DiSIT, Computer Science Institute
Università del Piemonte Orientale "A. Avogadro"
Viale Teresa Michel 11, 15121 Alessandria
http://www.di.unipmn.it

Minimum pattern length for short spaced seeds based on linear rulers

*L. Egidi, G. Manzini (lavinia.egidi@mfn.unipmn.it,
giovanni.manzini@mfn.unipmn.it)*

TECHNICAL REPORT TR-INF-2012-04-03-UNIPMN
(April 2012)

Research Technical Reports published by DiSIT, Computer Science Institute, Università del Piemonte Orientale are available via WWW at URL `http://www.di.unipmn.it/`.
Plain-text abstracts organized by year are available in the directory

Recent Titles from the TR-INF-UNIPMN Technical Report Series

2012-02 *Exploiting VM Migration for the Automated Power and Performance Management of Green Cloud Computing Systems*, C. Anglano, M. Canonico, M. Guazzone, April 2012.

2012-01 *Trace retrieval and clustering for business process monitoring*, G. Leonardi, S. Montani, March 2012.

2011-04 *Achieving completeness in bounded model checking of action theories in ASP*, L. Giordano, A. Martelli, D. Theseider Dupré, December 2011.

2011-03 *SAN models of a benchmark on dynamic reliability*, D. Codetta Raiteri, December 2011.

2011-02 *A new symbolic approach for network reliability analysis* , M. Beccuti, S. Donatelli, G. Franceschinis, R. Terruggia, June 2011.

2011-01 *Spaced Seeds Design Using Perfect Rulers*, L. Egidi, G. Manzini, June 2011.

2010-04 *ARPHA: an FDIR architecture for Autonomous Spacecrafts based on Dynamic Probabilistic Graphical Models*, D. Codetta Raiteri, L. Portinale, December 2010.

2010-03 *ICCBR 2010 Workshop Proceedings*, C. Marling, June 2010.

2010-02 *Verifying Business Process Compliance by Reasoning about Actions*, D. D'Aprile, L. Giordano, V. Gliozzi, A. Martelli, G. Pozzato, D. Theseider Dupré, May 2010.

2010-01 *A Case-based Approach to Business Process Monitoring*, G. Leonardi, S. Montani, March 2010.

2009-09 *Supporting Human Interaction and Human Resources Coordination in Distributed Clinical Guidelines*, A. Bottrighi, G. Molino, S. Montani, P. Terenziani, M. Torchio, December 2009.

2009-08 *Simulating the communication of commands and signals in a distribution grid*, D. Codetta Raiteri, R. Nai, December 2009.

2009-07 *A temporal relational data model for proposals and evaluations of updates*, L. Anselma, A. Bottrighi, S. Montani, P. Terenziani, September 2009.

2009-06 *Performance analysis of partially symmetric SWNs: efficiency characterization through some case studies*, S. Baarir, M. Beccuti, C. Dutheillet, G. Franceschinis, S. Haddad, July 2009.

2009-05 *SAN models of communication scenarios inside the Electrical Power System*, D. Codetta Raiteri, R. Nai, July 2009.

2009-04 *On-line Product Configuration using Fuzzy Retrieval and J2EE Technology*, M. Galandrino, L. Portinale, May 2009.

Minimum pattern length for short spaced seeds based on linear rulers

Lavinia Egidi and Giovanni Manzini

DiSIT, Sezione di Informatica
Università del Piemonte Orientale
{lavinia.egidi,giovanni.manzini}@unipmn.it

Abstract. We study the minimum pattern length for spaced seeds of the form $\mathbf{0}^{s_0} R_d \mathbf{0}^{s_1}$, with R_d a complete d-ruler and $\max(s_0, s_1) \le d$. We show how such minimum pattern length depends on the positions in which the integers $\le d$ are measured inside the ruler R_d.

1 Introduction

In this manuscript we analyze in detail the minimum pattern length of spaced seeds of the shape $\mathbf{0}^{s_0} R_d \mathbf{0}^{s_1}$, with R_d a complete linear d-ruler and $\max(s_0, s_1) \le d$. We show that these bounds depend heavily on the structure of the string R_d.

The results of this manuscript are a complement to the results in [1] to which the reader should refer for motivations and background.

