Combinatorics of Periods in Strings

Eric Rivals[1] and Sven Rahmann[2]

[1] L.I.R.M.M., CNRS U.M.R. 5506
161 rue Ada, F-34392 Montpellier Cedex 5, France,
`rivals@lirmm.fr`

[2] Max-Planck-Institut für Molekulare Genetik,
Dept. of Computational Molecular Biology,
Ihnestraße 73, D-14195 Berlin, Germany
`rahmann@molgen.mpg.de`

Abstract. We consider the set $\Gamma(n)$ 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 $\Gamma(n)$ is a lattice under set inclusion and does not satisfy the Jordan-Dedekind condition. We propose the first enumeration algorithm for $\Gamma(n)$ and improve upon the previously known asymptotic lower bounds on the cardinality of $\Gamma(n)$. Finally, we provide a new recurrence to compute the number of strings sharing a given period set.

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 [6], 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 $\Gamma(n)$, the set of all correlations of length n; nor has anyone proposed an efficient enumeration algorithm for $\Gamma(n)$.

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 $\Gamma(n)$ is a lattice under set inclusion and does not satisfy the Jordan-Dedekind condition. While $\Lambda(n)$, the set of all irreducible period sets, does satisfy that condition, it does not form a lattice (Section 3). We propose the first enumeration algorithm for $\Gamma(n)$ (Section 4) and improve upon the previously known asymptotic lower bounds for the cardinality of $\Gamma(n)$ (Section 5). Finally, we provide a new recurrence to compute the population sizes of correlations (Section 6).

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 [2] 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 [9]. In another work [8], 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.

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 $U := U_0 U_1 \ldots U_{n-1}$ by $|U|$. For any $0 \le i \le j < n$, $U_{i..j} := U_i \ldots U_j$ 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 (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: $\forall\, 0 \le i < n - p : U_i = U_{i+p}$.

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 [1, 7, 6]. Here, we need the Theorem of Fine and Wilf, also called the GCD-rule, and a useful corollary.

Theorem 1 (Fine and Wilf [4]). *Let $U \in \Sigma^n$. If U has periods p and q with $p \le q$ and $p + q \le n + \gcd(p, q)$, then $\gcd(p, q)$ is also a period.*

Lemma 1. *Let $U \in \Sigma^n$ with smallest non-null period $p \le \lfloor \frac{n}{2} \rfloor$. 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 $g \ge 1$. Therefore, $p + i - g \le n$, and Theorem 1 says that g is a period, contradicting the premise that p is the smallest non-null period. $\square$

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: $\forall\, 0 \le i < n$, $v_i = 1$ 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 $s \circ t$, and the k-fold concatenation of s with itself by s^k. So $10^k \circ w$ is the string starting with 1, followed by k 0s, 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 $\kappa(n)$. The autocorrelations in $\Gamma(n)$ can be partitioned according to their basic period; thus, for $0 \leq p < n$, 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 $\Gamma(n)$. 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 [6], Guibas and Odlyzko characterized the correlations of length n in terms of the Forward Propagation Rule (FPR), the Backward Propagation Rule (BPR), and by a recursive predicate Ξ. We review the main theorem and the definitions.

Theorem 2 (Characterization of Correlations [6]). *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 ≥ 2*
3. *$v_0 = 1$ and v satisfies the Forward and Backward Propagation Rules*
4. *v satisfies the predicate Ξ.*

Definition 2. FPR, BPR, Predicate Ξ. Let $v \in \{0,1\}^n$.

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

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

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: $p \leq \lfloor \frac{n}{2} \rfloor$

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: $p > \lfloor \frac{n}{2} \rfloor$

We have: $\forall j : 1 \leq j < p$, $v_j = 0$. 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 2 (see [6]).

Theorem 3. *Let v be a correlation of length n. Any substring $v_i \ldots v_j$ of v with $0 \le i \le j < n$ 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 $\Gamma(n)$ to $\Lambda(n)$, the set of all IPSs, is bijective. We also show how to compute the IPS from the period set, and conversely.

