Superiority of Spaced Seeds for Homology Search

Louxin Zhang
Department of Mathematics
National University of Singapore
2 Science Drive 2, Singapore 117543
E-mail: matzlx@nus.edu.sg

Abstract

In homology search, good spaced seeds have higher sensitivity for the same cost (weight). However, elucidating the mechanism that confers power to spaced seeds and characterizing optimal spaced seeds still remain unsolved. This paper investigates these two important open questions by formally analyzing the average number of non-overlapping hits and the hit probability of a spaced seed in the Bernoulli sequence model. We prove that when the length of a non-uniformly spaced seed is bounded above by an exponential function of the seed weight, the seed outperforms strictly the traditional consecutive seed of the same weight in both (i) the average number of non-overlapping hits and (ii) the asymptotic hit probability. This clearly answers the first problem mentioned above in the Bernoulli sequence model. The theoretical study in this paper also gives a new solution to finding long optimal seeds.

Keywords: Homology search, pattern matching, sequence alignment, spaced seeds, renewal theory, run statistics.

1 Introduction

Homology search is one of the most common activities in bioinformatics since the creation of DNA and protein sequence databases in the early of 1980s. For instance, BLAST program in NCBI database server processes over 100,000 queries per day. To meet the huge demand for searching homology in a fast and yet sensitive manner, various heuristic programs have been developed and improved over the years as the Smith-Waterman algorithm is too slow for the purpose. The readers are referred to [4] for the current status of homology search.

Starting with BLAST [1, 2], the seed (or index-based) approach has revolutionized sequence homology search. It is based on the principle of filtration, where alignment between two sequences is found by first identifying short identical matches between the query and target sequences, called *seed hits*, and then extending them on either side for approximate matches, called *local alignments*. The resulting alignments are scored for acceptance. BLASTN first finds perfect matches of contiguous k (usually 11) nucleotide bases between query and target DNA sequences, and then builds local alignments around the hits. One significant feature of this approach is the tradeoff between the specificity (or time) and sensitivity in detecting homology. Setting k larger increases speed and specificity, but reduces sensitivity; on the other hand, setting k smaller raises sensitivity, but decreases speed and specificity.

In PatterHunter (PH) [27], Ma, Tromp and Li introduced the idea of an optimized spaced pattern (called *spaced seed*) to trigger a local alignment in order to increase both speed and sensitivity. More specifically, PH of the first version looks for runs of 18 contiguous nucleotide bases in each sequence, in which the nucleotide matches are only required at the 11 positions specified by 1s in the string $111 * 1 * *1 * 1 * *11 * 111$. Such a spaced seed leads to surprisingly higher sensitivity. In addition, it is observed that sensitivity improvement can be further achieved by using multiple spaced seeds. Other earlier programs using the seed strategy include BLAT[19], WABA[20], SPLASH[9], and a program reported in [6]. But, only PatternHunter considers seed optimization. Recently, MegaBLAST, BLASTZ[32], next version of BLAST, and other alignment programs have also adopted the PH seeding approach.

Several important research problems arise from the PH seeding approach. The number of the matching positions in a spaced seed is a key factor in improving sensitivity and is defined to be its *weight*. Consider two spaced seeds of the same weight. As these two spaced seeds contain the same number of matching positions, the expected number of hits in a homologous region is almost the same no matter which one is used. However, the optimized spaced seed of weight 11 used by PH improves sensitivity as much as over 50% when two sequences having 70% similarity are compared [27].

The fact that spaced seeds generally outperform the consecutive seeds of the same weight is informally explained in terms of relaxing the correlations existing among contiguous hits. But, the theoretic analysis of this behavior is extremely challenging.

So far, only partial progress has been made in this aspect. For a given weight, the consecutive seed has larger hit probability than so-called uniformly spaced seeds, in which any two successive matching positions are separated by the same number of don't care positions, in any homology region [10, 7]; and hence not all spaced seeds are better in homology search. Non-uniformly spaced seeds are very likely to outperform consecutive seeds, but not much insight has been developed on their relative power. To elucidate the mechanism that confers power to spaced seeds, Buhler, Keich and Sun [7] proposed an asymptotic modeling of the problem; Preparata, Zhang and Choi [29] proposed a probability leakage model.

Since not all the spaced seeds are better than consecutive seeds, another important problem is to identify the most sensitive seeds for homology search. This problem has been extensively studied in the so called Bernoulli sequence model [12, 14, 18, 23, 26, 34, 21, 35] and more general Markov and HMM models [5, 7, 23]. All sequences are assumed to be generated by some Markov model on the same alphabet, typically nucleotides. In this paper, we restrict ourselves to the Bernoulli or zero-th Markov sequence model, that is, we assume the sequence symbols are independently and identically generated (i.i.d). (Most of our analysis in the current paper can generalize to higher-order Markov models.) Let $X = x_1x_2\cdots x_n$ and $Y = y_1y_2\cdots y_n$ be two aligned homologous sequences of length n, which may differ only through substitutions. Consider a spaced seed $Q = q_0q_1\cdots q_{L_Q-1}$ (where $q_i \in \{1, *\}$) of weight w_Q and length L_Q. We say that there is a *hit* of Q between X and Y occurring at position i ($L_Q \leq i \leq n$) if the two sequences are identical at all the w_Q positions $i-(L_Q-1)+k$, where ks are indices of the match positions in Q. Using 1s and 0s to represent matches and mismatches between X and Y, such a hit can be viewed as the detection of an occurrence of the spaced seed Q in a binary sequence of length n.

A direct approach to finding the most sensitive seeds is through exhaustive search after the hit probability of each spaced seed is calculated. The hit probability of a spaced seed can be computed by dynamic programming [18] (see also [5, 7, 23, 26, 34] for various generalizations) or a recurrent relation [10]. But this approach soon becomes impractical due to (i) the number of spaced seeds of length L and weight w is exponential in $\min\{w, L-w\}$, (ii) the worst-case time complexity of the dynamic programming algorithm or recurrent relation based method for computing the hit probability is exponential in $L-w$. Actually, computing the hit probability of spaced seeds is recently proved to be NP-hard [25]. One attempt to speed up the program is to use the hit probabilities of spaced seeds on a homology region of length $2L$ as an indication of their effectiveness [12]. Another solution is to reduce the seed search space by sampling [7, 26, 34]. The third approach is to use a simple, polynomial-time computable criteria to search for the most sensitive seeds [35, 21].

This paper focuses on the above open problems. Since the overlapping hits can only be extended into one local alignment between the two sequences, the sensitivity of a spaced seed depends largely on the expected number of non-overlapping hits. We prove that, for a *non-uniformly* spaced seed Q of weight w_Q and length L_Q, if

$L_Q < w_Q + (q/p)((1/p)^{w_Q-2} - 1)$, the expected number of its non-overlapping hits, μ_Q, in a Bernoulli sequence generated with probability p is strictly larger than the consecutive seed B of the same weight. This suggests strongly that non-uniformly spaced seeds Q are more sensitive than B.

Another indicator of sensitivity is the hit probability $Q_n(p)$ of a spaced seed Q in a length-n homologous region of similarity level p. The study of the hit probability of single or multiple patterns is rooted in run statistics and the renewal theory [3, 13, 17, 30, 31, 33]. The general run statistics focus on exact and asymptotic distribution of pattern occurrences. The theory developed here is about seed comparison in terms of hit probability. Using a general theorem of Nicodéme *et al.* [28], Buhler *et al.* proved that there are two constant numbers α_Q and λ_Q determined by Q and p, such that $\lim_{n\to\infty}(1 - Q_n(p))/(\alpha_Q\lambda_Q^n) = 1$ ([7]; see also [33]). Moreover, it is conjectured that λ_Q of a non-uniformly spaced seed Q is smaller than that of the consecutive seed B with the same weight [7]. In the current paper, we present tight lower and upper bounds on λ_Q in terms of μ_Q. Then, we prove that the conjecture is true if $L_Q < \frac{q}{p}(\frac{1}{p})^{w_Q-2} + \frac{1}{p}$. This implies that $Q_n(p) > B_n(p)$ when $L_Q < \frac{q}{p}(\frac{1}{p})^{w_Q-2} + \frac{1}{p}$ and n is large.

