IN SEARCH OF GOOD PREDICTORS FOR IDENTIFYING EFFECTIVE SPACED SEEDS IN HOMOLOGY SEARCH

LI JIANWEI

NATIONAL UNIVERSITY OF SINGAPORE

2005

IN SEARCH OF GOOD PREDICTORS FOR IDENTIFYING EFFECTIVE SPACED SEEDS IN HOMOLOGY SEARCH

LI JIANWEI

(B.Sc. Peking University, China)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2005

To my dearest family

ACKNOWLEDGEMENTS

For the completion of this thesis, I would like very much to express my heartfelt gratitude to my supervisor, Associate Professor Choi Kwok Pui, for all his invaluable advice and guidance, endless patience, kindness and encouragement during the mentor period in the Department of Statistics and Applied Probability of National University of Singapore. I have learned many things from him, especially regarding academic research and character building. I truly appreciate all the time and effort he has spent in helping me to solve the problems encountered even when he is in the midst of his work.

I also wish to express my sincere gratitude and appreciation to my other lecturers, namely Professors Bai Zhidong, Chen Zehua, Loh Wei Liem, etc, for imparting knowledge and techniques to me and their precious advice and help in my study.

It is a great pleasure to record my thanks to my dear friends: to Mr. Zhang Hao, Mr. Zhao Yudong, Ms. Liu Huixia and Ms. Zhu Min, who have given me much help in my study; to Ms. Qin Xuan, Mr. Guan Junwei and his wife Ms. Wang Yu, Ms Zou Huixiao, Ms Peng Qiao and Ms Chen Yan, who have colored my life in the past two years; to Mr. Cheng Xingzhi and Mr. Rong Guodong, who gave me suggestions on programming. Sincere thanks to all my friends who helped me in one way or another and for their friendship and encouragement.

Finally, I would like to attribute the completion of this thesis to other members and staff of the department for their help in various ways and providing such a pleasant working environment, especially to Jerrica Chua for administrative matters and Mrs. Yvonne Chow for advice in computing.

Special thanks to the website http://www.ctex.org for solving all my problems in LaTeX.

Li Jianwei

July 2005

CONTENTS

SUMMARY

It has been observed that the spaced seeds have better speed and sensitivity than the consecutive seeds with the same weight. Different spaced seeds have different sensitivities. To find the optimal spaced seed in the sense of sensitivity (hitting probability) is a very computationally challenging problem. For short spaced seeds, one can obtain the optimal seeds by exhaustive search. However, this is impractical, if not impossible, for long spaced seeds. To handle long seeds, we propose good predictors to reduce the computation and search space to identify the optimal spaced seed. We will introduce several predictors in this thesis. The predictors can be computed very quickly and the predicted optimal seeds are indeed optimal in sensitivity. Using these predictors, we can identify very effective long spaced seed which are impossible for in exhaustive search.

Although the predictors can be quickly computed, it also soon becomes more and

more demanding to handle longer and longer seeds. For very long spaced seeds, we cannot even calculate the predictors values exhaustively. In fact, it is never necessary to do calculation for every seeds, since many seeds are "bad" seeds. We then introduce some index variable to filter the spaced seeds, with which we need only to handle much less seeds but we can also obtain the effective seeds with a good speed.

For searching even longer seeds, we will introduce the sampling method, which needs very few seeds to handle. Combined with the method of predictors and filters, we can find effective seeds as fast as before.

LIST OF TABLES

LIST OF FIGURES

LIST OF NOTATIONS

$\mathbb{P}, \mathbb{E}, \mathbb{I}$	probability, expectation and indicator function
Q	spaced seed, a sequence of 1 and $*$ ("don't care" position)
L	total length of spaced seed Q
w	weight of spaced seed Q, i.e., number of 1's in Q
$\sigma(Q)$	collection of all realization of Q by filling $*$ by 0 or 1
$\mathscr{Q}_{L,w}$	collection of all spaced seeds with length L and weight w
$\left\|\mathscr{Q}_{L,w}\right\|$	the number of spaced seeds in $\mathscr{Q}_{L,w}$
S	(infinitely long) random sequence of 1 (with probability p) and 0 (with probability $q = 1 - p$)
$S[m:n]$	the substring of S from position m to n

A_i	the event that Q hit S at position n, i.e., any member of $\sigma(Q)$ occurs in $S[n-L+1:n]$
$\bar{A}_i$	complement of event A_i
$\bar{A}_{[i:j]}$	abbreviation for $\bar{A}_i\bar{A}_{i+1}\cdots\bar{A}_j$
$\mathbb{HP}_n(Q)$	probability that seed Q hits S at or before position n, i.e., any member of $\sigma(Q)$ occurs in $S[1:n]$
$\overline{\mathbb{HP}}_n(Q)$	$1-\mathbb{HP}_n(Q)$
$Q \gg i$	Q shifted to right by i positions, i.e., adding i 0 in front of Q
$\theta_Q^{(1)}(i)$	self-overlapping coefficient of order 1, defined in page 23
$\theta_Q^{(2)}(i,j)$	self-overlapping coefficient of order 2, defined in page 23
$\theta(i),\theta(i,j)$	abbreviations of $\theta_Q^{(1)}(i)$ and $\theta_Q^{(2)}(i,j)$
Σ_k	$\sum_{i_1\neq i_2\neq\cdots\neq i_k}\mathbb{P}(A_{i_1}\cdots A_{i_k})$
b	the number of blocks of 1's in a spaced seed Q
h	the number of 1's in the the first block of 1's in Q, h for head
t	the number of 1's in the the last block of 1's in Q, t for tail
Δw	the difference of the weight in the two halves of a spaced seed Q
$z_{\max}$	the maximal length of the blocks (runs) of $*$'s in Q
$u_{\max}$	the maximal length of the blocks (runs) of 1's in Q except the two blocks of 1's in the ends
λ_Q	the convergence rate of $\mathbb{HP}_n$ approaching to 1 as $n\to\infty$

CHAPTER 1

Introduction

1.1 Biological background

A common and yet powerful approach to discover biological functions and structures of a DNA sequence (or amino acid) is through sequence alignment with sequence in a database (Yeh *et al.* [2001], Delcher *et al.* [1999], Hardison *et al.* [1997], Li *et al.* [2001]). By comparing genomic sequences, information on translations, tandem and segment duplications can be easily inferred. It is usually done by aligning them using dynamic programming approach (Needleman and Wunsch [1970], Smith and Waterman [1981]). This stimulates unprecedented demand for long DNA sequence comparison, and poses a great challenge to alignment algorithm developers. Popular programs such as FASTA (Lipman and Pearson [1985]), BLAST (Altschul *et al.*

[1990], Altschul *et al.* [1997]), are too computationally demanding to analyze multimegabase sequence even in a modern computer (Gish [2001], Huang and Miller [1991]).

One of the most important techniques for designing faster algorithms for sequence comparison is the idea of filtration (Altschul *et al.* [1990], Altschul *et al.* [1997]). This idea involves a two-stage process. The first stage preselects a set of positions in which given sequences are potentially similar. The second stage verifies each of these possible positions using an accurate method rejecting those that do not satisfy the specified similarity criteria. For example, BLAST programs use this technique. Each of these programs first finds reasonably long exact matches (consecutive k bases) between a given sequence and a sequence in the database, and then extends these exact matches into local alignments. Based on statistical study, two sequences are likely to have high-scoring local alignments only if there are reasonably long exact matches between them. The value of k is usually set to 11 by considering tradeoff between search speed and the sensitivity. The larger the k is, the faster the program but the poorer its sensitivity.

In fact, employing the filtration technique for information retrieval/pattern matching in the computer science and for sequence comparison in computational molecular biology goes back almost two decades. It was first described by Rabin and Karp [1987] for the string matching problem.

Multiple spaced patterns are usually used for approximate matching and sequence comparison. Recently, a creative idea of using a single optimal spaced pattern (called

spaced seed) was introduced in designing a more efficient and sensitive program PatternHunter for sequence comparison by Ma *et al.* [2002]. PatternHunter uses a single optimal match pattern to improve the alignment sensitivity, which is important because the general sequence search aims to identify more homology sequences, and in this case, the mismatch positions are unknown. PatternHunter searches for runs of length 18 consecutive nucleotide bases in each sequence and requires matches at 11 positions. Even in a personal computer, PatternHunter is able to compare prokaryotic genomes in seconds, arabidopis chromosomes in minutes and human or mouse chromosomes in hours(Waterston *et al.* [2002], Scherer *et al.* [2003], Ureta-Vidal *et al.* [2003])

The spaced seeds idea in PatternHunter motivated the problems of identifying optimal spaced seeds in different sequence alignment models (Keith *et al.* [2002], Buhler [2001], Brejovà *et al.* [2003], Choi and Zhang [2004]). By assuming a Markov model, Buhler *et al.* [2003] calculated the sensitivity of a spaced seed adapting the dynamic programming technique in Keith *et al.* [2002]. From this, the optimal spaced seeds can be identified. Brejovà *et al.* [2003] worked on the optimal spaced seeds in the context of detecting homologous coding regions in unannotated genomic sequences. They modified the dynamic programming technique to calculate the sensitivity of spaced seeds in Keith *et al.* [2002] and identified the optimal spaced seeds for aligning coding regions. Choi and Zhang [2004] derived a set of recurrence relations to compute the sensitivity of a spaced seed by assuming a zero-th Markov model of the target sequence.

Although progress has been made to efficiently find the optimal spaced seeds, the current methods are still not fast enough to meet the practical requirement for long

spaced seeds. Some researchers now are trying to find predictors and other techniques so as to improve the speed without miss of effective spaced seeds. Kong [2004] proposed some quantities as predictors of effective spaced seeds. Preparata *et al.* [2005] proposed a sampling trick to reduce the number of seeds of consideration.

1.2 Concepts and notations

Homology search

Two sequences are said to be **homologous** if they share a common ancestry. Given a query sequence s, we want to search the database to find sequences or sub-sequences that are as similar as possible to s, and then use the sequences we find to predict the functions or structure of the new sequence s. The search precess is called **homology search**.

Sequence alignment and matches

In homology search, we align the query sequence s and the target sequence S to find the positions of exact match. For example, if the query sequence s = TAGC, the target sequence S = AATGTAGCGCA, we can align s and S together and shift s from left to right along S to find the exact match as follows:

S: A A T G T A G C G C A
s: T A G C

Spaced seed

If the query sequence s is very long, since S is very long, it is computationally demanding to do the exact homology search, so we use a short segment of s to find identical match in S. This short segment of the query sequence is called a **seed**. If the seed occurs in some position of S, we say that the seed **hits** S at this position. For example, if we treat s itself as a seed in the above alignment, then it hits S at positions $5 \sim 8$. We will use the last position of the segment identical with the seed in S as the hitting position, so we will say that s hits S at position 8.

Further, we can use a 0,1 sequence to denote the alignment between s and S, since we generally only care about match or mismatch. We use 1 for match and 0 for mismatch. This can be illustrated as:

S:	A	A	T	G	T	A	G	C	G	C	A
s:	T	A	G	C							
	0	1	0	0							

S:	A	A	T	G	T	A	G	C	G	C	A
s:							T	A	G	C	
							0	0	1	1	

We also call the 0,1 sequence a **seed**, denoted by Q. Thus, to find the identical match of a seed is equivalent to set the seed to be all 1's (i.e. consecutive seed) with the same length of the seed.

A **spaced seed** is a specified seed of 1 and $*$. Here we use $*$ to denote a "don't care" position to allow match or mismatch on this position. For example if we let

$$Q = 1*11***1*111*11, \; s = \text{ATGTCCACTGATCCT}, \; S = \text{ACGTAACTCCGATCCT},$$

then s will hit S as:

S:	A	C	G	T	A	C	T	C	C	G	A	T	C	C	T
s:	A	T	G	T	C	C	A	C	T	G	A	T	C	C	T
Q	1	$*$	1	1	$*$	$*$	$*$	1	$*$	1	1	1	$*$	1	1

We call the number of 1's in a spaced seed the **weight** of this seed, and the total number of 1's and $*$'s the **length**. We can always assume a spaced seed of length L to start and end with 1's, otherwise, we can simply cut off those $*$'s beyond the 1's in the two ends without loss of information.

Hitting probability

We use **similarity** to name the probability that a match occurs at one particular position. Apparently, the similarity is a kind of average of the probability of the matches of A-A, T-T, C-C and G-G. It measures how similar the query sequence and the target sequence are. We generally use p to denote the similarity. In practice, p is always set around 0.7.

The **hitting probability** or **sensitivity** is the probability that a spaced seed Q hits an independently and identically distributed (i.i.d.) Bernoulli random sequence S of 0 and 1; 1 occurs in S with the probability p, the similarity. We use $\mathbb{HP}_n(Q)$ to denote the hitting probability of spaces seed Q hitting S (with the similarity p) at or before position n.

A simple fact is that, if Q' is the reverse of Q, then we have $\mathbb{HP}_n(Q') = \mathbb{HP}_n(Q)$, because we can simply reverse the target random sequence S to be hit by Q', then the reverse of S is equivalent to S itself since different positions of S are totally independent 0-1 variables.

Obviously, there are many spaced seeds with the same length and same weight. Since we know that the hitting probability of Q and its reverse is the same, we can simply use one of them. Specifically, we always choose the spaced seed that is tail-heavy, which means the weight in the rear half is at least one half of the total weight.We use $\mathscr{Q}_{L,w}$ to denote the collection of all tail-heavy spaced seeds with length L and weight w.

1.3 Main objectives of this thesis

We start with a nested recursive algorithm of Choi and Zhang [2004] to calculate the hitting probability of a given spaced seed Q at any n. Theoretically, one can find the optimal spaced seeds (that is, seeds with the highest hitting probabilities) among all spaced seeds with the same length L and the same weight w. There are two main objectives of this thesis:

(1) to explore some simple but effective predictors for identifying effective spaced seeds;

(2) to introduce good seeds filters to reduce the number of spaced seeds which need to be considered substantially small, hence, improving the identification process more efficiently; and

(3) to estimate the convergence rate of the hitting probability to 1 as n goes to infinity.

In this thesis, we will discuss several indicators for good spaced seeds, which include

(1) the hitting probabilities at smaller n, i.e., the probabilities of early hits

(2) lower bounds or upper bounds of the hitting probabilities including

- Cauchy-Schwartz lower bound
- Bonferroni-type lower bound
- Bonferroni-type upper bound

Although calculating these indicators are much faster than calculating the hitting probabilities, the problem of identifying effective spaced seeds is that the number of spaced seeds with the length L and weight w increases exponentially with L. Therefore, another important issue is to find some simple seeds filter, which is inherently simple and is efficient to distinguish effective spaced seeds from the ineffective ones so as to reduce the total number of spaced seeds need to deal with.

We examine the following seeds filters in the thesis:

- the number of blocks of $*$'s in a spaced seed
- the difference in the number of 1's in the two halves
- the number of 1's in the front and in the tail
- the maximal length of runs of 1's and $*$'s

1.4 Organization of this thesis

We organize this thesis into five chapters. In the next chapter, chapter two, we give the recursive relation to calculate the hitting probability at n, and discuss some characteristics of the hitting probabilities, for example, what is the distribution of the

hitting probabilities over all the spaced seeds in $\mathcal{Q}_{L,w}$, and how does the hitting probability change with n,..., etc. In chapter three, we introduce and evaluate a number of predictors for good spaced. In chapter four, we propose and discuss the essential features of some seeds filters in order to reduce the number of seeds for consideration before we apply our prediction for seeds with larger L and w. In the last chapter, chapter five, we use some quantities to estimate the convergence rate of the hitting probabilities to 1 as n approaches infinity.

CHAPTER 2

Calculating the Hitting Probability

To find the optimal spaced seeds with the highest hitting probabilities, we have to know how to calculate the hitting probability. Previous research has established some recursive formula to calculate this. We first start with the simplest case.