2 Definitions and known results

The notion of *perfect ruler*, has been studied by mathematicians for more than sixty years [2, 4, 6] (in earlier works rulers were called *difference bases*). Here we recall the basic definitions using modern terminology [5]. We base the definition of rulers on the concept of *measure*:

Definition 1 (Measure). *Let U be a binary string. For any positive integer δ we say that U* measures *δ if there exist i, j, $0 \le i < j < |U|$, such that $j - i = \delta$ and $U[i] = U[j] = \mathbf{1}$. The pair (i, j) is said to be a* measure *of δ in U.* □

Definition 2 (Complete ruler). *Let R be a binary string of length $d+1$ such that $R[0] = \mathbf{1}$, $R[d] = \mathbf{1}$, and such that for any integer $\delta, 0 \le \delta \le d$, R measures δ. The string R is said to be a* complete d-ruler*, or simply a* complete ruler *when the length of R is clear from the context.* □

Intuitively, using the **1**'s as marks, with a complete d-ruler we can measure all distances between 1 and d. For example, the string **110101** is a complete 5-ruler. Note that even the string $\mathbf{1}^6 = \mathbf{111111}$ is a complete 5-ruler, but not an interesting one: the challenge of rule design is to find complete d-rulers with as few marks as possible. This notion is captured by the following definition.

Definition 3 (Perfect ruler). *Let R be a complete d-ruler containing ℓ $\mathbf{1}$'s. If there exists no complete d-ruler with less than ℓ $\mathbf{1}$'s then R is said to be a* perfect d-ruler*.* □

Tables of all perfect rulers of size up to 101 are available on the net [5].

The structure of complete rulers naturally suggests their use for the design of spaced seeds. Given a d-ruler R, if we replace each $\mathbf{0}$ with a '#' symbol and each $\mathbf{1}$ with a '-' symbol we obtain a seed in which there is a pair of don't care symbols at distance δ for $\delta = 1, \ldots, d$. This seed solves the $(m, 2)$-problem for $m \geq 2d + 1$. However, this is not the only seed we can derive from R. For any pair s_0, s_1 the seed derived from the string $\mathbf{0}^{s_0} R \mathbf{0}^{s_1}$ also has pairs of don't care symbols at distance δ for $\delta = 1, \ldots, d$. Hence, it solves the $(m, 2)$-problem for a sufficiently large m. Clearly there is a trade-off here: the larger are s_0 and s_1 the higher is the weight of the corresponding seed (a good thing) and the larger is the value m for which the seed solves the $(m, 2)$-problem (a bad thing).

To evaluate to what extent rulers are useful for seed design it is clearly necessary to investigate this trade-off. In this section we give upper bounds to the minimum m for which the seed associated to the string $\mathbf{0}^{s_0} R \mathbf{0}^{s_1}$ solves the $(m, 2)$-problem. The results of this section are valid for any complete d-ruler R.

Since the main object of our study are rulers, for simplicity we will only work with strings over the alphabet $\{\mathbf{0}, \mathbf{1}\}$, with the *implicit* associations[1] $\mathbf{0} \rightarrow$ `'#'`, $\mathbf{1} \rightarrow$ `'-'`. We introduce Definition 4 and Lemma 1 that essentially restate known properties of seeds in the language of strings over the alphabet $\{\mathbf{0}, \mathbf{1}\}$. In the following we state these properties for any k, even if in this manuscript we are only concerned with the case

Definition 4 (Completeness). *A binary string P is (m, k)-complete if, for any length-m binary string V containing exactly k $\mathbf{1}$'s, there exists at least an index t, with $0 \leq t \leq |V| - |P|$, such that for $i = 0, \ldots, |V| - 1$, it is*

$$V[i] = \mathbf{1} \quad \Longrightarrow \quad (i - t < 0) \ \vee \ (i - t \geq |P|) \ \vee \ (P[i - t] = \mathbf{1}). \tag{1}$$

If (1) holds we say that $P + t$ matches in V, or that P shifted by t matches in V. □

Note that $P + t$ matches in V if the $\mathbf{1}$'s in V are either outside $P + t$ or correspond to a $\mathbf{1}$ in $P + t$. Equivalently, there is no $\mathbf{1}$ in V corresponding to a $\mathbf{0}$ in $P + t$.

