Dipartimento di Informatica
Università del Piemonte Orientale "A. Avogadro"
Viale Teresa Michel 11, 15121 Alessandria
http://www.di.unipmn.it

Spaced Seeds Design Using Perfect Rulers

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

TECHNICAL REPORT TR-INF-2011-06-01-UNIPMN
(June 2011)

The University of Piemonte Orientale Department of Computer Science Research Technical Reports 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

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.

2009-03 *A GSPN Semantics for Continuous Time Bayesian Networks with Immediate Nodes*, D. Codetta Raiteri, L. Portinale, March 2009.

2009-02 *The TAAROA Project Specification*, C. Anglano, M. Canonico, M. Guazzone, M. Zola, February 2009.

2009-01 *Knowledge-Free Scheduling Algorithms for Multiple Bag-of-Task Applications on Desktop Grids*, C. Anglano, M. Canonico, February 2009.

2008-09 *Case-based management of exceptions to business processes: an approach exploiting prototypes*, S. Montani, December 2008.

2008-08 *The ShareGrid Portal: an easy way to submit jobs on computational Grids*, C. Anglano, M. Canonico, M. Guazzone, October 2008.

2008-07 *BuzzChecker: Exploiting the Web to Better Understand Society*, M. Furini, S. Montangero, July 2008.

Spaced Seeds Design Using Perfect Rulers*

Lavinia Egidi and Giovanni Manzini

Dipartimento di Informatica, Università del Piemonte Orientale, Italy.
{lavinia.egidi,giovanni.manzini}@mfn.unipmn.it

Abstract. We consider the problem of lossless spaced seed design for approximate pattern matching. We show that, using mathematical objects known as perfect rulers, we can derive a family of spaced seeds for matching with up to two errors. We analyze these seeds with respect to the trade-off they offer between seed weight and the minimum length of the pattern to be matched. We prove that for patterns of length up to a few hundreds our seeds have a larger weight, hence a better filtration efficiency, than the ones known in the literature. In this context, we study in depth the specific case of *Wichmann rulers* and prove some preliminary results on the generalization of our approach to the larger class of *unrestricted rulers*.

1 Introduction

The use of spaced seeds for approximate pattern matching has been introduced in [1, 11] and since then has received considerable attention. Spaced seeds are used to quickly filter-out highly dissimilar regions, and they are a fundamental tool, for example, for mapping to a reference genome the millions of reads produced by modern sequencing technologies (see [8] and references therein).

We consider the problem of designing spaced seeds to be used for detecting whether two strings of length m are at Hamming distance at most k; in the literature this is known as the (m, k)-detection problem. In particular we are interested in the design of lossless seeds, i.e., seeds that find *all* matches with the above properties. Spaced seeds consists of *solid* positions and *don't care* positions. The number of solid positions is called the seed *weight*. For a given pair of values (m, k) we want to find a seed with the largest possible weight since, under standard assumptions, this maximizes the filtration efficiency.

For the problem of lossless seed design, an important breakthrough has been obtained in [3] where, for any given pair (m, k), the authors provide a spaced seed with an asymptotically optimal weight. Although this result essentially solves the problem from the theoretical point of view, it remains open the problem of finding optimal, i.e. weight-maximal, seeds for the pattern lengths m used in practice (i.e. up to a few hundreds): the seeds in [3] are asymptotically optimal as $m \to \infty$, but we have no guarantees on their quality for small m. For a given

* This research is founded by the BioBITS Project *Converging Technologies* 2007, area: Biotechnology-ICT, Regione Piemonte.

pair (m, k) one can find an optimal seed using a combinatorial search algorithm, but this problem is known to be a hard one [4, 5, 7, 10, 13] so the problem of designing whole families of (suboptimal) seeds of practical interest is still open.

Our starting point is the observation that we can derive a family of lossless spaced seeds using mathematical objects known as *perfect rulers* (sometimes also called *difference bases*) [2, 6, 14]. Informally, a perfect d-ruler is a binary string with a minimal number of **1**'s with the property that for any positive $\delta \leq d$ there exist two **1**'s at distance δ (see Section 2 for further details). This structural property makes them suitable to design spaced seeds able to detect strings at Hamming distance at most 2. The study of the properties of these spaced seeds is the main objective of the paper.

As a first step, in Section 3 we analyze the seeds obtained from perfect rulers with respect to the tradeoff they offer between seed weight and the minimum m^* for which the seed is guaranteed to solve the $(m, 2)$-detection problem for all $m \geq m^*$. In Theorem 1 we establish an upper bound for m^* for all "interesting" seeds derived from perfect rules. This upper bound suffices to establish that for m up to 498 the seeds derived from perfect rulers have a larger weight than the asymptotically optimal seeds defined in [3] and therefore justifies an in-depth study of this family of seeds.

In Section 4 we refine our analysis by establishing *lower bounds* on the minimum pattern length m^*. First, we prove that the upper bound of Theorem 1 is tight for the seed of maximal weight derived from a given ruler (Corollary 1). Then, we introduce the concept of skewness of a ruler and use it to derive general lower bounds for m^* (Theorem 3).

In Section 5, we analyze the special case of Wichmann rulers [14] which are a family of rulers particularly important since they can be easily derived by a "generating function", whereas other perfect rulers are usually found by trial and error. For Wichmann rulers we show that the upper bound of Theorem 1 is almost tight applying the results of Section 4 (Theorem 4) and with an ad-hoc analysis (Theorem 5).

Finally, in Section 6 we consider spaced seeds obtained from *unrestricted* rulers [6], which are a natural generalization of perfect rulers. We show that some of the results of the previous sections can be applied to unrestricted rulers as well. Although we do not provide a complete analysis, our preliminary results show that, somewhat counterintuitively, spaced seeds derived from unrestricted rulers are less effective that the ones derived from perfect rulers.

