Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Combinatorial Theory, Series A 104 (2003) 95–113
Journal of Combinatorial Theory Series A
http://www.elsevier.com/locate/jcta

Combinatorics of periods in strings

Eric Rivals[a] and Sven Rahmann[b]

[a] *LIRMM, CNRS U.M.R. 5506, 161 rue Ada, F-34392 Montpellier, Cedex 5, France*
[b] *Max-Planck-Institut für Molekulare Genetik, Dept. of Computational Molecular Biology, Ihnestraße 73, D-14195 Berlin, Germany*

Received 9 July 2002

Abstract

We consider the set Γ_n of all period sets of strings of length n over a finite alphabet. We show that there is redundancy in period sets and introduce the notion of an irreducible period set. We prove that Γ_n is a lattice under set inclusion and does not satisfy the Jordan–Dedekind condition. We propose the first efficient enumeration algorithm for Γ_n and improve upon the previously known asymptotic lower bounds on the cardinality of Γ_n. Finally, we provide a new recurrence to compute the number of strings sharing a given period set, and exhibit an algorithm to sample uniformly period sets through irreducible period set.

Keywords: Autocorrelation; Strings; Periods; Lattice

1. Introduction

We consider the period sets of strings of length n over a finite alphabet, and specific representations of them, (*auto*)*correlations*, which are binary vectors of length n indicating the periods. Among the possible 2^n bit vectors, only a small subset are valid autocorrelations. In [9], Guibas and Odlyzko provide characterizations of correlations, asymptotic bounds on their number, and a recurrence for the *population size* of a correlation, i.e., the number of strings sharing a given correlation. However, until now, no one has investigated the combinatorial structure of Γ_n, the set of all correlations of length n; nor has anyone proposed an efficient

E-mail addresses: rivals@lirmm.fr (E. Rivals), sven.rahmann@molgen.mpg.de (S. Rahmann).

0097-3165/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0097-3165(03)00123-7

enumeration algorithm for Γ_n. Note that Γ_n can be enumerated by a brute force algorithm that computes the period sets for all possible strings over a given alphabet.

In this paper, we show that there is redundancy in period sets, introduce the notion of an *irreducible period set*, and show how to efficiently convert between the two representations (Section 2). We prove that Γ_n is a lattice under set inclusion and does not satisfy the Jordan–Dedekind condition. While Λ_n, the set of all irreducible period sets, does satisfy that condition, it does not form a lattice (Section 3). We propose the first efficient enumeration algorithm for Γ_n (Section 4) and improve upon the previously known asymptotic lower bounds for the cardinality of Γ_n (Section 5). We provide a new recurrence to compute the population sizes of correlations (Section 6). Finally, we exhibit a Markov chain algorithm to uniformly sample period sets using properties of irreducible period sets (Section 7). This article is a augmented version of an abstract [20].

Periods of strings have proven useful mainly in two areas of research. First, in pattern matching, several off-line algorithms take advantage of the periods of the pattern to speed up the search for its occurrences in a text (see [4] for a review). Second, several statistics of pattern occurrences have been investigated which take into account the pattern's periodicity. For instance, the probability of a pattern's absence in a Bernoulli text depends on its correlation [21]. In another work [18,19], we investigate the number of missing words in a random text and the number of common words between two random texts. Computing their expectation requires the enumeration of all correlations and the calculation of their population sizes. This has applications in the analysis of approximate pattern matching, in computational molecular biology, and in the testing of random number generators (RNG). Hereunder, we give some details on these applications.

Among numerous empirical tests designed to check RNGs (see [12] for a comprehensive list) are the monkey tests [14,16]. Each call to the RNG is used to choose a symbol in an alphabet and n successive calls yield a sequence of length n. If for numerous generated sequences, the number of words of length k (with $k \ll n$) that do not occur in the sequence is significantly different from the expected number of missing words of a random sequence, the monkey test rejects the RNG. The distribution of the number of missing words is conjectured to be Gaussian [19]. The first method to computes its expectation was presented in [19] and requires the enumeration of all autocorrelations of size k.

In the domain of approximate pattern matching, some algorithms first filter uninteresting regions of the text to be searched and then apply a dynamic programming algorithm on remaining regions that may contain an approximate match [11,15,22]. For a chosen word length k, the filtration steps work by comparing the vocabulary of a region and of the pattern. The average filtration efficiency on random texts is related to the number of missing words and can be assessed using the above-mentioned method. This suggests practical rules to choose the parameters of the method. Such filtration algorithms are applied in the field of computational biology where large sequence databases are searched [2,10,17].

1.1. Notations, definitions and elementary properties

Let Σ be a finite alphabet of size σ. A sequence of n letters of Σ indexed from 0 to $n-1$ is called a *word* or a *string* of length n over Σ. We denote the *length* of a word $U := U_0 U_1 \dots U_{n-1}$ by $|U|$. For any $0 \leqslant i \leqslant j < n$, $U_{i..j} := U_i \dots U_j$ is called a *substring* of U. Moreover, $U_{0..j}$ is a *prefix* and $U_{i..n-1}$ is a *suffix* of U. We denote by Σ^*, respectively by Σ^n, the set of all finite words, resp. of all words of length n, over Σ.

Definition 1.1 (Period). Let $U \in \Sigma^n$ and let p be a non-negative integer with $p < n$. Then p is a *period* of U iff: $\forall 0 \leqslant i < n-p : U_i = U_{i+p}$.

In other words, p is a period iff another copy of U shifted p positions to the right over the original matches in the overlapping positions, or equivalently, iff the prefix and suffix of U of length $n-p$ are equal. By convention, any word has the trivial null period, 0.

Some properties of periods are: If p is a period then any multiple of p lower than n is also period. If p is a period and the suffix of length $n-p$ has period q, then U has period $p+q$, and conversely. For an in-depth study, we refer the reader to [3,9,13]. Here, we need the Theorem of Fine and Wilf, also called the GCD-rule, and a useful corollary.

Theorem 1.1 (Fine and Wilf [7]). *Let $U \in \Sigma^n$. If U has periods p and q with $p \leqslant q$ and $p + q \leqslant n + \gcd(p,q)$, then $\gcd(p,q)$ is also a period.*

Lemma 1.1. *Let $U \in \Sigma^n$ with smallest non-null period $p \leqslant \lfloor \frac{n}{2} \rfloor$. If $i < n - p + 2$ is a period of U, then it is a multiple of p.*

Proof. Assume that $p \nmid i$. Then $g := \gcd(p,i) < p$, and trivially $g \geqslant 1$. Therefore, $p + i - g \leqslant n$, and Theorem 1.1 says that g is a period, contradicting the premise that p is the smallest non-null period. □

Sets of periods and autocorrelations: Let $U \in \Sigma^n$. We denote the *set of all periods* of U by $P(U)$. We have that $P(U) \subseteq [0, n-1]$. The *autocorrelation* v of U is a representation of $P(U)$. It is a binary vector of length n such that: $\forall 0 \leqslant i < n$, $v_i = 1$ iff $i \in P(U)$, and $v_i = 0$ otherwise. As v and $P(U)$ represent the same set, we use them interchangeably and write $P(U) = v$. We use both $i \in v$ and $v_i = 1$ to express that i is a period of a word U with autocorrelation v. We also write that *i is a period of v*. The smallest non-null period of U or of v is called its *basic period* and is denoted by $\pi(U)$ or $\pi(v)$.