The rest of this paper is divided into five parts. Section 2 introduces basic notations and formulas that are used in this paper. Section 3 gives two formulas for computing and estimating the average distance between non-overlapping hits of a spaced seed. Section 4 discusses the average number of non-overlapping hits in a homology region and elucidates why spaced seeds are more sensitive in homology search. Section 5 studies asymptotically the hit probability of a spaced seed in the Bernoulli sequence model. Section 6 justifies why a simple criteria used in [35, 21] is so effective for filtering out poor spaced seeds. Finally, we conclude this paper with several remarks.

2 Basic Formulas and Inequalities

In this section, we present basic definitions, known formulas and two useful inequalities that are used in the rest of paper.

Let S be an infinite Bernoulli random binary sequence in which 1 is generated with probability p in each position. We use $S[k]$ to denote the kth symbol of S and $S[0, k-1]$ the length-k prefix of S for $k = 0, 1, 2, \cdots$. Let Q be a spaced seed of length L_Q and weight w_Q given by the following relative match position set [8]:

$$\mathcal{RP}(Q) = \{i_1 = 0, i_2, \cdots, i_{w_Q} = L_Q - 1\}. \tag{1}$$

The seed Q is said to *hit* S at position k if

$$S[k - L_Q + i_1 + 1] = S[k - L_Q + i_2 + 1] = \cdots = S[k - L_Q + i_{w_Q} + 1] = 1$$

where $k - L_Q + i_{w_Q} + 1 = k$. Here, we use the *ending position* as the hit position following the renewal theory [13]. Finally, we let $Q_n := Q_n(p)$ denote the probability that Q hits $S[0, n-1]$.

2.1 Basic formulas

Computing the hit probability of a spaced seed is much more involved than the consecutive seed $B = 11\cdots1$. Let Q be the spaced seed specified by its relative match position set in Formula (1). To calculate Q_n, we define A_j to be the event that Q hits S at position j and $\bar{A}_j$ the complement of A_j for every $0 \le j \le n-1$. Then,

$$Q_n = P[\cup_{0\le j\le n-1} A_j].$$

Trivially, $A_i = \phi$ for $0 \le i \le L_Q - 2$, and hence $Q_i = 0$ if $0 \le i \le L_Q - 1$. Equivalently,

$$\bar{Q}_n := 1 - Q_n = P[\cap_{0\le j\le n-1} \bar{A}_j].$$

We call $\bar{Q}_n$ the *non-hit probability*. Moreover, the probability that Q first hits S at position $n-1$ is

$$f_n = P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-2}A_{n-1}].$$

We call f_n the *first-hit probability*. Obviously, $f_{L_Q} = p^w$ and $f_j = 0,\ \ 1 \le j < L_Q - 1$. It is easy to see that

$$Q_n = Q_{n-1} + f_n = \sum_{1\le i\le n} f_i, \tag{2}$$

or equivalently

$$\bar{Q}_n = \bar{Q}_{n-1} - f_n = \sum_{n<i<\infty} f_i. \tag{3}$$

Let $m = 2^{L_Q - w_Q}$. We define $\mathcal{W}_Q = \{W_1, W_2, \cdots, W_m\}$ to be the set of all m distinct strings obtained from the seed Q by filling 0 or 1 in the 'don't' positions. For example, if $Q = 1*11*1$, we have $\mathcal{W}_Q = \{101101, 101111, 111101, 111111\}$. The seed Q hits at position n if and only if there is some $W_j \in \mathcal{W}_Q$ that occurs at the position. For each j, we use $A_n^{(j)}$ to denote the event that the pattern W_j occurs at the position n. It is easy to see that $A_n = \cup_{1\le j\le m} A_n^{(j)}$, and $A_n^{(j)}$'s are disjoint for fixed n. Setting

$$f_n^{(j)} = P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-2}A_{n-1}^{(j)}],\ \ 1 \le j \le m,$$

we have $f_n = \sum_{1\le j\le m} f_n^{(j)}$ and that Formula (3) now becomes

$$\bar{Q}_n = \bar{Q}_{n-1} - f_n^{(1)} - f_n^{(2)} - \cdots - f_n^{(m)}. \tag{4}$$

Given a string s on alphabet $\{0, 1\}$, we use $|s|$ to denote its length, $|s|_0$ the number of 0's in s, and $|s|_1$ the number of 1's in s. For any $0 \le a < b \le s-1$, $s[a, b]$ is defined

to be the substring of s from position a to position b inclusively. Finally, we use $P[s]$ to denote the probability that s occurs at a position ($> |s|$) in the random sequence S. Obviously, $P[s] = p^{|s|_1}(1-p)^{|s|_0}$.

For any i, j and k such that $1 \le i, j \le m$, $1 \le k \le L_Q$, define

$$p_k^{(ij)} = \begin{cases} P[W_j[k, L_Q - 1]] & \text{if } k \le L_Q - 1 \ \& \ W_i[L_Q - k, L_Q - 1] = W_j[0, k-1] \\ 1 & k = L_Q \ \& \ i = j \\ 0 & \text{otherwise.} \end{cases}$$

Using these notations, we have the following recurrence relations [10] (see also (3.2) in [15]).

Theorem 2.1 *Let p_j be the probability that the pattern $W_j \in \mathcal{W}_Q$ occurs at position, $1 \le j \le m$. Then,*

$$\bar{Q}_n p_j = \sum_{i=1}^{m} \sum_{k=1}^{L_Q - 1} f_{n+k}^{(i)} p_k^{(ij)} + f_{n+L_Q}^{(j)} = \sum_{i=1}^{m} \sum_{k=1}^{L_Q} f_{n+k}^{(i)} p_k^{(ij)}, \ \ 1 \le j \le m. \tag{5}$$

2.2 Two inequalities

Recall that $\bar{A}_i$ denotes the event that the spaced seed Q does not hit the Bernoulli random sequence S at position i. For any two different $i < j$, if $j - i > L_Q$, then $\bar{A}_i$ and $\bar{A}_j$ are independent events; if $j - i \le L_Q$, they are positively correlated to each other and hence $P[\bar{A}_i \bar{A}_j] \ge P[\bar{A}_i]P[\bar{A}_j] = (1 - p^{w_Q})^2$. In general, we have the following inequalities on sensitivity.

Theorem 2.2 *Let Q be a spaced seed. Then,*
(i) For any $2L_Q - 1 \le k \le n$, $f_k \bar{Q}_{n-k+L_Q-1} \le f_n \le f_k \bar{Q}_{n-k}$.
(ii) For any $1 \le k \le n$, $\bar{Q}_k \bar{Q}_{n-k+L_Q-1} \le \bar{Q}_n \le \bar{Q}_k \bar{Q}_{n-k}$.

Theorem 2.2 (i) was proved by Choi and Zhang in [10]. The second inequality in both (i) and (ii) can be derived from dependence just mentioned. However, the proof of the first inequality in each case is tricky. It was obtained by using Chebyshev Inequality on two similarly ordered real number sequences (page 43, [16]). For $2L_Q - 1 \le k \le n$, the inequality in Theorem 2.2 (ii) can also be proved from Theorem 2.2 (i) using Formula (3).

3 Distance Between Non-overlapping Hits

Renewal theory studies certain recurrent events connected with repeated trials. Roughly speaking, an event $\mathcal{E}$ qualifies for the theory if after each occurrence of $\mathcal{E}$, the trials start from scratch [13]. Therefore, the numbers of trials between successive occurrences of $\mathcal{E}$ are jointly independent random variables with the identical distribution.