2 Notation

For spaced seeds we follow the notation introduced in [1, 5]. A *spaced seed* is a string over the alphabet $\{\texttt{\#}, \texttt{-}\}$; the symbol '`#`' represents a solid position, the symbol '`-`' a don't care position. Informally, a spaced seed defines a set of non-contiguous positions in which we require two sequences to match. We say that a spaced seed S solves the (m, k)-problem if for any pair of strings σ_1, σ_2 of length

m and Hamming distance k, there exists an index i such that

$$S[j] = \texttt{\#} \qquad \Longrightarrow \qquad \sigma_1[i+j] = \sigma_2[i+j]. \tag{1}$$

In other words, we require that, starting from position i, the strings σ_1, σ_2 contain the same symbols in every position corresponding to a `'#'` in S, while we tolerate mismatches in positions corresponding to `'-'` in S.

In the context of approximate string matching, if a seed solves the (m, k) problem it can be used as a filter to quickly discard regions which *are not* at Hamming distance at most k (see [1, 5] for further details). Note however, that (1) can hold for a given i even if σ_1 and σ_2 are not at Hamming distance k. These events are called *false positive matches* and it is desirable to reduce their number as much as possibile. The *weight* of a seed is defined as the number of `#`'s in it. Under standard assumptions, see [3, Sect. 1.1] the number of false positives decreases exponentially with the seed weight. Thus, it is desirable to solve the (m, k) problem with a seed with the largest possible weight.

The notion of *perfect ruler*, has been studied by mathematicians for more than sixty years [2, 6, 14] (in earlier works rulers were called *difference bases*). Here we recall the basic definitions using modern terminology [9]. 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, $0 \leq i < j < |U|$, 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 $\delta, 0 \leq \delta \leq d$, 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 **1**'s. If there exists no complete d-ruler with less than ℓ **1**'s then R is said to be a* perfect d-ruler. □

Let $\ell(d)$ denote the number of **1**'s in a perfect d-ruler. Table 1 reports the values $\ell(d)$ for $d = 10, \ldots, 90$. In [14] it is proven that $\lim_{d\to\infty}(\ell^2(d)/d)$ exists and that such limit is between 2.434 and 3. Perfect rulers are not easy to find: they are usually generated by exhaustive search procedures. An important exception are Wichman rulers which are discussed in Section 5. Tables of all perfect rulers of size up to 101 are available on the net [9].

Perfect rulers should not be confused with Golomb rulers that measure each integer at most once (they are not necessarily complete). Golomb rulers have been used in [12, 13] in relation to seed design, but with the totally different aim of analyzing the hardness of seed optimization.

d	10–13	14–17	18–23	24–29	30–36	37–43	44–50	51–58	59–68	69–79	80–90
$\ell(d)$	6	7	8	9	10	11	12	13	14	15	16

Table 1. Number of **1**'s $\ell(d)$ in a perfect d-ruler for $d = 10, \ldots, 90$.

3 From rulers to spaced seeds

The structure of complete rulers naturally suggests their use for the design of spaced seeds. Given a d-ruler R, if we replace each **0** with a '`#`' symbol and each **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. However, since seeds of higher weight are preferable, it is natural to derive seeds from rulers with the minimum number of **1**'s, that is, from *perfect rulers*.

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 $\mathbf{0} \rightarrow$ '`#`', $\mathbf{1} \rightarrow$ '`-`'. We introduce Definition 4 and Theorem 1 that essentially restate known properties of seeds in the language of strings over the alphabet $\{\mathbf{0}, \mathbf{1}\}$.

Definition 4 (Completeness). *A binary string P is (m, k)-complete if, for any length-m binary string V containing exactly k **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*

$$V[i] = \mathbf{1} \quad \Longrightarrow \quad (i - t < 0) \ \vee \ (i - t \geq |P|) \ \vee \ (P[i - t] = \mathbf{1}). \tag{2}$$

If (2) 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 **1**'s in V are either outside $P + t$ or correspond to a **1** in $P + t$. Equivalently, there is no **1** in V corresponding to a **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 the paper 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.

Definition 5 (Minimum length m_P^*). *Given a binary string P we denote by m_P^* the smallest integer m such that P is $(m,2)$-complete.*[1] □

The following theorem provides an upper bound for m_P^* for $P = \mathbf{0}^{s_0} R \mathbf{0}^{s_1}$ when R is a complete d-ruler with ℓ $\mathbf{1}$'s, and $\max(s_0, s_1) \leq d$. Since the seed associated to $\mathbf{0}^{s_0} R \mathbf{0}^{s_1}$ has weight $s_0 + s_1 + d + 1 - \ell$ the theorem establishes a trade-off between seed weight and minimum pattern length m_P^*.

Theorem 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 $m_P^* \leq 2|P| - 1 - \min(s_0, s_1)$.*

Proof. To prove the theorem we show that P is $(m,2)$ complete for $m = 2|P| - 1 - \min(s_0, s_1)$. Without loss of generality we assume that $s_0 \geq s_1$ (if this is not the case consider P^R, i.e. the string P reversed).

Let V any length-m binary string containing exactly two ones, in the positions v_1, v_2 ($0 \leq v_1 < v_2 \leq m-1$). We need to show that for any such pair v_1, v_2 we can find a shift t, $0 \leq t \leq m - |P|$, such that $P + t$ matches in V. By construction we have $|P| = d+1+s_0+s_1$ and $m = 2(d+1+s_0+s_1)-1-s_1 = 2d+2s_0+s_1+1$. Hence the admissible range for t is $0 \leq t \leq m - |P| = d + s_0$.