Lemma 1. *The binary string P is (m, k)-complete if and only if the spaced seed obtained with the map $\mathbf{0} \rightarrow$ `'#'`, $\mathbf{1} \rightarrow$ `'-'` solves the (m, k)-problem.* □

Having stated Lemma 1, in the rest of this manuscript most of the results will simply establish that certain binary strings are, or are not, (m, k)-complete, without even mentioning the immediate consequence that the corresponding seeds solve, or do not solve, the (m, k)-problem.

[1] Unfortunately, this is the opposite of [3], where $\mathbf{0}$ corresponds to a don't care symbol.

Definition 5 (Minimum pattern length m_P^*). *Given a binary string P we denote by m_P^* the smallest integer m such that P is $(m, 2)$-complete.*[2] □

In [1] it is proven the following upper bound on the minimum pattern length for a seed P obtained from a d-ruler R_d.

Theorem 1 (see [1]). *Let $P = \mathbf{0}^{s_0} R \mathbf{0}^{s_1}$ where R is a complete d-ruler. If $\max(s_0, s_1) \leq d$, then $m_P^* \leq 2|P| - 1 - \min(s_0, s_1)$.*

We introduce a specific notation for the upper bound of Theorem 1, for future reference:

Definition 6 (Upper bound m_P). *For any string $P = \mathbf{0}^{s_0} U \mathbf{0}^{s_1}$, we denote by m_P the value $m_P = 2|P| - 1 - \min(s_0, s_1)$.* □

The above upper bound is valid for any d-ruler R_d. In this paper we address the question of whether this upper bound is tight.

3 Analysis of the minimum pattern length

Table 1 shows that the upper bound of Theorem 1 is not always tight. In Table 1 we compare it with the actual minimum pattern length for patterns $P = \mathbf{0}^s R_d \mathbf{0}^s$, for some values of d and some $s \leq d$. These values of the minimum pattern length are computed by direct inspection. The first column in the table gives the value of d, the second specifies the ruler for which the minimum pattern length is computed, the third column gives the value of s, the fourth reports the upper bound m_P from Theorem 1, the fifth gives the minimum pattern length m_P^*, and the last one gives the difference $m_P - m_P^*$ for quick reference. In the table we only report m_P^* for the values of $s \leq d$ for which $m_P^* < m_P$. For values of $s \leq d$ larger than those reported for each d, $m_P^* = m_P$.

In view of the values of m_P^* in Table 1, it is interesting to establish in which cases the upper bound m_P is tight, and whether there are seeds of the form $P = \mathbf{0}^{s_0} R_d \mathbf{0}^{s_1}$ for which m_P^* is significantly smaller than m_P. In [1] lower bounds for m_P^* are established on the basis on a property of the ruler R_d called its skewness, which is based on the positions of measures of integers δ in R_d.

In the next section we prove an exact relation between the positions of $\mathbf{1}$'s in a ruler and the minimum pattern length for the derived seed.

We need an extended notion of measure of a given integer δ. We will consider, along with proper measures of δ also additional ordered pairs $(a, a + \delta)$ that have one or even both endpoints ouside of P. We will then show that the maximum distance, taken over all δ's, between two consecutive "measures" (in this extended sense) of a δ, determines the minimum pattern length.

[2] m_P^* also depends on k, but since in this manuscript we treat uniquely the case $k = 2$, k does not appear in m_P^* to make the notation less cumbersome.

d	R_d	s	m_P	m^*_P	$m_P - m^*_P$
6	**1100101**	0	13	12	1
11	**110000110101**	0	23	21	2
		1	26	23	3
		2	29	27	2
		3	32	30	2
12	**1100000110101**	0	25	23	2
		1	28	25	3
		2	31	29	2
		3	34	32	2
13	**11100010001001**	0	27	26	1
		1	30	29	1
14	**110001001010101**	0	29	28	1
15	**1100000011010101**	0	31	29	2
		1	34	31	3
		2	37	33	4
		3	40	36	4
		4	43	41	2
		5	46	44	2
16	**11000000011010101**	0	33	31	2
		1	36	33	3
		2	39	35	4
		3	42	38	4
		4	45	43	2
		5	48	46	2
17	**110000001001010101**	0	35	34	1

Table 1. A comparison of the upper bound m_P from Theorem 1 and of the actual minimum pattern length m^*_P, computed by direct inspection, for $P = \mathbf{0}^s R_d \mathbf{0}^s$, for some values of d and $s \leq d$.