It is easy to see that a non-overlapping hit of a spaced seed Q is a recurrent event with the following convention: If a hit at position i is selected as a *non-overlapping hit*, then the next non-overlapping hit is the first hit at or after position $i + L_Q$.

In this section, we shall focus on the average distance, μ_Q, between two successive non-overlapping hits of a spaced seed. We first give a formula for computing μ_Q using the generating function approach pioneered in [15]; then, we give a simple upperbound on μ_Q.

3.1 A formula for computing μ_Q

To find the average distance μ_Q between non-overlapping hits of Q, we define the generating functions

$$U(x) = \sum_{n=0}^{\infty} \bar{Q}_n x^n,$$

$$F_i(x) = \sum_{n=0}^{\infty} f_n^{(i)} x^n, \;\; 1 \leq i \leq m = 2^{L_Q - w_Q},$$

where recall that $f_n^{(i)}$ is the probability that the seed does not hit the random sequence S before $n-1$ but $W_i \in \mathcal{W}_Q$ hits S at position $n-1$ (see Section 2.1).

By definition, $\mu_Q = \sum_{j \geq L_Q} j f_j$. Applying Formula (3), we obtain

$$\mu_Q = L_Q + \sum_{j \geq L_Q} \bar{Q}_j = U(1)$$

and both $U(x)$ and $F_i(x)$'s converge when $x \in [0, 1]$. Multiplying (4) by x^{n-1} and summing on n, we obtain

$$(1-x)U(x) + F_1(x) + F_2(x) + \cdots + F_m(x) = 1. \tag{6}$$

Similarly, by (5), we obtain

$$-x^{L_Q} p_j U(x) + C_{1j}(x)F_1(x) + C_{2j}(x)F_2(x) + \cdots + C_{mj}(x)F_m(x) = 0, \; 1 \leq j \leq m \tag{7}$$

where $C_{ij}(x) = \sum_{k=1}^{L_Q} p_k^{(ij)} x^{L_Q - k}$.

Theorem 3.1 *Let*

$$A_Q = \begin{bmatrix} C_{11}(1) & C_{21}(1) & \cdots & C_{m1}(1) \\ C_{12}(1) & C_{22}(1) & \cdots & C_{m2}(1) \\ \vdots & \vdots & \cdots & \vdots \\ C_{1m}(1) & C_{2m}(1) & \cdots & C_{mm}(1) \end{bmatrix}$$

and

$$M_Q = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ -p_1 & C_{11}(1) & C_{21}(1) & \cdots & C_{m1}(1) \\ -p_2 & C_{12}(1) & C_{22}(1) & \cdots & C_{m2}(1) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ -p_m & C_{1m}(1) & C_{2m}(1) & \cdots & C_{mm}(1) \end{bmatrix}.$$

Then, $\mu_Q = \det(A_Q)/\det(M_Q)$.

Proof. Setting $x = 1$ in (6) and (7), we obtain

$$\begin{cases} & F_1(1) & + & F_2(1) & + & \cdots & + & F_m(1) & = & 1 \\ -p_1U(1) & + & C_{11}(1)F_1(1) & + & C_{21}(1)F_2(1) & + & \cdots & + & C_{m1}(1)F_m(1) & = & 0 \\ & & & & & & \cdots & & & & \\ -p_mU(1) & + & C_{1m}(1)F_1(1) & + & C_{2m}(1)F_2(1) & + & \cdots & + & C_{mm}(1)F_m(1) & = & 0 \end{cases}$$

Solving the above linear equation system, we obtain $\mu_Q = U(1) = \det(A_Q)/\det(M_Q)$.
Example (a) Using the above theorem, one can easily show that, for the consecutive seed B of weight w,

$$A_B = [\sum_{i=0}^{w-1} p^i], \quad M_B = \begin{bmatrix} 0 & 1 \\ -p^w & \sum_{i=0}^{w-1} p^i \end{bmatrix}.$$

and hence $\mu_B = \sum_{i=1}^{w} (\frac{1}{p})^i$.

(b) For spaced seed $Q = 1^a * 1^b$, $a \geq b \geq 1$, $\mathcal{W}_Q = \{W_1, W_2\} = \{1^a01^b, 1^{a+b+1}\}$ and

$$A_Q = \begin{bmatrix} \sum_{i=0}^{b-1} p^{a+i}q + 1 & \sum_{i=0}^{a-1} p^{b+i}q \\ \\ \sum_{i=0}^{b-1} p^{a+1+i} & \sum_{i=0}^{a+b} p^i \end{bmatrix}.$$

Therefore, $\mu_Q = \frac{\sum_{i=0}^{a+b} p^i + \sum_{i=0}^{b} \sum_{j=0}^{b-1} p^{a+i+j}q}{p^{a+b}(1+\sum_{i=1}^{b} p^i q)}$.

3.2 A tight upper bound for μ_Q

For any $0 \leq j \leq L_Q - 1$, we define

$$\mathcal{RP}(Q) + j = \{i_1, i_2, \cdots, i_{w_Q}\} + j = \{i_1 + j, i_2 + j, \cdots, i_{w_Q} + j\}$$

and let $o_Q(j) = |\mathcal{RP}(Q) \cap (\mathcal{RP}(Q) + j)|$. In other words, $o_Q(j)$ is the number of 1's that coincide between the seed and the j-th shifted version of it. Trivially, $o_Q(0) = w_Q$ and $o_Q(L_Q - 1) = 1$ for any seed Q. The following lemma was proved by Keith *et al.* using a martingale approach. Here, we give another short elementary proof.

Lemma 3.1 *([18]) With notations as above,* $\mu_Q \leq \sum_{i=0}^{L_Q-1} (\frac{1}{p})^{o_Q(i)}$.

Proof. Notice that since we use the ending position of a hit as the hit position, our μ_Q is L_Q more than the average first position the seed hits that was studied in [18]. Let $m_Q(j) = w_Q - o_Q(j)$. Recall that A_j denotes the event that the seed Q hits the random sequence at position j and $\bar{A}_j$ the complement of A_j. For any $1 \le j \le L_Q - 1$ and $n \ge L_Q$, by conditioning, we have

$$\begin{aligned}
& P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-j-2}A_{n-j-1}A_{n-1}] \\
= \; & P[A_{n-j-1}A_{n-1}]P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-j-2}|A_{n-j-1}A_{n-1}] \\
= \; & P[A_{n-1}|A_{n-j-1}]P[A_{n-j-1}]P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-j-2}|A_{n-j-1}A_{n-1}] \\
\le \; & P[A_{n-1}|A_{n-j-1}]P[A_{n-j-1}]P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-j-2}|A_{n-j-1}] \\
= \; & P[A_{n-1}|A_{n-j-1}]P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-j-2}A_{n-j-1}] \\
= \; & p^{m_Q(L_Q-j)}f_{n-j}
\end{aligned}$$

where the inequality is due to that A_{n-1} is negatively correlated with the joint event $\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-j-2}$. Applying

$$P[\bar{A}_0\bar{A}_1\cdots\bar{A}_iA_{n-1}] = P[\bar{A}_0\bar{A}_1\cdots\bar{A}_i\bar{A}_{i+1}A_{n-1}] + P[\bar{A}_0\bar{A}_1\cdots\bar{A}_iA_{i+1}A_{n-1}]$$

for $i = n - L_Q - 1, n - L_Q, \cdots, n - 3$ iteratively, we have,

$$\begin{aligned}
& \bar{Q}_{n-L_Q}p^{w_Q} \\
= \; & P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-L_Q-1}A_{n-1}] \\
= \; & P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-2}A_{n-1}] + \textstyle\sum_{i=1}^{L_Q-1} P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-L_Q+i-2}A_{n-L_Q+i-1}A_{n-1}] \\
\le \; & f_n + \textstyle\sum_{i=1}^{L_Q-1} f_{n-L_Q+i}p^{m_Q(i)}
\end{aligned}$$