Note that in our setting the condition in Definition 4 is equivalent to require that there exists a shift t, $0 \leq t \leq d + s_0$ such that for $i = 1, 2$

$$(v_i < t) \ \vee \ (v_i \geq t + |P|) \ \vee \ (P[v_i - t] = \mathbf{1}) \tag{3}$$

We consider the following four cases, according to the size of $\Delta v = v_2 - v_1$.

Case $\Delta v > |P|$. In this case $t = v_1 + 1$ is in the admissible range, since $t + |P| = v_1 + 1 + |P| \leq v_2 < m$; moreover, $v_1 < t$ and $v_2 \geq |P|$.

Case $\Delta v \leq d$. The positions of v_1 and v_2 in V define here three subcases:

If $v_1 \geq |P|$, $t = 0$ satisfies (3), since both $v_1, v_2 \geq t + |P|$.

Symmetrically, if $v_1 < s_0 + (d - \Delta v)$, then $t = v_2 + 1$ satisfies (3).

If $s_0 + (d - \Delta v) \leq v_1 < |P|$, then $V[v_1]$ and $V[v_2]$ can fall within the d-ruler inside P, but P must be shifted so that both $V[v_i]$'s match $\mathbf{1}$'s in P . By hypothesis, P contains a measure (m_1, m_2) of $\Delta v \leq d$, with $s_0 \leq m_1 < m_2 < |P| - s_1$. Then (3) holds for $t = v_1 - m_1$, since $v_1 - t = m_1$ and $v_2 - t = m_2$. Notice that $m_1 \leq v_1$ since $m_2 \leq s_0 + d + 1$ and $m_2 - m_1 = \Delta v$. Finally, t is in the admissible range since $t = v_1 - m_1 \leq |P| - 1 - s_0 = d + s_1 \leq d + s_0 = m - |P|$.

Case $d + s_0 < \Delta v \leq |P|$.

If $v_1 \geq s_0$, (3) is satisfied by $t = v_1 - s_0$.

If $v_1 < s_0$, then $m - v_2 > d \geq s_1$ and $t = v_2 - (s_0 + d)$ satisfies (3).

Case $d < \Delta v \leq d + s_0$. Since the distance between v_1 and v_2 is at most $d + s_0$, the only possible alignments include: matching $V[v_1]$ with a $\mathbf{1}$ strictly after the first one in P or, symmetrically, $V[v_2]$ with a $\mathbf{1}$ strictly before the last one, besides the trivial cases in which all of P fits before or after both $V[v_i]$'s.

If $v_1 \geq |P|$ then $t = 0$ satisfies (3); if $v_2 < m - |P|$ then $t = v_2 + 1$ does.

If $s_0 + d \leq v_1 < |P|$, then $t = v_1 - (s_0 + d)$ satisfies (3). See Fig. 1.(a).

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

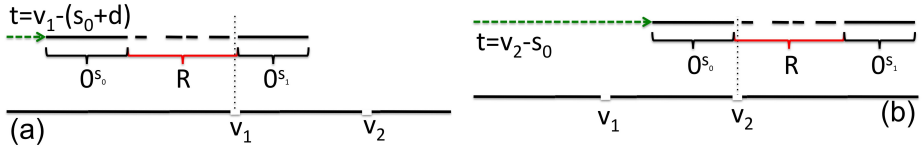

Fig. 1. Case $d < \Delta v \leq d+s_0$, (a): $s_0+d \leq v_1 < |P|$, (b): $m-|P| \leq v_2 < m-(s_1+d+1)$.

If $m - |P| \leq v_2 < m - (s_1 + d + 1)$, $t = v_2 - s_0$ satisfies (3) –see Fig. 1.(b).

The remaining case we must consider is $v_1 < s_0+d$ and $v_2 \geq m-(s_1+d+1)$, still with $d < \Delta v \leq d + s_0$. That is, $V[v_1]$ and $V[v_2]$ are so well centered, and enough spaced apart, that there is no room before $V[v_1]$ or after $V[v_2]$ to allow aligning either one of them with the first or last **1**'s in P.
Since $d < \Delta v \leq d + s_0$ there exists an r, $0 \leq r < s_0$, such that $\Delta v = d + s_0 - r$. Notice that $v_1 < s_0 + d$ together with the constraint on Δv implies $v_2 < m - (s_1 + 1)$, and $v_2 \geq m - (s_1 + d + 1)$ implies $v_1 \geq s_0$. Thus, $v_1 = s_0 + k$ and $v_2 = m - (s_1 + 1 + h)$ for some $0 \leq k < d$ and $0 < h \leq d$, with $h + k = d + r$.
Since R is a complete ruler, $r + 1 \leq d$ must have a measure (w_1, w_2) in R. Then, if $s_0 + k \geq s_0 + w_2$, $t = v_1 - (s_0 + w_2)$ is an admissible choice for t since $0 \leq v_1 - (s_0 + w_2) < s_0 + d \leq m - |P|$. Moreover (3) is satisfied since $v_1 - t = s_0 + w_2$ and $v_2 \geq t + |P|$ (since $|P| - (s_0 + w_2) < d + s_1 - r \leq \Delta v$). See Fig. 2.(a).

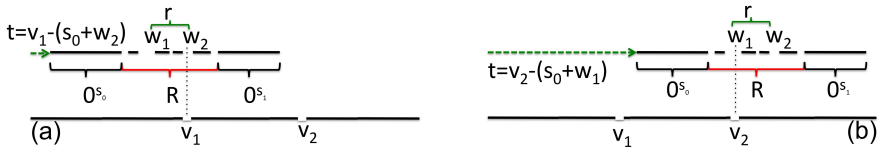

Fig. 2. Case $d < \Delta v \leq d + s_0$, $s_0 \leq v_1 < s_0 + d$, $m - (s_1 + d + 1) \leq v_2 < m - (s_1 + 1)$, and (a): $w_2 \leq v_1 - s_0$, (b): $w_2 > v_1 - s_0$.

