

Enumeration of Autocorrelations and Computation of Their Populations

Éric Rivals

LIRMM, Université Montpellier II

November 22, 1999

Summary by Pierre Nicodème

Abstract

This talk presents in a first part Guibas and Oldlyzko's characterization of autocorrelations and in a second part algorithms developed by Éric Rivals with Sven Rahmann (TBI, DKFZ, Heidelberg) to enumerate the autocorrelations and to simultaneously compute their populations.

1. Introduction

An interesting statistics about a random text of size N is the number of different words of a given size n it contains, or, equivalently, how many words of size n are missing in the random text. These statistics are closely linked with the autocorrelations of the words, that are sets of periods of the words. We consider here the enumeration of autocorrelations and the populations of the autocorrelations, originally studied by Guibas and Odlyzko [3]. The original motivation of Rivals and Rahmann comes from searching genomic databases with q -grams [1].

2. Definitions

We consider a finite *alphabet* Σ . Let $w = w_1w_2 \cdots w_n$ where $w_i \in \Sigma$. A *period* of w is an integer p such that for all i between 1 and $n - p$ we have $a_i = a_{i+p}$. As an example, the word **abracadabra** has for periods 0, 7, and 10. Its factor **abra** has for periods 0 and 1. The *autocorrelation vector* of a word w , denoted by $V(w)$, is the binary vector $V = (v_0, v_1, \dots, v_{n-1})$ such that v_i is equal to one if i is a period of w and to zero otherwise. Alternatively, the autocorrelation will be denoted by the corresponding binary word $v_0v_1 \cdots v_{n-1}$. We denote $\Pi(w)$ the set of autocorrelations of the word w .

We are interested in statistics about the whole set of words of size n and therefore denote $\Gamma(n)$ the set of autocorrelations of size n and $\kappa(n)$ its cardinality.

The periods have the following properties:

1. 0 is always a period;
2. if i is a period, then for all i in the range $(1, \lfloor n/p \rfloor]$ the integer ip is a period;
3. if p and q are periods of w , with $p < q$, then $q - p$ is a period of the prefix of length $n - p$ of w .

Theorem 1 (Fine and Wilf). *Let p and q be periods of a word w , with $p < q$. If $p+q \leq |w| + \gcd(p, q)$ then $\gcd(p, q)$ is a period of w .*

See [2, 3] for this theorem.

3. Periods in Strings

This section follows the lines of Guibas and Odlyzko [3].

In order to give equivalent characterizations of autocorrelations vectors within binary vectors in Theorem 2 below, we now give the definitions of the forward and backward propagation rules and of the Ξ predicate that are used in this theorem.

If $p < q$ are periods of a word w , then $q + (q - p)$ is also period. This gives the following rule.

Definition 1 (Forward Propagation Rule). A binary vector $V = (v_0, v_1, \dots, v_n)$ satisfies the forward propagation rule if, whenever we have $v_p = v_q = 1$ with $p < q$, we also have $v_t = 1$ for all t in $[p, n)$ such that $t = p + i(q - p)$ with $i = 0, 1, 2, \dots$.

The backward propagation rule asserts that if p and q are periods with $p < q$ and if $p - (q - p)$ is not a period, then none of the positive integers $p - i(q - p)$ may be a period.

Definition 2 (Backward Propagation Rule). A binary vector $V = (v_0, v_1, \dots, v_{n-1})$ satisfies the backward propagation rule if the following condition holds. Consider every p and q such that $p < q \leq 2p$ with $v_p = v_q = 1$, but $v_{2p-q} = 0$; then for all t in the range $[0, 2p - q]$ such that $t = p - i(q - p)$ and i belongs to the interval $\left[1, \left\lfloor \frac{n-p}{q-p} \right\rfloor\right]$ we have $v_t = 0$.

We now introduce a recursive predicate on binary vectors that is equivalent to the condition that the binary vector is an autocorrelation vector. In the following, we note the shortest period of a word v by $\pi(v)$.

Definition 3 (Recursive Predicate Ξ). Let $V = (v_0, v_1, \dots, v_{n-1})$ be a non-empty binary vector. Define $p = \pi(v_0 v_1 \dots v_{n-1})$. The vector V satisfies the predicate Ξ if and only if V is such that $v_0 = 1$ and V satisfies one of the following two conditions:

– *Case (A)*, $p \leq \left\lfloor \frac{n}{2} \right\rfloor$.

Let $r = n \bmod p$ and $q = p + r$ and let $w = w_1 \dots w_q$ be the suffix of $v_0 v_1 \dots v_{n-1}$ of length q .

Then:

1. for all j in the range $[1, n - q]$, $v_j = 1$ if $j = ip$ for some i , and $v_j = 0$ otherwise;
2. $w_p = 1$ or $r = 0$;
3. if $\pi(w) < p$ then $\pi(w) > (q - p) + \gcd(p, \pi(w))$;
4. the vector (w_1, \dots, w_q) satisfies predicate Ξ .

– *Case (B)*, $p > \left\lfloor \frac{n}{2} \right\rfloor$.

Let $w = w_1 \dots w_{n-p}$ be the suffix of $v_0 \dots v_{n-1}$ of length $n - p$. Then for all j in the range $[1, n - p]$ we have $v_j = 0$ and the vector (w_1, \dots, w_{n-p}) satisfies predicate Ξ .

The algorithmic check of the predicate Ξ requires $O(n)$ operations on a vector V of size n .

Theorem 2. Let $V = (v_0, v_1, \dots, v_n)$ be a non-empty binary vector. Then the following four statements are equivalents:

1. V is a correlation vector of a binary string;
2. V is a correlation vector of some string;
3. $v_0 = 1$ and V satisfies the forward and backward propagation rules;
4. V satisfies the predicate Ξ .

