_1 - u| + |a_2 - u| + + \ldots + |a_k - u|$ *is convex-cup. More precisely, if* k *is even the function is strictly decreasing up to* $u = a_{\frac{k}{2}}$*, constant on the interval* $\left[a_{\frac{k}{2}}, a_{\frac{k}{2}+1}\right]$*, and strictly increasing starting from* $u = a_{\frac{k}{2}+1}$*; if* k *is odd it is strictly decreasing up to* $u = \lfloor a_{\frac{k}{2}} \rfloor$ *and then strictly increasing.*

Proof. Use elementary calculus: the continuous function $f(u)$ is equal to $(2i-k)u + \text{const}$ on each open sub-interval $]a_i, a_{i+1}[$, $a_i \neq a_{i+1}$, and so its derivative there is negative, zero or positive, respectively, according whether i is strictly smaller, equal to, or strictly greater than $k/2$, respectively (the case $f'(u) = 0$ occurs only when k is even and $a_{k/2} \neq \neq a_{k/2+1}$). $\Diamond$

We are ready to state our main result; in practice, one has to check only *two* (compact) candidates to find a sequence at maximum rank distance from x:

Theorem 2. *Let* $x \in \mathcal{C}_i$ *be a binary sequence: then*

$$d(x,y) \leq \max\{d(x,\mathbf{01}), d(x,\mathbf{10})\}$$

for any sequence $y \in \mathcal{C}_i$.

Proof. Given sequences x and y, one can find a *parsing* of x and y into r corresponding infixes of lengths $\ell_1, \ell_2, \ldots, \ell_r$, $\sum_{1 \leq s \leq r} \ell_s = n$, such that for each couple of corresponding infixes the conditions of Lemma 2 are met. To see this, do as follows. Let the first position where x and y have unequal bits be j, say one has $x_j = 1$, $y_j = 0$. We proceed until we find a position h with $x_h = 0$, $y_h = 1$, with the latter 1 as yet unmatched (in other words, if ι is the index of that 1 in sequence y, position h, one has $\text{ord}(1_\iota|y) < \text{ord}(1_\iota|x)$; cf. Def. 1). As for the first infix set $\ell_1 = h - 1$. As for the second infix we proceed until we find a position h' with $x_{h'} = 1$ as yet unmatched, $y_{h'} = 0$; we set $\ell_1 + \ell_2 = h' - 1$; we proceed like this until the parsing is completed. Clearly, corresponding infixes have the same number of 1's and each of them verifies Lemma 2; the overall rank distance is additive over such infixes. We resort to an induction over the number r of infixes. If $r = 1$ Lemma 2 will do. Else, use the inductive assumption over the first $r - 1$ infixes and Lemma 2 on the last infix to obtain $d(x,z) \geq d(x,y)$ where z has one of the four patterns $\mathbf{0101}$, $\mathbf{1010}$, $\mathbf{010}$ or $\mathbf{101}$ (a boldface digit denotes an infix of the corresponding bit). As for the first two patterns, just use once more Lemma 2. Due to symmetry it will suffice to deal with the third pattern $z = \mathbf{010}$. Below the unknown u is an integer between 1 and $n-i$.

For $z = \mathbf{010}$, we have:

$$(2)\quad d(x,z) = 2\Big[\,|a_1 - u| + |a_2 - (u+1)| + \ldots + |a_i - (u+i)|\,\Big]$$

where the a_i are the positions occupied by the i bits 1 in x. Because of Lemma 3, the second side of (2) has a maximum either in $u = 1$ or in $u = n-1$. So, $d(x,y) \leq \max\{d(x,\mathbf{01}), d(x,\mathbf{10})\}$. $\Diamond$

As an example take the palindrome $x = 001100$; one has

$$d(x, 000011) = d(11000) = 8$$

and so this is the maximum for distances from x. However, there are other sequences at the same distance from x, just use Th. 1 to obtain by permutations inside the halves 000101, 000110, 101000 and 011000. As for the diameter $\delta(\mathcal{C})$ of $\mathcal{C}$, it is soon derived as a corollary, since $d(x,y) \leq d(x,\mathbf{01})$ (say) $\leq \max\{d(\mathbf{01},\mathbf{01}), d(\mathbf{01},\mathbf{10})\} = d(\mathbf{01},\mathbf{10})$:

Corollary 1. $\delta(\mathcal{C}) = 2n_0\,n_1 = 2n_1(n-n_1)$.