On the other hand, if $k < w_2$, it turns out that $h \geq d - w_1$ (since $h + k = d + r$ and $w_2 - w_1 = r + 1$) and thus $t = v_2 - (s_0 + w_1)$ is in the admissible range and satisfies (3). See Fig. 2.(b). □

As an immediate application of the theorem, for different pattern lengths m we computed the seed $\mathbf{0}^s R_d \mathbf{0}^s$ of maximal weight among those for which Theorem 1 guarantees $m_P^* \leq m$. The resulting maximal weights for some values of m are reported in Table 2, together with the weights of the asymptotically optimal seeds from [3]. We see that for m up to 498, seeds derived from perfect rulers have a larger weight. Hence, although such seeds are not asymptotically optimal, they are preferable for values of m which are of practical interest.

Theorem 1 assumes $\max(s_0, s_1) \leq d$. It turns out that if $\max(s_0, s_1) > d$ then $P = \mathbf{0}^{s_0} R \mathbf{0}^{s_1}$ has a much larger minimal pattern length m_P^*. The following

m	32	64	96	128	160	192	224	300	400	500
[3]	3	12	30	42	64	81	96	150	210	284
Th. 1	14	31	50	68	86	104	122	166	224	283
(d, s)	(8,5)	(15,11)	(19,19)	(26,25)	(33,31)	(40,37)	(47,43)	(61,59)	(81,79)	(101,99)

Table 2. Comparison of seed weight as a function of pattern length. The second row reports the weights of the asymptotically optimal seeds defined in [3, Th. 5]. The third row reports the maximal weights for the seeds of the form $\mathbf{0}^s R_d \mathbf{0}^s$ (the last row shows the values of d and s yielding the maximal weights). The weights of the two families are both equal to 282 for $m = 498$.

theorem proves that this is true even replacing R with an arbitrary binary string of length $d + 1$.

Theorem 2. *Let $P = \mathbf{0}^{s_0} U \mathbf{0}^{s_1}$, where U is any binary string of length $d + 1$. Then:*

$$\max(s_0, s_1) > d \quad \Longrightarrow \quad m^*_P \geq 2|P| \tag{4}$$

$$\min(s_0, s_1) > d \quad \Longrightarrow \quad m^*_P \geq 2|P| + \min(s_0, s_1). \tag{5}$$

Proof. Without loss of generality we can assume $s_0 \geq s_1$. Let $p = |P|$. To prove (4) we show that if $\max(s_0, s_1) > d$ then P is not $(2p - 1, 2)$-complete according to Definition 4. Let V denote the binary string of length $2p - 1$ with $\mathbf{1}$'s in positions $v_1 = p-1-(d+1)$ and $v_2 = p-1$. Since P does not measure $d+1$, $P+t$ can match V only if $v_1 - t < 0$ which implies $t > p-1-(d+1) = v_2-(d+1)$. Moreover, since $t \leq |V| - |P|$ must hold, then it must be $t \leq p - 1$. Hence, we must have $0 \leq v_2 - t < (d+1) \leq s_0$. This latter inequality implies $P[v_2 - t] = \mathbf{0}$ and (2) cannot hold. To prove (5), we show that if $\min(s_0, s_1) > d$ then P cannot be $(2p - 1 + \min(s_0, s_1), 2)$-complete by taking $v_1 = p - 1$ and $v_2 = p + s_1 - 1$ and reasoning as above. □

Theorem 2 implies that the seed $\mathbf{0}^{s_0} R \, \mathbf{0}^{s_1}$ is not interesting when $\max(s_0, s_1) > d$. To see this, compare for example $P = \mathbf{0}^{d+1} R \, \mathbf{0}^d$ with $P' = \mathbf{0}^{d+1} R' \, \mathbf{0}^{d+1}$ where R' is a complete $(d+1)$-ruler. We have $|P| = 3d+2$, $|P'| = 3d+4$, and P' has at least one $\mathbf{0}$ more than P. By Theorem 2 it is $m^*_P \geq 6d+4$ whereas by Theorem 1 it is $m^*_{P'} \leq 5d + 6$ which is preferable for $d > 2$.

Summing up, we have that among the seeds that we can derive from perfect rulers there is a family whose members are of practical interest since they offer a competitive trade-off between seed weight and minimum pattern length m^*_P. In the next sections we will further investigate the properties of these seeds. Since the upper bound established in Theorem 1 will play an important role in our analysis, we introduce a notation for it.

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 $m_P = 2|P| - 1 - \min(s_0, s_1)$.* □

4 Lower bounds on the minimum pattern length m_P^*

In this section we investigate whether the upper bound established in Theorem 1 is tight or there exist seeds of the form $\mathbf{0}^{s_0} R \mathbf{0}^{s_1}$ which are $(m, 2)$-complete for m significantly smaller than the upper bound of Theorem 1.

We begin our analysis with the case of the seed associated to $\mathbf{0}^d R \mathbf{0}^d$ where R is a complete d-ruler. This is an important case since in Section 3 we saw that the only interesting seeds are those with $\min(s_0, s_1) \leq d$. Among them, $\mathbf{0}^d R \mathbf{0}^d$ is the one with the largest weight. For $P = \mathbf{0}^d R \mathbf{0}^d$ Theorem 1 yields $m_P^* \leq 5d+1$. The next result shows that it is indeed $m_P^* = 5d+1$ and that such value is the best possible even if we replace R with an arbitrary binary string of length $d+1$.

