# Computation of the exact distributions of the words

Hayato Takahashi*
Random Data Lab. Inc. hayato.takahashi@ieee.org

Computing the distribution of the words is an important topic in DNA analysis. The distributions of the words are approximated by Poisson and CLT distributions. On the other hand there are several computing methods of the exact distributions of the words see [1–5]. The exact distributions of the overlapping words are compricated. This presentation focus on the simple case i.e., the computation of the exact distributions of non-overlapping words and demonstrate a simple formula for the distributions. We refere to the computer experiments for the human DNA size.

A word $w$ is called overlapping if there are prefixes $x, y$ of $w$ and $w = xy$. A word is called non-overlapping if it is not overlapping. For example 111 and 1010 are overlapping, and 100 and 110 are non-overlapping.

Let $X \in A^n$ with finite alphabet $A$ and $w \in A^*$. Let $|w|$ be the length of the word $w$ and $N_w$ be the number of $w$ in $X$. For example let $A = \{a, b, c, d\}$, $w = cda$, and $X = abcdacda$ then $|w| = 3, N_w = 2$.

The following is the special case of [3–5].

**Theorem 1 ( [3–5])** *Let $w$ be a non-overlapping word, $P$ be i.i.d. process on $A^n$, and $p := P(w)$. Let*

$$F_A(z) = \sum_k \binom{n - |w|k + k}{k} p^k z^k, \text{ and}$$
$$F_B(z) = \sum_k P(N_w = k) z^k.$$

*Then*

$$F_A(z) = F_B(z + 1).$$

**Corollary 1**

$$P(N_w = s) = \sum_{s \le k} \binom{n - |w|k + k}{s, k - s} (-1)^{k-s} p^k \quad (1)$$

*www.h-takahashi.sakura.ne.jp

**Theorem 2 ( [5])** *Let $w$ be a non-overlapping word.*

$$\forall t \; E(N_w^t) = \sum_{s=1}^{\min\{T,t\}} A_{t,s} \binom{n - s|w| + s}{s} P^s(w).$$

$$A_{t,s} = \sum_r \binom{s}{r} r^t (-1)^{s-r}, \; T = \max\{t \in \mathbb{N} \mid n - t|w| \ge 0\}.$$

In the above theorem, $A_{t,s}$ is the number of surjective functions from $\{1, 2, \ldots, t\} \to \{1, 2, \ldots, s\}$ for $t, s \in \mathbb{N}$ where $\mathbb{N}$ is the set of natural numbers.

**Computer experiments.** Let $A = \{a, b, c, d\}, |X| = 32 \cdot 10^8$, and the probability of each letter is $1/4$. $X$ is the human DNA size and $P(w) = 4^{-|w|}$. If $|w| \ge 14$, we identified that the exact distribution (1) is numerically almost same to Poisson distribution, $Po(N_w = k) = e^{-\lambda} \frac{\lambda^k}{k!}$ and $\lambda = E(N_w)$. In case that $|w|$ is small, the exact distributions will be well approximated with CLT. Our algorithm compute the exact distribution for human DNA size within few seconds with desktop computer. The distributions of the sparse pattern can be computed with similar manner [5].

# References

[1] L. Guibas and A. Odlyzko, "String overlaps, pattern matching, and nontransitive games," *J. Combin. Theory Ser. A*, vol. 30, pp. 183–208, 1981.

[2] M. Régnier and W. Szpankowski, "On pattern frequency occurrences in a markovian sequence," *Algorithmica*, vol. 22, no. 4, pp. 631–649, 1998.

[3] I. Goulden and D. Jackson, *Combinatorial Enumeration*. John Wiley, 1983.

[4] F. Bassino, J. Clément, and P. Micodème, "Counting occurrences for a finite set of words: combinatorial methods," *ACM Trans. Algor.*, vol. 9, no. 4, p. Article No. 31, 2010.

[5] H. Takahashi, "The distributions of sliding block patterns in finite samples and the inclusion-exclusion principles for partially ordered sets," Dec. 2018, probability Symposium, Kyoto Univ. Japan arxiv:1811.12037v2.