where we assume that $n \ge L_Q$ and $\bar{Q}_j = 1$ for $j \le L_Q - 1$, and hence,

$$\mu_Q p^{w_Q} = \sum_{n=0}^{\infty} \bar{Q}_n p^{w_Q} \le 1 + \sum_{i=1}^{L_Q-1} p^{m_Q(i)} = \sum_{i=0}^{L_Q-1} p^{m_Q(i)}$$

or

$$\mu_Q \le \sum_{i=0}^{L_Q-1} p^{m_Q(i)-w_Q} = \sum_{i=0}^{L_Q-1} (\frac{1}{p})^{o_Q(i)},$$

since $\sum_{i=L_Q}^{\infty} f_i = 1$. □

Remark. (i) In Lemma 3.1, the equality holds for the consecutive seed B as shown in the last example. For a general spaced seed Q, $p^{w_Q} = f_{L_Q}$ and $p^{w_Q} = f_{L_Q+1} + p^{m_Q(1)}f_{L_Q}$. However, the strict inequality $p^{w_Q} < f_{L_Q+2} + p^{m_Q(1)}f_{L_Q+1} + p^{m_Q(2)}f_{L_Q}$ usually holds. Thus, the above proof implies a slightly better upper bound on μ_Q for non-uniformly spaced seeds.

(ii) In [10], Choi and Zhang showed the following lower and upper bounds on μ_Q:

$$L_Q + \frac{\bar{Q}_{3L_Q-2}}{f_{2L_Q-1}} \le \mu_Q \le L_Q + \frac{\bar{Q}_{2L_Q-1}}{f_{2L_Q-1}},$$

in which the difference between the upper and lower bounds is at most $L_Q - 1$. Our numerical computation indicates that $L_Q + \frac{\bar{Q}_{2L_Q-1}}{f_{2L_Q-1}}$ is smaller than the upper bound given in Lemma 3.1. It is interesting to know whether a simple, better upper bound can be derived from $L_Q + \frac{\bar{Q}_{2L_Q-1}}{f_{2L_Q-1}}$ or not.

A spaced seed Q is said to be *uniform* if its matching positions form an arithmetic sequence. For example, $1 * *1 * *1$ is uniform with relative matching position set $\{0, 3, 6\}$ in which the difference between two successive positions is 3. Obviously, any spaced seed of weight 2 is uniform. Recall that B denotes the consecutive seed of the same weight as Q. If Q is uniform, we have know that $Q_n \leq B_n$ for any n [7, 10]. Therefore, we are not interested in uniformly spaced seeds for homology search purpose. Assume Q is non-uniform. By definition, we have that $\mu_Q \geq L_Q$. Thus, for any fixed p, $\mu_Q > \sum_{i=1}^{w_Q}(1/p)^i = \mu_B$ when $L_Q > \sum_{i=1}^{w_Q}(1/p)^i$. Now, using the above lemma, we show that when L_Q is small, μ_Q is smaller than μ_B indeed.

Theorem 3.2 *For any non-uniformly spaced seed Q,*

$$\mu_Q \leq \sum_{i=1}^{w_Q}(\frac{1}{p})^i + (L_Q - w_Q) - \frac{q}{p}[(\frac{1}{p})^{w_Q-2} - 1]$$

Proof. By Lemma 3.1, we only need to prove that

$$\sum_{i=0}^{L_Q-1}(\frac{1}{p})^{o_Q(i)} \leq \sum_{i=1}^{w_Q}(\frac{1}{p})^i + (L_Q - w_Q) - \frac{q}{p}[(\frac{1}{p})^{w_Q-2} - 1] \tag{8}$$

for any non-uniformly spaced seed Q by induction on the weight w_Q. Note that $w_Q \geq 3$ for a non-uniformed spaced seed Q.

When $w_Q = 3$, we may assume $Q = 1 *^r 1 *^s 1$, $r > s$. Then, we have

$$o_Q(i) = \begin{cases} 3 & \text{if } i = 0, \\ 1 & \text{if } i = s+1, r+1, L_Q - 1, \\ 0 & \text{otherwise.} \end{cases}$$

Hence,

$$\begin{array}{rl} & \sum_{i=1}^{w_Q}(\frac{1}{p})^i + (L_Q - w_Q) - \frac{q}{p}[(\frac{1}{p})^{w_Q-2} - 1] \\ = & \sum_{1}^{3}(\frac{1}{p})^i + (L_Q - 3) - \frac{q}{p}(\frac{1}{p} - 1) \\ = & \sum_{1}^{3}(\frac{1}{p})^i + (L_Q - 3) - (\frac{q}{p})^2 \\ = & (\frac{1}{p})^3 + (\frac{1}{p}) + (L_Q - 4) + (\frac{1}{p})^2 + 1 - (\frac{q}{p})^2 \\ = & (\frac{1}{p})^3 + (\frac{1}{p}) + (L_Q - 4) + \frac{1+p^2-q^2}{p^2} \\ = & (\frac{1}{p})^3 + (\frac{1}{p}) + (L_Q - 4) + 2\frac{1}{p} \\ = & \sum_{i=0}^{L_Q-1}(\frac{1}{p})^{o_Q(i)} \end{array}$$

and the inequality holds.

In general, if $w_Q > 3$, we assume $\mathcal{RS}(Q) = \{l_1 = 0, l_2, \cdots, l_{w_Q} = L_Q - 1\}$ and $Q = 1 *^r Q'$, where $r = l_2 - 1 \geq 0$ and Q' is the length-$(L_Q - r - 1)$ suffix of Q. By assumption, Q' starts and ends with 1. Now, we consider Q' as a spaced seed. Obviously, $L_{Q'} = L_Q - r - 1$ and $w_{Q'} = w_Q - 1$. Assume there are k 1s in the length-$(L_Q - r - 1)$ prefix of Q. Then, there are $w_Q - k$ 1s (match positions) and $r + 1 + k - w_Q$ $*$s (don't care positions) in $Q[L_Q - r - 1, L_Q - 1]$, the length-$(r+1)$ suffix of Q. Hence, as shown in Table 1, we have, for $0 \leq i \leq L_Q - r - 2$,

$$o_Q(i) = \begin{cases} o_{Q'}(i) + 1 & i = l_1, l_2, \cdots, l_k \\ o_{Q'}(i) & \text{otherwise} \end{cases} \tag{9}$$

and for $L_Q - r - 1 \leq i \leq L_Q - 1$,

$$o_Q(i) = \begin{cases} 1 & i = l_{k+1}, l_{k+2}, \cdots, l_w \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Since $l_1 = 0$, $o_Q(l_1) = w_Q = 1 + w_{Q'} = 1 + o_{Q'}(l_1)$.