For every $n \in \mathbb{N}$, we define a function FC_n, the *Forward Closure*, from $2^{[0, n-1]}$ to $2^{[0, n-1]}$. Intuitively, FC_n repeatedly applies the FPR to all pairs of elements until closure is reached. Note that the order in which pairs of elements are considered does not matter, and therefore FC_n is well defined.

Definition 3 (Irreducible Period Set). Let $T \in \Gamma(n)$ be a period set. A subset $S := \{p_0, \ldots, p_l\}$ 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., $\mathrm{FC}_n(S) = T$,
2. For all triples (h, i, j) satisfying $0 \le h < i < j \le l$ we have $\forall\, k \in \mathbb{N}^+ : p_j \ne p_i + k(p_i - p_h)$

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 $\mathrm{FC}_n(S) = T$. 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 4. $R : \Gamma(n) \to \Lambda(n), \ P \mapsto R(P)$ *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 $P = \mathrm{FC}_n(R(P)) = \mathrm{FC}_n(R(Q)) = Q$ by condition (1) of Definition 3. $\qquad \square$

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 $I[0] := P[0]$; $\delta := n$; $i := 1$; $k := 1$; $S := \emptyset$;
2 while $((i < t)$ *and* $(\delta > 1))$ **do**
3 $\quad$ $\delta := P[i] - P[i-1]$; $size := n - P[i-1]$; $mul := \lfloor \frac{size}{\delta} \rfloor$;
4 $\quad$ **if** $P[i] \notin S$ **then**
5 $\quad\quad$ $I[k] := P[i]$; $k := k+1$;
6 $\quad\quad$ **if** $mul = 2$ **then**
7 $\quad\quad\quad$ **if** $\mathrm{mod}(size, \delta) \neq 0$ **then** $S.insert(P[i] + \delta)$;
8 $\quad\quad$ **else if** $mul > 2$ **then** $S.insert(P[i-1] + mul \times \delta)$; $i := i + mul - 2$;
9 $\quad$ $i := i + 1$;
10 return I;

By Theorem 4, R^{-1} exists; indeed, it is FC_n restricted to $\Lambda(n)$. Algorithm 1 is an efficient implementation of R. The next theorem claims that R runs in a time sublinear in the input size (which may be as large as $\Theta(n)$) because $|R(P)| = O(\log n)$ (We omit the proof and the algorithm R^{-1}.) This is achieved by exploiting the known structure of period sets; the algorithm does not need to examine the whole input array P (cf. line 8 of R).

Theorem 5. *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 $mul \geq 2$. If $mul = 2$ and $\mathrm{mod}(size, \delta) \neq 0$, the forward propagation generates only $P[i] + \delta$ which is inserted into S (lines 6 and 7). If $mul > 2$, Lemma 1 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. $\square$

3 Structural Properties of $\Gamma(n)$ and $\Lambda(n)$

3.1 $\Gamma(n)$ is a Lattice Under Inclusion.

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

Lemma 2. *If $u, v \in \Gamma(n)$, then $(u \cap v) \in \Gamma(n)$.*

Proof. Let $u, v \in \Gamma(n)$ and $w := u \cap v$. The results hold when $n = 1$. If $w = \{0\}$ we are done. Otherwise assume that for all $q < n, u', v' \in \Gamma(q)$ we have $(u' \cap v') \in \Gamma(q)$. Let p be the smallest common non-null period of u and v. So p is the smallest non-null period of w.

Case $p \leq \lfloor \frac{n}{2} \rfloor$: Let $i := \lfloor \frac{n}{p} \rfloor$. We have that multiples of p are periods of u and v: $\forall\, 1 \leq j \leq i$; $u_{p \cdot j} = v_{p \cdot j} = 1$ and so $w_{p \cdot j} = 1$. Moreover, we have $\forall\, 0 < k < n - p$, $k \neq jp$: $u_k \neq v_k$, otherwise p would not be the smallest common period of u and v. Hence, for all such k: $w_k = 0$. Consider the suffixes u', v', w' of length $n' := n - (i - 1)p$ of u, v and w respectively. We know that $w = (10^{p-1})^{i-1} \circ w'$, $u'_0 = v'_0 = u'_p = v'_p = 1$, and $w' = u' \cap v'$. As from Theorem 3, we know that $u', v' \in \Gamma(n')$, we have by induction that $w' \in \Gamma(n')$. Because w' satisfies the Theorem of Fine and Wilf, we have $\pi(w') + p > n' + \gcd(\pi(w'), p)$. Hence, w satisfies Predicate Ξ, i.e., $w \in \Gamma(n)$.