Definition 7 (Maximum gap Γ_δ). *For any binary string P and $1 \leq \delta \leq |P|+1$, let $0 \leq a_1 < a_2 < \ldots < a_{k-1} < |P| - \delta$ ($k \geq 1$) be such that $\langle a_i, a_i + \delta \rangle$ are all the linear measures of δ in P. If P doesn't measure δ, for example when $\delta > |P|$, then $k = 1$, and no a_i defined as above exists.*

In addition, let $a_{-\ell+1} < a_{-\ell+2} < \ldots < a_0 < 0$ ($\ell \geq 0$) be all the integers $-\delta \leq a_i < 0$ such that $P[a_i + \delta] = \mathbf{1}$, and let $a_{-\ell} = -\delta - 1$. If no $a_i < \delta$ such that $P[a_i] = \mathbf{1}$ exists, then $\ell = 0$ and $a_0 = -\delta - 1$.

Finally, let $a_k < a_{k+1} < \ldots < a_{k+r-1}$ ($r \geq 0$) be all the integers a_i, $|P| - \delta \leq a_i < |P|$ such that $P[a_i] = \mathbf{1}$, and let $a_{k+r} = |P|$. If no $|P| - \delta < a_i < |P|$ with $P[j] = \mathbf{1}$ exists, then $r = 0$ and $a_k = |P|$.

For each $1 \leq \delta \leq m$, define the maximum gap between measures as

$$\Gamma_\delta = \max_{-\ell \leq i \leq k+r-1} (a_{i+1} - a_i).$$

□

Informally, $(a_i, a_i + \delta)$ for $0 < i < k$ are all the usual measures of δ with both endpoints inside P, and we consider $(a_i, a_i + \delta)$, for $i \leq 0$ and $i \geq k$, as additional measures of δ with one or both endpoints outside P. The maximum gap Γ_δ denotes the maximum distance of two measures of δ in P. Notice that $a_{-\ell} = -\delta - 1$ and $a_{k+r} = |P|$ always exist, even if δ is not measured in P. Therefore, Γ_δ is well defined for any δ. Using the values Γ_δ for $\delta = 1, \ldots, |P|+1$, we can compute the exact value of the minimum pattern length m^*_P.

Theorem 2. *Let P be any binary string. Let Γ_δ be defined for each $1 \leq \delta \leq |P|+1$ as in Definition 7. Then, the minimum m such that P is $(m, 2)$-complete is*

$$m^*_P = |P| + \max_{1 \leq \delta \leq |P|+1} \Gamma_\delta - 1.$$

Proof. We first prove that P is $(m^*_P, 2)$-complete. Let V be a binary string of length m^*_P and with two $\mathbf{1}$'s in the positions v_1 and v_2. Let $\delta = v_2 - v_1$. If $\delta > |P| + 1$ we have that $P + t$ matches in V for $t = v_1 + 1$.

If $\delta \leq |P| + 1$, let $\{a_i | i = -\ell, \ldots, k + r\}$ denote the starting points of the measures of δ defined as in Definition 7. If $v_1 = a_i$ for some $-\ell \leq i \leq h + r$, then each v_i is either outside P or $P[v_i] = \mathbf{1}$, by definition of the a_i's. Then P matches in V. Similarly, if $v_1 > a_{h+r} = |P|$ then also $v_2 > |P|$, and P matches in V. Finally, if there exists a_i such that $a_i < v_1 < a_{i+1}$ we have that $P + t$ matches in V for $t = v_1 - a_i$. Since $|V| = |P| + \max_{1 \leq \delta \leq |P|+1} \Gamma_\delta - 1$, and $t \leq a_{i+1} - a_i - 1 \leq \Gamma_\delta - 1$, then $t \leq |V| - |P|$ so t is admissible.

In order to prove minimality, let δ' be such that $m^*_P = |P| + \Gamma_{\delta'} - 1$. let $\{a_i | i = -\ell, \ldots, k + r\}$ denote the starting points of the measures of δ' defined as in Definition 7. Let j be such that $a_{j+1} - a_j = \Gamma_{\delta'}$.