Actually, this result might have been derived also directly, because the contribution brought about by i bits 1 to the rank distance cannot exceed $i(n-i)$, and this does happen when the i bits 1 are compacted at the beginning of one sequence and at the end of the other sequence, respectively.

4. General sequences: a conjecture

Analogy to the binary case and computer simulations lead us to put forward the following conjecture (unfortunately, our proof for the binary case does not seem to generalise in a straightforward way):

Conjecture 1. *Given a sequence x in the composition class $\mathcal{C}$, if $\mathcal{P} \subset \mathcal{C}$ is the subset of compact sequences, then $\forall y \in \mathcal{C}$, $d(x,y) \leq \max_{z\in\mathcal{P}} d(x,z)$.*

Were the conjecture true, one would have to check at most $|\mathcal{A}|!$ candidates to find a sequence at maximum rank distance from x (at most, because some $\mathcal{A}$-letters might be lacking in x); we stress that the number $|\mathcal{A}|!$ does *not* depend on the sequence length n.

Now we move to the second problem, i.e. to diameters $\delta(C)$. After recalling that for n odd we have agreed that the letter in the central position is in neither of the two halves of length $\lfloor \frac{n}{2} \rfloor$, we give a definition:

Definition 4. The pivot of a compact sequence is the run of the same letter which occurs on both halves of the sequence.

Clearly, a composition class is balanced iff it contains a compact sequence with a *void* pivot. For example, the pivots of the compact sequences *aaabbb*, *aaabbbb*, *aaacbbb* are all void, and so the corresponding composition classes are all balanced. If x is a compact sequence, we denote by n_P the length of its pivot, by n_L and n_R the lengths of the prefix which precedes the pivot and the suffix which follows the pivot, respectively; $n_L + n_P + n_R = n$ (L stands for Left, R for Right). We set n_m and n_M equal to the minimum and the maximum out of n_L and n_R: $n_m = \min\{n_L, n_R\}, n_M = \max\{n_L, n_R\}$; $0 \le n_m \le \left\lfloor \dfrac{n}{2} \right\rfloor, 0 \le n_M \le \left\lceil \dfrac{n}{2} \right\rceil$. As an example, let x be a compact sequence, and let x^* be its mirror image; then:

$$
\begin{aligned}
(3) \qquad d(x, x^*) &= n_L(n_R + n_P) + n_R(n_L + n_P) + n_P|n_L - n_R| = \\
&= 2n_L n_R + n_P\big(n_L + n_R + |n_L - n_R|\big) = \\
&= 2n_M(n_m + n_P) = 2n_M(n - n_M).
\end{aligned}
$$

The last side in (3) is an increasing function of n_M for $n_M \le n/2$, and so the largest $d(x, x^*)$ is achieved by n_M as large as possible, without breaking the constraint $n_M \le \lceil \frac{n}{2} \rceil$. This soon gives for the diameter $\delta(\mathcal{C})$ of $\mathcal{C}$:

$$
(4) \qquad 2n^*(n - n^*) \;\le\; \delta(\mathcal{C}) \;\le\; \frac{n^2 - \langle n \text{ odd}\rangle}{2}\ , \quad n^* = \max n_M .
$$

A straightforward corollary of Conj. 1 would be that the diameter of a composition class $\mathcal{C}$ is exactly equal to the lower bound in (4), precisely as happens with binary sequences, cf. Cor. 1. Actually, the tightness of the lower bound, whatever the size of $\mathcal{A}$, can be proved also directly by generalising the alternative proof of Cor. 1 hinted at at the end of Sec. 3. Thus, the diameter $\delta(\mathcal{C})$ is achieved by compact sequences as in (3) with largest n_M as in (4), $n_M = n^*$.

