ICECA
International Conference
Enumerative Combinatorics and Applications
University of Haifa – Virtual – August 26-28, 2024

SUBSEQUENCE FREQUENCY IN BINARY WORDS

KRISHNA MENON

Abstract The numbers we study are of the form $B_{n,p}(k)$, which is the number of binary words of length n that contain the word p (as a subsequence) exactly k times. Our motivation comes from the analogous study of pattern containment in permutations, especially [1]. In our first set of results, we obtain explicit expressions for $B_{n,p}(k)$ for small values of k. We then focus on words p with at most 3 runs and study the maximum number of occurrences of p that a word of length n can have. We also study the internal zeros in the sequence $(B_{n,p}(k))_{k\geq 0}$ for fixed n and discuss the unimodality and log-concavity of such sequences. This is based on joint work with Anurag Singh [2].

Preliminaries. A *binary word* is a finite sequence $w = w_1w_2\cdots w_n$ where $w_i \in \{\mathtt{0},\mathtt{1}\}$ for all $i \in [n]$. Here n is called the *length* of w. An *occurrence* of a binary word $p = p_1p_2\cdots p_l$ in a binary word $w = w_1w_2\cdots w_n$ is a subsequence of w that matches p, i.e., a choice of indices $1 \leq i_1 < i_2 < \cdots < i_l \leq n$ such that $w_{i_1}w_{i_2}\cdots w_{i_l} = p$. In this context, we usually call p a *pattern*. Analogous to the notation of [1], we denote the number of occurrences of the pattern p in w by $c_p(w)$. For example, $c_{\mathtt{10}}(\mathtt{10010}) = 4$.

For any binary word p and $n, k \geq 0$, we define $B_{n,p}(k)$ to be the number of binary words of length n that have exactly k occurrences of the pattern p. That is,

$$B_{n,p}(k) = \#\{w \in \{\mathtt{0},\mathtt{1}\}^n \mid c_p(w) = k\}.$$

Just as for permutation patterns, we say that two patterns are *trivially equivalent* if one can be obtained from the other using reversal and complementation operations. For example, the words $\mathtt{100}, \mathtt{001}, \mathtt{011}$, and $\mathtt{110}$ are all trivially equivalent One can check that if p and q are trivially equivalent, then $B_{n,p}(k) = B_{n,q}(k)$ for all $n, k \geq 0$.

For use in the sequel, we also recall the following definitions. A sequence of non-negative integers $(a_k)_{k=0}^m$ is said to have an *internal zero* if there exist $0 \leq k_1 < k_2 < k_3 \leq m$ such that $a_{k_1}, a_{k_3} \neq 0$ but $a_{k_2} = 0$. The sequence is said to be *unimodal* if there exists an $i \in [0, m]$ such that $a_0 \leq a_1 \leq \cdots \leq a_i \geq a_{i+1} \geq \cdots \geq a_m$. The sequence is said to be *log-concave* if $a_i^2 \geq a_{i-1}a_{i+1}$ for all $i \in [m-1]$.

Few occurrences of a pattern. A *run* in a binary word is a maximal subsequence of consecutive terms that are equal. For instance, the word $\mathtt{11100001} = \mathtt{1}^3\mathtt{0}^4\mathtt{1}^1$ has three runs, which are of sizes 3, 4, and 1 respectively.

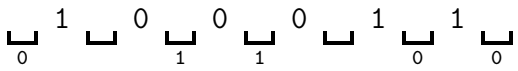
Let p be a binary word of length l that has r runs, r_i of which are of size i for each $i \geq 1$. We have the following expressions for $B_{n,p}(k)$ for small values of k.

Proposition 1. *For any $n \geq 0$, we have the following.*

- $B_{n,p}(0) = \sum_{j=0}^{l-1} \binom{n}{j}$.
- $B_{n,p}(1) = \binom{n-r+1}{l-r+1}$.
- $B_{n,p}(2) = r_1 \binom{n-r}{l-r+1}$ *if* $l \geq 2$, *and* $B_{n,p}(2) = \binom{n}{2}$ *if* $l = 1$.

We obtain the results for $k \geq 1$ by considering the spaces between the letters of the pattern p as *slots* and studying how inserting letters into these slots affects the number of occurrences of p. Similar ideas can be used to obtain expressions for $B_{n,p}(3)$ and $B_{n,p}(4)$ as well [2].

For example, if $p = \mathtt{100011}$ then all words w such that $c_p(w) = 1$ can be obtained by adding appropriate letters in the slot diagram below.



Here the letters under the slots represent what types of letters can be inserted in them. For example, inserting appropriate letters, we get the word $\mathtt{00}\mathtt{100}\mathtt{1}\mathtt{01}\mathtt{00}\mathtt{1}$ that contains exactly one occurrence of p (which has been highlighted).