We denote the concatenation of two binary strings s and t by st, and the k-fold concatenation of s with itself by s^k. So $\mathtt{10}^k w$ is the string starting with `1`, followed by k `0`s, and ending with the string w.

Let $\Gamma_n := \{v \in \{0,1\}^n \mid \exists U \in \Sigma^n : v = P(U)\}$ be the set of all autocorrelations of strings in Σ^n. We denote its cardinality by κ_n. The autocorrelations in Γ_n can be

partitioned according to their basic period; thus, for $0 \leqslant p < n$, we denote by $\Gamma_{n,p}$ the subset of autocorrelations whose basic period is p, and by $\kappa_{n,p}$ the cardinality of this set. The set inclusion defines a partial order on elements of Γ_n. For $u, v \in \Gamma_n$, we denote by $u \subseteq v$, resp. by $u \subset v$, the inclusion, resp. the strict inclusion, of u in v. We write $v \succ u$ if v covers u in the inclusion relationship, i.e., if $u \subset v$, and $u \subseteq y \subset v$ implies $y = u$.

1.2. Characterization of correlations

In [9], Guibas and Odlyzko characterized the correlations of length n in terms of the forward propagation rule (FPR), the backward propagation rule (BPR), and also by a recursive predicate Ξ. We review the main theorem and the definitions.

Theorem 1.2 (Characterization of correlations [9]). *Let* $v \in \{0,1\}^n$. *The following statements are equivalent*:

1. v *is the correlation of a binary word.*
2. v *is the correlation of a word over an alphabet of size* $\geqslant 2$.
3. $v_0 = 1$ *and* v *satisfies the forward and backward propagation rules.*
4. v *satisfies the predicate* Ξ.

Let $v \in \{0,1\}^n$. We define the FPR, BPR and give predicate Ξ.

Definition 1.1. v satisfies the FPR iff for all pairs (p,q) satisfying $0 \leqslant p < q < n$ and $v_p = v_q = 1$, it follows that $v_{p+i(q-p)} = 1$ for all $i = 2, \ldots, \lfloor (n-p)/(q-p) \rfloor$.

Definition 1.2. v satisfies the BPR iff for all pairs (p,q) satisfying $0 \leqslant p < q < 2p$, $v_p = v_q = 1$, and $v_{2p-q} = 0$, it follows that $v_{p-i(q-p)} = 0$ for all $i = 2, \ldots, \min(\lfloor p/(q-p) \rfloor, \lfloor (n-p)/(q-p) \rfloor)$.

Predicate Ξ: v satisfies Ξ iff $v_0 = 1$ and, if p is the basic period of v, one of the following conditions is satisfied:

Case a: $p \leqslant \lfloor n/2 \rfloor$. Let $r := \mathrm{mod}\,(n,p)$, $q := p + r$ and w the suffix of v of length q. Then for all j in $[1, n-q]$ $v_j = 1$ if $j = ip$ for some i, and $v_j = 0$ otherwise; and the following conditions hold:

1. $r = 0$ or $w_p = 1$,
2. if $\pi(w) < p$ then $\pi(w) + p > q + \gcd(\pi(w), p)$,
3. w satisfies predicate Ξ.

Case b: $p > \lfloor n/2 \rfloor$. We have $\forall j$: $1 \leqslant j < p$, $v_j = 0$. Let w be the suffix of v of length $n - p$, then w satisfies predicate Ξ.

Guibas and Odlyzko proved that verifying the predicate requires $O(n)$ time. Note that Ξ is recursive on the length of the binary vector. When v is tested, Ξ is

recursively applied to a unique suffix of v denoted w (in case a, $|w| = p + r$; in case b, $|w| = n - p$). We call the corresponding w the *nested autocorrelation* of v. The following theorem is a consequence of the FPR and BPR, and of characterization 3 in Theorem 1.2 (see [9]).

Theorem 1.3. *Let v be a correlation of length n. Any substring $v_i \dots v_j$ of v with $0 \leqslant i \leqslant j < n$ such that $v_i = 1$ is a correlation of length $j - i + 1$.*

2. Irreducible periods

We show that the period set of a word is in one-to-one correspondence with a smaller set which we call its associated *irreducible period set* (*IPS* for short).

A full period set contains redundancies since some periods are deducible from others as specified by the forward propagation rule (FPR, see Section 1.2). For example with $n = 12$, in the period set $\{0, 7, 9, 11\}$, 11 can be obtained from 7 and 9 using the FPR ($11 = 9 + 1(9 - 7)$) and is the only deducible period. The IPS is thus $\{0, 7, 9\}$. In this section, we formally define the notion of IPS and we prove that the mapping R from Γ_n to Λ_n, the set of all IPSs, is bijective. We also show how to compute the IPS from the period set, and conversely.

Let $n \in \mathbb{N}$. Define the map $FC'_n : 2^{[0,n-1]} \to 2^{[0,n-1]}, v \mapsto v'$ by $v'_j := 1$ if there exist indices $p \leqslant q$ such that $v_p = v_q = 1$ and $j = p + k(q - p)$ for some $k \geqslant 0$, and $v'_j = 0$ otherwise. Now the *forward closure* FC_n is the repeated application of FC'_n until closure is reached.

Definition 2.2 (Irreducible period set). Let $T \in \Gamma_n$ be a period set. A subset $S := \{p_0, \dots, p_l\}$ of T is an associated irreducible period set (IPS) of T iff it satisfies both following conditions:

1. T is the forward closure of S, i.e., $FC_n(S) = T$,
2. For all triples (h, i, j) satisfying $0 \leqslant h < i < j \leqslant l$ we have $\forall k \in \mathbb{N}^+ : p_j \neq p_i + k(p_i - p_h)$.

Condition (2) expresses formally the fact that in an IPS no period can be obtained from smaller periods with the FPR. It is equivalent to saying that S is the *smallest* subset of T such that $FC_n(S) = T$. In other words, S is an IPS of T if it is the intersection of all sets whose forward closure is T. From this, one can see that the associated IPS exists and is unique. Therefore, we can define a function R that maps a period set to its associated IPS. Now, we define $\Lambda_n := R(\Gamma_n)$ and prove that the correspondence between period sets and IPSs is one-to-one.