2.1 Simple formula for consecutive seeds

We call a spaced seed Q which consist of only 1's without any $*$'s a **consecutive seed**. For example, 111111 is a consecutive seed with length 6 and weight 6. We let B denote the consecutive seed with weight w. Let $\mathbb{HP}_n(B)$ be the probability that the seed B hits a random sequence S at or before position n, and $\overline{\mathbb{HP}}_n(B) = 1 - \mathbb{HP}_n(B)$ be

the probability that B only hits S after n. Then we can simply have

$$\begin{aligned} \mathbb{HP}_n(B) &= 0, \text{ for } n = 0, 1, \dots w-1, \\ \mathbb{HP}_w(B) &= p^w. \end{aligned} \tag{2.1}$$

To derive this formula for $n \geq w+1$, we study the event that B first hits S at position n, which has probability

$$\mathbb{HP}_n(B) - \mathbb{HP}_{n-1}(B) = \overline{\mathbb{HP}}_{n-1}(B) - \overline{\mathbb{HP}}_n(B).$$

This event occurs if and only if $S[n-L+1:n]$ are all 1's, $S[n-L]$ is 0,and there are no hits in $S[1:n-L-1]$. In this case, S must be like:

$$S: \quad \underbrace{\text{✗}\cdots\cdots\text{✗✗✗}}_{n-w-1} 0 \underbrace{11\cdots11}_{w}$$

where ✗ denote no hit at that position. We can easily get the probability

$$\overline{\mathbb{HP}}_{n-1}(B) - \overline{\mathbb{HP}}_n(B) = p^w q \overline{\mathbb{HP}}_{n-w-1}(B),$$

which leads to the recursive relation as:

$$\overline{\mathbb{HP}}_n(B) = \overline{\mathbb{HP}}_{n-1}(B) - p^w q \overline{\mathbb{HP}}_{n-w-1}(B),$$

or

$$\mathbb{HP}_n(B) = \mathbb{HP}_{n-1}(B) + p^w q \left[1 - \mathbb{HP}_{n-w-1}(B)\right]. \tag{2.2}$$

Using the initial value given in (2.1), we easily get $\mathbb{HP}_n(B)$ for $w \leq n \leq 2w+1$:

$$\begin{aligned} \mathbb{HP}_n(B) &= p^w + (n-w)p^w q, \text{ for } w \leq n \leq 2w \\ \mathbb{HP}_{2w+1}(B) &= p^w + (w+1)p^w q - p^{2w} q \end{aligned}$$

We can calculate the hitting probabilities of larger n recursively by (2.2).

2.2 Formula for general spaced seed

Choi and Zhang [2004] derived a nested relation to compute the hitting probability of general spaced seeds recursively. For completeness of discussion, we include the dirivation here.

To calculate the hitting probability of spaced seed Q at position n, we let A_j be the event that Q hits S at position j, and $\bar{A}_j$ be the complement of A_j. We use $A_{[i:j]}$ for abbreviation of $A_i A_{i+1} \cdots A_j$ for $i < j$, and similarly $\bar{A}_{[i:j]} \triangleq \bar{A}_i \bar{A}_{i+1} \cdots \bar{A}_j$, then we have

$$\mathbb{HP}_n(Q) = \mathbb{P}\left(\bigcup_{L \le i \le n} A_i \right).$$

We define f_n as the probability that Q first hits S at n, that is

$$f_n = \mathbb{P}(\bar{A}_{[L:n-1]} A_n). \tag{2.3}$$

Let $\sigma(Q) = \{Q_1, Q_2 \cdots, Q_m\}$ be the set of all $m = 2^{L-w}$ distinct realizations of Q by replacing the "don't care" positions by 0 or 1. For example, if $Q = 1 * 1 * 1$ then

$$\sigma(Q) = \{10101, 11101, 10111, 11111\}.$$

We let $A_n^{(j)}$ be the event that the word Q_j occurs in S at n, then $A_n = \bigcup_{1 \le j \le m} A_n^{(j)}$ and $A_n^{(j)}$ are all disjoint. We let $f_n^{(j)} = \mathbb{P}(\bar{A}_{[L:n-1]} A_n^{(j)})$ be the probability that Q_j first occurs in S at n. Then we have the following theorem.

Theorem 2.1 (Choi and Zhang 2004) *We can calculate* $\mathbb{HP}_n$ *by the following relations:*

$$\mathbb{HP}_n = \sum_{i=1}^{n} f_i \tag{2.4}$$

$$f_n = \sum_{1 \le j \le m} f_n^{(j)} \tag{2.5}$$

$$f_n^{(j)} = \mathbb{P}(Q_j)\overline{\mathbb{HP}}_{n-L} - \sum_{i=1}^{L-1}\left[\left(\sum_{k\in\Gamma_{i,j}} f_{n-i}^{(k)}\right)\mathbb{P}(Q_j[L-i+1:L])\right] \tag{2.6}$$

with the following initial values

$$\mathbb{HP}_n = f_n = 0,\ 1 \le n < L$$

$$\mathbb{HP}_n = f_n = p^w,\ n = L$$

Here $P(Q_j)$ *is the probability of the word* Q_j *occurs and*

$$\Gamma_{i,j} = \left\{k | Q_k[i+1:L] = Q_j[1:L-i]\right\}.$$

Proof: It is easy to see that(2.4), (2.5) and the initial values hold. For equation (2.6), we notice that

$$\bar{A}_{[L:n-1]} = \bar{A}_{[L:n-L]} \Big\backslash \bigcup_{i=1}^{L-1} \bar{A}_{[L:n-i-1]} A_{n-i}\ , \tag{2.7}$$

which is simply corresponding to

$$\overline{\mathbb{HP}}_{n-1} = \overline{\mathbb{HP}}_{n-L} - \sum_{i=n-L+1}^{n-1} f_i,$$

we intersect with $A_n^{(j)}$ on each event in (2.7) and get

$$\begin{aligned}\bar{A}_{[L:n-1]}A_n^{(j)} &= \bar{A}_{[L:n-L]}A_n^{(j)} \Big\backslash \bigcup_{i=1}^{L-1} \bar{A}_{[L:n-i-1]}A_{n-i}A_n^{(j)} \\ &= \bar{A}_{[L:n-L]}A_n^{(j)} \Big\backslash \bigcup_{i=1}^{L-1}\left(\bigcup_{k=1}^{m} \bar{A}_{[L:n-i-1]}A_{n-i}^{(k)}A_n^{(j)}\right).\end{aligned}$$

The event $A^{(k)}_{n-i}A^{(j)}_n$ occurs if and only if the substring $Q_k[i+1:L]$ and $Q_j[1:L-i]$ are identical. In the event $\bar{A}_{[1:n-L]}A^{(j)}_n$, $\bar{A}_{[1:n-L]}$ and $A^{(j)}_n$ are independent because they involve totally separate part $S[1:n-L]$ and $S[n-L+1:n]$ of S. If we observe that the events in the union are all independent, then the above equation naturally leads to (2.6). ■

2.3 Computational results of exact calculation

Table 2.1 (on page 15) shows the top 10 seeds together with their hitting probabilities at position $n = 64$ of $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ and $\mathcal{Q}_{20,13}$ for $p = 0.5, 0.7, 0.9$.

From this table, we observe that the $\mathbb{HP}_{64}$ of the top 10 spaced seeds of one $\mathcal{Q}_{L,w}$ do not vary much, and the differences among them become smaller and smaller as L and w increase. For example, for $\mathcal{Q}_{20,13}$, which have 15912 spaced seeds, the largest hitting probability at $p = 0.7$ is 0.26475018; the 1000-th largest is 0.25809995; the 10000-th largest is 0.24613015; the 100-th smallest is 0.21659947; the smallest is 0.16495660.

To see the distribution of $\mathbb{HP}_n$ over all spaced seeds clearer, we may refer to the density plot in Figure 2.1 (on page 16). We can observe that the distribution of $\mathbb{HP}_n$ is very skewed. A large part of seeds have good sensitivities.

Hence, in practice, we may only need to find very good spaced seeds instead of the best one, because

(1) the hitting probabilities of very good spaced seeds differ slightly,

Table 2.1 *Top 10 seeds of $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}, \mathcal{Q}_{20,13}$ for different p*

$\mathcal{Q}_{L,w}$	$p = 0.5$	$\mathbb{HP}_{64}$	$p = 0.7$	$\mathbb{HP}_{64}$	$p = 0.9$	$\mathbb{HP}_{64}$
$\mathcal{Q}_{15,9}$	111***1*1*11*11	0.0835314	111***1*1*11*11	0.7291560	111***1*1*11*11	0.9999117
	111**1**1*1*111	0.0835138	111*1***11*1*11	0.7285212	111**1**1*1*111	0.9999089
	111*1***11*1*11	0.0835065	111**1**1*1*111	0.7284156	111*1***11*1*11	0.9999088
	11*11**1*1**111	0.0834830	11*11**1*1**111	0.7283361	11*11**1*1**111	0.9999073
	11**1*1*1**1111	0.0833132	11**1*1*1**1111	0.7271766	11**1*1*1**1111	0.9999071
	111**1**11*1*11	0.0832590	11**1**1*1*1111	0.7262585	11**1**1*1*1111	0.9999050
	11**11*1**1*111	0.0832450	111**1**11*1*11	0.7259705	1*1*1**11**1111	0.9999027
	11**1**1*1*1111	0.0831087	11**11*1**1*111	0.7257927	1*1*11*11***111	0.9999019
	111*1**1**1*111	0.0830764	1*1*1**11**1111	0.7254126	1**11**1*1*1111	0.9999016
	11*1*1**1**1111	0.0830667	11*1*1**1**1111	0.7252475	11*11***11*1*11	0.9999012
$\mathcal{Q}_{18,12}$	111*1*11*1**11*111	0.0107008	111*1*11*1**11*111	0.3564296	111*1*11*1**11*111	0.9958336
	111*1**11*1*11*111	0.0106887	111*1**11*1*11*111	0.3556505	111*1**11*1*11*111	0.9957644
	11*11*1*1*11**1111	0.0106783	11*11*1*1*11**1111	0.3545175	111*1*1**111*11*11	0.9956795
	111*1*1**11*11*111	0.0106697	111*1*1**11*11*111	0.3544993	11*1*111*1**111*11	0.9956546
	111**11*11*1*1*111	0.0106603	111*1*1**111*11*11	0.3541413	111**11*1*1**11111	0.9956131
	111*1*1**111*11*11	0.0106565	1111**11**1*1*1111	0.3538696	11*1*1*11**11*1111	0.9956102
	111*1*11**1*11*111	0.0106552	111**11*1*1**11111	0.3538638	111*1*1**11*11*111	0.9956097
	11*1*1*11**11*1111	0.0106545	11*1*1*11**11*1111	0.3537460	1111*1***111*11*11	0.9955834
	111*11**1*1*11*111	0.0106526	111**11*11*1*1*111	0.3533500	11*11*1*1*11**1111	0.9955396
	11*111**1*11*1*111	0.0106503	111*11**11*1**1111	0.3530935	11**111*1**1*11111	0.9955339
$\mathcal{Q}_{20,13}$	111*1*11**11**1*1111	0.0052289	111*1*11**11**1*1111	0.2647502	111*1**11*1**111*111	0.9906267
	1111*1*1**11*11**111	0.0052265	111*1**11*1**111*111	0.2645119	111*1*11**11**1*1111	0.9904919
	111*1**11*1**111*111	0.0052242	1111*1*1**11*11**111	0.2644288	111*1*1**1*11**11111	0.9904793
	111*11**1*1*11**1111	0.0052216	111*1*1**1*11**11111	0.2640164	1111*1*1**11*11**111	0.9904206
	111*11**1*11*1*1*111	0.0052209	111*11**1*1*11**1111	0.2637489	111*1**1*11**111*111	0.9902543
	1111**11**1*1*11*111	0.0052195	1111*1**1**111*1*111	0.2634269	1111*1*1**111**11*11	0.9902031
	111*11*1**11*1*1*111	0.0052190	1111**11**1*1*1*1111	0.2634076	1111*1**1**111*1*111	0.9901883
	111*11**1*1*11*1*111	0.0052189	1111**1*1*1**11*1111	0.2633813	111*1*1***11*11*1111	0.9901600
	111*1*11**1*1*11*111	0.0052185	111*11*1**11*1*1*111	0.2633607	111*11**1*1*11**1111	0.9901581
	1111*1*1**111**11*11	0.0052169	111*11**1*1*11*1*111	0.2633077	111**11*1*1*1**11111	0.9901399

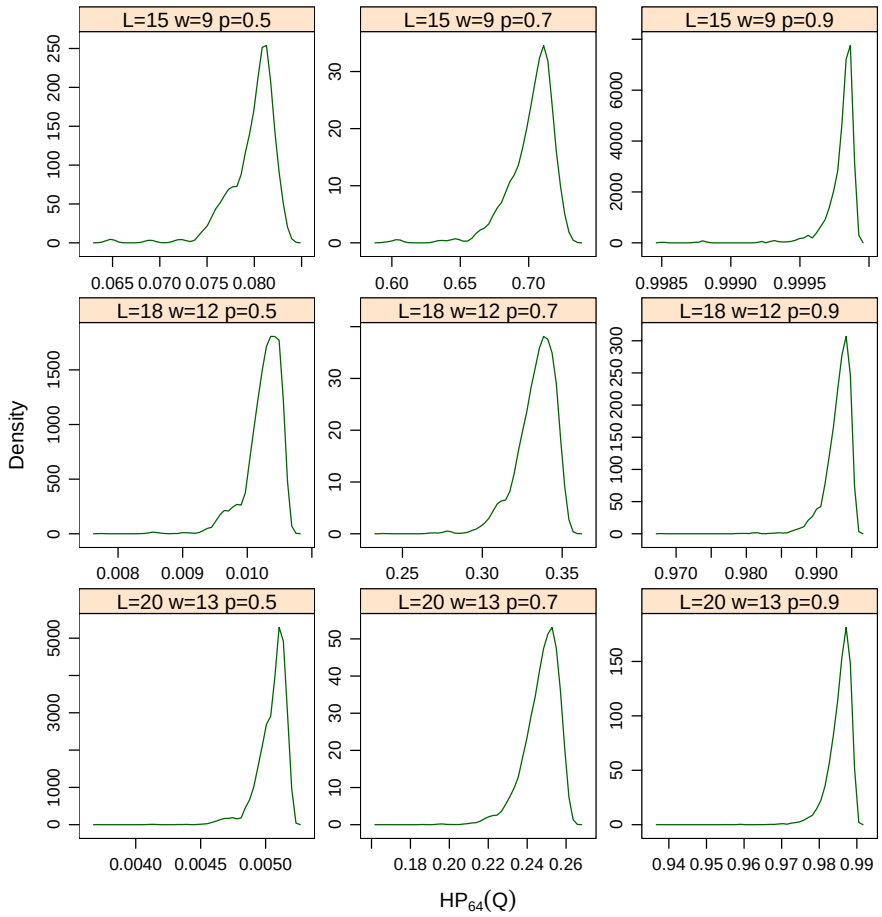

Figure 2.1 *Kernel density plots of* $\mathbb{HP}_n(Q)$ *of* $\mathscr{Q}_{15,9}, \mathscr{Q}_{18,12}, \mathscr{Q}_{20,13}$.

(2) the optimal spaced seed for one p may not be the best for another p. For example, in Table 2.1 (on page 15), the optimal seed of $\mathscr{Q}_{20,13}$ at $p = 0.7$ is only the second best for the case $p = 0.9$. Thus, when we have no idea of the precise p value, we need not know which seed is the best.

In Figure 2.2 (on page 17), the relation between $\mathbb{HP}_n$ and n are illustrated for four spaced seeds of $\mathscr{Q}_{20,13}$, in which $111*1*11**11**1*1111$ and $1*******111111111111$ are respectively the optimal seed and worst seed when $p = 0.7$. We can observe the

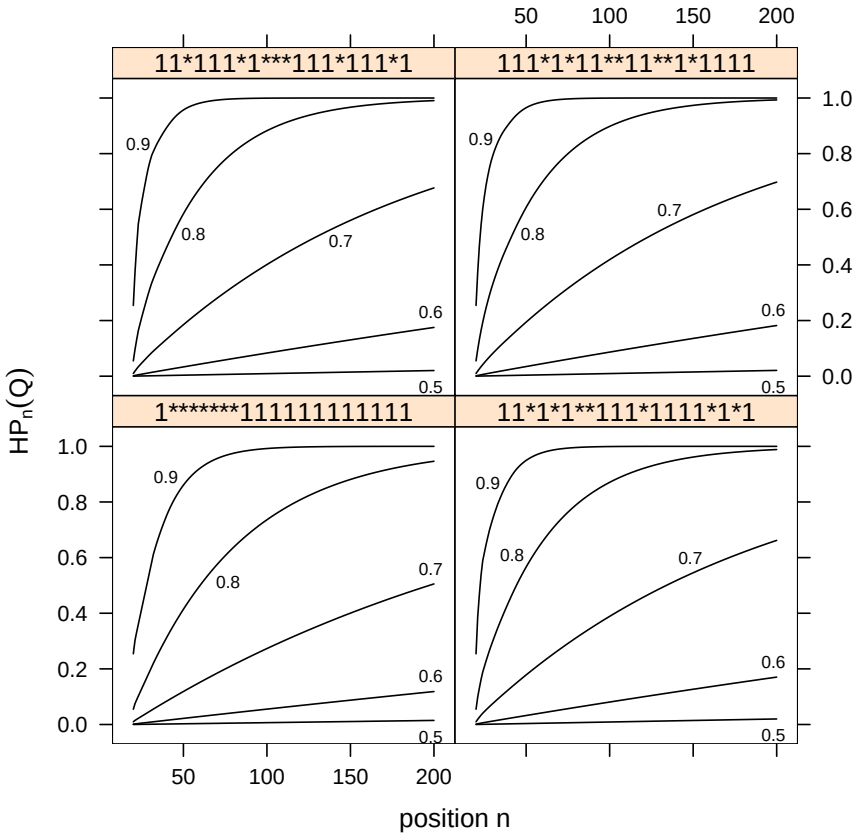

Figure 2.2 *Plots of* $\mathbb{HP}_n(Q)$ *vs n for four spaced seeds of* $\mathcal{Q}_{20,13}$*, in which, according to their* $\mathbb{HP}_{64}(Q)$ *at* $p = 0.7$*,* $111*1*11**11**1*1111$ *is the optimal seed of* $\mathcal{Q}_{20,13}$ *and* $1*******111111111111$ *the worst seed of* $\mathcal{Q}_{20,13}$*. The 5 lines from bottom to top in each sub-plot are hitting probabilities for* $p = 0.5 \sim 0.9$*. The x-axis, which stands for n, is from 20 to 200.*

hitting probability is quite proportional to the position n for small p (the lower lines). For p close to 1, e.g. 0.9 (the top curve), the hitting probability will soon increase close to 1.