Consider the binary string V of length $m^*_P - 1 = |P| + \Gamma_{\delta'} - 2$, with exactly two $\mathbf{1}$'s, in positions $v_1 = a_{j+1} - 1$ and $v_2 = v_1 + \delta'$. By construction, the minimum value of t for which $P + t$ matches in V would be $t = a_{j+1} - a_j - 1 = \Gamma_{\delta'} - 1$. But since $|V| = |P| + \Gamma_{\delta'} - 2$, such value of t is not admissible, and therefore P is not $(m^*_P - 1, 2)$-complete. □

Theorem 2 gives an insight on the positions of $\mathbf{1}$'s in seeds that have specific completeness properties. We first notice how it implies that the bounds on s_0 and s_1 are necessary in Theorem 1:

Corollary 1. *Let $P = \mathbf{0}^{s_0} R_d \mathbf{0}^{s_1}$, with R_d a complete d-ruler. If $\min(s_0, s_1) > d$, then $m^*_P \geq 2|P| + \min(s_0, s_1)$. If, on the other hand, $\min(s_0, s_1) \leq d$ but $max(s_0, s_1) > d$, then $m^*_P \geq 2|P|$.*

Proof. Without loss of generality, let $s_0 \geq s_1$.

If both $s_0 \geq s_1 > d$, then consider a_0 and a_1 defined as in Definition 7 for $\delta = s_1$ (since $\delta > d$ there are no measures of δ inside P). Since $a_0 = -s_1 - 1$ and $a_1 = |P|$, $\Gamma_{s_1} \geq a_1 - a_0 = |P| + s_1 + 1$ and, by Theorem 2, $m^*_P \geq 2|P| + s_1$ as claimed.

On the other hand, if $s_0 > d \geq s_1$, then a_0 and a_1 defined according Definition 7 for $\delta = d + 1$ are $a_0 = -d - 2$ and $a_1 = s_0 + d$. Then, $\Gamma_d \geq a_1 - a_0 = s_0 + 2d + 2$. Therefore, $m^*_P \geq |P| + s_0 + 2d + 1 \geq 2|P|$, since $s_1 \leq d$. □

As another consequence of Theorem 2, we show that the upper bound m_P is tight for $P = \mathbf{0}^d R_d \mathbf{0}^d$:

Corollary 2. *Let $P = \mathbf{0}^d R_d \mathbf{0}^d$, with R_d a complete d-ruler. Then $m_P^* = m_P$.*

Proof. In the hypotheses given, a_0 and a_1 for $\delta = d$ are $a_0 = -d-1$ and $a_1 = d$. Then, $\Gamma_d \geq a_1 - a_0 = 2d+1$ and, by Theorem 2, $m_P^* \geq |P| + 2d$. The thesis follows since by Theorem 1 it is $m_P^* \leq 2|P| - 1 - d = |P| + 2d$. □

The next result shows that the upper bound of Theorem 1 is tight also if a small integer has a unique measure in R_d, which is at one endpoint of R_d:

Corollary 3. *Let $P = \mathbf{0}^d R_d \mathbf{0}^d$, with R_d a complete d-ruler.*

If $s_1 = \min(s_0, s_1)$ and there exists $\delta \leq s_0$ that has in R_d the unique measure $(d-\delta, d)$, then $m_P^* = m_P$.

If $s_0 = \min(s_0, s_1)$, and there exists $\delta \leq s_1$ that has in R_d the unique measure $(0, \delta)$, then $m_P^* = m_P$.

Proof. Let $s_1 = \min(s_0, s_1)$, and δ have in R_d the unique measure $(d-\delta, d)$. Applying Definition 7 to δ, it is $a_0 = -\delta - 1$ and $a_1 = s_0 + d - \delta$. Then, $\Gamma_\delta \geq a_1 - a_0 = s_0 + d + 1$ and, by Theorem 2, $m_P^* \geq |P| + s_0 + d = 2|P| - 1 - \min(s_0, s_1) = m_P$.

Similarly, if $s_0 = \min(s_0, s_1)$ and the unique measure of a $\delta \leq s_1$ is $(0, \delta)$, applying Definition 7 to δ, it is $a_1 = s_0$ and $a_2 = |P|$. Then, $\Gamma_\delta \geq |P| - s_0$ and, by Theorem 2, $m_P^* \geq 2|P| - s_0 - 1 = 2|P| - 1 - \min(s_0, s_1) = m_P$. □

In view of these results, let us analyze some of the data from Table 1.