Theorem 2.1. *$R : \Gamma_n \to \Lambda_n, P \mapsto R(P)$ is bijective.*

Proof. By definition, R is surjective. To prove that R is injective we need to show that $R(P) = R(Q)$ implies $P = Q$. If $R(P) = R(Q)$ then $P = FC_n(R(P)) = FC_n(R(Q)) = Q$ by condition (1) of Definition 2.2. □

Algorithm 1: R

Input : Word length n, array P of periods in increasing order, size t of P
Output: Associated IPS $R(P)$ as an array I;
Variable: S: a sorted set;

```
1  I[0] := P[0]; δ := n; i := 1; k := 1; S := ∅;
2  while ((i < t) and (δ > 1)) do
3  |  δ := P[i] − P[i − 1]; size := n − P[i − 1]; mul := ⌊size/δ⌋;
4  |  if P[i] ∉ S then
5  |  |  I[k] := P[i]; k := k + 1;
6  |  |  if mul = 2 then
7  |  |  |  if mod(size, δ) ≠ 0 then S.insert(P[i] + δ);
8  |  |  else
9  |  |  |  if mul > 2 then
10 |  |  |  |  S.insert(P[i − 1] + mul × δ);
11 |  |  |  |  i := i + mul − 2;
12 |  |  i := i + 1;
13 return I;
```

By Theorem 2.1, R^{-1} exists; indeed, it is FC_n restricted to Λ_n. Algorithm 1 is an efficient implementation of R. We omit the algorithm for R^{-1}. The next theorem claims that R runs in a time sublinear in the input size (which may be as large as $\Theta(n)$) because $|R(P)| = O(\log n)$, as proved in Lemma 2.2. This is achieved by exploiting the known structure of period sets; the algorithm does not need to examine the whole input array P (cf. lines 9–11 of R).

Theorem 2.2. *For a given word length n and* $P \in \Gamma_n$, *Algorithm* 1 *correctly computes* $R(P)$ *in* $O(|R(P)| \log(|R(P)|))$ *time.*

Proof. R considers the periods of P in increasing order and uses the sorted set S to store the forthcoming deducible periods. For each $P[i]$, R tests whether it is an irreducible period (line 4). If it is not, it is skipped; otherwise it is copied into I (line 5), and we are either in case (a) or (b) of Predicate Ξ. In case (b), no deducible periods are induced by $P[i]$, so nothing else is done. In case (a), we have $mul \geqslant 2$. If $mul = 2$ and mod $(size, \delta) \neq 0$, the forward propagation generates only $P[i] + \delta$ which is inserted into S (lines 6 and 7). If $mul > 2$, Lemma 1.1 allows to skip the periods in the range $[P[i], P[i] + (mul - 2) \times \delta]$ and insert only $P[i - 1] + mul \times \delta$, which is done on line 8. This proves the correctness.

We now prove that the running time is $O(|R(P)|\log|R(P)|)$. We claim that the while loop is executed at most $2\cdot(R(P)-1)$ times. Indeed, in each iteration, either an element is inserted into I and possibly into S, or nothing happens; the latter case arises only when the current $P[i]$ is in S. But at most $R(P)-1$ elements are ever inserted into S and I, as after termination $|I|=|R(P)|$. Clearly, every operation in the loop takes constant time, except the operations on S, which take $O(\log|S|)$ time when S is implemented as a balanced tree. □

Moreover, we prove that the size of $R(P)$ is less than logarithmic in n.

Lemma 2.2. *If $P\in\Gamma_n$ with $n\geqslant 1$, then $|R(P)|\leqslant 1+\lfloor\log_{3/2} n\rfloor$.*

Proof. By induction, we formally prove that $|R(P)|\leqslant 1+\log_{3/2} n$. The lemma follows because $|R(P)|$ is an integer. The statement is true for $1\leqslant n\leqslant 8$ by direct inspection. Now assume $n\geqslant 9$, and let p be the basic period in P.

If $p>n/2$, we know from case (b) of Predicate Ξ that P', the nested correlation of P, belongs to $\Gamma_{(n-p)}$ with $n-p<n/2<2n/3$. Let R' denote the associated IPS of P'. We have $|R'|\leqslant 1+\log_{3/2}(2n/3)=\log_{3/2} n$ by induction hypothesis. Since $R(P)=\{0\}\cup\{p+r\colon r\in R'\}$, we have $|R(P)|=1+|R'|$ and are done with this case.

If $p\leqslant n/2$, we consider P', the nested autocorrelation of P as specified by case (a) of Predicate Ξ. P' starts at a multiple of p, say kp, and satisfies $|P'|\leqslant 2n/3$. Therefore its associated IPS, R', satisfies $|R'|\leqslant\log_{3/2} n$ by hypothesis. Now $R(P)$ contains $0, p$, and atmost all elements from $kp+R'$, except kp (the 0 of R'), as kp is deducible from 0 and p. Thus, $|R(P)|\leqslant 2+(|R'|-1)\leqslant 1+\log_{3/2} n$. □

3. Structural properties of Γ_n and Λ_n

3.1. Γ_n is a lattice under inclusion

First, we prove that the intersection of two period sets is a period set.

Lemma 3.3. *If $u,v\in\Gamma_n$, then $(u\cap v)\in\Gamma_n$.*

Proof. Let $u,v\in\Gamma_n$ and U,V be in Σ^n such that U has period set u and V has period set v. Such strings exist by definition of Γ_n. Let W be the string of $(\Sigma\times\Sigma)^n$ defined by $W_i:=(U_i,V_i)$ for any $0\leqslant i<n$. A period of W is necessarily a period of U and of V. It follows that W has period set w. Thus, as periods sets are independent of the alphabet by Theorem 1.2, w belongs to Γ_n. □

Lemma 3.4. *$(\Gamma_n,\subseteq)$ has a null element, $\mathtt{10}^{n-1}$, and a universal element, $\mathtt{1}^n$.*

Theorem 3.1. *$(\Gamma_n,\subseteq)$ is a lattice.*

Proof. From Lemma 3.3, we know that Γ_n is closed under intersection. Therefore, the meet $u \wedge v$ of $u, v \in \Gamma_n$ is their intersection, and the join $u \vee v$ is the intersection of all elements containing both u and v. The existence of a universal element ensures that this intersection is not empty. □

We present a constructive proof of the existence of the join.

Theorem 3.2. *Let $u, v \in \Gamma_n$. The join under inclusion of u and v exists and is unique in Γ_n.*

Proof. Let $u, v \in \Gamma_n$. We define a function named FW_n for Fine and Wilf, from $2^{[0,n-1]}$ to $2^{[0,n-1]}$. It adds to the input set all periods strictly lower than n required by Theorem 1.1 (it performs the test recursively for each pair of successive periods in decreasing order.) We claim that $w \coloneqq FC_n(FW_n(u \cup v))$ is a period set and the unique join of u and v. To prove this, we need to show that $w \in \Gamma_n$ and that w is minimal.

Let us show that $w \in \Gamma_n$. As $u \cup v \subset w$, $0 \in w$. By construction, w satisfies the FPR. According to characterization 3 of Theorem 1.2, we must show that w satisfies the BPR. Assume that w violates the BPR for some periods p, q such that $p < q < 2p$. Thus, $w_p = w_q = 1, w_{2p-q} = 0$ and it exists i such that $2 \leqslant i \leqslant \min(\lfloor p/(q-p) \rfloor, \lfloor (n-p)/(q-p) \rfloor)$ and $w_{p-i(q-p)} = 1$. Consider w' the suffix of w starting at position $p - i(q-p)$. w' has length $n - p + i(q-p)$. We have $w'_{i(q-p)} = w'_{(i+1)(q-p)} = 1$ corresponding to periods p and q of w. We show that $i(q-p) + (i+1)(q-p) < n - p + i(q-p) + \gcd(i(q-p), (i+1)(q-p))$, which means that periods $i(q-p)$ and $(i+1)(q-p)$ make w' violate Theorem 1.1. We have

$$\begin{aligned} n > p + i(q-p) &\Leftrightarrow n - p > i(q-p) \\ &\Leftrightarrow n - p + i(q-p) > 2i(q-p) \end{aligned}$$

and

$$\begin{aligned} 2i(q-p) &= 2(i+1)(q-p) - (q-p) \\ &= i(q-p) + (i+1)(q-p) - (q-p) \\ &= i(q-p) + (i+1)(q-p) - \gcd(i(q-p), (i+1)(q-p)). \end{aligned}$$

This implies that w' violates Theorem 1.1, which is impossible because of the construction of w. Therefore, it contradicts the hypothesis that w violates the BPR and we have proven that $w \in \Gamma_n$.