Lemma 2. *Let $P = \mathbf{0}^d U \mathbf{0}^d$, where U is any binary string of length $d+1$. Then $m_P^* \geq 5d+1$.*

Proof. We prove that P is not $(5d, 2)$-complete reasoning as in the proof of Theorem 2. Let V denote the binary string of length $5d$ with $\mathbf{1}$'s in positions $v_1 = d-1$ and $v_2 = 4d$. First notice that admissible values for t are in the range $0 \leq t \leq 2d-1$. For $t = 0, \ldots, d-1$ it is $0 \leq v_1 - t \leq d-1$ and $P[v_1 - t] = \mathbf{0}$ since $P[v_1 - t]$ is inside the d leading $\mathbf{0}$'s. For $t = d, \ldots, 2d-1$ it is $2d+1 \leq v_2 - t \leq 3d$ and $P[v_2 - t] = \mathbf{0}$ since $P[v_2 - t]$ is inside the d trailing $\mathbf{0}$'s. Therefore no choice of t satisfies (2). □

Corollary 1. *If $P = \mathbf{0}^d R \mathbf{0}^d$, where R is a complete d-ruler, then $m_P^* = 5d+1$.* □

For the general case $P = \mathbf{0}^{s_0} R \mathbf{0}^{s_1}$ we provide lower bounds which depend on the distributions of the $\mathbf{1}$'s in R. We make use of the following technical lemma.

Lemma 3. *Assume $P = \mathbf{0}^{s_0} U \mathbf{0}^{s_1}$ is $(m, 2)$-complete where U is an arbitrary binary string. For any δ that is measurable in U, let $(x, x+\delta)$ and $(y, y+\delta)$ denote respectively the leftmost and rightmost measures of δ in P. We have*

$$\delta \leq s_0 \implies x + \delta \leq m - |P| \tag{6}$$

$$\delta \leq s_1 \implies |P| - 1 - y \leq m - |P| \tag{7}$$

Proof. We prove the thesis considering a length-m binary string V with $\mathbf{1}$'s in appropriate positions v_1 and v_2 and showing that there exists t such that $P + t$ matches in V only if (6) and (7) hold.

Let $p = |P|$. If $\delta \leq s_0$ let $v_1 = x - 1$, $v_2 = v_1 + \delta = x + \delta - 1$. Since $(x, x+\delta)$ is the leftmost measure of δ, $P + t$ can match in V only for $t \geq v_1 + 1$. But, since $\delta \leq s_0$, for $v_1 + 1 \leq t \leq v_2$, $P[v_2 - t] = 0$ since it falls inside $\mathbf{0}^{s_0}$. Therefore it must be $t > v_2$. This implies $m - p \geq v_2 + 1 = x + \delta$, and $x + \delta \leq m - p$ as claimed.

If $\delta \leq s_1$, let $v_1 = m - p + y + 1$ and $v_2 = v_1 + \delta$. With arguments analogous to the case $\delta \leq s_1$, we have that $P + t$ can match in V only if $v_1 - t \geq p$, which implies $p \leq v_1 = m - p + y + 1$ and $p - y - 1 \leq m - p$ as claimed. □

A fundamental notion in our analysis is the one of (λ, σ)-skewness. Informally, a string is (λ, σ)-skew for small values of λ and σ if there are small integers that are measured only near the endpoints of the string. The latter implies that the minimum length m_P^* is close to the upper bound m_P (see Theorem 3).

Definition 7 (Skewness). *Let U denote a binary string of length u. We say that U is (λ, σ)-skew if there exist δ_L and δ_R, not necessarily distinct, such that $(u_L, u_L + \delta_L)$ (resp. $(u_R, u_R + \delta_r)$) is the only measure of δ_L (resp. δ_R) in U, and the conditions*

$$\max(u_L, u - 1 - u_R - \delta_R) \leq \lambda \qquad \textit{and} \qquad \max(\delta_L, \delta_R) \leq \sigma \tag{8}$$

hold. □

Note that (8) implies that the range $(u_L, u_L + \delta_L)$ is entirely within the first $\lambda + \sigma$ positions of U and starts within the first λ positions of U. Symmetrically, the range $(u_R, u_R + \delta_R)$ is within the last $\lambda + \sigma$ positions of U and ends within the last λ positions of U.

Example 1. Let $U = \mathbf{1100110000111}$. The only measure of $\delta_L = 3$ is $(1, 4)$, and the only measure of $\delta_R = 2$ is $(10, 12)$. Since $|U| = 13$, U is $(1, 3)$-skew. □

Let $P = \mathbf{0}^{s_0} U \mathbf{0}^{s_1}$ where U is (λ, σ)-skew. The next theorem establishes a lower bound for m_P^*, given in terms of the upper bound m_P, under the assumption that $\min(s_0, s_1)$ is larger than the integers that are measured only near the endpoints of U. The assumption is not restrictive for practical applications, since more **0**'s in P translates to a spaced seed with larger weight.

Theorem 3. *For any string U, let $P = \mathbf{0}^{s_0} U \mathbf{0}^{s_1}$. If U is (λ, σ)-skew and $\min(s_0, s_1) \geq \sigma$, then $m_P^* \geq m_P - \lambda$.*

Proof. We prove the theorem showing that, if P is $(m, 2)$-complete with $m \leq m_P - \lambda$, then necessarily $m \geq m_P - \lambda$. Let $p = |P|$, $u = |U|$.

Since U is (λ, σ)-skew, for δ_L and δ_R as in Definition 7 it is $\delta_L, \delta_R \leq \sigma \leq \min(s_0, s_1)$. Hence, Lemma 3 can be applied to both δ_L and δ_R.