Now, we are ready to estimate $\sum_{i=0}^{L_Q-1}(\frac{1}{p})^{o_Q(i)}$ as follows. Using (9), (10) and $\frac{1}{p} = 1 + \frac{q}{p}$, we first have

$$\begin{aligned}
& \textstyle\sum_{i=0}^{L_Q-1}(\frac{1}{p})^{o_Q(i)} \\
= \; & \textstyle\sum_{i=0}^{L_Q-r-2}(\frac{1}{p})^{o_Q(i)} + \sum_{i=L_Q-r-1}^{L_Q-1}(\frac{1}{p})^{o_Q(i)} \\
= \; & [\textstyle\sum_{j=1}^{k}(\frac{1}{p})^{o_{Q'}(l_j)+1} + \sum_{i=0,i\notin\mathcal{RS}(Q)}^{L_Q-r-2}(\frac{1}{p})^{o_{Q'}(i)}] + [\textstyle\sum_{j=k+1}^{w_Q}\frac{1}{p} + \sum_{i=L_Q-r-1,i\notin\mathcal{RS}(Q)}^{L_Q-1}(\frac{1}{p})^0] \\
= \; & [\textstyle\sum_{j=1}^{k}(\frac{1}{p})^{o_{Q'}(l_j)+1} + \sum_{i=0,i\notin\mathcal{RS}(Q)}^{L_Q-r-2}(\frac{1}{p})^{o_{Q'}(i)}] + [(w_Q - k)\frac{1}{p} + r + 1 + (k - w_Q)] \\
= \; & [(1 + \tfrac{q}{p})\textstyle\sum_{j=1}^{k}(\frac{1}{p})^{o_{Q'}(l_j)} + \sum_{i=0,i\notin\mathcal{RS}(Q)}^{L_Q-r-2}(\frac{1}{p})^{o_{Q'}(i)}] + (w_Q - k)\frac{q}{p} + (r+1) \\
= \; & [\textstyle\sum_{j=1}^{k}(\frac{1}{p})^{o_{Q'}(l_j)} + \sum_{i=0,i\notin\mathcal{RS}(Q)}^{L_Q-r-2}(\frac{1}{p})^{o_{Q'}(i)}] + \frac{q}{p}\textstyle\sum_{j=1}^{k}(\frac{1}{p})^{o_{Q'}(l_j)} + (w_Q - k)\frac{q}{p} + (r+1) \\
= \; & \textstyle\sum_{i=0}^{L_Q-r-2}(\frac{1}{p})^{o_{Q'}(i)} + \frac{q}{p}\sum_{j=1}^{k}(\frac{1}{p})^{o_{Q'}(l_j)} + (w_Q - k)\frac{q}{p} + r + 1
\end{aligned}$$

Replacing x by $\frac{1}{p}$ in the formula $x^{w_Q} - 1 = (x-1)(1 + x + x^2 + \cdots + x^{w_Q - 1})$, we obtain $(\frac{1}{p})^{w_Q} = 1 + \frac{q}{p}\sum_{i=0}^{w_Q-1}(\frac{1}{p})^i$. Replacing 1 in the last term by $(\frac{1}{p})^{w_Q} - \frac{q}{p}\sum_{i=0}^{w_Q-1}(\frac{1}{p})^i$ and grouping the terms having $\frac{q}{p}$ together, we further have

$$\begin{aligned}
& \textstyle\sum_{i=0}^{L_Q-1}(\frac{1}{p})^{o_Q(i)} \\
= \; & \textstyle\sum_{i=0}^{L_Q-r-2}(\frac{1}{p})^{o_{Q'}(i)} + \frac{q}{p}\sum_{j=1}^{k}(\frac{1}{p})^{o_{Q'}(l_j)} + (w_Q - k)\frac{q}{p} + r + [(\frac{1}{p})^{w_Q} - \frac{q}{p}\textstyle\sum_{i=0}^{w_Q-1}(\frac{1}{p})^i] \\
= \; & (\frac{1}{p})^{w_Q} + \textstyle\sum_{i=0}^{L_Q-r-2}(\frac{1}{p})^{o_{Q'}(i)} + r + \frac{q}{p}[\sum_{j=1}^{k}(\frac{1}{p})^{o_{Q'}(l_j)} + (w_Q - k) - \sum_{i=0}^{w_Q-1}(\frac{1}{p})^i] \\
= \; & (\frac{1}{p})^{w_Q} + \textstyle\sum_{i=0}^{L_{Q'}-1}(\frac{1}{p})^{o_{Q'}(i)} + r + \frac{q}{p}[\sum_{j=2}^{k}(\frac{1}{p})^{o_{Q'}(l_j)} + (w_Q - k) - \sum_{i=0}^{w_Q-2}(\frac{1}{p})^i].
\end{aligned}$$

Now, we consider the following two cases.

Case 1. The seed Q' is uniform. Assume the matching positions of Q' form an arithmetic sequence with difference s. Since Q is non-uniform, $r \neq s$, and hence,

$o_{Q'}(l_j) \leq w_Q - j - 1$ for $j = 2, 3, \cdots, k$. Therefore,

$$\sum_{j=2}^{k} (\frac{1}{p})^{o_{Q'}(l_j)} + (w_Q - k) - \sum_{i=0}^{w_Q-2} (\frac{1}{p})^i \leq -(\frac{1}{p})^{w_Q-2} + 1.$$

Again, since Q' is uniform,

$$\sum_{i=0}^{L_{Q'}-1} (\frac{1}{p})^{o_{Q'}(i)} = \sum_{i=1}^{w_{Q'}} (\frac{1}{p})^i + L_{Q'} - w_{Q'} = \sum_{i=1}^{w_Q-1} (\frac{1}{p})^i + L_Q - w_Q - r.$$

Hence,

$$\sum_{i=0}^{L_Q-1} (\frac{1}{p})^{o_Q(i)} \leq \sum_{i=1}^{w_Q} (\frac{1}{p})^i + (L_Q - w_Q) - \frac{q}{p}[(\frac{1}{p})^{w_Q-2} - 1].$$

Case 2. The seed Q' is non-uniform. By induction,

$$\begin{array}{rl} & \sum_{i=0}^{L_{Q'}-1} (\frac{1}{p})^{o_{Q'}(i)} \\ \leq & \sum_{i=1}^{w_{Q'}} (\frac{1}{p})^i + L_{Q'} - w_{Q'} - \frac{q}{p}[(\frac{1}{p})^{w_{Q'}-2} - 1] \\ = & \sum_{i=1}^{w_Q-1} (\frac{1}{p})^i + L_Q - w_Q - r - \frac{q}{p}[(\frac{1}{p})^{w_Q-3} - 1]. \end{array}$$

Since Q' is non-uniform, $o_{Q'}(l_2) = o_Q(l_2) - 1 \leq w_Q - 2 - 1 = w_Q - 3$ and $o_{Q'}(l_j) \leq w_Q - j$, $j \geq 3$. Hence,

$$\begin{array}{rl} & \sum_{i=0}^{L_Q-1} (\frac{1}{p})^{o_Q(i)} \\ \leq & \sum_{i=1}^{w_Q} (\frac{1}{p})^i + (L_Q - w_Q) + \frac{q}{p}[1 - (\frac{1}{p})^{w_Q-3} + \sum_{j=2}^{k} (\frac{1}{p})^{o_{Q'}(l_j)} + (w_Q - k) - \sum_{i=0}^{w_Q-2} (\frac{1}{p})^i] \\ \leq & \sum_{i=1}^{w_Q} (\frac{1}{p})^i + (L_Q - w_Q) - \frac{q}{p}[(\frac{1}{p})^{w_Q-2} - 1]. \end{array}$$

This finishes the proof. □

4 Count Non-overlapping Hits

Recall, for the consecutive seed B of weight w, $\mu_B = \sum_{i=1}^{w} (\frac{1}{p})^i$, by Theorem 3.2, we have

Theorem 4.1 *Let Q be a non-uniformly spaced seed and B the consecutive seed of the same weight. If $L_Q < w_Q + \frac{q}{p}[(\frac{1}{p})^{w_Q-2} - 1]$, then, $\mu_Q < \mu_B$.*

The values of the upper bound in Theorem 4.1 are calculated and recorded in Table 2 for $w = 10, 11, 12, 13, 14$ and $p = 0.6, 0.7, 0.8, 0.9$.

Renewal theory shows that the number of the non-overlapping hits of a spaced seed Q in a long Bernoulli random sequence of length N has, approximately, a normal distribution with mean $\frac{N}{\mu_Q}$ (see for example [30]). This theoretic result can also be

validated by simulation. For each $p = 0.6, 0.7, 0.8$, we generated 800 random binary sequences of length 1000. As shown in Table 3, the observed average number of non-overlapping hits on the generated sequences is quite close to $\frac{N}{\mu_Q}$ for the three spaced seeds reported in [12] and the consecutive seeds of the same weight.