Let us now prove that w is minimal. Assume $\exists y \in \Gamma_n$: $(u \cup v) \subset y \subsetneq w$. Let us denote by i the smallest index such that $y_i = 0$ and $w_i = 1$. Either was i added by the procedure FC_n or by the procedure FW_n. It means that y violates the FPR or Theorem 1.1, in the first and second case respectively. This contradicts $y \in \Gamma_n$ and we have proven the theorem. □

3.2. Γ_n does not satisfy the Jordan–Dedekind condition

The Jordan–Dedekind condition requires that all maximal chains between the same elements have the same length. We demonstrate that Γ_n does not satisfy the Jordan–Dedekind condition, implying that it is neither modular, distributive, nor a matroid. In a partially ordered set or *poset*, a *chain* is defined as a subset of completely ordered elements, an *antichain* as a subset in which any two elements are uncomparable. The length of a chain is its number of elements minus one. The next lemma proves the existence of a specific maximal chain between 1^n and 10^{n-1} in Γ_n.

Lemma 3.5. *Let* $n \in \mathbb{N}$ *and* $p := \lfloor \frac{n}{2} \rfloor + 1$. *The following chain exists in* Γ_n:

$$1^n \succ 10^{p-1}1^{n-p}, \tag{1}$$

$$\forall p \geqslant i \geqslant n-2 : 10^{i-1}1^{n-i} \succ 10^i 1^{n-i-1}, \tag{2}$$

$$10^{n-2}1 \succ 10^{n-1}. \tag{3}$$

Moreover, this chain is maximal and has length $\lceil n/2 \rceil$.

Proof. We prove (1). Obviously, $1^n \supset 10^{p-1}1^{n-p}$. We must show that: if $1^n \supset x \supseteq 10^{p-1}1^{n-p}$ then $x = 10^{p-1}1^{n-p}$. Assume that such an x exists and is different from $10^{p-1}1^{n-p}$. Then $0 < \pi(x) < p$ and $x_{\pi(x)} = 1$. By Lemma 1.1, we have $\forall j < n - \pi(x) + 2$, $x_j = 0$ iff $\pi(x) \nmid j$. Thus, for some $p \leqslant k < n$, $x_k = 0$ and $x \not\supseteq 10^{p-1}1^{n-p}$, which is a contradiction.

The autocorrelations involved in (2) and (3) exist by predicate Ξ and only differ from each other by one period. This implies (2) and (3) and proves that the chain is maximal. By counting the links of the chain, one gets $n - p + 1 = \lceil n/2 \rceil$. □

With $p := \lfloor \frac{n}{2} \rfloor + 1$ as above, consider $\Gamma_{n,p}$ and its associated sub-lattice in Γ_n. From predicate Ξ, we have that $\Gamma_{n,p} = \{10^{p-1}\}\Gamma_{n-p}$. So the structure of the sub-lattice defined by $\Gamma_{n,p}$ is exactly the one of the lattice of Γ_{n-p}. Using the previous lemma, we deduce the existence of an induced maximal chain between $10^{p-1}1^{n-p}$ and $10^{p-1}10^{n-p-1}$ in Γ_n. Combining this with Eq. (1) and $10^{p-1}10^{n-p-1} \succ 10^{n-1}$, we obtain another maximal chain between 1^n and 10^{n-1} in Γ_n. This proves the following lemma.

Lemma 3.6. *Let* $n > 8$ *and* $p := \lfloor \frac{n}{2} \rfloor + 1$ *be integers. The chain going from* 1^n *to* $10^{p-1}1^{n-p}$, *from there to* $10^{p-1}10^{n-p-1}$ *through the induced maximal chain over* $\Gamma_{n,p}$, *and then to* 10^{n-1} *is a maximal chain of* Γ_n. *Its length is* $\lceil (\lceil n/2 \rceil - 1)/2 \rceil + 2$.

Hand inspection for $n := 1, \ldots, 6$ shows that Γ_n satisfies the Jordan–Dedekind condition. We now demonstrate it is not the case when $n > 6$. The representation of Γ_9 given in Fig. 1 illustrates the two maximal chains and the next theorem.

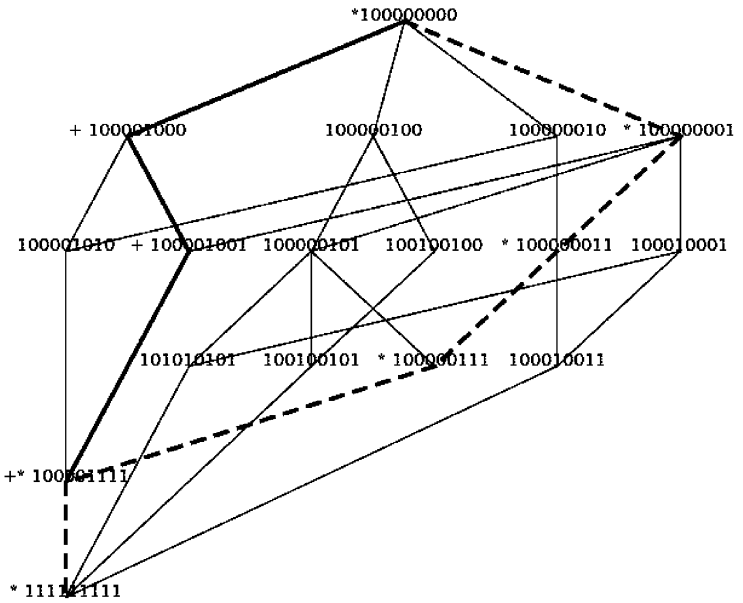

Fig. 1. A representation of the lattice $\Gamma(9)$. The bold-edges and dashed-edges paths shows two maximal chains of different lengths between 111111111 and 100000000. The correlations on these paths are marked with a $+$ or a $*$, respectively.