Case $p > \lfloor \frac{n}{2} \rfloor$: Let u', respectively v', be the suffix of length $n - p$ of u, resp. of v. By Theorem 3, u', v' are autocorrelations of size $n - p$. As $w = 10^{p-1} \circ (u' \cap v')$, by induction it fulfills Predicate Ξ. $\qquad\square$

Lemma 3. *$(\Gamma(n), \subseteq)$ has a null element, 10^{n-1}, and a universal element, 1^n.*

Theorem 6. *$(\Gamma(n), \subseteq)$ is a lattice.*

Proof. From Lemma 2, we know that $\Gamma(n)$ 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. $\qquad\square$

3.2 $\Gamma(n)$ does not satisfy the Jordan-Dedekind condition.

We demonstrate that $\Gamma(n)$ does not satisfy the Jordan-Dedekind condition, implying that it is neither modular, distributive, nor a matroid. The next lemma proves the existence of a specific maximal chain[1] between 1^n and 10^{n-1} in $\Gamma(n)$.

Lemma 4. *Let $n \in \mathbb{N}$ and $p := \lfloor \frac{n}{2} \rfloor + 1$. The following chain exists in $\Gamma(n)$:*

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

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

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

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

---

[1] 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.

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, we have $\forall\, j < n - \pi(x) + 2$, $x_j = 0$ iff $\pi(x) \nmid j$. Thus, for some $p \leq k < n$, $x_k = 0$ 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 $n - p + 1 = \lceil \frac{n}{2} \rceil$. □

With $p := \lfloor \frac{n}{2} \rfloor + 1$ as above, consider $\Gamma(n, p)$ and its associated sub-lattice in $\Gamma(n)$. From Predicate Ξ, we have that $\Gamma(n, p) = \{10^{p-1}\} \circ \Gamma(n-p)$. So the structure of the sub-lattice defined by $\Gamma(n, p)$ is exactly the one of the lattice of $\Gamma(n - p)$. 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 $\Gamma(n)$. Combining this with Equation 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 $\Gamma(n)$. This proves the following lemma.

Lemma 5. *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 $\Gamma(n)$. Its length is $\lceil \frac{\lceil \frac{n}{2} \rceil - 1}{2} \rceil + 2$.*

Hand inspection for $n := 1, \ldots, 6$ shows that $\Gamma(n)$ satisfies the Jordan-Dedekind condition, i.e., all maximal chains between the same elements have the same length. We now demonstrate it is not the case when $n > 6$.

Theorem 7. *For $n > 6$, $\Gamma(n)$ does not satisfy the Jordan-Dedekind condition.*

Proof. From lemmas 4 and 5, we obtain the existence between 1^n and 10^{n-1} of two maximal chains of lengths $\lceil \frac{n}{2} \rceil$ and $\lceil \frac{\lceil \frac{n}{2} \rceil - 1}{2} \rceil + 2$. Clearly, for $n > 8$ these are different. Moreover, hand inspection of $\Gamma(7)$ and $\Gamma(8)$ shows that they also do not fulfill the Jordan-Dedekind condition. □

3.3 The poset $(\Lambda(n), \subseteq)$ satisfies the Jordan-Dedekind condition.

For $n \geq 3$, $(\Lambda(n), \subseteq)$ is not a lattice ($\{0, 1\}$ and $\{0, 2\}$ never have a join). On the other hand, in contrast to $\Gamma(n)$, we have the stronger result that any subset of an IPS containing 0 is an IPS.