If $L_Q < w_Q + \frac{q}{p}[(\frac{1}{p})^{w_Q-2} - 1]$, by Theorem 4.1 and the above known fact, Q has on average more non-overlapping hits than B in a long homologous region with sequence similarity p in the Bernoulli model. Since overlapping hits can only be extended into one local alignment, the above fact suggests why a homology search program with a good spaced seed is much more sensitive than with the consecutive seed (of the same weight) especially for genome-genome comparison.

5 Hit Probability Analysis

5.1 Approximating the hit probability

Because of its larger span, in terms of hit probability, a spaced seed Q usually lag behind the consecutive seed B of the same weight for small n and then surpass B when n is large enough. To compare spaced seed efficiently, Buhler *et al.* proposed the asymptotic analysis of spaced seeds. In [7], they proved that for any spaced seed Q, there are two constants α_Q and λ_Q that do not depend on n such that $\lim_{n\to\infty} \bar{Q}_n/(\alpha\lambda_Q^n) = 1$ (see also [28]), where λ_Q is the largest eigenvalue of some transition matrix of a Markov Chain model constructed for computing the sensitivity of Q. Independently, Solov'ev proved a similar result in the more general setting four decades ago [33].

Solov'ev analyzed asymptotically the first occurrence of an event C that may or not occur in a position in a Bernoulli random sequence with the following assumption:

(i) Let C_n denote the occurrence of C at the n-th position and $\bar{C}_n$ the complement event. For any $i_1, i_2, \cdots, i_n$, the probability $P[C_{i_1+j}C_{i_2+j}\cdots C_{i_n+j}]$ does not depend on $j \geq 0$.

(ii) There exists a constant L such that, for any $i_1, i_2, \cdots i_n$, and $j_i, j_2, \cdots, j_m$, the joint events $C_{i_1}C_{i_2}\cdots C_{i_n}$ and $C_{j_1}C_{j_2}\cdots C_{j_n}$ are independent if $i_k - j_l > L$ for all $k \leq n$ and $l \leq m$.

Applying his main result to spaced seeds directly, we obtain the following theorem. Recall that f_n denotes the probability that the spaced seed Q first hits the random sequence at the n-th position.

Theorem 5.1 *There exists constants α_Q and λ_Q that do not depend on n such that*
(1) $f_n = \alpha_Q(1-\lambda_Q)\lambda_Q^n + r_n$, where $|r_n| = O((p^{2w_Q/(5L_Q-1)})^n)$.
(2) $\bar{Q}_n = \alpha_Q\lambda_Q^{n+1} + r'_n$, where $|r'_n| \leq O(p^{2w_Q/(5L_Q-1)^{n+1}}/(1-p^{2w_Q/(5L_Q-1)})$.

As one can see from the estimated bounds for r_n and r'_n, the remainder terms are very small if p is not close to 1, since the probability p is taken to a large power. Thus,

the formulas for Q_n and f_n in the theorem are very exact and at the same time are convenient for computation. Moreover, it is easy to see that the λ_Q in Theorem 5.1 is identical to that in the asymptotic result of Buhler *et al.*. In addition, Ma and Li estimated the error terms using the difference between the first and second largest eigenvalues of a positive matrix constructed from Q [25].

5.2 Asymptotic analysis

Interestingly, λ_Q is closely related to μ_Q studied in Section 3 as proved below.

Theorem 5.2 *(1) For the consecutive seed B of weight w, λ_B in Theorem 5.1 satisfies the following upper and lower bounds:*

$$1 - \frac{1}{\sum_{i=1}^{w}(1/p)^i - w + 1} \leq \lambda_B \leq 1 - \frac{1}{\sum_{i=1}^{w}(1/p)^i - w + \sum_{i=0}^{w-1} p^i}.$$

(2) In general, for any (uniformly and non-uniformly) spaced seed Q, λ_Q in Theorem 5.1 satisfies the following upper and lower bounds:

$$1 - \frac{1}{\mu_Q - L_Q + 1} \leq \lambda_Q \leq 1 - \frac{1}{\mu_Q}.$$

Proof. (1). Noticing that $\mu_B = \sum_{i=1}^{w}(1/p)^i$, we can derive the first inequality from the corresponding one in (2). Now, we prove the second one.

Recall that A_i denotes the event that the seed B hits at position i and $\bar{A}_i$ the complement of A_i. Since the length of B is w,

$$\begin{array}{rl} & P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-w-1}\bar{A}_{n-1}] \\ = & P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-w-1}\bar{A}_{n-w}\bar{A}_{n-1}] + P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-w-1}A_{n-w}\bar{A}_{n-1}] \\ = & P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-2}\bar{A}_{n-1}] + \sum_{i=1}^{w-1} P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-w+i-2}A_{n-w+i-1}\bar{A}_{n-1}] \end{array}$$

and so

$$\begin{array}{rl} & \bar{B}_n \\ = & P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-2}\bar{A}_{n-1}] \\ = & P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-w-1}\bar{A}_{n-1}] - \sum_{j=1}^{w-1} P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-w+j-2}A_{n-w+j-1}\bar{A}_{n-1}] \\ = & \bar{B}_{n-w}\bar{B}_w - \sum_{j=1}^{w-1} P[A_{n-w+j-1}]P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-w+j-2}|A_{n-w+j-1}]P[\bar{A}_{n-1}|A_{n-w+j-1}] \end{array}$$

Since B is consecutive and of weight w, $P[A_{n-w+j-1}] = p^w$, $P[\bar{A}_{n-1}|A_{n-w+j-1}] = 1 - p^{w-j}$, and $P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-w+j-2}|A_{n-w+j-1}] = q\bar{B}_{n-2w+j-1}$ for any $1 \leq j \leq w-1$. Let $b_{n+1} = P[\bar{A}_0\bar{A}_1\cdots\bar{A}_{n-1}A_n]$. Then, $b_{n+1} = p^w q\bar{B}_{n-w}$ since that A_n and $\bar{A}_n$ occur implies the last w positions have 1 and the position $w+1$ from the end has 0.

Therefore,

$$\begin{aligned}
& \bar{B}_n \\
= \;& \bar{B}_{n-w}\bar{B}_w - p^w q \textstyle\sum_{j=1}^{w-1} \bar{B}_{n-2w+j-1}(1-p^{w-j}) \\
\leq \;& \bar{B}_{n-w}\bar{B}_w - p^w q \textstyle\sum_{j=1}^{w-1} \bar{B}_{n-w}(1-p^{w-j}) \\
= \;& \bar{B}_{n-w}[\bar{B}_w - p^w q \textstyle\sum_{j=1}^{w-1}(1-p^{w-j})] \\
= \;& \bar{B}_{n-w} p^w q[\bar{B}_w/(p^w q) + 1 - w + \textstyle\sum_{i=1}^{w-1} p^i] \\
= \;& b_{n+1}[(1-p^w)/(p^w q) + 1 - w + \textstyle\sum_{i=1}^{w-1} p^i] \\
= \;& b_{n+1}[\textstyle\sum_{i=1}^{w}(1/p)^i - w + \sum_{i=0}^{w-1} p^i].
\end{aligned}$$

Taking limit and using Formula (3), we obtain

$$\lambda_B = \lim_{n\to\infty} \frac{\bar{B}_{n+1}}{\bar{B}_n} = 1 - \lim_{n\to\infty} \frac{b_{n+1}}{\bar{B}_n} \leq 1 - \frac{1}{\sum_{i=1}^{w}(1/p)^i - w + \sum_{i=0}^{w-1} p^i}.$$