2.4 Complexity of the exact calculation

It can be shown that the complexity of this algorithm is $O(Ln2^{2(L-w)})$, which means it will increase exponentially with $L-w$ and linearly with L and n. For spaced seeds with relatively small L and $L-w$, it is feasible to run the exact calculation to compute their hitting probabilities. For example, for a given p and $n = 64$, it may takes less than one hour in a microcomputer (with Pentium® IV 2.4GH CPU) to exhaustively compute the hitting probability of all the spaced seeds of $\mathscr{Q}_{18,12}$, but it takes about one day to exhaustively calculate the $\mathbb{HP}_{128}$ of $\mathscr{Q}_{23,15}$ for a specified p.

Since the exhaustive search is so time-consuming, we have to find some other quantities which can be calculated relatively easily to predict the best spaced seeds. In the next chapter, we will introduce some predictors for best spaced seeds.

However, it is still meaningful to search the optimal spaced seed exhaustively for small L and w, since the optimal spaced seeds will provide us important information on what the effective spaced seeds would probably look like, and from this we are able to formulate some heuristic methods to predict effective spaced seeds for large L and w. In addition, this algorithm enables us to check whether the spaced seeds we predict are really better than some others.

CHAPTER 3

Predictors for Effective Spaced Seeds

Recall that the complexity of the algorithm for exact calculation of the hitting probability will increase very exponentially with $L-w$ and linearly with L and n. This implies that we cannot identify the optimal seeds by exhaustive search for large L and w. For example, it will take years to calculate $\mathbb{HP}_{128}$ of $\mathscr{Q}_{35,22}$. Another important reason is the number of seeds of $\mathscr{Q}_{L,w}$ increases tremendously with L, we will talk about this later in chapter 4). Thus, it is necessary to find some indicators which can be easily computed to predict the optimal spaced seeds or at least very good spaced seeds.

3.1 Predict using hitting probability $\mathbb{HP}_{2L-1}$

A simple and also efficient method is to use the hitting probability at small n to predict those at large n as was exploited by Choi *et al* [2004]. Figure 2.2 (on page 17) shows the relation between $\mathbb{HP}_n(Q)$ and n for four selected spaced seeds of $\mathcal{Q}_{20,13}$. We can see from the figure that, when p is not very close to 1, $\mathbb{HP}_n(Q)$ is quite proportional to n for moderate n, when p is close to 1, there will be a curve relation between them. Among these four seeds, $111*1*11**11**1*1111$ and $1*******111111111111$ are respectively the best and worst seeds of $\mathcal{Q}_{20,13}$ for $n = 64, p = 0.7$. The other two is about the 33 and 66 percentile of the ranked spaced seeds of $\mathcal{Q}_{20,13}$. So we may expect all the member of $\mathcal{Q}_{20,13}$ and other $\mathcal{Q}_{L,w}$ will possess this linearity feature, and we do find that this feature also shown on other spaced seeds. Therefore, we expect that $\mathbb{HP}$ at small n forms a good predictor of $\mathbb{HP}_n$ at larger n.

Figure 3.1 (on page 21) illustrate the strong correlation as we expected between $\mathbb{HP}_n$ and $\mathbb{HP}_{2L-1}$ of $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ and $\mathcal{Q}_{23,15}$ for $p = 0.5, 0.7, 0.9$. We also computed the Pearson correlation coefficients and Spearman rank correlation between $\mathbb{HP}_n$ and $\mathbb{HP}_{2L-1}$ for the nine cases in this figure (not shown here), all the nine values are greater that 0.97, which gives strong evidence of the predictability of $\mathbb{HP}_{2L-1}$.

We choose $\mathbb{HP}_{2L-1}$ instead of other early $\mathbb{HP}$ are mainly based on the following two reasons:

(1) Since the proposition of the concept of spaced seeds is to beat the consecutive seeds, we will want the hitting probabilities of spaced seed being greater than those of the consecutive seeds. However, as the consecutive seed is shorter in

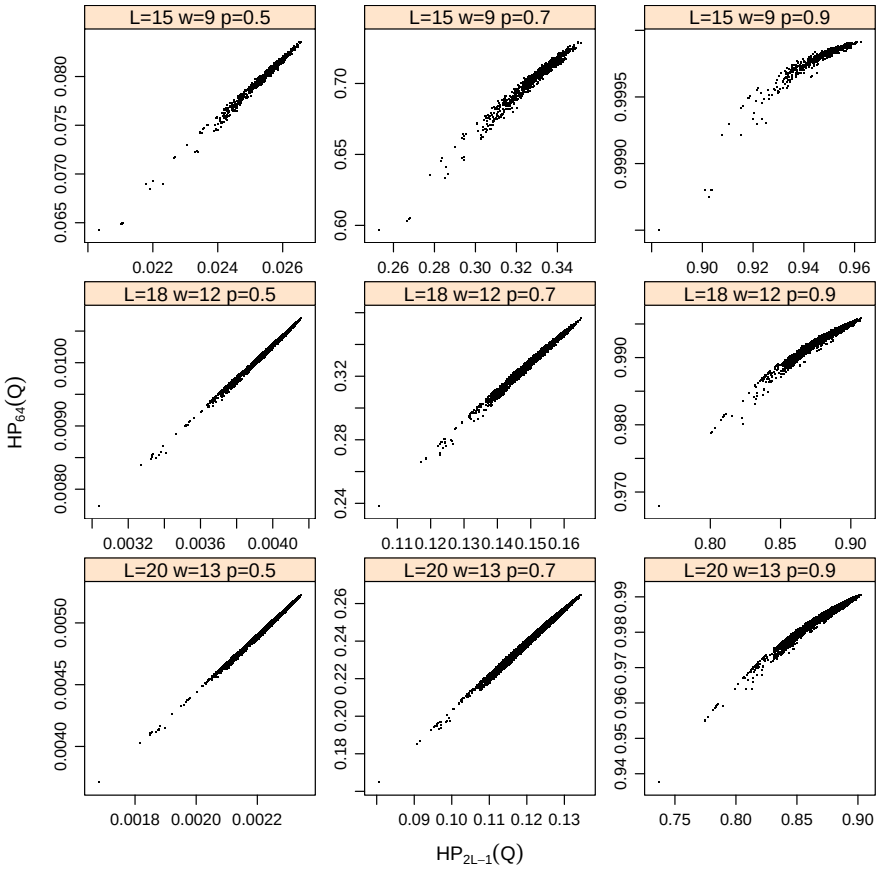

Figure 3.1 *Plots of $\mathbb{HP}_n(\mathcal{Q})$ vs $\mathbb{HP}_{2L-1}(\mathcal{Q})$ for $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ and $\mathcal{Q}_{20,13}$ (rows from top to bottom) for $p = 0.5, 0.7, 0.9$ (columns from left to right).*

length, it has the priority at the early hitting, but soon it will be caught up with by the spaced seeds in the hitting probability. Choi and Zhang [2004] showed that when comparing with consecutive seeds, the hitting probabilities of good spaced seeds have already caught up with the consecutive seed well before $2L$. This consists a reason for us to consider $\mathbb{HP}_{2L-1}$.

(2) Research has shown that the information of overlaps of spaced seed with itself plays an important role in the hitting problem, and the indicators we will introduce below is also concerned with the overlapping of the spaced seeds.

The following theorem implies the calculation of $\mathbb{HP}_{2L-1}$ takes account of all possible overlapping structure of a spaced seed with itself.

Theorem 3.1 (Choi and Zhang) *For a spaced seed Q with length L and weight w, we have*

$$\mathbb{HP}_{2L-1} = Lp^w - (L-1)\mathbb{P}(A_L A_{L+1}) - \sum_{k=2}^{L-1}(L-k)\mathbb{P}(A_L \bar{A}_{[L+1:L+k]} A_{L+k}) \tag{3.1}$$

where A_j defined as section 2.2.

Proof: Consider

$$\begin{aligned}\mathbb{HP}_{2L-1} &= \mathbb{HP}_L + (L-1)f_L - \sum_{j=L+1}^{2L-1}(f_L - f_j) \\ &= L\mathbb{HP}_L - \sum_{j=L+1}^{2L-1}\sum_{k=L}^{j-1}(f_k - f_{k+1}) \\ &= Lp^w - \sum_{k=L}^{2L-2}(2L-1-k)(f_k - f_{k+1}).\end{aligned}$$

Observe that

$$f_n = \mathbb{P}(\bar{A}_{[L:n-1]}A_n) = \mathbb{P}(\bar{A}_{[L:n-1]}) - \mathbb{P}(\bar{A}_{[L:n]}) = \mathbb{P}(\bar{A}_{[L+1:n]}) - \mathbb{P}(\bar{A}_{[L:n]}) = \mathbb{P}(A_L\bar{A}_{[L+1:n]}),$$

we have

$$f_L - f_{L+1} = \mathbb{P}(A_L A_{L+1})$$

for $k \geq L+1$,

$$\begin{aligned}f_k - f_{k+1} &= \mathbb{P}(A_L\bar{A}_{[L+1:k]}) - \mathbb{P}(A_L\bar{A}_{[L+1:k]} \\ &= \mathbb{P}(A_L\bar{A}_{[L+1:k]}A_{k+1}).\end{aligned}$$

Substituting these into above equation gives us the result. ■

In equation (3.1), the events $A_L A_{L+1}$ and $A_L \bar{A}_{[L+1:L+k]} A_{L+k}$ involve all the possible overlapping of spaced seed with the translation of itself.

3.2 Predictors using upper or lower bounds of $\mathbb{HP}_n$

Besides using the hitting probability itself, we can also use some estimations of $\mathbb{HP}_n$. Applying some known inequalities, we are able to derive lower or upper bounds of $\mathbb{HP}_n$. We explore whether these bounds will form good indicators of the effectiveness of spaced seeds.

We need to introduce the notation of **self-overlapping index of order 1**, $\theta_Q^{(1)}(i)$, which will be abbreviated as $\theta(i)$ if it is clear from the context. When the spaced seed Q is written in a vector Q of 0 and 1 with length L (we fill the "don't-care" position with 0 now), we always set $Q[i] = 0$ for $i < 1$ or $i > L$(e.g., if $L = 5$, $Q[6] = Q[-2] = 0$). We use $Q \gg i$ to denote the sequence of Q shifted to the right by i positions, or the vector of Q with i zeros added in front. For example, if $Q = 10101$, then $Q \gg 2 = 0010101$. We define $Q \gg 0 = Q$. Now we can give the definition of $\theta_Q^{(1)}(i)$ as

$$\theta_Q^{(1)}(i) \triangleq \sum_{j=1}^{L} Q[j] \cdot (Q \gg i)[j] \tag{3.2}$$

which is actually equivalent to the number of common 1's when Q and $Q \gg i$ are aligned together. We use $\theta(i)$ for abbreviation of $\theta_Q^{(1)}(i)$.

Similarly, we define **self-overlapping index of order 2**, $\theta_Q^{(2)}(i, j)$, as

$$\theta_Q^{(2)}(i, j) \triangleq \sum_{k=1}^{L} Q[k] \cdot (Q \gg i)[k] \cdot (Q \gg i + j)[k]$$

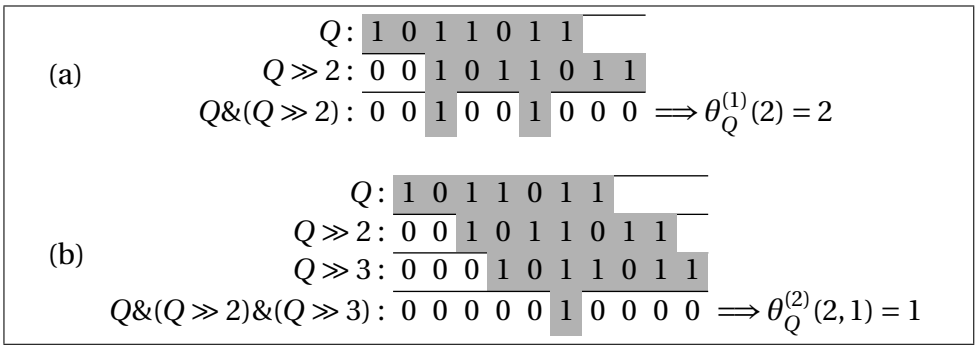

Figure 3.2 *(a) illustrates $\theta_Q^{(1)}(2)$ for $Q = 1011011$. (b) illustrates $\theta_Q^{(2)}(2,1)$ for $Q = 1011011$. The shaded cells in the first 2 rows of (a) and first 3 rows of (b) highlight the spaced seed Q, the shaded cells in the last rows highlight the common 1's of Q and the shifted Qs.*

which is equal to the number of common 1's when Q, $Q \gg i$ and $Q \gg i + j$ are aligned together. We use $\theta(i, j)$ to abbreviate $\theta_Q^{(2)}(i, j)$.

Obviously, $\theta(i) = 0$ if $i \geq L$, and similarly, $\theta(i, j) = 0$ if $i + j \geq L$. Figure 3.2 (on page 24) illustrates the calculation of $\theta_Q(2)$ and $\theta_Q(2,1)$ for $Q = 1011011$. Now we introduce the following three bounds of $\mathbb{HP}_n$.

3.2.1 Lower bound by Cauchy-Schwartz inequality

Let H_n denote the number of hits of Q in $S[1:n]$, Cauchy-Schwartz inequality gives us

$$[\mathbb{E}(H_n)]^2 = \left[\mathbb{E}(H_n \mathbb{I}_{H_n \geq 1})\right]^2 \leq \mathbb{E}(H_n^2)\mathbb{P}(H_n \geq 1) = \mathbb{E}(H_n^2)\mathbb{HP}_n,$$

The last equation is because the event $\{H_n \geq 1\}$ is equivalent to Q hitting S at or before position n. So we get

$$\mathbb{HP}_n \geq \frac{(\mathbb{E}H_n)^2}{\mathbb{E}(H_n^2)}. \tag{3.3}$$

Because we know that $H_n = \sum_{i=L}^{n} \mathbb{I}_{A_i}$, where A_i defined as section 2.2 and $\mathbb{I}_{A_i}$ is the indicator of whether event A_i occurs, we can calculate $\mathbb{E}(H_n)$ as

$$\mathbb{E}(H_n) = \mathbb{E}\left(\sum_{i=L}^{n} \mathbb{I}_{A_i}\right) = \sum_{i=L}^{n} \mathbb{P}(A_i) = (n-L+1)\,p^w.$$

Similarly,

$$\mathbb{E}(H_n^2) = \mathbb{E}\left(\sum_{i=L}^{n} \mathbb{I}_{A_i}\right)^2 = \mathbb{E}\left(\sum_{i=L}^{n} \mathbb{I}_{A_i} + \sum_{i\neq j} \mathbb{I}_{A_i}\mathbb{I}_{A_j}\right) = (n-L+1)\,p^w + \sum_{i\neq j} \mathbb{P}(A_iA_j).$$

To calculate $\mathbb{P}(A_iA_j)$, we only need to count the number of 1's in the sequence $(Q \gg i) \bigcup (Q \gg j)$. Note that the numbers of 1's in $Q \gg i$ and $Q \gg j$ are both equal to the weight w, and that the common number of 1's of $Q \gg i$ and $Q \gg j$ is $\theta(j-i)$, so $\mathbb{P}(A_iA_j) = p^{2w-\theta(j-i)}$. Now

$$\begin{aligned}
\sum_{i\neq j} \mathbb{P}(A_iA_j) &= 2\sum_{i<j} \mathbb{P}(A_iA_j) = 2\sum_{j=L+1}^{n}\sum_{i=L}^{j-1} \mathbb{P}(A_iA_j) = 2\sum_{j=L+1}^{n}\sum_{i=L}^{j-1} p^{2w-\theta(j-i)} \\
&= 2\sum_{d=1}^{n-L}\sum_{i=L}^{n-d} p^{2w-\theta(d)} = 2\sum_{d=1}^{n-L}(n-L-d+1)\,p^{2w-\theta(d)} \\
&= 2\sum_{d=1}^{L-1}(n-L-d+1)\,p^{2w-\theta(d)} + (n-2L+1)(n-2L+2)\,p^{2w}. \qquad (3.4)
\end{aligned}$$

Thus, we can now express the lower bound of $\mathbb{HP}_n$ in (3.3) as

$$\frac{(n-L+1)^2\,p^{2w}}{(n-L+1)\,p^w + (n-2L+1)(n-2L+2)\,p^{2w} + 2\sum_{d=1}^{L-1}(n-L-d+1)\,p^{2w-\theta(d)}}.$$

According to this, we are able to calculate the Cauchy-Schwartz lower bound of each spaced seed.