Example 1. As remarked above, in Table 1, we only listed for each ruler those values of s for which $m_P^* < m_P$. This means that for the specific d-rulers listed for $d = 6$, $d = 14$ and $d = 17$, it is $m_P^* = m_P$ already for $s_0 = s_1 = 1$. Indeed notice that in these three rulers $\delta = 1$ has only one measure at the very beginning of the ruler. Therefore, by Corollary 3, for $s_1 \geq 1$, $m_P^* = m_P$. In all other rulers listed in Table 1, the value $\delta = 1$ has more than one measure. □

Example 2. The 13-ruler of Table 1 has a unique measure for $\delta = 2$ in $(0, 2)$. Accordinglyi, by Corollary 3, it is $m_P^* = m_P$ for $s_1 \geq 2$. □

Example 3. Consider the 11-ruler $R_{11} = \mathbf{110000110101}$. The smallest integer that has a unique measure at one endpoint of R_{11} is $\delta = 4$, whose measure is $(7, 11)$.

For $s_0 = s_1 = 0$, the extended measures for $\delta = 4$ are $a_{-2} = -5$, $a_{-1} = -4$, $a_0 = -3$, $a_1 = 7$, $a_2 = 9$, $a_3 = 11$ and $a_4 = 12$. Then, $\Gamma_4 \geq 10$ and $m_P^* \geq |P| + \Gamma_4 - 1 \geq 21$. It can be checked by direct inspection that indeed $\Gamma_4 = \max_{1 \leq \delta \leq |P|+1} \Gamma_\delta$.

For $s_0 = s_1 = 1$, Γ_4 is again the maximum among all Γ_δ's and $\Gamma_4 = a_1 - a_0 = 10$ since $a_{-2} = -5$, $a_{-1} = -3$, $a_0 = -2$, $a_1 = 8$, $a_2 = 10$, $a_3 = 12$ and $a_4 = 14$. Here $|P| = 14$ and thus $m_P^* = 23$.

For $s_0 = s_1 = 2$, $\Gamma_4 = 10$ once again, but this time it is not the maximal Γ_δ. The reason is that, with two leading zeroes, the extended measures of $\delta = 2$ are now further apart: $a_0 = -3$, $a_1 = 9$, $a_2 = 11$ and $a_3 = 16$. Therefore, $\Gamma_2 = 12$, and since $|P| = 16$, $m_P^* = 27$. Yet, since 2 has two measures, again Corollary 3 does not apply.

For $s_0 = s_1 = 3$, the argument is similar as for $s_0 = s_1 = 3$: $\Gamma_4 = \Gamma_2 = 13$ are maximal, but not large enough to have $m_P^* = m_P$, because 4 is still larger than s_0 and 2 has two measures.

Finally, for $s_0, s_1 \geq 4$, by Corollary 3, $m_P^* = m_P$. Indeed, 4 is now smaller than s_0 and has a unique measure at the end of R_{11}. Now $a_0 = -5$ and $a_1 = 11$ are the starting points of the extended measures of 4 that are furthest apart, so $\Gamma_4 = 16$. Notice that now $a_1 - a_0 = s_0 + d$. □

References

1. Lavinia Egidi and Giovanni Manzini. Spaced seeds design using perfect rulers. In Roberto Grossi, Fabrizio Sebastiani, and Fabrizio Silvestri, editors, *SPIRE*, volume 7024 of *Lecture Notes in Computer Science*, pages 32–43. Springer, 2011.
2. P. Erdós and I. S. Gál. On the representation of $1, 2, \ldots, n$ by differences. *Indagationes Math.*, 10:379–382, 1948.
3. Martin Farach-Colton, Gad M. Landau, Süleyman Cenk Sahinalp, and Dekel Tsur. Optimal spaced seeds for faster approximate string matching. *J. Comput. Syst. Sci.*, 73(7):1035–1044, 2007.
4. J. Leech. On the representation of $1, 2, \ldots, n$ by differences. *J. London Math. Soc.*, 31:160–169, 1956.
5. Peter Luschny. Perfect and optimal rulers, 2003. http://www.luschny.de/math/rulers/prulers.html.
6. B. Wichmann. A note on restricted difference bases. *J. London Math. Soc.*, 38:465–466, 1962.