Proposition 2. $\delta(\mathcal{C}) = 2n^*(n - n^*)$.

We sketch a proof for non-balanced classes, else there is nothing to prove. We assume n even, but this would soon be fixed. First: take a batch of complete letter-runs whose overall length ℓ fits into one of the two halves, $\ell \le n/2$. Then the maximal additive contribution that the corresponding letters can bring to $d(x, y)$ is obtained when those runs are at opposite ends of x and y, and is equal to $\ell(n-\ell)$; this is the maximum contribution ℓ letters can bring to $d(x,y)$, even if they are all distinct as in the "unconstrained case" of Th. 1. This is the case of the letter-runs corresponding to n_M; we choose n_M as large as possible, i.e. equal to n^*. Second: in two compact sequences as in (3), one of the two halves, the one corresponding to n_m, brings a contribution equal to $n^2/4$, which is as high as in the unconstrained case. Third, thinking of the other half, the one corresponding to n_M, the $n/2 - n_M$ letters of the pivot are unfortunately in the same half of x and x^*, but this is unavoidable, as an obvious arithmetic check soon shows. One soon checks that the best situation is when these $n/2 - n_M$ letters are as near the centre of the sequence as possible.

The following straightforward corollary of Prop. 2 is of independent interest (cf. also Prop. 1). Below x is a compact sequence as in (3) with largest n_M as in (4); a and b are any two distinct letters in $\mathcal{A}$:

Corollary 2. *Diameters $\delta(C)$ can be re-obtained with binary sequences: in a compact sequence x achieving the diameter, just replace the letters corresponding to $n_M = n^*$ by letter a and replace the letters corresponding to n_m and n_P by letter b. One has $d(\mathbf{ab}, \mathbf{ba}) = 2n^*(n - n^*)$.*

Acknowledgement. Research supported by FIRB-LIBI Bioinformatics, PRIN Large-scale development and MEdC-ANCS PNII-Idei project 228.

References

[1] CORMEN, Th. H., LEISERSON, Ch. E., RIVEST, R. and STEIN, Cl.: Introduction to Algorithms, 2nd ed., MIT Press and McGraw Hill, 2001.

[2] CSISZÁR, I. and KÖRNER, J.: Information Theory, Akadémiai Kiadó (Budapest) and Academic Press (New York), 1981.

[3] DIACONIS, P. and GRAHAM, R. L.: Spearman Footrule as a Measure of Disarray, *Journal of Royal Statistical Society. Series B Methodological* **39** (2) (1977), 262–268.

[4] DINU, L. P.: On the Classification and Aggregation of Hierarchies with Different Constitutive Elements, *Fundamenta Informaticae* **55** (1) (2003), 39–50.

[5] DINU, L. P.: Rank Distance with Applications in Similarity of Natural Languages, *Fundamenta Informaticae* **64** (1-4) (2005), 135–149.

[6] DINU, A. and DINU, L. P.: On the Syllabic Similarities of Romance Languages, in: A. Gelbukh (ed.), CICLing 2005. LNCS 3406, Lecture Notes in Computer Science 3406, 785–788, 2005.

[7] DINU, L. P. and MANEA, F.: An Efficient Approach for the Rank Aggregation Problem, *Theoretical Computer Science* **359** (1-3) (2006), 455–461.

[8] DINU, L. P. and SGARRO, A.: A Low-complexity Distance for DNA Strings, *Fundamenta Informaticae* **73** (3) (2006), 361–372.

[9] DINU, L. P., POPESCU, M. and DINU, A.: Authorship Identification of Romanian Texts with Controversial Paternity, *Proceedings 6-th LREC 2008*, Marrakech, Morocco, 2008.

[10] GAREY, M. R. and JOHNSON, D. S.: Computers and Intractability, W. H. Freeman and Company, New York, 2000.

[11] HIGUERA, C. de la and CASACUBERTA, F.: Topology of Strings: Median String is NP-complete, *Theoretical Computer Science* **230** (2000), 39–48.