Theorem 3.3. *For $n>6$, Γ_n does not satisfy the Jordan–Dedekind condition.*

Proof. From Lemmas 3.5 and 3.6, we obtain the existence between 1^n and 10^{n-1} of two maximal chains of lengths $\lceil n/2 \rceil$ and $\lceil (\lceil n/2 \rceil - 1)/2 \rceil + 2$. Clearly, for $n>8$ these are different. Moreover, hand inspection of Γ_7 and Γ_8 shows that they also do not fulfill the Jordan–Dedekind condition. □

3.3. The poset $(\Lambda_n, \subseteq)$ satisfies the Jordan–Dedekind condition

For $n \geqslant 3$, $(\Lambda_n, \subseteq)$ is not a lattice ($\{0,1\}$ and $\{0,2\}$ never have a join). On the other hand, in contrast to Γ_n, we have the stronger result that any subset of an IPS containing 0 is an IPS. This implies that if we define $\Lambda'_n := \{I - \{0\}\colon I \in \Lambda_n\}$, then Λ'_n is a subset family.

Lemma 3.7. *Let $R \in \Lambda_n$ and let $\{0\} \subset Q \subset R$, then $Q \in \Lambda_n$.*

Proof. Let $P := FC_n(R) \in \Gamma_n$. We must show that $P' := FC_n(Q) \in \Gamma_n$, and that no element of Q is deducible from others by the FPR. The latter property follows from the minimality of R. To show $P' \in \Gamma_n$, we only need to consider the special case where $R = Q \dot\cup \{t\}$, i.e., where Q contains exactly one element less than R. The general case follows by repeated application of the special case.

For a contradiction, assume $P' \notin \Gamma_n$. Since P' satisfies the FPR, it must violate the BPR (see characterization 3 of Theorem 1.2). So let $0<p<q<n$ with $\delta := q-p$ such that $p-\delta \notin P'$, but $p-i\delta \in P'$ for some $i \in \{2, \ldots, \min(\lfloor p/(q-p) \rfloor, \lfloor (n-p)/(q-p) \rfloor)\}$. Since P does satisfy the BPR, we must have that $p-\delta \in P$, and this must be a result of adding t to Q and propagating it. From this, we conclude that one of the supposedly non-deducible elements of Q, and hence of R, is in fact deducible from t. So R is not an IPS, a contradiction. □

Theorem 3.4. *The poset $(\Lambda_n, \subseteq)$ satisfies the Jordan–Dedekind condition.*

Proof. From Lemma 3.7, for all pairs $P, Q \in \Lambda_n$: $P \succ Q$ iff $P = Q \cup \{q\}$ for some q in $[1, n-1]$. □

As a corollary of Lemma 3.7, the intersection of two IPSs is an IPS, but the intersections of two IPSs is not the IPS of the intersection of their respective period sets. Neither Γ_n nor Λ_n are closed under union. The union of two IPSs may recursively violate Theorem 1.1 several times, as in the following example: $u := \{0,5,7\}$, $v := \{0,5,8,9\}$, $u \cup v = \{0,5,7,8,9\}$ ($(7,8)$ require 6 in the suffix of length 5, and $(5,6)$ require 1 in the whole $u \cup v$).

4. Enumeration of all autocorrelations of length n

In this section, we present the first efficient enumeration algorithm for string autocorrelations of length n. Another brute force algorithm (in addition to the one mentioned in the Introduction) is to apply predicate Ξ to each of the 2^n possible binary vectors and retain those that satisfy Ξ. This is exponential in n and not practical. The recursive structure of Ξ permits the use of Ξ as the basis of a dynamic programming algorithm that efficiently computes Γ_n from $\Gamma_{m,p}$ with $m<2n/3$ and $1 \leqslant p \leqslant m$. $\Gamma_{n,1} = \{1^n\}$ and $\Gamma_{n,n} = \{10^{n-1}\}$ for all n. Below is the algorithm to compute $\Gamma_{n,p}$ for $n \geqslant 3$ and $2 \leqslant p \leqslant (n-1)$. We assume that all necessary $\Gamma_{m,p}$ with $m<2n/3$ have already been computed.

Case (a) $[2 \leqslant p \leqslant n/2]$: Let $r' := n \bmod p$ and $r := r' + p$. Then $p \leqslant r < 2p$, and there are two sub-cases. In each of them, $\Gamma_{n,p}$ can be constructed from a subset of Γ_r. Let $s_{n,p} := (10^{p-1})^{\lfloor n/p \rfloor - 1}$; every correlation in $\Gamma_{n,p}$ is of the form $s_{n,p}w$ with $w \in \Gamma_r$ chosen as follows:

1. Case $r = p$:

$$\Gamma_{n,p} = \{s_{n,p}w \mid w \in \Gamma_{r,p'}; r' + \gcd(p,p') < p' < p\} \tag{4}$$

2. Case $p < r < 2p$:

$$\Gamma_{n,p} = \{s_{n,p}w \mid w \in \Gamma_{r,p}\} \dot{\bigcup} \{s_{n,p}w \mid w \in \Gamma_{r,p'}; r' + \gcd(p,p') < p' < p; w_p = 1\} \tag{5}$$

In (4) and (5), the inequality $(r' + \gcd(p, p') < p' < p)$ implies that p' does not divide p.

Case (b) $[(n/2) < p \leqslant (n-1)]$: $\Gamma_{n,p}$ is constructed from $\Gamma_{(n-p)}$.

$$\Gamma_{n,p} = \{1 0^{p-1} w \mid w \in \Gamma_{(n-p)}\} \tag{6}$$

Proof (Correctness). Comparison with Ξ reveals that every element that is included in $\Gamma_{n,p}$ according to each of (4), (5), or (6) fulfills Ξ. (Case (a) of Ξ has been further subdivided into $r = p$ and $p < r < 2p$.) It remains to be shown that every vector satisfying Ξ is included in the appropriate $\Gamma_{n,p}$. If this is not the case, let v be a vector of minimal length n that is an autocorrelation, but that is not included in $\Gamma_{n,p}$ where $p = \pi(v)$. The only way this could happen would be if the r-suffix of v were already not contained in its appropriate $\Gamma_{r,p'}$. But this would contradict the minimality of n. □

Improvements. Two improvements increase the efficiency and allow computation up to $n = 450$.

1. For given values of n and p, all autocorrelations in $\Gamma_{n,p}$ have the same prefix. The prefix length is p for $p > n/2$ and $p(\lfloor n/p \rfloor - 1)$ for $p \leqslant n/2$. This prefix is immediately available, and need not be stored explicitly.

2. In case (a), $\Gamma_{n,p}$ is obtained from autocorrelations $w \in \Gamma_r$ with $r \geqslant p$. By Lemma 1.1, such w must satisfy $\pi(w) > (n \bmod p)$, and therefore it is possible to construct $\Gamma_{n,p}$ from the sets Γ_s with $s < p$. Hence, to obtain $\Gamma_{n,p}$, in both cases (a) and (b), only the sets $\Gamma_{m,p'}$ with $m \leqslant \lfloor n/2 \rfloor$, $1 \leqslant p' \leqslant m$ are needed. For example, to compute Γ_{200}, we only need to know $\Gamma_1, \ldots, \Gamma_{100}$ and their respective subsets, but not $\Gamma_{101}, \ldots, \Gamma_{133}$.