Figure 3.3 (on page 26) shows the correlation between $\mathbb{HP}_n$ and its Cauchy-Schwartz lower bound, we can see from this figure that when p is not close to 1, then $\mathbb{HP}$ and the Cauchy-Schwartz lower bound have a fairly good linear relationship. Although this

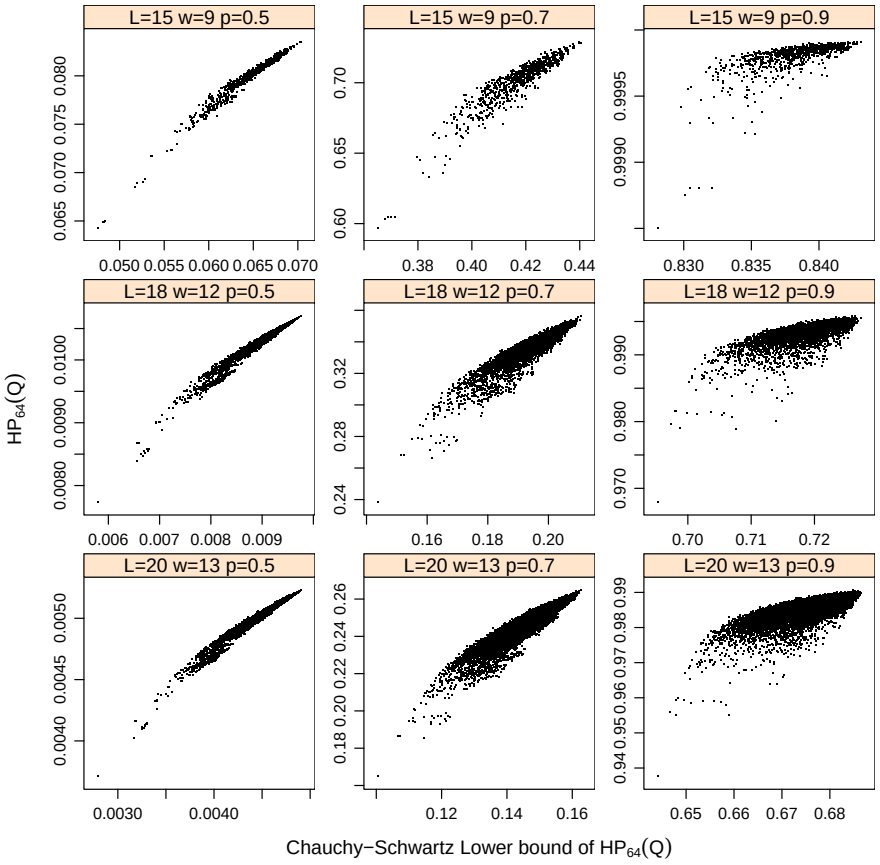

Figure 3.3 *Plots of $\mathbb{HP}_n(\mathcal{Q})$ vs its Cauchy-Schwartz lower bound of $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ and $\mathcal{Q}_{20,13}$ (rows from top to bottom) for $p = 0.5, 0.7, 0.9$ (columns from left to right).*

may change as p becoming close to 1, we can also observe that there is also strong rank correlation between them, so we can conclude that the Cauchy-Schwartz lower bound of $\mathbb{HP}_n$ turns out to be a fairly good indicator.

3.2.2 Lower bound by a Bonferroni-type inequality

We start with a well known Bonferroni-type inequality, which can be found, for example, in Galambos J. and Simonelli I. [1996].

Theorem 3.2 *For a set of event* $\{E_i\}_{i=1}^n$, *we have*

$$\mathbb{P}(\bigcup_{i=1}^n E_i) \geq \frac{2}{k}\sum_{i=1}^n \mathbb{P}(E_i) - \frac{1}{k(k-1)}\sum_{i\neq j}\mathbb{P}(E_iE_j)$$

where $k = \left\lfloor \frac{\sum_{i\neq j}\mathbb{P}(E_iE_j)}{\sum_{i=1}^n \mathbb{P}(E_i)} \right\rfloor + 2$.

When we apply this inequality in the hitting probability problem, we will have

$$\mathbb{HP}_n = \mathbb{P}(\bigcup_{i=L}^n A_i) \geq \frac{2\Sigma_1}{k} - \frac{\Sigma_2}{k(k-1)}$$

where

$$\Sigma_1 = \sum_{i=1}^n \mathbb{P}(A_i) = (n-L+1)p^w, \qquad \Sigma_2 = \sum_{i\neq j}\mathbb{P}(A_iA_j), \qquad k = \left\lfloor \frac{\Sigma_2}{\Sigma_1} \right\rfloor + 2. \tag{3.5}$$

In fact, Σ_2 has been calculated in (3.4).

Figure 3.4 (on page 28) shows the scatter plot of $\mathbb{HP}(\mathcal{Q})$ with this Bonferroni-type lower bounds. We may observe that this figure seems very similar with Figure 3.3 (on page 26) for the Cauchy-Schwartz lower bounds.

3.2.3 Upper bound by Bonferroni inequality

We recall the well-known Bonferroni inequalities.

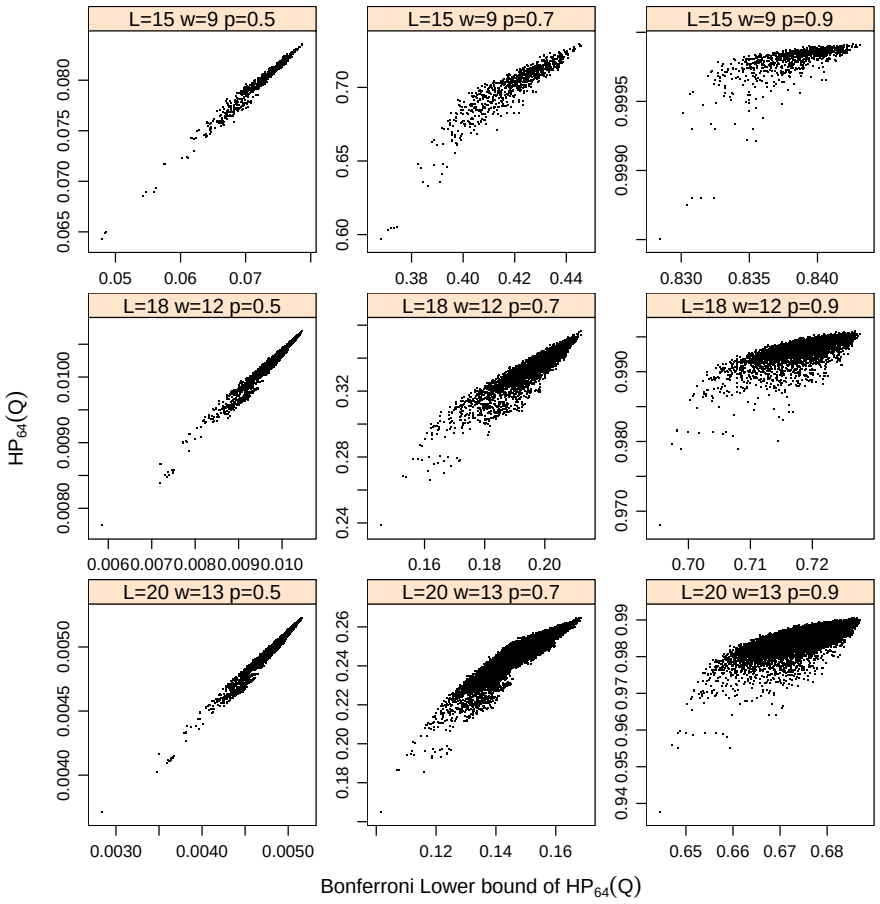

Figure 3.4 *Plots of $\mathbb{HP}_n(\mathcal{Q})$ vs its Bonferroni lower bound of $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ and $\mathcal{Q}_{20,13}$ (rows from top to bottom) for $p = 0.5, 0.7, 0.9$ (columns from left to right).*

Theorem 3.3 (Bonferroni) *For a set of event $\{E_i\}_{i=1}^n$, if we let $\sigma_1 = \sum_{i=1}^n \mathbb{P}(E_i)$, $\sigma_2 = \sum_{i\neq j} \mathbb{P}(E_i E_j), \ldots, \sigma_k = \sum_{i_1\neq i_2\neq\cdots\neq i_k} \mathbb{P}(E_{i_1}\cdots E_{i_k})$, then we have*

$$\mathbb{P}(\bigcup_{i=1}^n E_i) \leq \sigma_1 - \sigma_2 + \cdots + (-1)^{k+1}\sigma_k, \text{ when } k \text{ is odd} \tag{3.6}$$

$$\mathbb{P}(\bigcup_{i=1}^n E_i) \geq \sigma_1 - \sigma_2 + \cdots + (-1)^{k+1}\sigma_k, \text{ when } k \text{ is even} \tag{3.7}$$

We now apply inequality (3.6) for the case $k = 3$ to get

$$\mathbb{HP}_n \le \Sigma_1 - \Sigma_2 + \Sigma_3 \tag{3.8}$$

where Σ_1, Σ_2 have been defined in (3.5), and

$$\Sigma_3 \triangleq \sum_{i \neq j \neq k} \mathbb{P}(A_i A_j A_k). \tag{3.9}$$

After some derivation (see Appendix A), we will finally have

$$\begin{aligned}\Sigma_3 = \frac{1}{6}&(n-3L+1)(n-3L+2)(n-3L+3)p^{3w} \\ &+ \sum_{i=1}^{L-1}(n-2L-i+1)(n-2L-i+2)p^{3w-\theta(i)} \\ &+ \sum_{i=1}^{L-1}\sum_{j=L-i}^{L-1}(n-i-j-L+1)p^{3w-\theta(i)-\theta(j)} \\ &+ \sum_{i=1}^{L-1}\sum_{j=1}^{L-i-1}(n-i-j-L+1)p^{3w-\theta(i)-\theta(j)-\theta(i+j)+\theta(i,j)}\end{aligned} \tag{3.10}$$

Figure 3.5 (on page 30) shows the scatter plot of $\mathbb{HP}(\mathcal{Q})$ with the Bonferroni upper bounds in (3.8). From the plots in this figure that for small p like 0.5 and large p close to 1, the Bonferroni upper bound predict fairly well for $\mathbb{HP}_n$. However, for moderate p, it performs relatively bad because there seem to be a transition period at these p for the correlation between $\mathbb{HP}_n$ and the Bonferroni upper bound from positively proportional to negatively proportional. Further inspection of the numerical values of the Bonferroni upper bound for $p = 0.5$ shows that the Bonferroni upper bound is very close to the real $\mathbb{HP}_n$. This upper bound, in general, is much closer to $\mathbb{HP}_n$ than the two lower bounds. Indeed, this upper bound becomes closer and closer to the real $\mathbb{HP}_n$ as L gets larger and larger. So for longer spaced seeds, we may use the Bonferroni upper bound as a fairly good estimation of $\mathbb{HP}_n$ when p is small.

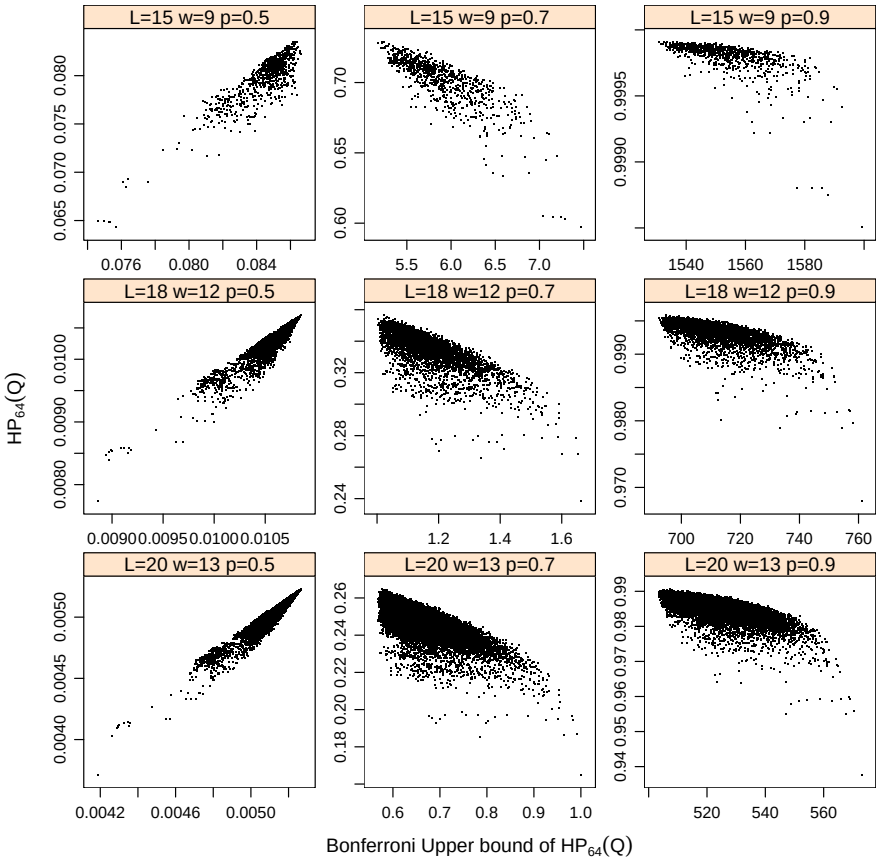

Figure 3.5 *Plots of $\mathbb{HP}_n(\mathcal{Q})$ vs its Bonferroni upper bound of $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ and $\mathcal{Q}_{20,13}$ (rows from top to bottom) for $p = 0.5, 0.7, 0.9$ (columns from left to right).*

3.3 Compare the predictability of the above predictors

3.3.1 Discussion on the predictors

From the above figures (Figures 3.1 – 3.5), we may conclude that, among the above indicators, $\mathbb{HP}_{2L-1}$ has the best correlation with $\mathbb{HP}_n$, followed by the Cauchy-Schwartz

lower bound and Bonferroni-type lower bound. Further we find:

(1) Although $\mathbb{HP}_{2L-1}$ is the best predictor of $\mathbb{HP}_n$, it does not constitute a practical predictor of $\mathbb{HP}_n$, because the calculation of $\mathbb{HP}_{2L-1}$ itself is also very time-consuming. From section 2.2, we know that the exact calculation has the complexity of $O(Ln2^{2(L-w)})$, so the time to compute $\mathbb{HP}_{2L-1}$ is only about L/n times that of $\mathbb{HP}_n$. Generally, L/n is between 1/3 and 1/7, so there is no essential improvement on the computing time. So the predictor $\mathbb{HP}_{2L-1}$ may be used to predict effective spaced seeds of moderate length, not practical for very long spaced seeds.