Lemma 6. *Let $R \in \Lambda(n)$ and let $\{0\} \subset Q \subset R$, then $Q \in \Lambda(n)$.*

Proof. Let $P := \mathrm{FC}_n(R) \in \Gamma(n)$. We must show that $P' := \mathrm{FC}_n(Q) \in \Gamma(n)$, 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 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

$i \in \{2, \ldots, \min(\lfloor \frac{p}{q-p} \rfloor, \lfloor \frac{n-p}{q-p} \rfloor)\}$. 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. $\square$

Theorem 8. *The set $\Lambda(n)$ of all Irreducible Period Sets is partially ordered and satisfies the Jordan-Dedekind condition with respect to set inclusion.*

Proof. Clearly, set inclusion induces a partial order on $\Lambda(n)$. From Lemma 6, for all pairs $P, Q \in \Lambda(n)$: $P \succ Q$ iff $P = Q \cup \{q\}$ for some q in $[1, n-1]$. Thus, any two maximal chains between the same element have the same length. $\square$

As a corollary of Lemma 6, 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 $\Gamma(n)$ nor $\Lambda(n)$ are closed under union. The union of two IPSs may recursively violate Theorem 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 enumeration algorithm for string autocorrelations of length n. A brute force algorithm 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 $\Gamma(n)$ from $\Gamma(m, p)$ with $m < 2n/3$ and $1 \leq p \leq m$. $\Gamma(n, 1) = \{1^n\}$ and $\Gamma(n, n) = \{10^{n-1}\}$ for all n. Below is the algorithm to compute $\Gamma(n, p)$ for $n \geq 3$ and $2 \leq p \leq (n - 1)$. We assume that all necessary $\Gamma(m, p)$ with $m < 2n/3$ have already been computed.

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

1. Case $r = p$:

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

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

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

In (4) and (5) : $(r' + \gcd(p, p') < p' < p) \Rightarrow p' \nmid p$.

Case (b) $[\frac{n}{2} < p \le (n-1)]$: $\Gamma(n, p)$ is constructed from $\Gamma(n-p)$.

$$\Gamma(n, p) = \{10^{p-1} \circ w \mid w \in \Gamma(n-p)\} \tag{6}$$

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 $p > \frac{n}{2}$ and $p (\lfloor n/p \rfloor - 1)$ for $p \le \frac{n}{2}$. 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 $r \ge p$. By Lemma 1, such w must satisfy $\pi(w) > (n \bmod p)$, and therefore it is possible to construct $\Gamma(n, p)$ from the sets $\Gamma(s)$ with $s < p$. Hence, to obtain $\Gamma(n, p)$, in both cases (a) and (b), only the sets $\Gamma(m, p')$ with $m \le \lfloor \frac{n}{2} \rfloor$, $1 \le p' \le m$ are needed. For example, to compute $\Gamma(200)$, 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 $\kappa(n)$ of different autocorrelations of length n grows with n. From Theorem 2, we know that $\kappa(n)$ is independent of the alphabet size. In [6], it is shown that as $n \to \infty$,

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

As shown in Figure 1, these bounds are rather loose. In fact, for small n, the actual value of $\kappa(n)$ 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 Figure 1.

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 $\kappa_n \ge \sum_{n/2 < p \le n} \kappa_{n-p} = \sum_{i=0}^{\lceil n/2 \rceil - 1} \kappa_i$. We define $L_0 := 1$, $L_1 := 1$, and, for $n \ge 2$, $L_n := \sum_{i=0}^{\lceil n/2 \rceil - 1} L_i$. By induction, $L_n \le \kappa_n$ for all $n \ge 0$. From the definition of L_n, we deduce that for $n \ge 2$, $L_n = L_{n-1}$ for n even, and $L_n = L_{n-2} + L_{\frac{n-1}{2}}$ for n odd.
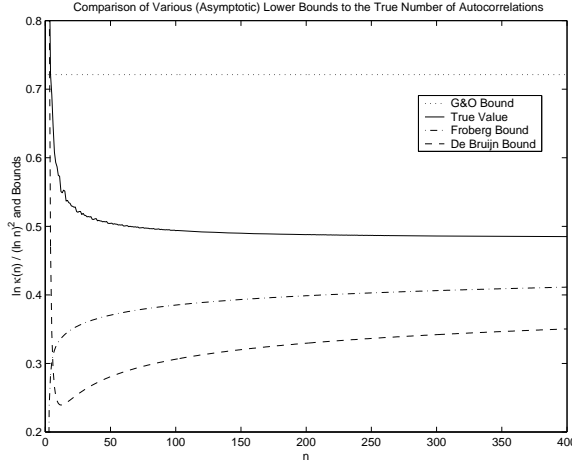
Fig. 1. True values of $\ln \kappa_n/(\ln n)^2$ for $n \leq 400$, compared to Guibas & Odlyzko's (G&O) asymptotic lower bound, the improved asymptotic bound from Theorem 9 (ii) derived from DeBruijn's results, and the non-asymptotic lower bound from Theorem 9 (i) based on Fröberg's work. Both of these 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.