5. Bounds on the number of autocorrelations

In this section, we investigate how the number κ_n of different autocorrelations of length n grows with n. From Theorem 1.2, we know that κ_n is independent of the alphabet size. In [9], it is shown that as $n \to \infty$,

$$\frac{1}{2 \ln 2} + o(1) \leqslant \frac{\ln \kappa_n}{(\ln n)^2} \leqslant \frac{1}{2 \ln(3/2)} + o(1). \tag{7}$$

As shown in Fig. 2, these bounds are rather loose. In fact, for small n, the actual value of κ_n is below its asymptotic lower bound. While we conjecture that $\lim_{n \to \infty} \frac{\ln \kappa_n}{(\ln n)^2} = \frac{1}{2 \ln 2}$, it remains an open problem to derive a tight upper bound and prove this conjecture. Our contribution is that a good lower bound for κ_n is closely related to the number of binary partitions of an integer. Both improved bounds we derive from this relationship are also shown in Fig. 2.

We have $\kappa_0 = 1$, $\kappa_1 = 1$, and $\kappa_2 = 2$. Considering only the correlations given by case (b) of predicate Ξ, we have

$$\kappa_n \geqslant \sum_{n/2 < p \leqslant n} \kappa_{n-p} = \sum_{i=0}^{\lceil n/2 \rceil - 1} \kappa_i. \tag{8}$$

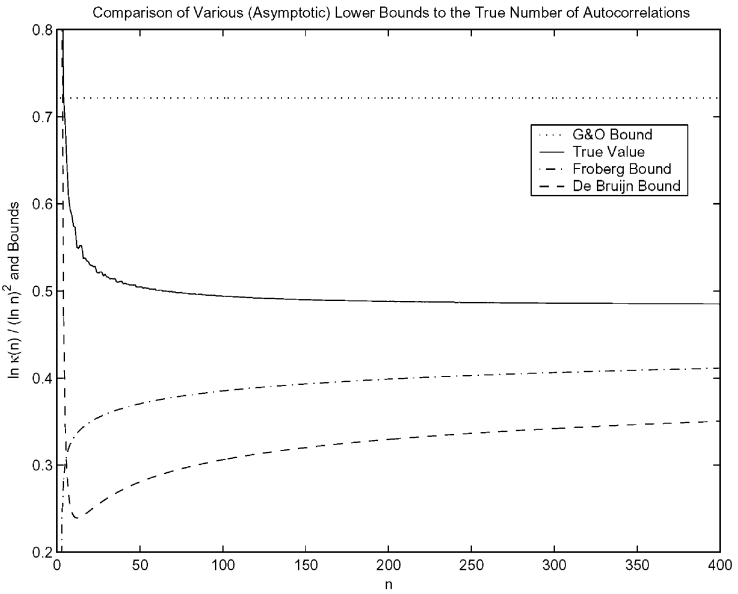

Fig. 2. True values of $\ln \kappa_n/(\ln n)^2$ for $n \leqslant 400$, compared to Guibas and Odlyzko's (G&O) asymptotic lower bound, the improved asymptotic bound from Theorem 5.1(ii) derived from DeBruijn's results, and the non-asymptotic lower bound from Theorem 5.1(i) based on Fröberg's work. Both of these bounds converge to the G&O asymptotic value of $1/(2\ln 2)$ for $n \to \infty$. The upper bound of G&O, corresponding to the line $y = 1/(2\ln(3/2)) \approx 1.23$, is not visible on the figure.

We define $L_0 := 1$, $L_1 := 1$, and, for $n \geqslant 2$, $L_n := \sum_{i=0}^{\lceil n/2 \rceil - 1} L_i$. By induction, $L_n \leqslant \kappa_n$ for all $n \geqslant 0$. From the definition of L_n, we deduce that for $n \geqslant 2$:

$$L_n = \begin{cases} L_{n-1}, & n \text{ even}, \\ L_{n-2} + L_{(n-1)/2}, & n \text{ odd}. \end{cases} \tag{9}$$

Now we consider a related sequence: the number of binary partitions B_n of an integer $n \geqslant 0$, i.e., the number of ways to write n as a sum of powers of 2 where the order of summands does not matter. For example, 6 can be written as such a sum in 6 different ways: $4+2$, $4+1+1$, $2+2+2$, $2+2+1+1$, $2+1+1+1+1$, $1+1+1+1+1+1$. Therefore $B_6 = 6$. By convention, $B_0 = 1$; furthermore $B_1 = 1$. Binary partitions have been extensively studied; for example, see [5,8]. For $n \geqslant 2$, they satisfy the recursion

$$B_n = \begin{cases} B_{n-2} + B_{n/2}, & n \text{ even}, \\ B_{n-1}, & n \text{ odd}. \end{cases}$$

The following lemma states the close relation between the lower bound L_n for κ_n and the number of binary partitions B_n.

Lemma 5.8. *For* $n \geqslant 1$, $L_n = 1/2 \cdot B_{n+1}$.

Proof. The proof is by induction. For $n = 1$, we have $L_1 = 1 = 1/2 \cdot B_2$. If $n \geqslant 2$ is even, $L_n = L_{n-1} = B_{(n-1)+1}/2 = B_{n+1}/2$, as $(n+1)$ is then odd. If $n \geqslant 3$ is odd,

$$L_n = L_{n-2} + L_{\frac{n-1}{2}} = \frac{1}{2}(B_{n-1} + B_{\frac{n+1}{2}}) = \frac{1}{2} \cdot B_{n+1},$$

by the recursion for B_{n+1} for even $(n+1)$. □

We state some known properties of B_n from Fröberg [8] and De Bruijn [5].

Results on the number of binary partitions, B_n: In [8], Fröberg proves the following: Define

$$F(n) := \sum_{k=0}^{\infty} \frac{n^k}{2^{\frac{k(k+1)}{2}} k!}. \tag{10}$$

Then $B_n = C_n \cdot F(n)$, where (C_n) is a sequence bounded between $0.63722 < C_n < 1.920114$ for all $n \geqslant 0$. It is estimated (but unproven) that C_n tends to a limit $C: \approx 0.923307 \pm 0.000001$ as $n \to \infty$.

De Bruijn [5] shows that for an even integer $n = 2m$,

$$\ln B_{2m} = \frac{(\ln m - \ln\ln m)^2}{2\ln 2} + \left(\frac{1}{2} + \frac{1+\ln\ln 2}{\ln 2}\right)\ln m - \left(1 + \frac{\ln\ln 2}{\ln 2}\right)\ln\ln m + O(1). \tag{11}$$

Combining Lemma 5.8, Fröberg's and De Bruijn's results allows us to derive good lower bounds on κ_n in the next theorem.