(2) The Cauchy-Schwartz and the Bonferroni-type lower bound perform very similar. Of course, this is not a coincidence. This is because the two lower bound are both based on the quantity of Σ_2 which defined in (3.5).Typically, Σ_2 is generally much greater than Σ_1. Generally we have the ratio $\frac{\Sigma_2}{\Sigma_1} \geq 5$, and it increases with L. We observe that for L larger than 20, the ratio $\frac{\Sigma_2}{\Sigma_1}$ is greater than 100. Hence $k = \left\lfloor \frac{\Sigma_2}{\Sigma_1} \right\rfloor + 2 \approx \frac{\Sigma_2}{\Sigma_1} + 2$, then the Bonferroni lower bound

$$\text{BLB} = \frac{2\Sigma_1}{k} - \frac{\Sigma_2}{k(k-1)} \approx \frac{\Sigma_1^2}{\Sigma_1 + \Sigma_2} = \text{CSLB}.$$

Therefore we can show that in their value range, the two types of bounds are approximately equal.

We may simply use Σ_2 (which will be negatively correlated with $\mathbb{HP}_n$) for indicator of good spaced seeds instead of this two lower bounds.

(3) From their correlation with $\mathbb{HP}_n$, the Bonferroni upper bounds seem to perform worse than the two lower bounds. However, we may observe from Figure 3.5 (on page 30) the better and better performance of the Bonferroni upper

bound for small p as L gets larger and larger . If this holds as true, then the Bonferroni upper bound will be a very good predictor, even better Σ_2, for long spaced seeds when p is small. Table 3.1 (on page 33) and Table 3.2 (on page 34) verifies this conjecture. $\Sigma_2 - \Sigma_3$, which is equivalent to the bonferroni upper bound, really predicts well for large L.

We conclude that for moderate L and w, $\mathbb{HP}_{2L-1}$ is the best indicator, however, for large L, we recommend Σ_2 and/or $\Sigma_2 - \Sigma_3$ as predictors.

3.3.2 Further comparison of the predictability of Σ_2 and $\Sigma_2 - \Sigma_3$

We notice that if a spaced seed performs well for one p, it will also perform well (in general) for other p (this may be seen in Table 3.1 (on page 33)). Hence we may just predict the top spaced seeds for one particular p and give one set of top seeds for all p. So we can simply choose a p value under which the predictors has the best predictability. It seems that using smaller p is the wise choice. So in the following tables of predicted best seeds, we only give one set of best seeds which are predicted by the indicators under $p = 0.5$.

Table 3.1 (on page 33) shows the predicted top 10 spaced seed of $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ and $\mathcal{Q}_{20,13}$ together with their rank under $\mathbb{HP}_{64}$. We observe that $\Sigma_2 - \Sigma_3$ predicts better and better as L gets larger and larger, and it outperforms Σ_2 for the case of $\mathcal{Q}_{20,13}$.

Table 3.2 (on page 34) lists the top 10 spaced seeds predicted by $\Sigma_2 - \Sigma_3$ and Σ_2 for $\mathcal{Q}_{23,15}, \mathcal{Q}_{24,16}, \mathcal{Q}_{29,17}$ and $\mathcal{Q}_{33,20}$ (we choose L and w value with the ratio w/L staying around $p = 0.7$). In this figure, we also give the relative ranks (based on $\mathbb{HP}_{128}(Q)$)

Table 3.1 *Predicted top 10 seeds of $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}, \mathcal{Q}_{20,13}$*

$\mathcal{Q}_{L,w}$	$\Sigma_2-\Sigma_3$	0.5	0.7	0.9	Σ_2	0.5	0.7	0.9
$\mathcal{Q}_{15,9}$	1**111**1*111*1	65	55	30	111**1**1*1*111	2	3	2
	11**1**1111*1*1	56	42	17	11*11**1*1**111	4	4	4
	1**1*1*11**1111	57	46	21	111*1***11*1*11	3	2	3
	1**11**1*1*1111	46	25	9	111***1*1*11*11	1	1	1
	1***11*11*1*111	156	151	112	111**1**11*1*11	6	7	15
	1*1*1**11**1111	13	9	7	11**11*1**1*111	7	8	20
	1***111111*1**1	700	688	623	11**1*1*1**1111	5	5	5
	111***1*1*11*11	1	1	1	111*1**1**1*111	9	11	22
	1*1*11*11***111	25	16	8	111**1*1**1*111	11	13	27
	11*1***111*11*1	44	38	38	11*11***1*1*111	12	14	23
$\mathcal{Q}_{18,12}$	111*1*11*1**11*111	1	1	1	111*1*11*1**11*111	1	1	1
	111*1**11*1*11*111	2	2	2	111*1**11*1*11*111	2	2	2
	11*11*1*1*11**1111	3	3	9	11*11*1*1*11**1111	3	3	9
	111*1*1**111*11*11	6	5	3	111*1*1**11*11*111	4	4	7
	11*1*111*1**111*11	15	14	4	111**11*11*1*1*111	5	9	15
	111**11*11*1*1*111	5	9	15	1111**11**1*1*1111	12	6	18
	11*1*1*11**11*1111	8	8	6	111*1*11**1*11*111	7	12	28
	111*1*1**11*11*111	4	4	7	111*11**1*1*11*111	9	21	73
	1*111**11*11*1*111	32	66	39	111**11*1*1*11*111	11	22	31
	11*111*1*11*1**111	17	27	19	11*111**1*11*1*111	10	15	21
$\mathcal{Q}_{20,13}$	111*1*11**11**1*1111	1	1	2	111*1*11**11**1*1111	1	1	2
	1111*1*1**11*11**111	2	3	4	1111*1*1**11*11**111	2	3	4
	111*1**11*1**111*111	3	2	1	111*11**1*11*1*1*111	5	15	30
	111*11**1*1*11**1111	4	5	9	111*11**1*1*11**1111	4	5	9
	111*11**1*11*1*1*111	5	15	30	1111**11**1*1*11*111	6	16	39
	1111*1*1**111**11*11	10	11	6	111*11*1**11*1*1*111	7	9	12
	11*1*111**11**1*1111	16	11	18	111*1*11**1*1*11*111	9	19	37
	111*11**1*1*11*1*111	8	10	17	111*11**1*1*11*1*111	8	10	17
	111*11*1**11*1*1*111	7	9	12	111*1**11*1**111*111	3	2	1
	111*1**1*11**111*111	11	13	5	1111**1*1*11**11*111	13	24	49

The first column of spaced seeds is predicted by $\Sigma_2-\Sigma_3$ at $p=0.5$ (same as the Bonferroni upper bound). The second column of spaced seeds is predicted by Σ_2 at $p=0.5$ (same as the Cauchy-Schwartz lower bound and approximately same as the the Bonferroni-type lower bound). The columns following the seeds are the rank under their hitting probabilities at $n=64, \mathbb{HP}_{64}(Q)$.

among these seeds, and we calculate sums of these ranks. In most cases, the top 10 seeds predicted by the two quantities overlap substantially. Most seeds are in both predicted top 10 seeds. The ✗ signs mark the seeds not belonging to $\Sigma_2-\Sigma_3$ (column A) or Σ_2 (column B). Because we rank the best seeds by 1, the predictor is better with small rank sums. We can see in most cases, the rank sums before the slashes, which are the rank sums corresponding to $\Sigma_2-\Sigma_3$, are smaller, so in the four $\mathcal{Q}$, $\Sigma_2-\Sigma_3$ turns out to be the better predictor.

Table 3.2 *Predicted top 10 seeds of $\mathcal{Q}_{23,15}, \mathcal{Q}_{24,16}, \mathcal{Q}_{29,17}, \mathcal{Q}_{33,20}, \mathcal{Q}_{35,22}$*

$\mathcal{Q}_{L,w}$	No	Seed	A	B	$p=0.5$	$p=0.7$	$p=0.9$
$\mathcal{Q}_{23,15}$	1	1111**1*1*1*11**11*1111			1	1	2
	2	111*1**11*1*1**111*1111		✗	5	5	9
	3	111*11**11*1**11*1*1111			3	3	4
	4	111*11**1*1**111*1*1111			7	7	6
	5	1111**11*1*1*1**11*1111			2	2	1
	6	1111*1*1**11**11*1*1111			4	4	3
	7	1111**11**1*1*1*11*1111			6	6	11
	8	11*11*11***111*1*1*1111		✗	13	13	10
	9	111*111**1*11**1*1*1111		✗	9	9	12
	10	111*1*11*1**11**11*1111			8	8	5
	11	1111**1*1*11**11*1*1111	✗		10	10	8
	12	111*1*1*11*1**11**11111	✗		11	11	13
	13	1111*1**11**11*1*1*1111	✗		12	12	7
					58/64	58/64	63/60
$\mathcal{Q}_{24,16}$	1	1111**11*1*1*11**11*1111			1	1	1
	2	1111*1*11**11**11*1*1111			2	3	5
	3	111*1*11*1*11**11**11111			3	2	2
	4	111*11*11***111*1*11*111		✗	5	5	3
	5	111*11*11***111*11*1*111		✗	10	7	9
	6	111*11*1**11**111*1*1111			4	4	4
	7	111*11*11**111*1*1**1111			9	11	7
	8	111*11*11**1*1*111**1111			7	8	11
	9	111*11*11**111**1*1*1111			8	9	8
	10	1111**11**11*1*1*11*1111			6	6	6
	11	1111**11**11*11*1*1*1111	✗		12	12	10
	12	111*111**1*1*11*11**1111	✗		11	10	12
					55/63	56/66	56/66
$\mathcal{Q}_{29,17}$	1	1111**1*1**11**11**1*1*1*1111			1	1	2
	2	111*11***11*1*1*1**11*1**1111			2	9	5
	3	111*11***11**1*11**1*1*1*1111			4	2	4
	4	1111*1*1*1**1**1*11**11**1111			5	7	9
	5	111*11*1*1***11**1*11**1*1111			6	5	8
	6	1111**11**1*1*1**1*11**1*1111			3	3	1
	7	1111**1*1**11**1**11*1*1*1111			7	6	7
	8	1111**11*1***1*11*1*1**11*111			9	8	6
	9	1111**1*1*1*1**11**11**1*1111			8	4	3
	10	1111**11*1*1**1**11**1*1*1111			10	10	10
					55/55	55/55	55/55
$\mathcal{Q}_{33,20}$	1	1111*1**111**11**1*11**1*1*1*1111			1	4	6
	2	1111*1*1**11*1*1**11**1*11**11111			2	2	5
	3	1111**11*1*1**11**11*1**1*1*11111			3	1	2
	4	1111*1**11**111**1*1*1*11**1*1111			4	7	4
	5	1111*1*1**111**1**11*1**11*1*1111			5	3	3
	6	1111**11**111*1**1*11**1*1*1*1111			6	9	9
	7	1111*1*1*11**11**1*11**1*1**11111			7	5	1
	8	1111**11*11***11*1*11**1*1*1*1111			8	8	7
	9	1111*1*1*1**11*1**1*11**11**11111			9	6	10
	10	1111*1**11*1**11**111**1*1*1*1111			10	10	8
					55/55	55/55	55/55
$\mathcal{Q}_{35,22}$	1	1111*1*11**1*11*1*1*1*11**11**11111			1	3	3
	2	1111*1*1*11**111**1*11**11**1*11111			2	6	11
	3	1111*1**11*1*1*11*1**11**111**11111			3	1	2
	4	1111*1*1*111***11*11**11**1*11*1111			4	4	6
	5	1111**111*1*1*1*11**11**11*1**1*11111			5	7	5
	6	1111*1*1*111**11**1*11*1**11**11111			7	8	10
	7	1111**111**1*1*11**11*11**1*1*11111		✗	8	9	7
	8	1111*1*1**11*1**11**11**1*111**11111			6	2	1
	9	1111*1*1*1**11*11*1**111**11**11111			9	11	9
	10	1111*1*1*111**11**11*1**1*11**11111			10	5	4
	11	1111*1*11*1**11*1*1*1*11**11**11111	✗		11	10	8
					55/58	56/57	58/57

The seeds for each $\mathcal{Q}$ are the union of top 10 seeds predicted by $\Sigma_2 - \Sigma_3$ and Σ_2. The ✗signs in column A *and* B *mark out the seeds not belonging to the top 10 of $\Sigma_2 - \Sigma_3(A)$ and $\Sigma_2(B)$ respectively. The column of $p = 0.5$ is the internal ranks (by $\mathbb{HP}_{128}(Q)$) of the seeds among themselves, and the last row in each $\mathcal{Q}$ is the sums of the rank of the top 10 seeds predicted by $\Sigma_2 - \Sigma_3$ and Σ_2. Similar for others.*

However, since the predicted top seeds of $\Sigma_2 - \Sigma_3$ and Σ_2 overlap so much, and the sums of the ranks becomes closer as L increases, we expect the two predictors performs almost the same for larger L. Since the calculation of Σ_2 is faster than $\Sigma_2 - \Sigma_3$, **we recommend to use Σ_2 for larger L.**

CHAPTER 4

Features for Good Spaced Seeds

Generally, the calculation time of Σ_2 or $\Sigma_2 - \Sigma_3$ is much less than that of the $\mathbb{HP}_n$ using the recursive relation given in Theorem 2.1. For example, we just take several minutes to run the computation of both the the indicators of Σ_2 and $\Sigma_2 - \Sigma_3$ of all the spaced seeds of $\mathscr{Q}_{23,15}$ for one p, but the exact calculation of the hitting probability may take about one day.

However, the number of seeds will increase very rapidly. We have the following lemma on the number of spaced seeds.

Table 4.1 *Number of spaced seeds in* $\mathcal{Q}$

$\mathcal{Q}_{L,w}$	$\left\|\mathcal{Q}_{L,w}\right\|$	$\mathcal{Q}_{L,w}$	$\left\|\mathcal{Q}_{L,w}\right\|$
$\mathcal{Q}_{15,9}$	868	$\mathcal{Q}_{24,16}$	160,050
$\mathcal{Q}_{18,12}$	4,032	$\mathcal{Q}_{29,17}$	8,692,788
$\mathcal{Q}_{20,13}$	15,912	$\mathcal{Q}_{33,20}$	103,129,040
$\mathcal{Q}_{23,15}$	101,850	$\mathcal{Q}_{35,22}$	286,587,224

Lemma 4.1 *Let* $\left\|\mathcal{Q}_{L,w}\right\|$ *be the number of spaced seeds belonging to* $\mathcal{Q}_{L,w}$*, then we have*

$$\left\|\mathcal{Q}_{L,w}\right\| = \begin{cases} \frac{1}{2}\binom{L-2}{w-2} & L \text{ is even, } w \text{ is odd;} \\ \frac{1}{2}\left[\binom{L-2}{w-2} + \binom{\lfloor \frac{L}{2} \rfloor - 1}{\lfloor \frac{w}{2} \rfloor - 1}\right] & \text{otherwise.} \end{cases} \tag{4.1}$$

Proof: Recall that $\mathcal{Q}_{L,w}$ is the collection of all spaced seeds with heavy tail. When L is even and w is odd, there does not exist any symmetric seeds. Among all the $\binom{L-2}{w-2}$ candidate spaced seeds, exactly one half will be discarded.

In other cases, the weight can be evenly divided into the two halves. In this case, we cannot discard those Q with its reverse being the same as Q itself, namely, those Q being symmetric about its center. The number of these symmetric Q is just

$$\binom{\lfloor \frac{L}{2} \rfloor - 1}{\lfloor \frac{w}{2} \rfloor - 1}.$$

So the total number of spaced seeds now is as (4.1). ■

Table 4.1 (on page 37) shows the number of seeds in $\mathcal{Q}_{L,w}$ that appears in the previous chapters.

Thus, even though the computation of the indicators for a single spaced seeds is very fast, it will become very time-consuming to compute them for all spaced seeds

in $\mathcal{Q}_{L,w}$ for very large L. For example, the computation of both Σ_2 and $\Sigma_2 - \Sigma_3$ of $\mathcal{Q}_{35,22}$ takes less than one day. Therefore, we need to find even simpler index to reduce the magnitude of the total number of spaced seeds falling into our consideration.

4.1 Number of blocks of $*$'s in Q

From Table 2.1 (on page 15), we notice that the number of blocks (runs) of $*$'s in the best spaced seeds are about the same. For example, for the case that $\mathcal{Q}_{15,9}$, all the best spaced seeds listed have 4 blocks of $*$'s despite of what value p takes. We also checked the top 100 seeds of $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}, \mathcal{Q}_{20,13}$ for $p = 0.5, 0.7, 0.9$, and found that the number of blocks of $*$'s really remains very robust with variation less than 2.

If we use b to denote the number of blocks of $*$'s in Q, then we can show that

$$b = w - \theta(1) - 1,$$

where w is the weight of Q, and $\theta(1)$ is the first order self-overlapping index at 1 as defined in (3.2) (on page 23). Recall that $\theta(1)$ is the common number of 1's in Q and $Q \gg 1$. In order to a common 1 occur in position k of Q, it is necessary and sufficient to having 1's occur in positions $k-1$ and k. Thus, $\theta(1)$ is just the number of 11's in Q, which equals the total number of 1's minus the number of $1*$'s, then minus 1 (for the last 1 of Q). Apparently, the number of $1*$'s is just b. So b, which is equivalent to $\theta(1)$ is also a measure of self overlapping, and it is the simplest one.