Let, as in Definition 7, $(s_0 + u_L, s_0 + u_L + \delta_L)$ be the unique measure of δ_L in P. By (7) of Lemma 3 and (8) it is

$$m \geq 2p - 1 - (s_0 + u_L) \geq 2p - 1 - s_0 - \lambda. \tag{9}$$

Similarly, let $(s_0 + u_R, s_0 + u_R + \delta_R)$ be the unique measure of δ_R in P. By (6) of Lemma 3, it is $s_0 + u_R + \delta_R \leq m - p$. Applying (8) and finally recalling that $p = u + s_0 + s_1$, we get

$$m \geq p + s_0 + u_R + \delta_R \geq p + s_0 + u - 1 - \lambda = 2p - 1 - s_1 - \lambda. \tag{10}$$

From (9) and (10) we get, as claimed, $m \geq 2p - 1 - \min(s_0, s_1) - \lambda = m_P - \lambda$. □

The above theorem holds for any (λ, σ)-skew string U. For a complete d-ruler, combined with Theorem 1, it yields the following result.

Corollary 2. *Let $P = \mathbf{0}^{s_0} R \mathbf{0}^{s_1}$ where R is a (λ, σ)-skew complete d-ruler. If $\max(s_0, s_1) \leq d$ and $\min(s_0, s_1) \geq \sigma$, then $m_P - \lambda \leq m_P^* \leq m_P$.* □

5 Wichmann rulers

As we mentioned in Section 3, perfect rulers are difficult to find. The exception is the family of Wichmann rulers [14] which have a sort of "generating function". The Wichmann ruler $W_{r,s}$ is the binary string defined by

$$W_{r,s} = \mathbf{1}^{r+1}\ \mathbf{0}^r\mathbf{1}\ (\mathbf{0}^{2r}\mathbf{1})^r\ (\mathbf{0}^{4r+2}\mathbf{1})^s\ (\mathbf{0}^{2r+1}\mathbf{1})^{r+1}\ \mathbf{1}^r.$$

$W_{r,s}$ has length $w_{r,s} = 4(r+1)^2 + s(4r+3)$ and contains exactly $4r+s+3$ $\mathbf{1}$'s. It is a classical result that $W_{r,s}$ is a complete $(w_{r,s}-1)$-ruler [14].

In this section we consider the seeds of the form $P = \mathbf{0}^{s_0}W_{r,s}\mathbf{0}^{s_1}$ and we analyze how tight is Theorem 1 for these seeds. Our first result proves that, if the number of leading and trailing $\mathbf{0}$'s in P is large enough, then the upper bound m_P of Theorem 1 is very accurate. Recall that large s_0 and s_1 are required to obtain seeds with large weight.

Theorem 4. *Let $P = \mathbf{0}^{s_0}W_{r,s}\mathbf{0}^{s_1}$ with $r > 1$. We have*

$$\min(s_0, s_1) \geq 3r+1 \implies m_P^* \geq m_P - 1 \tag{11}$$
$$\min(s_0, s_1) \geq 2r+3 \implies m_P^* \geq m_P - r + 1 \tag{12}$$

Proof. We prove (11) showing that, for $r > 1$, $W_{r,s}$ is $(1, 3r+1)$-skew according to Definition 7. To see this consider $\delta_L = 2r$ and $\delta_R = 3r+1$. It is straightforward to verify that the only measure for $2r$ in $W_{r,s}$ is $(1, 1+2r)$, and the only measure for $3r+1$ is $(w_{r,s}-3r-3, w_{r,s}-2)$. Having established that $W_{r,s}$ is $(1, 3r+1)$-skew, the thesis follows by Theorem 3.

To prove (12) we show that for $r > 1$, $W_{r,s}$ is $(r-1, 2r+3)$-skew considering $\delta_L = r+2$ and $\delta_R = 2r+3$. The only measure for $r+1$ in $W_{r,s}$ is $(r, 2r+1)$, and the only measure for $2r+3$ is $(w_{r,s}-3r-3, w_{r,s}-r)$. The thesis follows again by Theorem 3. □

We now establish lower bounds for $P = \mathbf{0}^{s_0}W_{r,s}\mathbf{0}^{s_1}$ with no constraints on s_0 and s_1. These results are not based on the concept of skewness, but on the specific structure of Wichmann rulers.

Lemma 4. *Let $P = \mathbf{0}^{s_0}W_{r,s}\mathbf{0}^{s_1}$ with $r > 1$. If $\max(s_0, s_1) \geq 2r+3$ then $m_P^* \geq m_P - (2r+2)$.*

Proof. We prove the lemma assuming $m_P^* \leq m_P - (2r+3)$ and obtaining a contradiction. Let $p = |P|$. Notice that, since $W_{r,s}$ is a complete $(w_{r,s}-1)$-ruler, it measures all integers $1 \leq \delta \leq w_{r,s} - 1$.

Assume first that $s_1 = \min(s_0, s_1)$. Then $s_0 = \max(s_0, s_1) \geq 2r+3$, and we can apply Lemma 3 to $\delta = 2r+3 \leq s_0$. Since the only measure of $2r+3$ in P is $(p - s_1 - 3r - 3, p - s_1 - r)$, by (6) we get $p - s_1 - r \leq m_P^* - p$ which implies

$$m_P^* \geq 2p - s_1 - r = 2p - \min(s_0, s_1) - r = m_P + 1 - r$$

which is a contradiction.