Theorem 5.1 (Lower bounds on κ_n). *Define* $F(n)$ *as in Eq.* (10). (i) *For all* $n \geqslant 1$, $\kappa_n \geqslant 0.31861 F(n+1)$. (ii) *Asymptotically* (*with approximated constants*),

$$\frac{\ln \kappa_n}{(\ln n)^2} \geqslant \frac{1}{2\ln 2}\left(1 - \frac{\ln\ln n}{\ln n}\right)^2 + \frac{0.4139}{\ln n} - \frac{1.47123 \ln\ln n}{(\ln n)^2} + O\left(\frac{1}{(\ln n)^2}\right).$$

Proof. We know $\kappa_n \geqslant L_n = B_{n+1}/2$ for all $n \geqslant 1$. The first bound follows directly from Fröbergs results. To derive the second, note that (11) also holds for odd integers $m = 2n+1$, since $\ln(m+1) = \ln m + O(1/m)$. Then, replacing $n = m/2$ in (11), and re-sorting the terms, we obtain

$$\ln B_n = \frac{1}{2\ln 2}(\ln n - \ln\ln n)^2 + \left(\frac{1}{\ln 2} + \frac{\ln\ln 2}{\ln 2} - \frac{1}{2}\right)\ln n - \left(2 + \frac{\ln\ln 2}{\ln 2}\right)\ln\ln n + O(1).$$

Since $\ln \kappa_n \geqslant \ln L_n = \ln(1/2) + \ln B_{n+1} = \ln B_n + O(1)$, the same asymptotic formula yields a lower bound for κ_n. Dividing by $(\ln n)^2$ and approximating the lower-order constants by their numerical values proves the second bound. □

6. Computing the size of populations

The correlation of a string depends on its self-overlapping structure, but is not directly related to its characters. Hence, different strings share the same correlation. For instance over the alphabet $\{a, b\}$, take *abbabba* and *babbabb*. The *population* of a correlation v is the set of strings over Σ whose correlation is v. We wish to compute the *size of the population* of a given correlation, and by extension of all correlations.

In [9], Guibas and Odlyzko exhibit a recurrence linking the population sizes of a correlation and of its nested correlation. Here, we exhibit another recurrence which links the population size of an autocorrelation v to the population sizes of the autocorrelations it is included in. The recurrence depends on the *number of free characters* (nfc for short) of v, to be defined next.

Definition 6.3 (Number of free characters). The nfc of a correlation v is the maximum number of positions in a string U with $P(U) = v$ that are not determined by the periods.

To illustrate this definition, note that a correlation represents a set of equalities between the characters of a string. For example, take $v := 100001001 \in \Gamma_9$. A string $U = u_0 \ldots u_8$ with $P(U) = v$ must satisfy the following set of equations: $\{u_0 = u_3 = u_5 = u_8, u_1 = u_6, u_2 = u_7\}$. Thus we can write any word U as $u_0u_1u_2u_0u_4u_0u_1u_2u_0$ for some $u_0, u_1, u_2, u_4 \in \Sigma$. So the nfc of v is 4.

The nfc is independent of Σ and can be computed from v alone. Given a correlation v and its length n, Algorithm 2 (NFC), computes the nfc of v. NFC follows the recursive structure of predicate Ξ and requires $\Theta(n)$ time.

Algorithm 2: *NFC*

Input: $n \in \mathbb{N}$, $v \in \Gamma_n$; **Output**: the number of free characters of v;

```
1 i := 1;
2 while (i < n) and (v_i ≠ 1) do i := i+1;// search for the basic period;
3 if i = n then return n;// no basic period ;
4 if i = 1 then return 1;
5 if (i ≤ ⌊n/2⌋) then
6   return NFC(i + mod(n,i), v[n − i − mod(n,i)..n − 1]);
7 else
8   return 2 × i − n + NFC(n − i, v[i..n − 1]);
```

We now state our recurrence on the population sizes.

Theorem 6.1. *Let $n \in \mathbb{N}$ and let v_k be the kth ($k = 1, \ldots, \kappa_n$) autocorrelation of Γ_n. Let ρ_k denote the number of free characters of v_k, and N_k be its population size. We have*

$$N_k = \sigma^{\rho_k} - \sum_{j: v_k \subset v_j} N_j.$$

Proof. For any word U with $P(U) = v_k$ there are ρ_k free positions. For each of the σ^{ρ_k} combinations of ρ_k characters from Σ, we construct a word V satisfying the character equalities associated with v_k, and have $v_k \subseteq P(V)$. We do not necessarily have $v_k = P(V)$, because V may in fact satisfy additional character equalities. Conversely, every word V with $v_k \subseteq P(V)$ is obtained in this way. Therefore

$$\sigma^{\rho_k} = \sum_{j: v_k \subseteq v_j} N_j = N_k + \sum_{j: v_k \subset v_j} N_j,$$

which proves the theorem. □

7. Application: uniform random sampling of period sets

In this section we show how the notion of IPS can be used to uniformly sample from Γ_n without enumerating Γ_n or knowing κ_n. A consequence of Lemma 3.7 is that $\Lambda'_n := \{I - \{0\}: I \in \Lambda_n\}$ is a subset family. This observation leads to a simple Markov chain algorithm for uniform random sampling from Γ_n.

The state space of the Markov chain (X_t) is the set Λ'_n of IPSs (without zeros) for string length n. The chain starts deterministically at $X_0 := \{\}$, and moves from IPS to IPS according to the rules given below. After a sufficiently large number T of steps, the algorithm outputs the forward closure of the current IPS, i.e., $FC_n(X_T \cup \{0\})$, which is in Γ_n.

To make a one-step transition, i.e., to move from X_t to X_{t+1}, the following is done:

1. Draw a random variable $R \in \{1, 2, \ldots, n-1\}$ according to the uniform distribution on $\{1, 2, \ldots, n-1\}$.

2. If $R \in X_t$, then $X_{t+1} := X_t \backslash \{R\}$; otherwise, if $X_t \cup \{R\} \in \Lambda'_n$, then $X_{t+1} := X_t \cup \{R\}$; otherwise do nothing, i.e. $X_{t+1} := X_t$.

To prove that this procedure works, we first show that the Markov chain never leaves the set of valid IPSs.

Lemma 7.1. *In the Markov chain defined above, $X_t \in \Lambda'_n$ for all $t \geqslant 0$. Every step of the chain takes $O(n)$ time.*

Proof. Clearly, $X_0 = \{\} \in \Lambda'_n$. Assume that $X_t \in \Lambda'_n$ for some $t \geqslant 0$. We face three alternatives: $X_{t+1} = X_t \backslash \{R\}$ for some $R \in X_t$, $X_{t+1} = X_t$, or $X_{t+1} = X_t \cup \{R\}$ for some $R \notin X_t$. In the first case, $X_{t+1} \in \Lambda'_n$ since Λ'_n is a subset family. In the second case, there is nothing to prove and in the third case, $X_{t+1} \in \Lambda'_n$ is explicitly checked, and this check requires linear time by Lemma 2.2. □

To show that the proposed Markov chain indeed solves the uniform random sampling problem, we will prove that it is ergodic (i.e., irreducible and aperiodic) and converges to the uniform distribution on Λ'_n. To that end, we shall examine the structure of the transition matrix $(P_{ST})_{S,T\in\Lambda'_n}$. Note that an important point of the algorithm is that the transition matrix never has to be computed explicitly.

Let $I_1, \ldots, I_{\kappa_n}$ be an arbitrary enumeration of the elements of Λ'_n, with $I_1 := \{\}$. For $i \neq j$, the entry P_{ij}, i.e., the probability to move from I_i to I_j in one step, is $1/(n-1)$ if either $I_i = I_j \cup \{k\}$ or $I_j = I_i \cup \{k\}$ for some k, and zero otherwise. The diagonal entries P_{ii} are defined by the requirement that the rows of P sum to one. This shows that the transition matrix is symmetric.