Figure 4.1 (on page 39) shows the box-plots of $\mathbb{HP}_{64}$ vs b. We can see clearly that the distribution of $\mathbb{HP}_{64}$ is very different for different b values. For example, for $\mathcal{Q}_{15,9}$

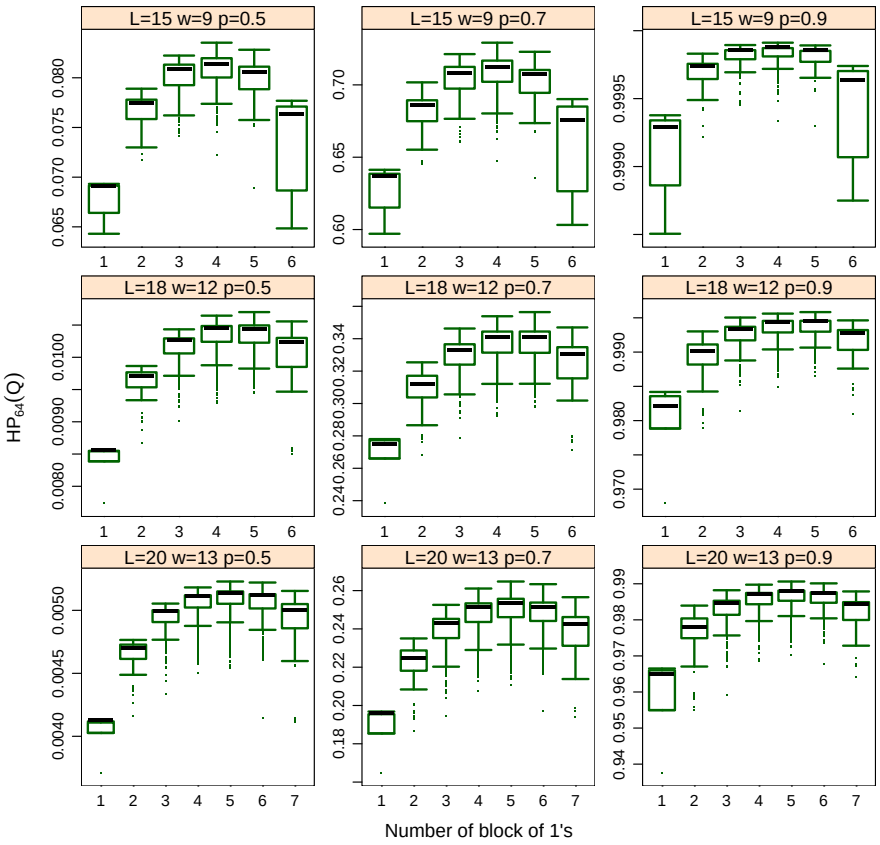

Figure 4.1 *Box-plots of $\mathbb{HP}_n(Q)$ vs b, the number of block of $*$'s of $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ and $\mathcal{Q}_{20,13}$ (rows from top to bottom) for different $p = 0.5, 0.7, 0.9$ (columns from left to right).*

the spaced seeds with $b = 4$ as a whole have the highest hitting probabilities. From the figure, we can observe that the spaced seeds with very small number (like 1,2) of blocks of $*$'s always have lower hitting probabilities.

Table 4.2 (on page 40) lists the optimal b values for some $\mathcal{Q}_{L,w}$. For the cases of $\mathcal{Q}_{23,15}, \mathcal{Q}_{24,16}, \mathcal{Q}_{29,17}, \mathcal{Q}_{33,20}, \mathcal{Q}_{35,22}$, we simply use the predictor, $\Sigma_2 - \Sigma_3$, instead of the exact hitting probability of the seeds. But we conjecture that the optimal spaced seeds have the same b values as given in the table.

Table 4.2 *Optimal b values of different* $\mathcal{Q}_{L,w}$

L	14	15	15	16	16	17	15	17	18	18	20	20	23	24	29	33	35
w	11	12	11	12	11	12	9	11	12	11	13	12	15	16	17	20	22
b	3	3	4	4	4	4	4	4	5	5	5	5	6	6	8	9	10
$\hat{b}$	3.1	3.4	3.5	3.8	3.9	4.2	3.9	4.3	4.6	4.7	5.2	5.3	6.0	6.3	8.1	9.2	9.6

The top spaced seeds for $\mathcal{Q}_{23,15}, \mathcal{Q}_{24,16}, \mathcal{Q}_{29,17}, \mathcal{Q}_{33,20}, \mathcal{Q}_{35,22}$ *are predicted by* $\Sigma_2 - \Sigma_3$.

We may observe that, at least in our range, b has a strong linear relation with L, w. The regression line is

$$b = -0.578 + 0.397L - 0.168w.$$

In Table 4.2 (on page 40), the row $\hat{b}$ records the fitted value of b using the above regression line (with one efficient digit). We excitedly find that the fit is very good. So in practice, we may simply use the rounded integer of the above value to filter out the rest of spaced seeds.

4.2 Weight difference of two halves of Q

Another observation about the features of the best spaced seeds given in Table 2.1 (on page 15), we find that the 1's distribute very evenly in the seeds. Then we can expect good spaced seeds to be balanced, so the weight difference between the right half and the left half cannot exceed some number. We use Δw to denote the difference (always be positive) of the number of 1's in the left half and the right half of a spaced seed Q. For example, if $Q =$ `11*11*|1111*1`, then there are four 1's in the left six positions, and five 1's in the right six positions, so $\Delta w = 1$; if $Q =$ `11*11*`1`1111*1`,

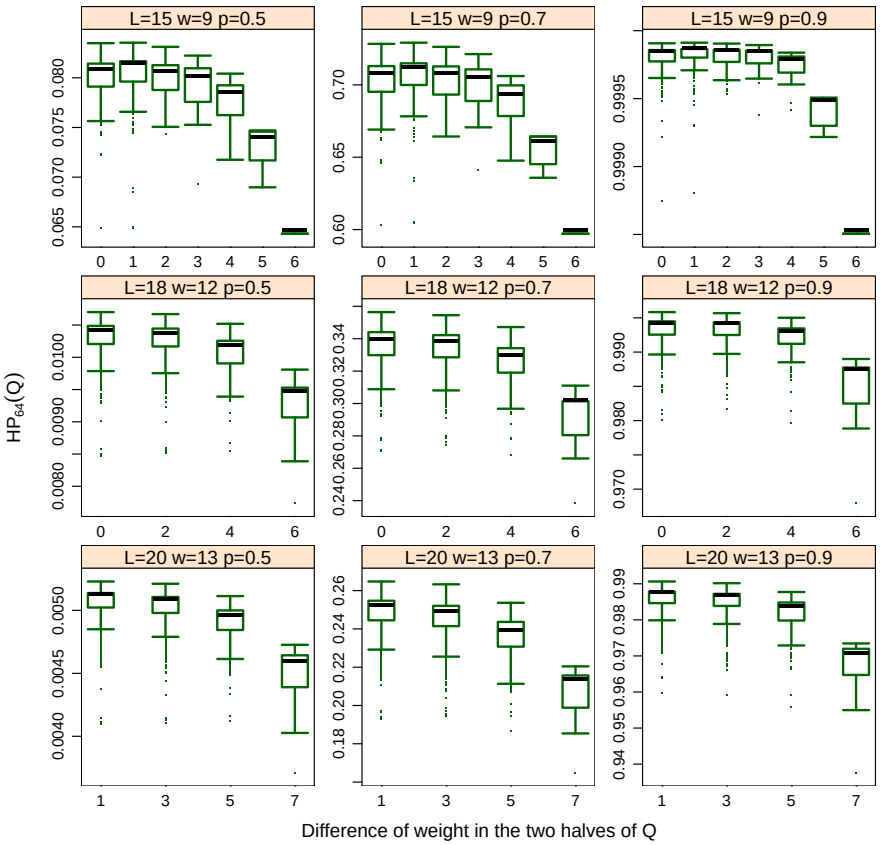

Figure 4.2 *Box-plots of $\mathbb{HP}_n(Q)$ vs Δw, the difference of weight of the two halves of a spaced seed Q, of $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ and $\mathcal{Q}_{20,13}$ (rows from top to bottom) for different $p = 0.5, 0.7, 0.9$ (columns from left to right).*

then there are four 1's in the left six positions[1], and five 1's in the right six positions, so $\Delta w = 1$.

To see the distribution of the hitting probabilities for different Δw, we refer to the box-plots in Figure 4.2 (on page 41). We observe clearly that the hitting probabilities with small Δw values are generally larger. In this figure, the highest hitting probabilities occur only when Δw is 0 or 1. If we refer back to Table 3.2 (on page 34), we may count the Δw for the predicted spaced seed as listed in Table 4.3 (on page 42). In this

1. Easy to see that it is equivalent to define the left half to be position 1 to $\lfloor \frac{L}{2} \rfloor$ or 1 to $\lceil \frac{L}{2} \rceil$.

Table 4.3 *Δw of the predicted top 10 spaced seeds*

$\mathcal{Q}_{L,w}$	1	2	3	4	5	6	7	8	9	10
$\mathcal{Q}_{23,15}$	1	1	0	1	0	0	1	2	0	1
	1	0	0	0	0	1	0	1	1	1
$\mathcal{Q}_{24,16}$	0	0	0	2	2	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
$\mathcal{Q}_{29,17}$	1	0	1	1	0	0	1	1	1	0
	1	0	0	1	0	1	1	0	1	1
$\mathcal{Q}_{33,20}$	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
$\mathcal{Q}_{35,22}$	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0

For each $\mathcal{Q}_{L,w}$, the first row is the Δw value of the top 10 spaced seeds predicted by $\Sigma_2 - \Sigma_3$, the second row is predicted by Σ_2. The column with title i is the i-th predicted top

table, the Δw value are all less than or equal to 2, and most of them are 0's.

Therefore, we conjecture that the value of Δw of the best spaced seed is a small integer, generally no more than 2. Further, we may imply from Table 4.3 that if either one of L and w is even, then we strongly prefer the seeds with $\Delta w = 0$, at most we give some consideration for those with $\Delta w = 2$; if the both L and w are odd, then we prefer the seeds with $\Delta w = 0$ or 1.

4.3 Number of 1's in head and tail of Q

When studying the best seeds given in Table 2.1 (on page 15) and the predicted best seeds given in Table 3.1 (on page 33) and Table 3.2 (on page 34), we may notice that the number of 1's in the front of Q and in the tail of Q also remain stable with very small variation, and the number of 1's in head and tail covers a large part of the total

number of 1's. This is a common phenomenon in all $\mathcal{Q}_{L,w}$ we have examined.

We let h_Q denote the number of consecutive 1's in the head of Q, and t_Q denote the number of consecutive 1's in the tail. For example, if $Q = 111 * 11 * 11$, then $h_Q = 3$, $t_Q = 2$ as there are three 1's in the first block of 1's and two 1's in the last block of 1's.

We are often interested in the total weight $h + t$ in the head and tail and the difference of weight $|h - t|$ in the head and tail. Since these two quantities are the same for Q and the reverse of Q, therefore, it does not matter whether we choose Q or its reverse.

Figure 4.3 (on page 44) and Figure 4.4 (on page 45) show us the box plots of $\mathbb{HP}_n$ to the two indices. From these two figures, we see that the hitting probabilities do vary with different $h+t$ or $|h-t|$ values, and we can roughly see the optimal values of $h+t$ occur in the middle of its range, and the optimal value of $|h-t|$ occur at the lower end in its range.

Table 4.4 (on page 46) shows us the $h+t$ and $|h-t|$ values of the top seeds or the predicted top seeds for some $\mathcal{Q}_{L,w}$. Similar to the results of Δw, we find that the $|h-t|$ value of the good spaced seeds is always a small integer less than (or sometimes equal to) 2.

We can find $h+t$ has a good linear relation with L and w. The estimated regression line is

$$h + t = 3.001 - 0.156L + 0.533w. \tag{4.2}$$

The $\widehat{h+t}$ values (with one efficient digit) in Table 4.4 (on page 46) are the fitted values of the above regression line. In many cases, the estimation is fairly good. Generally,

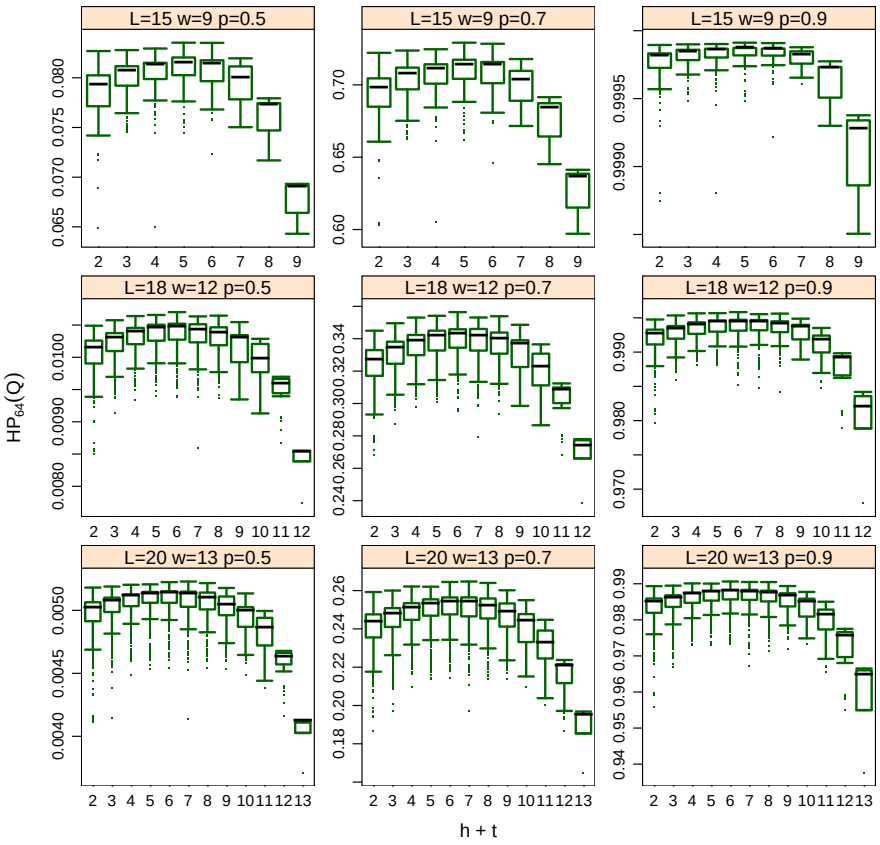

Figure 4.3 *Box-plots of $\mathbb{HP}_n(Q)$ vs $h+t$, the number of 1's in the head and tail of a spaced seed Q of $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ and $\mathcal{Q}_{20,13}$ (rows from top to bottom) for $p = 0.5, 0.7, 0.9$ (columns from left to right).*

the optimal $h+t$ values are the floor or ceiling integer of the fitted value by L and w.

In summary, we may prefer the seeds with $|h-t|$ value less than 2. For $h+t$, we can infer according to the formula (4.2) and allow a variation less than 2.