Now we consider a related sequence: the number of binary partitions B_n of an integer $n \geq 0$, 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 [3, 5]. For $n \geq 2$, they satisfy the recursion $B_n = B_{n-2} + B_{\frac{n}{2}}$ for n even and $B_n = B_{n-1}$ for n odd. The following lemma states the close relation between the lower bound L_n for $\kappa(n)$ and the number of binary partitions B_n.

Lemma 7. *For $n \geq 1$, $L_n = 1/2 \cdot B_{n+1}$.*

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

Fröberg [5] and De Bruijn [3] give some bounds on B_n. Combining Lemma 7, Fröberg's and De Bruijn's results allows us to derive good lower bounds on $\kappa(n)$ in the next Theorem (The proof is omitted).

Theorem 9 (Lower Bounds on $\kappa(n)$). *Define*

$$F(n) := \sum_{k=0}^{\infty} \frac{n^k}{2^{\frac{k(k+1)}{2}} \cdot k!}. \tag{8}$$

i/ For all $n \geq 1$, $\kappa_n \geq 0.31861 \cdot F(n+1)$. ii/ Asymptotically (with approximated constants),

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

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 [6], 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 4 (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 $u_0 u_1 u_2 u_0 u_4 u_0 u_1 u_2 u_0$ 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.

We now state our recurrence on the population sizes.

Theorem 10. *Let $n \in \mathbb{N}$ and let v_k be the k-th $(k = 1, \ldots, \kappa(n))$ autocorrelation of $\Gamma(n)$. Let ρ_k denote the number of free characters of v_k, and N_k be its population size. We have:*

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

Algorithm 2: *NFC*

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

1 $i := 1$; **while** $(i < n)$ *and* $(v_i \neq 1)$ **do** $i := i + 1$; // search for the basic period ;
2 **if** $i = n$ **then return** n; // no basic period ;
3 **if** $i = 1$ **then return** *1*;
4 **if** $(i \leq \lfloor \frac{n}{2} \rfloor)$ **then** **return** $NFC(i + \mathrm{mod}(n, i), v[n - i - \mathrm{mod}(n, i)..n - 1])$;
5 **else** **return** $2 \times i - n + NFC(n - i, v[i..n - 1])$;

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. □

References

1. C. Choffrut and J. Karhumäki. Combinatorics of words. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, pages 329–438. Springer-Verlag, 1997.
2. M. Crochemore and W. Rytter. *Text Algorithms*. Oxford University Press, 1994.
3. N. G. DeBruijn. On Mahler's partition problem. *Proc. Akad. Wet. Amsterdam*, 51:659–669, 1948.
4. N. J. Fine and H. S. Wilf. Uniqueness theorems for periodic functions. *Proc. Amer. Math. Soc.*, 16:109–114, 1965.
5. C.-E. Fröberg. Accurate estimation of the number of binary partitions. *BIT*, 17:386–391, 1977.
6. L. J. Guibas and A. M. Odlyzko. Periods in strings. *Journal of Combinatorial Theory*, Series A, 30:19–42, 1981.
7. M. Lothaire. *Algebraic Combinatorics on Words*. in preparation, 1999. URL: http://www-igm.univ-mlv.fr/~berstel/Lothaire/index.html.
8. S. Rahmann and E. Rivals. Exact and Efficient Computation of the Expected Number of Missing and Common Words in Random Texts. In R. Giancarlo and D. Sankoff, editors, *Proc. of the 11th Symposium on Combinatorial Pattern Matching*, number 1848 in LNCS, pages 375–387, Montréal, Canada, 2000. Springer-Verlag, Berlin.
9. R. Sedgewick and P. Flajolet. *Analysis of Algorithms*. Addison-Wesley, Reading, MA, 1996.