(2) For any $n \geq 2L_Q$ and $k \geq 2$, by the first inequality in Theorem 2.2 (i), $f_{n+k} \geq f_{n+1}\bar{Q}_{L_Q+k-2}$. Therefore,

$$\frac{f_{n+1}}{\bar{Q}_n} = \frac{f_{n+1}}{\sum_{i=1}^{\infty} f_{n+i}} \leq \frac{f_{n+1}}{f_{n+1} + f_{n+1}\sum_{i=0}^{\infty} \bar{Q}_{L+i}} = \frac{1}{1+\mu_Q - L_Q},$$

and so

$$\lambda_Q = \lim_{n\to\infty} \frac{\bar{Q}_{n+1}}{\bar{Q}_n} = 1 - \lim_{n\to\infty} \frac{f_{n+1}}{\bar{Q}_n} \geq 1 - \frac{1}{\mu_Q - L_Q + 1}.$$

Similarly, by the second inequality in Theorem 2.2 (i), $f_{n+1+j} \leq f_{n+1}\bar{Q}_j$ for any $j \geq L_Q$. Therefore,

$$\frac{f_{n+1}}{\bar{Q}_n} = \frac{f_{n+1}}{\sum_{i=1}^{\infty} f_{n+i}} \geq \frac{f_{n+1}}{\sum_{j=1}^{L_Q} f_{n+j} + f_{n+1}\sum_{i=0}^{\infty} \bar{Q}_{L_Q+i}} \geq \frac{f_{n+1}}{f_{n+1}(L_Q + \sum_{i=0}^{\infty} \bar{Q}_{L_Q+i})} = \frac{1}{\mu_Q},$$

and so

$$\lambda_Q \leq 1 - \frac{1}{\mu_Q}.$$

This conclude the proof. □

Theorem 5.3 *Let Q be a non-uniformly spaced seed. If $L_Q < \frac{q}{p}(\frac{1}{p})^{w_Q-2} + \frac{1}{p}$, the hit probability Q_n of Q is larger than that of the consecutive seed B of the same weight when n is large enough.*

Proof. If $L_Q < \frac{q}{p}(\frac{1}{p})^{w_Q-2} + \frac{1}{p} = \frac{q}{p}[(\frac{1}{p})^{w_Q-2} - 1] + 1$, by Theorems 3.2 and 5.2, $\lambda_Q \leq 1 - \frac{1}{\mu_Q} < 1 - \frac{1}{\mu_B - w_Q + 1} \leq \lambda_B$. Hence,

$$\lim_{n\to\infty} \frac{\bar{Q}_n}{\bar{B}_n} = \lim_{n\to\infty} \frac{\alpha_Q \lambda_Q^n}{\alpha_B \lambda_B^n} = \lim_{n\to\infty} \frac{\alpha_Q}{\alpha_B}(\frac{\lambda_Q}{\lambda_B})^n = 0,$$

and there exists a large integer N such that, for any $n \geq N$, $\frac{\bar{Q}_n}{\bar{B}_n} < 1$ or $Q_n > B_n$. $\square$

Remark. For a uniformly spaced seed Q of length L_Q and w_Q. Let $l = \frac{L_Q - 1}{w_Q - 1}$. Then, l is the greatest common divisor of the indices of the match positions in Q. It was proved in [10] that for any n, $\bar{Q}_n = (\bar{B}_k)^{l-r}\bar{B}_{k+1}^r$, where B is the consecutive seed of the same weight w_Q, $k = \lfloor n/l \rfloor$ and $r = n - kl$. This implies that $\lambda_Q = \lambda_B$ [11].

6 Identify Optimal Spaced Seeds

The *sensitivity* of a spaced seed Q over the homologous region of length n is defined to be $Q_n(p)$ given the similarity level p [27]. The PatternHunter default spaced seed $111*1**1*1**11*111$ was selected due to the fact that it has the maximum value of $Q_{64}(0.70)$ over all the spaced seeds of weight 11. This raises two questions: first, whether there is a spaced seed that is optimal for every $0 < p < 1$ when n is fixed, and second, whether the optimal spaced seed with $n = 64$ is optimal for other $n > 64$ when p is fixed.

The study in [12] indicates that the ranking of a spaced seed changes with p when n is fixed. We define the optimum span of a spaced seed as the similarity interval in which it is optimal over all the spaced seeds of the same weight. It was reported in [12] that, when $n = 64$, the PatternHunter default seed is only optimum in the similarity interval $[61\%, 73\%]$, but the spaced seed $111*1*11*1**11*111$ of weight 12 has rather large optimum span $[59\%, 96\%]$. When the length and weight of a spaced seed are large, its sensitivity could fluctuate greatly with the similarity level p. Hence, the larger the weight and length of a spaced seed, the narrower its optimum span.

Theorem 5.2 demonstrates that μ_Q is asymptotically the dominant factor of the sensitivity of a spaced seed Q. By Theorem 3.1, μ_Q depends only on the similarity level p and the seed's structure. Hence, the rankings of the spaced seeds are quite stable when the length of the considered homologous region changes, especially when the length is large. This is also consistent with our numerical computation.

A straightforward approach to identifying good spaced seeds is through exhaustive search after the sensitivity ($Q_n(p)$ for some n and p) is computed. As we have known, this approach, however, soon becomes impractical due to the exponentially larger number of spaced seeds and the NP-hardness of computing the sensitivity [25]. Our asymptotic analysis suggests that μ_Q can be used to select good spaced seeds. Therefore, any close but simple approximation of μ_Q, such as the upper bound in Lemma 3.1 can be used to filter out poor spaced seeds. Actually, it was already proposed in [35] and [21] even without too much theoretical justification. Since $\frac{Q_{2L_Q-1}}{f_{2L_Q-1}} \leq \mu_Q \leq \frac{Q_{2L_Q-1}}{f_{2L_Q-1}} + L_Q$ as proved in [10], $\frac{Q_{2L_Q-1}}{f_{2L_Q-1}}$ is another good filter for identifying good spaced seeds.

7 Conclusion

To analyze the sensitivity of spaced seeds for homology search, we have studied the average distance μ_Q between non-overlapping hits and the hit probability Q_n of a spaced seed Q under the Bernoulli sequence model. We give a formula to calculate μ_Q as well as a tight bound of it. We also establish a close relationship between μ_Q and Q_n. Using these results, we have proved that when the length of a non-uniformly spaced seed Q is not too large, it has more non-overlapping hits and larger hit probability in a large homologous region. The suggests why good spaced seeds are more sensitive in homology search.

Many results in this paper can directly generalize to the multiple spaced seeds case. However, it is not clear how to generate these results to a higher order Markov sequence model. In addition, the past research in run statistics mainly focuses on consecutive patterns. Therefore, this work has great potential in finding more applications in quality control and weather forecast where run statistics are applied.

8 Acknowledgment

The author would like to thank K.P. Choi, F. Preparata and Y. Kong for valuable discussions on different aspects of spaced seeds. In particular, Theorem 3.1 is obtained by discussion with Y. Kong. He would also thank the two anonymous referees for helpful suggestions for the revision of Section 4 and J. Yang for help in simulation data analysis. The extended abstract of this work appeared in the Proceedings of SODA'06. This work was partially supported by BMRC Research Grant BMRC01/1/21/19/140 and ARF grant 146-000-068-112.

References

[1] S.F. Altschul *et al.*, Basic local alignment search tool, *J. Mol. Biol.* **215**(1990), pp. 403-410.

[2] S. F. Altschul *et al.*, Gapped Blast and Psi-Blast: a new generation of protein database search programs, Nucleic Acids Res. **25** (1997), pp. 3389-3402.

[3] N. Balakrishnan and M.V. Koutras, Runs and Scans with Applications, John Wiley & Sons, U.S.A., 2002.

[4] S. Batzoglou, The many faces of sequence alignment, Briefings in Bioinformatics **6** (2005). pp. 6-22.

[5] B. Brejovà, D. Brown, and T. Vinař, Optimal spaced seeds for homologous coding regions. *J. Bioinf. and Comp. Biol.* **1**(2004), pp. 595-610. Early version appeared in CPM 2003.