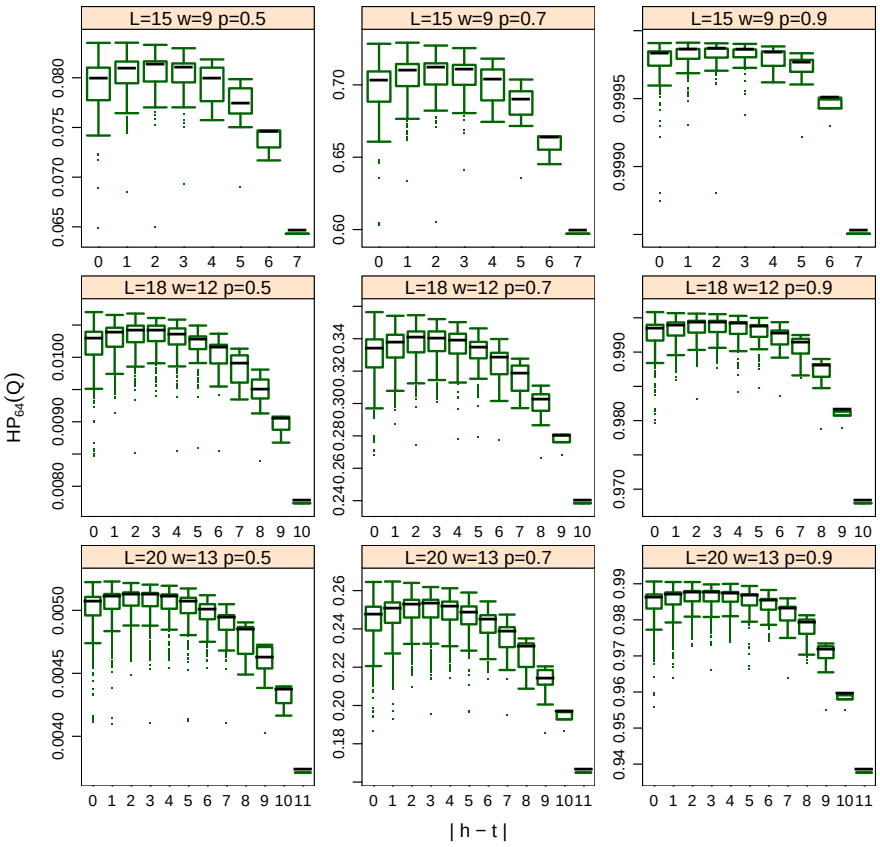

Figure 4.4 *Box-plots of* $\mathbb{HP}_n(Q)$ *vs* $|h-t|$*, the difference of the number of 1's in the head and in the tail of a spaced seed* Q *of* $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ *and* $\mathcal{Q}_{20,13}$ *(rows from top to bottom) for* $p = 0.5, 0.7, 0.9$ *(columns from left to right).*

4.4 Maximal length of the blocks of 1's and ∗'s

Besides the filters we discussed above, there are other possible good filters, for example, the maximal length of runs of ∗'s or 1's (except the first and last runs of 1's). If we let $z_{\max}$ and $u_{\max}$ denote the maximal length of the runs of ∗'s and 1's(except the first and the last run), then generally we have $z_{\max}, u_{\max} = 2$ or 3, which can be seen in Table 4.5 (on page 46).

Table 4.4 *$h+t$ and $|h-t|$ of the top spaced seeds*

L	14	15	15	16	16	17	15	17	18	18	20	20	23	24	29	33	35
w	11	12	11	12	11	12	9	11	12	11	13	12	15	16	17	20	22
$h+t$	8	7	6	6	6	7	5	6	6	6	7	7	8	8	8	8	9
$\widehat{h+t}$	6.7	7.0	6.5	6.9	6.4	6.7	5.5	6.2	6.6	6.0	6.8	6.3	7.4	7.8	7.5	8.5	9.2
$\|h-t\|$	2	1	0	2	0	1	1	0	0	0	1	1	0	0	0	0	1

The top spaced seeds for $\mathcal{Q}_{23,15}, \mathcal{Q}_{24,16}, \mathcal{Q}_{29,17}, \mathcal{Q}_{33,20}, \mathcal{Q}_{35,22}$ are predicted by $\Sigma_2 - \Sigma_3$.

Table 4.5 *Optimal $z_{\max}$ and $u_{\max}$ values*

$\mathcal{Q}$	$\mathcal{Q}_{15,9}$	$\mathcal{Q}_{18,12}$	$\mathcal{Q}_{20,13}$	$\mathcal{Q}_{23,15}$	$\mathcal{Q}_{24,16}$	$\mathcal{Q}_{29,17}$	$\mathcal{Q}_{33,20}$	$\mathcal{Q}_{35,22}$
$z_{\max}$	3	2	2	2	2	2	2	2
$u_{\max}$	2	2	2	2	2	2	3	2

Optimal $z_{\max}$ and $u_{\max}$ values for $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}, \mathcal{Q}_{20,13}$ are values of the optimal seeds. The other $z_{\max}$ and $u_{\max}$ values are values of the predicted top seeds by Σ_2.

Figure 4.5 (on page 47) and Figure 4.6 (on page 48) show the box plot of $\mathbb{HP}_{64}$ vs these two filters. From this figure, we can see clearly that 2 or 3 are optimal values for $z_{\max}$ and $u_{\max}$.

4.5 Separability and filterability of seeds filters

To measure the goodness of a seeds filter, we may use two index: separability and filterability.

- Separability measures the capability of a filter to separate the seed according to their hitting probabilities. Higher separability indicates that, as a whole,the seeds with good filter value(s) have higher hitting probabilities than seeds with

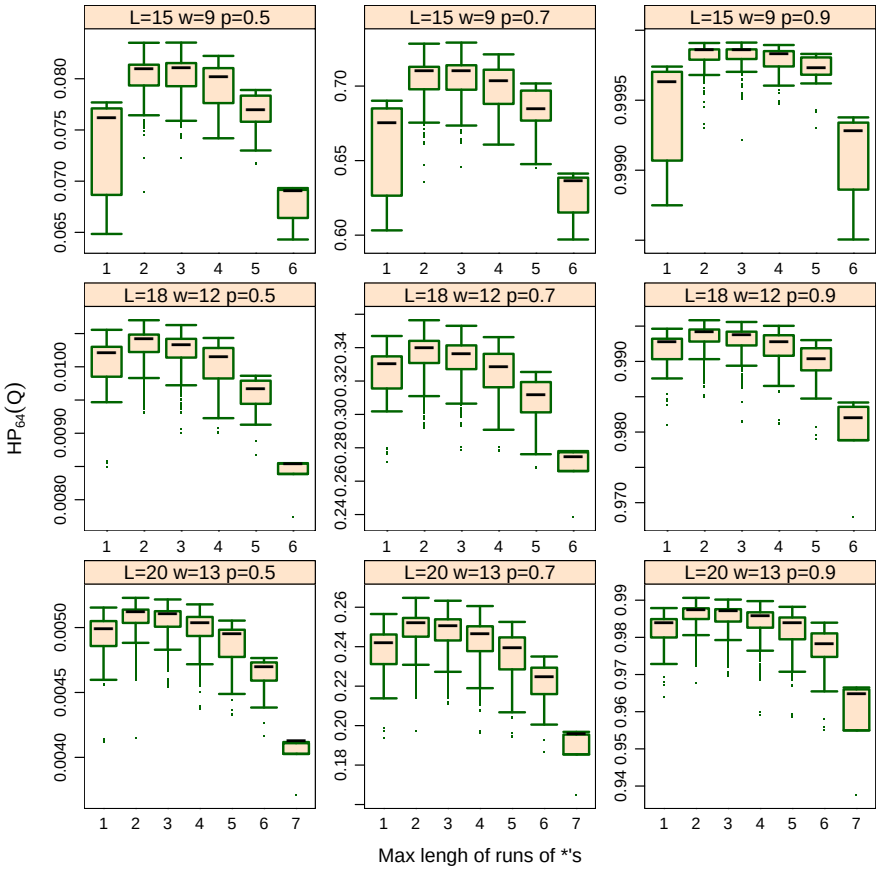

Figure 4.5 *Box-plots of* $\mathbb{HP}_n(Q)$ *vs* $z_{\max}$ *(left) and* $u_{\max}$ *(right) of* $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ *and* $\mathcal{Q}_{20,13}$ *(rows from top to bottom) for different* $p = 0.5, 0.7, 0.9$ *(columns from left to right).*

bad filter values. The ideal filter partitions $\mathbb{HP}_n$ into several categories with the hitting probability of one category always no less than that of the other categories. Using the ideal filter, we can totally reduce the seeds to those with the optimal filter value. However, it is not clear whether such a filter exists. None of the filters we have proposed are ideal in this sense.

- Filterability refers to the filtration ability of a filter, i.e., the proportion of seeds that are filtered out by using the filter. Obviously, the higher the proportion is,

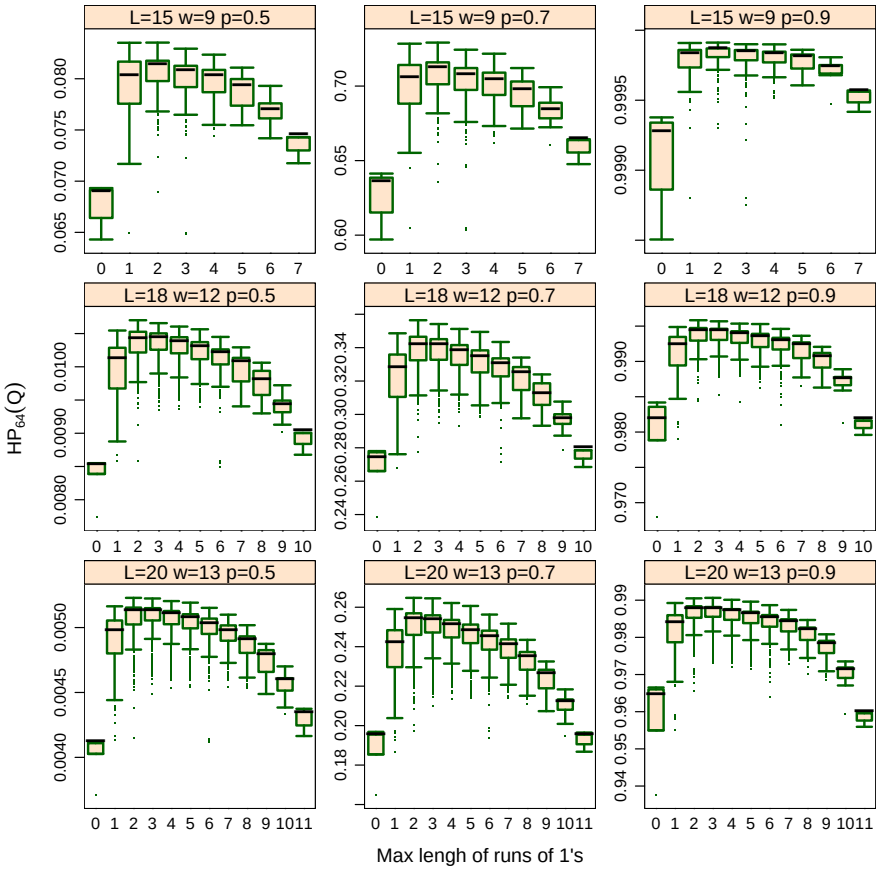

Figure 4.6 *Box-plots of* $\mathbb{HP}_n(Q)$ *vs* $u_{\max}$ *of* $\mathscr{Q}_{15,9}, \mathscr{Q}_{18,12}$ *and* $\mathscr{Q}_{20,13}$ *(rows from top to bottom) for different* $p = 0.5, 0.7, 0.9$ *(columns from left to right).*

the more efficient the filter.

In summary, the more selective, the less seeds left to handle; the more sensitive, the higher probability the remaining seeds have. Of course, we hope to have a filter with both high separability and filterability, but it is not an easy problem.

From Figure 4.1 (on page 39) to Figure 4.4 (on page 45), we may find the separabilities of the filter we proposed, including $b, \Delta w, |h+t|$ and $|h-t|$, are about the same level. Generally speaking, we can only exclude the seeds with the very bad filter

values.

On the filterability, we have the following lemma (for proof, see Appendix B).

Lemma 4.2 *For all the spaced seeds with length L and weight w in $\mathcal{Q}_{L,w}$, we have*

(1) the number of seeds with $b(1 \le b \le L-w)$ blocks of $*$'s is

$$\begin{cases} \frac{1}{2}\binom{w-1}{b}\binom{L-w-1}{b-1}, & \text{if } \begin{cases} L \text{ even, } w \text{ odd, or} \\ L \text{ odd, } w+b \text{ even} \end{cases} \\ \frac{1}{2}\left[\binom{w-1}{b}\binom{L-w-1}{b-1} + \binom{\lceil \frac{w}{2} \rceil - 1}{\lfloor \frac{b}{2} \rfloor}\binom{\lfloor \frac{L-w}{2} \rfloor - 1}{\lceil \frac{b}{2} \rceil - 1}\right], & \text{otherwise.} \end{cases}$$

(2) the number of seeds with $\Delta w(0 \le \Delta w \le L-2)$ is

$$\begin{cases} 0, & \text{if } L \text{ even and } w+\Delta w \text{ odd,} \\ \binom{\lfloor \frac{L}{2} \rfloor - 1}{\lceil \frac{w+\Delta w-1}{2} \rceil - 1}\binom{\lfloor \frac{L}{2} \rfloor - 1}{\lceil \frac{w-\Delta w-1}{2} \rceil - 1}, & \text{otherwise and } \Delta w \neq 0, \\ \frac{1}{2}\left[\binom{\lfloor \frac{L}{2} \rfloor - 1}{\lceil \frac{w-1}{2} \rceil - 1}^2 + \binom{\lfloor \frac{L}{2} \rfloor - 1}{\lceil \frac{w-1}{2} \rceil - 1}\right], & \text{otherwise and } \Delta w = 0. \end{cases}$$

(3) the number of seeds with $h+t=k\,(2 \le k \le w)$ is

$$\begin{cases} \frac{k-1}{2}\binom{L-k-2}{w-k}, & \text{if } \begin{cases} k \text{ is odd, or} \\ L \text{ even, } w \text{ odd} \end{cases} \\ \frac{k-1}{2}\binom{L-k-2}{w-k} + \frac{1}{2}\binom{\lfloor \frac{L}{2} \rfloor - \frac{k}{2} - 1}{\lfloor \frac{w}{2} \rfloor - \frac{k}{2}}, & \text{otherwise.} \end{cases}$$

(4) the number of seeds with $|h-t| = k(0 \le k \le w-2)$ is

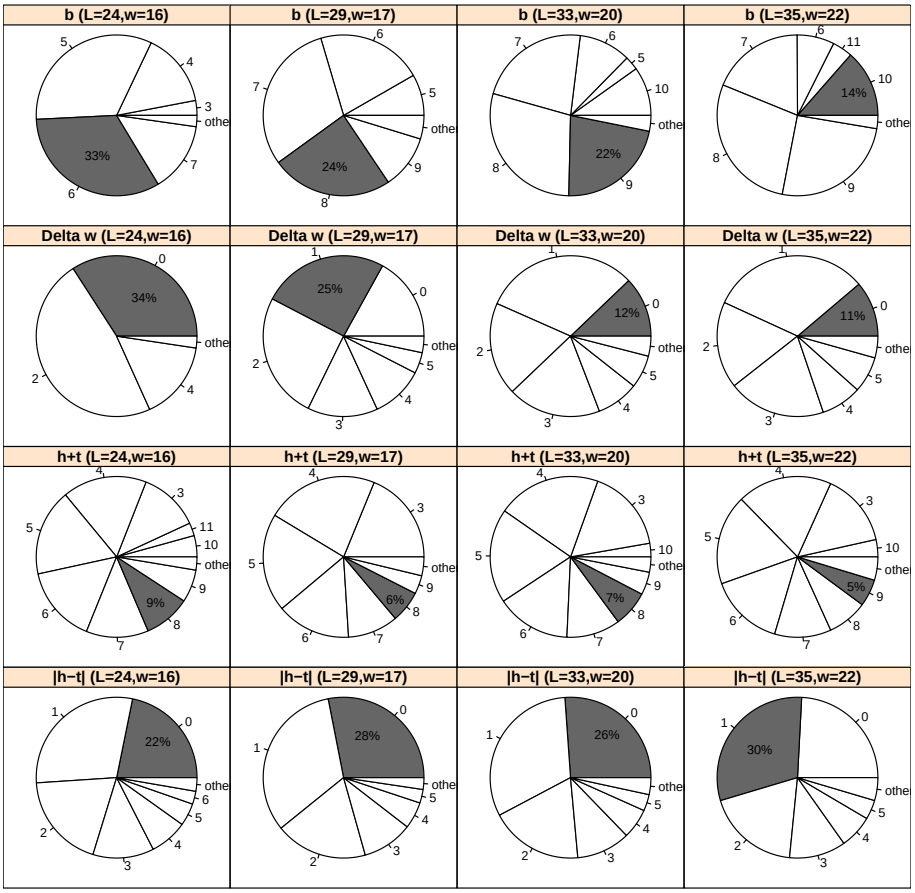

Figure 4.7 *Pie chart of the filterability of the seeds filters $b, \Delta w, h+t, |h-t|$ (rows from top to bottom) for $\mathcal{Q}_{24,16}, \mathcal{Q}_{29,17}, \mathcal{Q}_{33,20}, \mathcal{Q}_{35,22}$ (columns from left to right). The shaded sectors highlight the proportions in which the predicted best seeds falls.*

$$\begin{cases} \displaystyle\sum_{h=1}^{\left\lfloor \frac{w-k}{2} \right\rfloor} \binom{L-2h+k-2}{w-2h+k}, & k \neq 0, \\ \displaystyle\frac{1}{2} \sum_{h=1}^{\left\lfloor \frac{w}{2} \right\rfloor} \binom{L-2h-2}{w-2h}, & k = 0,\ L \text{ is even and } w \text{ is odd}, \\ \displaystyle\frac{1}{2} \sum_{h=1}^{\left\lfloor \frac{w}{2} \right\rfloor} \left[\binom{L-2h-2}{w-2h} + \binom{\left\lfloor \frac{L}{2} \right\rfloor - h - 1}{\left\lfloor \frac{w}{2} \right\rfloor - h} \right], & \text{otherwise.} \end{cases}$$