We also note the following result which is easy to verify by studying how many occurrences of p there are in a word obtained by adding a letter to p.

Lemma 2. *For any $k \geq 2$, we have $B_{l+1,p}(k) = r_{k-1}$.*

Definition 3. We say that two patterns p, q are *strong Wilf-equivalent* if $B_{n,p}(k) = B_{n,q}(k)$ for all $n, k \geq 0$.

Clearly, strong Wilf-equivalent patterns must have the same length. A consequence of Lemma 2 is that two strong Wilf-equivalent patterns must have the same number of runs of each size. We have already noted that trivially equivalent patterns are strong Wilf-equivalent. Computations suggest that these are the only strong-Wilf equivalences.

Conjecture 4. The patterns p and q are strong Wilf-equivalent if and only if they are trivially equivalent.

We have verified the above conjecture for patterns of length up to 13 using Sage [3].

Maximum occurrences and internal zeroes. Given $n \geq 0$ and a pattern p, we set $M_{n,p}$ to be maximum possible number of occurrences of p in a binary word of length n. Hence,

$$M_{n,p} = \max\{c_p(w) \mid w \in \{\mathtt{0},\mathtt{1}\}^n\} = \max\{k \mid B_{n,p}(k) \neq 0\}.$$

A binary word w of length n is said to be *p-optimal* if $c_p(w) = M_{n,p}$. We have the following result for patterns that have at most 3 runs. Note that any such pattern is trivially equivalent to a pattern of the form mentioned below.

Theorem 5. *Let $p = \mathtt{1}^i\mathtt{0}^j\mathtt{1}^k$ for some $i, k \geq 0$ and $j \geq 1$.*

- *For any $n \geq 0$, there exists a p-optimal word of length n that has the same number of runs as p. Hence, we have*
$$M_{n,p} = \max\left\{ \binom{a}{i}\binom{b}{j}\binom{c}{k} \mid a+b+c = n \right\}.$$
- *The sequence $(M_{n,p})_{n\geq 0}$ is log-concave.*
- *If $\mathtt{1}^a\mathtt{0}^b\mathtt{1}^c$ is p-optimal, so is at least one word in $\{\mathtt{1}^{a+1}\mathtt{0}^b\mathtt{1}^c, \mathtt{1}^a\mathtt{0}^{b+1}\mathtt{1}^c, \mathtt{1}^a\mathtt{0}^b\mathtt{1}^{c+1}\}$.*

The key points behind the proof of the above theorem are that 0s play a special role when finding occurrences of p and that the binomial coefficients are log-concave.

Definition 6. A binary word p is said to have an *internal zero* at n if the sequence $(B_{n,p}(k))_{k\geq 0}$ has an internal zero.

For example, Lemma 2 shows that if p is of length l with maximum run size i, then p does not have an internal zero at $l+1$ if and only if p has a run of size j for all $j \in [i]$. We have the following result for patterns with at most 3 runs, where we say that a binary word is *alternating* if all its runs are of size 1.

Theorem 7. *Let p be a binary word of length l with at most 3 runs.*

- *If p is alternating, then p does not have an internal zero at any $n \geq 7$.*
- *If p is not alternating, then p has an internal zero at all $n \geq l+3$. In fact, we have $B_{n,p}(M_{n,p} - 1) = 0$.*

Since sequences that have internal zeroes cannot be unimodal, when p has at most 3 runs, the sequence $(B_{n,p}(k))_{k\geq 0}$ can be unimodal only if p is alternating. If $p = \mathtt{0}$ or $\mathtt{1}$, this sequence is just $(\binom{n}{k})_{k\geq 0}$ which is not only unimodal, but log-concave. However, if p is an alternating pattern of length 2 or 3, using Proposition 1 we have $B_{n,p}(0) > B_{n,p}(1) < B_{n,p}(2)$ for $n \geq 4$ and hence the sequence $(B_{n,p}(k))_{k\geq 0}$ is not unimodal.

References

[1] M. Bóna, B. E. Sagan, and V. R. Vatter. Pattern frequency sequences and internal zeros. *Adv. in Appl. Math.*, 28 (2002), p.395-420. Special issue in memory of Rodica Simion.

[2] K. Menon and A. Singh. Subsequence frequency in binary words. *Discrete Math.*, 347:5 (2024) Article 113928.

[3] The Sage Developers. *SageMath, the Sage Mathematics Software System (Version 9.5.0)*, 2022. `https://www.sagemath.org`

Chennai Mathematical Institute, India. Email: krishnamenon@cmi.ac.in