[6] J. Buhler, Efficient large-scale sequence comparison by local-sensitive hashing, *Bioinformatics* **17** (2001), pp. 419-428.

[7] J. Buhler, U. Keich, and Y. Sun, Designing seeds for similarity search in genomic DNA. *Proc. 7th Annual Int'l Conf. on Comput. Mol. Biol.* (RECOMB03), pp. 67-75, Berlin, Germany.

[8] S. Burkhardt and J. Karkkainen, Better Filtering with Gapped q-Grams, Fundamenta Informaticae XXIII (2003), 1001-1018.

[9] A. Califano and I. Rigoutsos, FLASH: fast look-up algorithm for string homology, Technical Report, IBM T.J. Watson Research Center, 1995.

[10] K.P. Choi, and L. Zhang, Sensitivity Analysis and efficient method for identifying optimal spaced seeds, *J. Comput. Sys. Sci.*, vol. 68, 2004, pp. 22-40.

[11] K.P. Choi, and L. Zhang, Analysis of spaced seed technique in sequence alignment, Cosmos **1**(2005), 57-73.

[12] K.P. Choi, F. Zeng, and L. Zhang, Good Spaced Seeds for Homology Search, *Bioinformatics* **20**(2004), pp. 1053-1059.

[13] W. Feller, *An introduction to Probability Theory and its Applications*, Vol. I (3rd edition), Wiley, New York.

[14] V. Gotea, V. Veeramachaneni, and W. Makalowski, "Mastering seeds for genomic size nucleotide BLAST searches," *Nucleic Acids Res.*, vol. 31, 2003, pp. 6935-6941.

[15] L.J. Guibas and A.M. Odlyzko, String Overlaps, Pattern Matching, and Nontransitive Games, *J. of Combin. Theory* (series A) **30** (1981), pp. 183-208.

[16] G.H. Hardy, J.E. Littlewood and G. Pólya, *Inequalities* (2nd Edition), Cambridge University Press, 1952.

[17] P. Jacquet and W. Szpankowski, Analytic Approach to Pattern Matching, In *Applied Combinatorics on Words* (Editor: M. Lothaire), Cambridge Press, 2005.

[18] U. Keich, M. Li, B. Ma, and J. Tromp, On Spaced Seeds for Similarity Search, *Discrete Appl. Math.*, **3** 2004, pp. 253-263.

[19] W.J. Kent, BLAT – the BLAST-like alignment tool. *Genome Res.* **12**(2002), pp. 656 -664.

[20] W.J. Kent, and A.M. Zahler, Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. Genome Res. **10** (2000), 1115-1125.

[21] Y. Kong, Methods to find optimal multiple spaced seeds for homology search using generalized correlations. *Manuscript*.

[22] L. Noé, and G. Kucherov, YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acid Research* **33** (2005), W540-W543.

[23] G. Kucherov, L. Noé, and M. Roytberg, A unifying framework for seed sensitivity and its application to subset seeds, In *Proc. WABI'05*. LNCS, vol. 3692, pp. 251-263, 2005.

[24] S.Y. Li, A martingale approach to the study of occurrence of sequence patterns in repeated experiments. *Ann Prob.* **8** (1980), pp. 1171-1176.

[25] M. Li and B. Ma, On the complexity of computing the sensitivity of spaced seeds. Manuscript.

[26] M. Li, B. Ma, D. Kisman, and J. Tromp, PatternHunterII: highly sensitive and fast homology search. *J. Bioinformatics and Comput. Biol.* **2** (2004), pp. 417-440.

[27] B. Ma, J. Tromp, and M. Li, PatternHunter-faster and more sensitive homology search *Bioinformatics* **18**(2002), pp. 440-445.

[28] P. Nicodéme, B. Salvy, and P. Flajolet, Motif Statistics. *Lecture Notes in Computer Sciences*, vol. 1643, pp. 194-211, 1999.

[29] F.P. Preparata, L. Zhang, and K.P. Choi, Quick, practical selection of effective seeds for homology search, *J. Comput. Biol.* **12** (2005), pp. 1137-1152.

[30] G. Reinert, S. Schbath, and M. Waterman, Probabilistic and statistical properties of words: An overview, J. Comput. Biol. **7** (2000), pp. 1-46.

[31] S.J. Schwager, Run Probabilities in Sequences of Markov-Dependent Trials, *J. Amer. Statist. Assoc.* **78** (1983), pp. 168-175.

[32] S. Schwartz *et al.* Human-Mouse alignment with BLASTZ, *Genome Res.***13** (2003), pp. 103-107.

[33] A.D. Solov'ev, A combinatorial identity and its application to the problem concerning the first occurences of a rare event, *Theory of Probab. and Appl.* **11** (1966), pp. 276-282.

[34] Y. Sun and J. Buhler, Designing multiple simultaneous seeds for DNA similarity search, in *Proc. of RECOMB'04*, 2004, pp. 76-85.

[35] I.-H. Yang, S.-H. Wang, Y.-H. Chen, P.-H. Huang, L. Ye, X. Huang, and K.-M. Chao, Efficient Methods for Generating Optimal Single and Multiple Spaced Seeds, In *Proc. IEEE 4th Symp. on Bioinfor. and Bioeng.*, pp. 411-418, 2004.

i	i-shift	$o_Q(i)$	$o_{Q'}(i)$
0	1 * * 1 * * * 1 * 1 1 * * 1 * * * 1 * 1	4	3
1	1 * * 1 * * * 1 * 1 1 * * 1 * * * 1 * 1	0	0
2	1 * * 1 * * * 1 * 1 1 * * 1 * * * 1 * 1	1	1
3	1 * * 1 * * * 1 * 1 1 * * 1 * * * 1 * 1	1	0
4	1 * * 1 * * * 1 * 1 1 * * 1 * * * 1 * 1	1	1
5	1 * * 1 * * * 1 * 1 1 * * 1 * * * 1 * 1	0	0
6	1 * * 1 * * * 1 * 1 1 * * 1 * * * 1 * 1	1	1
7	1 * * 1 * * * 1 * 1 1 * * 1 * * * 1 * 1	1	
8	1 * * 1 * * * 1 * 1 1 * * 1 * * * 1 * 1	0	
9	1 * * 1 * * * 1 * 1 1 * * 1 * * * 1 * 1	1	

Table 1: *The values of $o_Q(i)$s and $o_{Q'}(i)$s for $Q = 1{*}{*}1{*}{*}{*}1{*}1$ and $Q' = 1{*}{*}{*}1{*}1$. In this example, $L_Q = 10$, $r = 2$, and the relative matching positions of Q are 0, 3, 7, 9.*

p \ w	10	11	12	13	14
0.6	49.02	76.49	121.59	196.09	315.60
0.7	17.00	21.19	26.74	24.25	44.53
0.8	11.24	12.61	14.08	15.66	17.39
0.9	10.15	11.18	12.21	13.24	14.28

Table 2: The values of the upper bound in Theorem 4.1 for different w and p.

Weight	Spaced seeds	p	N/μ_Q	Observed mean
10	1111111111	0.6	2.43	2.56
		0.7	8.72	9.44
		0.8	24.06	24.63
	11*11***11*1*111	0.6	4.41	4.97
		0.7	14.30	14.43
		0.8	31.54	31.32
11	11111111111	0.6	1.46	1.52
		0.7	6.05	6.27
		0.8	18.79	18.97
	11*1*1*11**1**1111	0.6	2.79	3.01
		0.7	10.66	10.79
		0.8	26.18	25.54
12	111111111111	0.6	0.87	0.91
		0.7	4.21	4.08
		0.8	14.76	14.48
	111*1*11*1**11*111	0.6	1.73	1.87
		0.7	7.89	8.25
		0.8	22.41	22.35

Table 3: *Comparison of the theoretical mean* $\frac{N}{\mu_Q}$ *and the observed mean on simulated binary sequences for the selected spaced seeds when* $N = 1000$.