The following lemma shows one method how to reach any target set I_j from any starting set I_i in a sufficiently high given number of steps.

Lemma 7.2. *Let $n \geqslant 2$. Define $S_n := 1 + \lfloor \log_{3/2} n \rfloor$, the size bound of IPSs for string length n. For every $k \geqslant 0$, any set $I_j \in \Lambda'_n$ can be reached from any set $I_i \in \Lambda'_n$ in $2S_n + k$ steps with positive probability. In other words, if P is the one-step transition matrix of the Markov chain, then for every $k \geqslant 0$, the matrix P^{2S_n+k} has strictly positive entries.*

Proof. Fix some i and j. Let $s \leqslant S_n$ be the cardinality of I_i, and $t \leqslant S_n$ the cardinality of I_j. Let $x := 2S_n - (s+t) + k - 2$. In the first s steps, the elements of I_i are removed (starting with the largest, say) until $\{\}$ is reached. Then we move to $\{1\}$ and cycle there for x steps, move back to $\{\}$, and from there build up I_j in t steps. (The reason for the detour over $\{1\}$ is convenience; we cannot cycle in $\{\}$ to "wait" some number of steps, because to $\{\}$ any number in $[1, n-1]$ will be added successfully. The set $\{1\}$, however, always exists for $n \geqslant 2$ and is a typical "wait state": Once there, the probability of staying there is $(n-2)/(n-1)$, because the only valid move is back to $\{\}$.) Each of the indicated transitions has a positive probability, $1/(n-1)$. Therefore, we can move from any I_j to any I_j in $t + 1 + x + 1 + s = 2S_n + k$ steps with positive probability. □

Theorem 7.1 (Convergence to the uniform distribution on Λ'_n). *The Markov chain $(X_t)_{t\geqslant 0}$, as defined above, is ergodic. Its stationary distribution is the uniform distribution on Λ'_n, and the chain converges to its stationary distribution.*

Proof. The ergodicity (i.e., aperiodicity and irreducibility) follows from the existence of a power (e.g., $2S_n$) of the transition matrix P that has strictly positive entries. Therefore, there exists a unique stationary distribution π on the states that satisfies (written as a row vector) $\pi P = \pi$.

By the transition rules, P is symmetric, and hence doubly stochastic (each row and column sums to one). Therefore, the uniform distribution $\pi = (\frac{1}{\kappa_n}, \ldots, \frac{1}{\kappa_n})$ satisfies $\pi P = \pi$; so it is the stationary distribution of the Markov chain.

It is a classical result of Markov chain theory (e.g. see [1, Theorem 8.9]) that an ergodic Markov chain converges exponentially fast to its stationary distribution, independently of the start state. □

It is a difficult problem to decide for how long the Markov chain must run to come ε-close to the uniform distribution. It depends on the connectivity of the chain, or the number of paths from one state to another. While has been some remarkable progress towards the convergence analysis of Markov chains (e.g. see [6]), we could not establish a useful bound on the number of required steps in this case.

Acknowledgments

We thank D. Bryant, the groups of S. Schbath at INRA Jouy en Josas, and of Ph. Flajolet at INRIA Rocquencourt for helpful discussions, as well as an anonymous referee for his suggestions on the proof of Lemma 3.3. E.R. is supported by the CNRS, part of this work has been done while working at the DKFZ, in Heidelberg, Germany. S.R. is grateful to LIRMM for a travel grant.

References

[1] P. Billingsley, Probability and Measure, 3rd Edition, Wiley, New York, 1995.
[2] S. Burkhardt, A. Crauser, P. Ferragina, H.-P. Lenhof, E. Rivals, M. Vingron, q-Gram based database searching using a suffix array (QUASAR), in: Third Annual International Conference on Computational Molecular Biology, Lyon, France, 11–14 April 1999, ACM Press, New York, pp. 77–83.
[3] C. Choffrut, J. Karhumäki, Combinatorics of words, in: G. Rozenberg, A. Salomaa (Eds.), Handbook of Formal Languages, Vol. 1, Springer, Berlin, 1997, pp. 329–438.
[4] M. Crochemore, W. Rytter, Text Algorithms, Oxford University Press, Oxford, 1994.
[5] N.G. DeBruijn, On Mahler's partition problem, Proc. Akad. Wet. Amsterdam 51 (1948) 659–669.
[6] J.A. Fill, An interruptible algorithm for perfect sampling via markov chains, Ann. Appl. Probability 8 (1) (1998) 131–162.
[7] N.J. Fine, H.S. Wilf, Uniqueness theorems for periodic functions, Proc. Amer. Math. Soc. 16 (1965) 109–114.
[8] C.-E. Fröberg, Accurate estimation of the number of binary partitions, BIT 17 (1977) 386–391.
[9] L.J. Guibas, A.M. Odlyzko, Periods in strings, J. Combin. Theory Ser. A 30 (1981) 19–42.
[10] D. Gusfield, Algorithms on Strings, Trees and Sequences, Cambridge University Press, Cambridge, 1997.
[11] P. Jokinen, J. Tarhio, E. Ukkonen, A comparison of approximate string matching algorithms, Software Practice and Experience 26 (12) (1996) 1439–1458.
[12] D.E. Knuth, The Art of Computer Programming, Seminumerical Algorithms, Vol. 2, 3rd Edition, Addison-Wesley, Reading, MA, 1998.
[13] M. Lothaire, Algebraic Combinatorics on Words, Cambridge University Press, Cambridge, 1999 URL: http://www-igm.univ-mlv.fr/~berstel/Lothaire/index.html.
[14] G. Marsaglia, A. Zaman, Monkey tests for random number generators, Comput. Math. Appl. 26 (9) (1993) 1–10.
[15] O. Owolabi, D.R. McGregor, Fast approximate string matching, Software Practice Exp. 18 (4) (1988) 387–393.
[16] O.E. Percus, P.A. Whitlock, Theory and application of marsaglia's monkey test for pseudorandom number generators, ACM Trans. Modeling Comput. Simulation 5 (2) (1995) 87–100.
[17] P. Pevzner, Computational Molecular Biology, MIT Press, Cambridge, MA, 2000.
[18] S. Rahmann, E. Rivals, Exact and Efficient Computation of the expected number of missing and common words in random texts, in: R. Giancarlo, D. Sankoff (Eds.), Proceedings of the 11th

Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science, Vol. 1848, Springer, Berlin, 2000, pp. 375–387.
[19] S. Rahmann, E. Rivals, The number of missing words in random texts, Combin. Probab. Comput. 12 (2003) 73–87.
[20] E. Rivals, S. Rahmann, Combinatorics of periods in strings, in: F. Orejas, P. Spirakis, J. van Leuween (Eds.), Proceedings of the 28th ICALP, Lecture Notes in Computer Science, Vol. 2076, Springer, Berlin, 2001, pp. 615–626.
[21] R. Sedgewick, P. Flajolet, Analysis of Algorithms, Addison-Wesley, Reading, MA, 1996.
[22] E. Ukkonen, Approximate string-matching with q-grams and maximal matches, Theoret. Comput. Sci. 92 (1) (1992) 191–211.