Assume now $s_0 = \min(s_0, s_1)$. Then $s_1 = \max(s_0, s_1) \geq 2r + 3$, and we can apply Lemma 3 to $\delta = r + 2 \leq s_1$. Since the only measure of $r + 2$ in P is $(s_0 + r - 1, s_0 + 2r + 1)$, by (7) we get $p - 1 - (s_0 + r - 1) \leq m_P^* - p$ which implies

$$m_P^* \geq 2p - s_0 - r = 2p - \min(s_0, s_1) - r = m_P + 1 - r$$

which is again a contradiction. □

Lemma 5. *Let* $P = \mathbf{0}^{s_0} W_{r,s} \mathbf{0}^{s_1}$ *with* $r > 0$. *If* $\max(s_0, s_1) \leq 2r + 2$ *then* $m_P^* \geq m_P - (4r + 2)$.

Proof. Let $p = |P|$, and consider the binary string V of length m_P^* with $\mathbf{1}$'s in positions $v_1 = p - s_1 - r - 2$ and $v_2 = v_1 + 1$. Note that (v_1, v_2) are the positions immediately before the first pair of consecutive $\mathbf{1}$'s after the central block $(\mathbf{0}^{4r+2}\mathbf{1})^s$. The closest pair of consecutive $\mathbf{1}$'s in P to the left of (v_1, v_2) are those in positions $(s_0 + r - 1, s_0 + r)$. Hence $P + t$ matches in V only for $t \geq v_1 - (s_0 + r - 1)$. Since the largest admissible shift t is $m_P^* - p$ we must have $m_P^* - p \geq v_1 - (s_0 + r - 1)$, which implies

$$\begin{aligned} m_P^* &\geq p + (p - s_1 - r - 2) - (s_0 + r - 1) \\ &= 2p - s_1 - s_0 - 1 - 2r \\ &= m_P - \max(s_0, s1) - 2r \\ &\geq m_P - 4r - 2 \end{aligned}$$

as claimed. □

Combining Lemmas 4 and 5 we get the following theorem that provides a lower bound for any seed P based on a Wichmann ruler with $r > 1$.

Theorem 5. *Let* $P = \mathbf{0}^{s_0} W_{r,s} \mathbf{0}^{s_1}$ *with* $r > 1$. *It is* $m_P^* \geq m_P - (4r + 2)$. □

Note that $r = O(\sqrt{w_{r,s}})$, so Theorem 5 once more proves the estimate of Theorem 1 accurate. In the following example we present a specific case, in order to give a feeling of the values involved.

Example 2. The string $W_{2,1}$ is a complete 46-ruler. Consider the seed $P = \mathbf{0}^i W_{2,1} \mathbf{0}^i$. For $i =, 0, \ldots, 46$ it is $m_P = 94 + 3i - 1$. By Corollary 1, for $i = 46$ it is $m_P^* = m_P = 231$. By Theorem 4, for $7 \leq i \leq 45$ it is $m_P - 1 \leq m_P^* \leq m_P$. For example for $i = 7$ it is $113 \leq m_P^* \leq 114$. By Theorem 5, for $0 \leq i \leq 6$, it is $m_P - 10 \leq m_P^* \leq m_P$. For example, for $i = 0$ it is $83 \leq m_P^* \leq 93$. □

6 Restricted vs Unrestricted Rulers

The complete rulers defined in Section 2 are sometimes called *restricted* rulers since we require that the string R measuring all integers between 1 and d has length exactly $d+1$. In the literature [2, 6] there is also the notion of *unrestricted d-ruler* which is a binary string of arbitrary length that measures all integers between 1 and d. The next Lemma shows that every spaced seed must have an unrestricted ruler at its heart.

Lemma 6. *If P is $(m,2)$-complete, then it must measure any integer δ, $1 \leq \delta \leq 2|P| - m - 1$.*

Proof. Let $p = |P|$. We prove the lemma showing that if P does not measure δ then $\delta \geq 2p - m$. Consider the length-m binary string V with **1**'s in positions $v_1 = p - 1 - \delta$, and $v_2 = p - 1$. Since P is $(m,2)$-complete there must exist $t \leq m - p$ such that $P + t$ matches in V according to Definition 4. However, if P does not measure δ, $P + t$ does not match in V whenever $t \leq v_1$. Hence, we must have $t \geq v_1 + 1 = p - \delta$, which is possible only if $p - \delta \leq m - p$ which implies $\delta \geq 2p - m$ as claimed. □

The above result suggests that it could be worthwhile to analyze also spaced seeds of the form $P = \mathbf{0}^{s_0} U_d \mathbf{0}^{s_1}$, where U_d is an unrestricted d-ruler. This interest is motivated theoretically by the fact that an unrestricted d-ruler can contain less **1**'s than a perfect d-ruler.

Example 3. The minimum number of **1**'s in a restricted 18-ruler is eight [9]. The string $U_{18} = \mathbf{10000010011000000010000101}$ has length 25, seven **1**'s, and is an unrestricted 18-ruler since it measures all integers from 1 to 18. □

In the following we give some evidence that for seed design unrestricted rulers appear to be less effective than restricted rulers.

Let U_{18} denote the string defined in Example 3 and let m^*_Q denote the minimum m such that $Q = \mathbf{0}^s U_{18} \mathbf{0}^s$ is $(m,2)$-complete. The ruler U_{18} is $(0,6)$-skew, since 6 has the unique measure $(0,6)$, and 2 has the unique measure $(22,24)$. Theorem 3 implies that $m^*_Q \geq m_Q = 2|Q| - 1 - s$. Note however, that Theorem 1 has been proven only for restricted rulers. Indeed, a direct verification shows that it does not hold for unrestricted ones, since for $s \geq 13$ it is $m^*_Q > m_Q$.

Based on the above observations, we compare the completeness properties of a seed obtained from U_{18} to one obtained from a restricted d-ruler R_d.