Note that equivalence between statements 1 and 2 implies that the characterization of an autocorrelation vector is independent of the size of the alphabet.

```

Autocorrelations(n)
  if  $n = 1$  then return {1}
  elif  $n = 2$  then return {11, 10}
  else
     $\Gamma(n) := \{\}$ 
    # Case (A),  $p \leq \lfloor \frac{n}{2} \rfloor$ 
    for  $p$  for  $\lfloor \frac{n}{3} \rfloor$  to  $\lfloor \frac{n}{2} \rfloor$  do
       $r := n \bmod p$ 
       $q := p + r$ 
       $\Gamma(q) := \text{Autocorrelations}(q)$ 
       $j_0 := \min \{ j \mid j + p > q + \gcd(j, p) \}$ 
      for  $w$  in  $\Gamma(q)$  do
        if  $\pi(w) > j_0$  and  $p \bmod \pi(w) \neq 0$  then
           $\Gamma(q) := \Gamma(q) \cup \left\{ (10^{p-1})^{\lfloor \frac{n}{p} \rfloor - 1} w \right\}$ 
        fi
      od
    od
    # Case (B),  $p > \lfloor \frac{n}{2} \rfloor$ 
    for  $p$  for  $\lfloor \frac{n}{2} \rfloor$  to  $n$  do
       $\Gamma(n-p) := \text{Autocorrelations}(n-p)$ 
      for  $w$  in  $\Gamma(n-p)$  do
         $\Gamma(n) := \Gamma(n) \cup \{ 10^{p-1} w \}$ 
      od
    od
  return  $\Gamma(n)$ 
fi
end

 $u^p$  is the word  $u \dots u$  where  $u$  is repeated  $p$  times

```

FIGURE 1. Recursive algorithm **Autocorrelations**.

4. An Algorithm to Enumerate all Autocorrelations of Size n

We use the predicate Ξ to build a recursive bottom-up procedure that constructs autocorrelation vectors. To this end, note that the condition (2) of Case (A) of the predicate Ξ is equivalent to

$$\pi(w) \text{ does not divide } p \quad \text{and} \quad \pi(w) > j_0 = \min \{ j \mid j + p > q + \gcd(j, p) \}.$$

Algorithm **Autocorrelation** to enumerate all autocorrelations until size n is given in Figure 1.

Implementation. The autocorrelations are stored as binary vectors. The implementation has been done as an iterative procedure, although the algorithm presented in Figure 1 is recursive. Note that in Case (A) of the algorithm the tests of conditions (a) and (b) of the Ξ predicate can be done in $O(1)$ operations. Moreover only the valid subset of $\Gamma(q)$ is computed.

Complexity and optimality. Each bit of an autocorrelation is computed only once. The complexity is unknown, no close formula for the number of autocorrelations of size n being known.

Asymptotic bounds. Guibas and Odlyzko [3] give the following bounds for the logarithm of the number $\kappa(n)$ of autocorrelations of size n :

$$b_l = \left(\frac{1}{2 \log 2} + o(1) \right) \log^2 n \leq \log \kappa(n) \leq \left(\frac{1}{2 \log(3/2)} + o(1) \right) \log^2 n.$$

For numerical computations up to $n = 200$, Rivals and Rahmann obtain $\kappa(n) < b_l$. They conjecture that the asymptotic value of $\kappa(n)$ is b_l , the lower bound of Guibas and Odlyzko.

5. Computation of the Populations of Autocorrelations

In this section, the size n of the autocorrelations vectors is fixed.

Definition 4. The population N of an autocorrelation vector V is defined as

$$N(V) = \text{Card} \{ w \mid V \text{ is the autocorrelation vector of } w \}.$$

We define a partial order \preceq on the autocorrelation vectors by $V = v_0 v_1 \cdots v_{n-1} \preceq V' = v'_0 v'_1 \cdots v'_{n-1}$ if for all i in $[0, n-1]$, $v'_i = 1$ whenever $v_i = 1$. We also define the total order \leq by $V \leq V'$ if the word $v_0 v_1 \cdots v_{n-1}$ precedes lexicographically the word $v'_0 v'_1 \cdots v'_{n-1}$. Then $V \preceq V'$ implies $V \leq V'$. Autocorrelation vectors of size n are sorted along the total order \leq and numbered along this order from 1 to $\kappa(n)$. The notation V_k refers to the vector at rank k in this order.

Definition 5. The number ρ_k of free characters of the autocorrelation V_k is the number of characters that we can choose freely to build a word with the correlation V_k . The other characters are determined by the periods of the autocorrelation.

With an alphabet of size σ , for k from $\kappa (= \kappa(n))$ to 1, we get

$$N(V_k) = \sigma^{\rho_k} - \sum_{k < j < \kappa \text{ and } V_j \succ V_k} N(V_j).$$

The implementation is quadratic in $\kappa(n)$.

Bibliography

- [1] Burkhardt (S.), Crauser (A.), Ferragina (P.), Lenhof (H.-P.), Rivals (E.), and Vingron (M.). – q -gram based database searching using a suffix array (QUASAR). In *Third International Conference on Computational Biology*, pp. 77–83. – ACM-Press, 1999. S. Istrail, P. Pevzner and M. Waterman, editors.
- [2] Fine (N. J.) and Wilf (H. S.). – Uniqueness theorems for periodic functions. *Proceedings of the AMS*, vol. 16, 1965, pp. 109–114.
- [3] Guibas (Leo J.) and Odlyzko (Andrew M.). – Periods in strings. *Journal of Combinatorial Theory. Series A*, vol. 30, n° 1, 1981, pp. 19–42.