- Let $P = \mathbf{0}^{s_0} R_{18} \mathbf{0}^{s_1}$, where R_{18} is a restricted 18-ruler with eight **1**'s (see [9]). Let $s_0 = s + 4$ and $s_1 = s + 3$. Then $|P| = |Q| + 1$ and P and Q have the same weight. Since $m_P = 2(|Q| + 1) - 1 - (s + 3) = m_Q - 1$, then $m^*_Q \geq m_Q = m_P + 1 > m^*_P$, which implies $m^*_Q > m^*_P$.
- If we take $P = \mathbf{0}^{s+4} R_{18} \mathbf{0}^{s+4}$ with the same R_{18}, so that P is longer and has a larger weight than Q, we obtain $m_Q = m_P$ and therefore $m^*_Q \geq m^*_P$.
- Finally, let $P = \mathbf{0}^{s_0} R_{17} \mathbf{0}^{s_1}$ where R_{17} is a restricted 17-ruler with seven **1**'s as U_{18} (see [9]). Let $s_0 = s + 4$ and $s_1 = s + 3$. Then, Q and P have the same length and weight, and $m_Q = m_P + 3$, therefore $m^*_Q \geq m^*_P + 3$.

We consider now Wichmann rulers. We observe that in general a restricted d'-ruler with $d' > d$ is an unrestricted d-ruler. Hence, for $b > 0$, $W_{r+b,s}$ can be seen as an unrestricted $(w_{r,s} - 1)$-ruler. We show that a seed built out of the restricted ruler $W_{r,s}$ is better than one of the same length based on $W_{r+b,s}$.

Let $P = \mathbf{0}^{s_0} W_{r,s} \mathbf{0}^{s_1}$. To preserve length, in replacing $W_{r,s}$ with $W_{r+b,s}$ we reduce the number of leading and trailing zeroes. Since $|W_{r+b,s}| - |W_{r,s}| = 4b(b + 2r + s + 2)$, we let $\sigma_b = 2b(b + 2r + s + 2)$ and define $Q = \mathbf{0}^{s'_0} W_{r+b,s} \mathbf{0}^{s'_1}$

with $s'_i = s_i - \sigma_b$ $(i = 0, 1)$, so that $|Q| = |P|$. Notice that the following theorem holds, somewhat counterintuitively, even though P has a larger weight than Q.

Theorem 6. *Let $P = \mathbf{0}^{s_0} W_{r,s} \mathbf{0}^{s_1}$ and $Q = \mathbf{0}^{s'_0} W_{r+b,s} \mathbf{0}^{s'_1}$ as above, with $r > 1$. It is $m^*_Q \geq m^*_P$ for $b \geq 1$, and $m^*_Q > m^*_P$ for $b > 1$ or $s > 0$.*

Proof. Since $|Q| = |P|$, it is $m_Q = 2|Q| - 1 - \min(s_0 - \sigma_b, s_1 - \sigma_b) = m_P + \sigma_b$. From Theorem 5 applied to Q and the latter equality we get $m^*_Q \geq m_P + \sigma_b - 4(r+b) - 2$; the inequality is strict for $b > 1$ or $s > 0$ as claimed. □

7 Conclusions

In this paper we have shown how to derive a new family of lossless seeds from the mathematical objects known as perfect rulers. We have proven upper and lower bounds on the effectiveness of these seeds and shown that they are of practical interest. A natural extension of this work is to study their use in the context of multiseed filtration for the detection of more than 2 mismatches.

References

1. Stefan Burkhardt and Juha Kärkkäinen. Better filtering with gapped q-grams. *Fundam. Inform.*, 56(1-2):51–70, 2003.
2. P. Erdós and I. S. Gál. On the representation of $1, 2, \ldots, n$ by differences. *Idagationes 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. Uri Keich, Ming Li, Bin Ma, and John Tromp. On spaced seeds for similarity search. *Discrete Applied Mathematics*, 138(3):253–263, 2004.
5. Gregory Kucherov, Laurent Noé, and Mikhail A. Roytberg. Multiseed lossless filtration. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 2(1):51–61, 2005.
6. J. Leech. On the representation of $1, 2, \ldots, n$ by differences. *J. London Math. Soc.*, 31:160–169, 1956.
7. Ming Li, Bin Ma, Derek Kisman, and John Tromp. Patternhunter II: Highly sensitive and fast homology search. *J. Bioinformatics and Computational Biology*, 2(3):417–440, 2004.
8. Hao Lin, Zefeng Zhang, Michael Q. Zhang, Bin Ma, and Ming Li. Zoom! zillions of oligos mapped. *Bioinformatics*, 24(21):2431–2437, 2008.
9. Peter Luschny. Perfect and optimal rulers, 2003. http://www.luschny.de/math/rulers/prulers.html.
10. Bin Ma and Ming Li. On the complexity of the spaced seeds. *J. Comput. Syst. Sci.*, 73(7):1024–1034, 2007.
11. Bin Ma, John Tromp, and Ming Li. Patternhunter: faster and more sensitive homology search. *Bioinformatics*, 18(3):440–445, 2002.
12. Bin Ma and Hongyi Yao. Seed optimization is no easier than optimal Golomb ruler design. In Alvis Brazma, Satoru Miyano, and Tatsuya Akutsu, editors, *APBC*, volume 6 of *Advances in Bioinformatics and Computational Biology*, pages 133–144. Imperial College Press, 2008.

13. François Nicolas and Eric Rivals. Hardness of optimal spaced seed design. *J. Comput. Syst. Sci.*, 74(5):831–849, 2008.
14. B. Wichmann. A note on restricted difference bases. *J. London Math. Soc.*, 38:465–466, 1962.