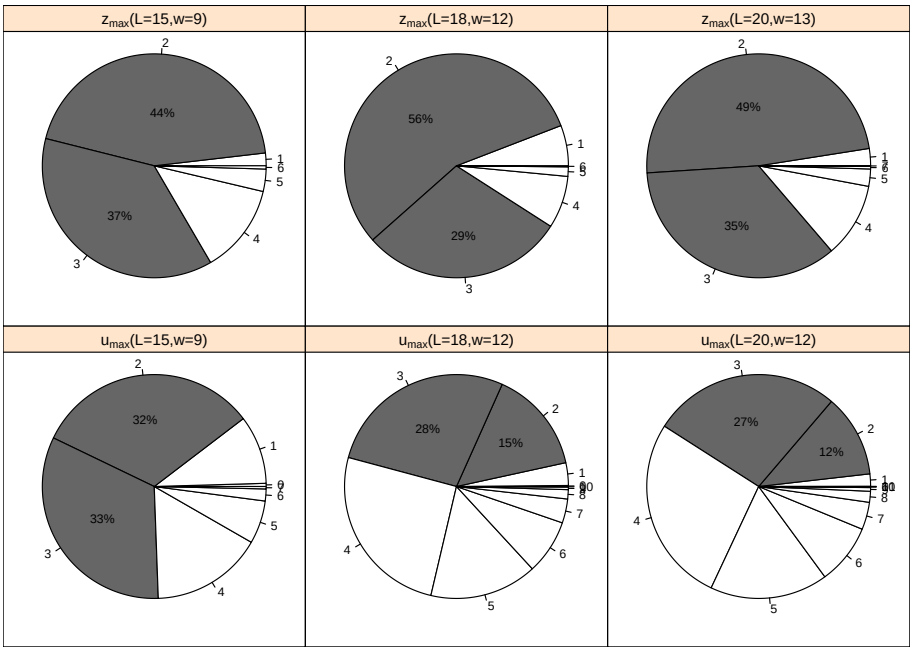

Figure 4.8 *Pie chart of the filterability of* $z_{\max}$ *and* $u_{\max}$ *for* $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ *and* $\mathcal{Q}_{20,13}$ *(columns from top to bottom).*

Figure 4.7 (on page 50) shows us the filterability of the filters $b, \Delta w, h+t, |h-t|$ for the cases of $\mathcal{Q}_{24,16}, \mathcal{Q}_{29,17}, \mathcal{Q}_{33,20}, \mathcal{Q}_{35,22}$. In each chart, the shaded sector is where the predicted best seeds fall in. From these charts, we can see, generally, the above filters can filter out 40% ~ 90% seeds. If we can combine several filters, then we can filter out more. However we have no idea what is the optimal combination of the filters. It would be an interesting direction to pursue in the future. But to achieve higher filterability, we may as well try all the possible filters.

Figure 4.8 (on page 51) shows the filterability of filters $z_{\max}$ and $u_{\max}$. It appears that the filterability of this two filters are lower than the other filters in Figure 4.7 (on page 50). Of course, for these two filters, we take two possible optimal values. This reduces their filterability.

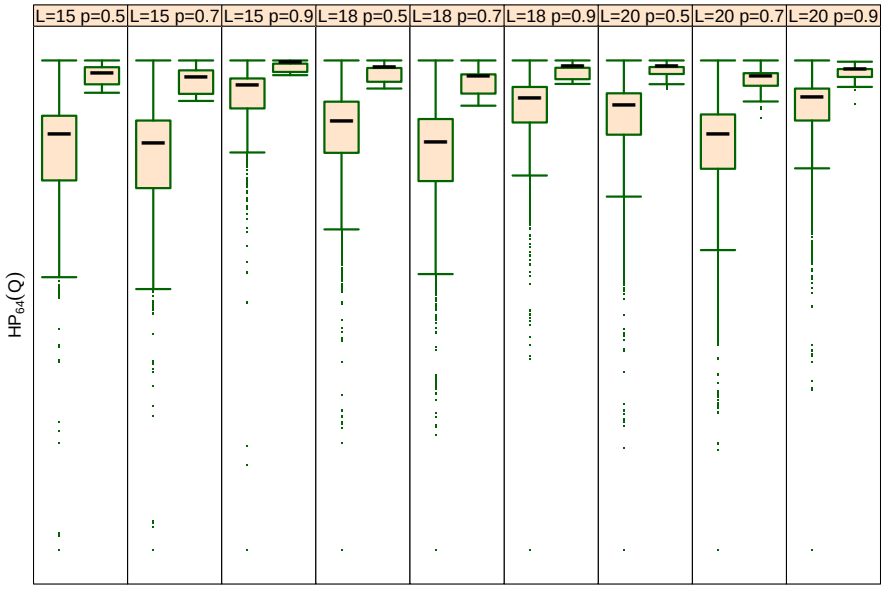

Figure 4.9 *Box plot of* $\mathbb{HP}_{64}$ *with optimal filter values of* $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}, \mathcal{Q}_{20,13}$ *for* $p = 0,5,0.7,0.9$. *The left box in each panel is the box plot for all the spaced seeds, the right box is for the spaced seeds with optimal* $b, \Delta w, h+t, |h-t|, z_{\max}, u_{\max}$ *values.*

From Figure 4.7 (on page 50), we notice that $h+t$ is the most efficient filter among the filters as long as we know the exact optimal choice of $h+t$, which in many case we cannot know. Filterability of other filters are about the same level. If we allow ± 1 error of the optimal choice of $h+t$, then the filterability of $h+t$ will fall to the same level as that of the others. Although the filterabilities of the other filters are not as good as $h+t$, they are more stable and easier to predict (the optimal values).

Table 4.6 (on page 53) shows us the filterability when we combine several filters together. The last 3 rows record the percentage of remaining seeds after filtration when we use exact optimal filter values. We can see that if we use several filters at the same time, then we can greatly deduce the number of seeds. Especially when we use

Table 4.6 *Filterability of the combinations of the filters for $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}, \mathcal{Q}_{20,13}$*

b	✓				✓	✓	✓				✓	✓	✓		✓
Δw		✓			✓			✓	✓		✓	✓		✓	✓
$h+t$			✓			✓		✓		✓	✓		✓	✓	✓
$\|h-t\|$				✓			✓		✓	✓		✓	✓	✓	✓
$\mathcal{Q}_{15,9}$	41.0	34.6	16.1	34.1	14.6	6.9	15.0	6.3	12.7	8.1	3.1	5.5	3.5	3.5	1.7
$\mathcal{Q}_{18,12}$	28.8	39.6	13.1	22.9	13.8	3.1	8.2	5.2	10.8	2.7	1.4	4.2	0.6	1.4	0.4
$\mathcal{Q}_{20,13}$	37.3	66.5	8.7	30.8	26.2	2.8	12.3	5.3	22.4	2.9	1.8	9.2	0.9	2.2	0.8

The ✓ signs mark out which filters (rows) are used. The numbers in the last 3 rows are percentage obtained by dividing the number of spaced seeds with the checked optimal filter values by the total number of seeds.

four filters, we need only handle less than 1% of the total seeds.

Figure 4.9 (on page 52) shows the separability when we use all the above filters. The box plot at left hand side in each panel is for all seeds of $\mathcal{Q}_{L,w}$, the right hand side is for those seeds with optimal $b, \Delta w, h+t, |h-t|, z_{\max}$ and $u+\max$ values. We can see from the figure that the hitting probabilities of the optimal seeds are almost among the top quarter of $\mathcal{Q}_{L,w}$. Since the hitting probability distribution are very skew, in fact, the hitting probability of the seeds with optimal filter values are generally very close to the maximal hitting probability.

4.6 Quick and practical search for effective spaced seeds

Having the predictors and seed filters, we can now follow the procedure to predict effective spaced seeds as follows:

(1) find the optimal filter values including $b, \Delta w, h+t, |h-t|, z_{\max}, u_{\max}$

(2) compute the predictors' value for the seeds with the above optimal filters values

(3) sort the predictors value and obtain the top spaced seeds as effective seeds

In procedure (1), we can use the suggested optimal values we discussed in the precious sections. Since this step is very important, we introduce the **sampling method** to determine and secure our selection (see Preparata *et al.* [2005]). The procedures are

(1) generate a set of sample seeds;

(2) for each seeds in this sample, calculate the hitting probability;

(3) choose the one with the largest hitting probability as an effective seed.

Now having the idea of predictors and seeds filters, we can further use the following procedures that is computationally faster:

(1) generate a set of sample seeds with different filters values;

(2) for each seeds in this sample, calculate the value of the predictors (e.g. Σ_2) for selected[2] p and n;

(3) choose the one with optimal predictor value to determine the optimal filters values;

(4) generate all or a sample[3] of the spaced seeds with the optimal filters values achieved from step (3), compute the predictor value for each seed;

(5) choose the spaced seed with the optimal predictor value in step (4) as a effective spaced seed.

2. We prefer small p, e.g. 0.5, for good predictability

3. Depending on the number of seeds with the optimal filters values

CHAPTER 5

Asymptotic Hitting Probability

When the length of the random sequence S is very very long, then it is impossible for us to compute $\mathbb{HP}_n$. In this case, in order to predict the behavior of different spaced seeds, we need to explore the asymptotic behavior of $\mathbb{HP}_n$, and to find what attribute of a spaced seed controls the asymptotic behavior.

For the asymptotic behavior of $\mathbb{HP}_n$, based on the work of deterministic finite state automata of Nicodéme *et al.* [2002], Buhler *et al.* [2003] derived the following theorem.

Theorem 5.1 *For spaced seed Q with length L and weight w, there exist $\beta_Q > 0, 0 < \lambda_Q < 1$ such that*

$$\overline{\mathbb{HP}}_n(Q) = \beta_Q \lambda_Q^n (1 + o(1)),$$

Having this theorem, we can study the asymptotic behavior through studying the behavior of β_Q and λ_Q. Easy to know that the smaller the λ_Q, the faster the $\overline{\mathbb{HP}}_n(Q)$ goes to 0, or the faster the $\mathbb{HP}_n(Q)$ goes to 1, that is, the better the spaced seed Q. Similarly, the smaller the β_Q, the better the spaced seed Q.

5.1 Bounds of λ_Q

Because λ_Q controls the convergence rate of the hitting probability, it is very important to estimate its value. We have the following bounds on λ_Q:

Theorem 5.2 *For λ_Q in Theorem 5.1, we have*

$$\max_{0\le i\le L-1}\left(\overline{\mathbb{HP}}_{L+i}\right)^{\frac{1}{i+1}} \le \lambda_Q \le \min_{0\le i\le L-1}\left(\overline{\mathbb{HP}}_{L+i}\right)^{\frac{1}{L+i}}.$$

To prove this theorem, we need the following lemma in Choi and Zhang [2004].

Lemma 5.3 (Choi and Zhang) *Let Q be a spaced seed with length L, then for any $2L-1 \le k \le n$,*

$$\overline{\mathbb{HP}}_k\overline{\mathbb{HP}}_{n-k+L-1} \le \overline{\mathbb{HP}}_n \le \overline{\mathbb{HP}}_k\overline{\mathbb{HP}}_{n-k}.$$

Proof of Theorem 5.2: For $0 \le i \le L$, applying the first inequality in Lemma 5.3, we have

$$\overline{\mathbb{HP}}_{L+i}\overline{\mathbb{HP}}_{n-i-1} \le \overline{\mathbb{HP}}_n.$$

Let $n \to \infty$, then from Theorem 5.1, we can deduce that

$$\overline{\mathbb{HP}}_{L+i} \le \frac{\overline{\mathbb{HP}}_n}{\overline{\mathbb{HP}}_{n-i-1}} = \frac{\beta_Q \lambda_Q^n (1+\varepsilon_n)}{\beta_Q \lambda_Q^{n-i-1}(1+\varepsilon_{n-i-1})} \to \lambda_Q^{i+1}, \quad n \to \infty$$

and this will imply

$$\lambda_Q \ge \left(\overline{\mathbb{HP}}_{L+i}\right)^{\frac{1}{i+1}}.$$

Taking the maximum for $i = 0, 1, \ldots, L-1$ yields

$$\lambda_Q \ge \max_{0 \le i \le L-1} \left(\overline{\mathbb{HP}}_{L+i}\right)^{\frac{1}{i+1}}.$$

Thus we prove the first inequality in Theorem 5.2. Similarly, we apply the second inequality in Lemma 5.3 to get the second inequality. ■

Since λ_Q controls the convergence rate of $\mathbb{HP}_n$, it will play an important role in the performance of $\mathbb{HP}_n$ for different spaced seeds Q at large n.

Figure 5.1 (on page 58) and Figure 5.2 (on page 59) show the relationship between $\mathbb{HP}_{64}$ and the lower or upper bound of λ_Q. These two figures exhibit a very strong correlation between the bounds of λ_Q and $\mathbb{HP}_{64}$, even though the position $n = 64$ is not large enough for the hitting probability approaching 1 for $p = 0.5$ and 0.7. This provides strong numerical, evidence that λ_Q controls the performance of $\mathbb{HP}_n$ for proper large n.

A pleasing feature from these bounds is that the bounds of λ_Q are very tight. The difference of the two bounds are only as small as $0.001 \sim 0.08$. Tighter bounds occur when p is moderate (not close to 1).

According to this, we can use the lower bound or upper bound of λ_Q as a predictor of $\mathbb{HP}_n$. But since the calculation of the two bounds involves calculating the hitting

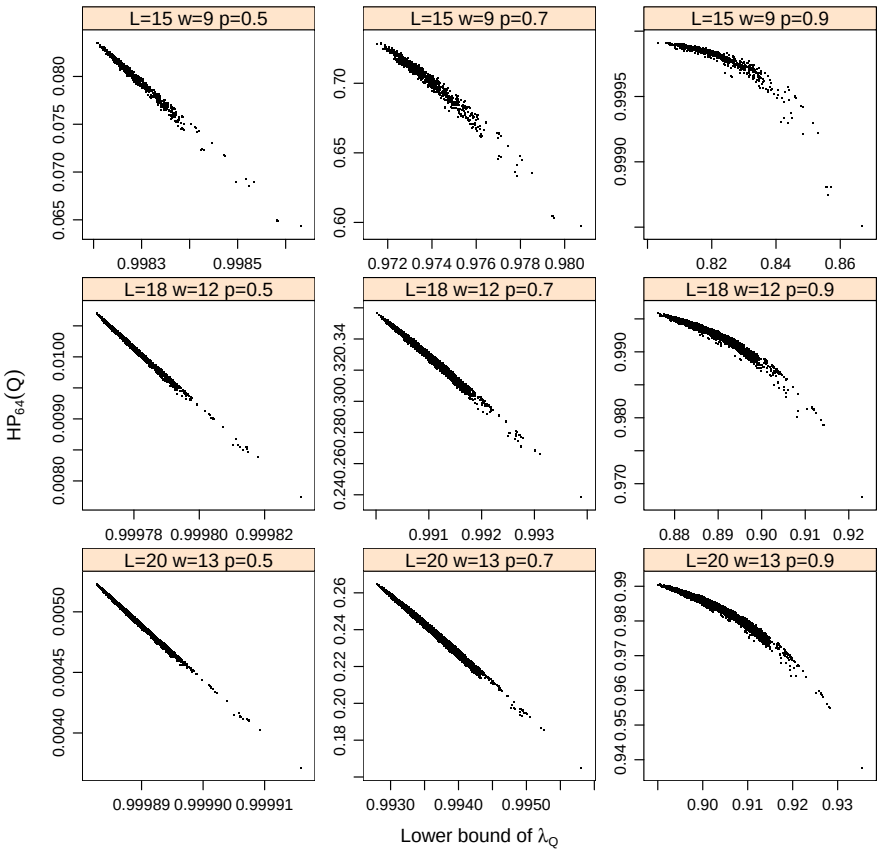

Figure 5.1 *Plots of $\mathbb{HP}_n(\mathcal{Q})$ vs the lower bound of λ_Q for $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ and $\mathcal{Q}_{20,13}$ (rows from top to bottom) for $p = 0.5, 0.7, 0.9$ (columns from left to right).*

probabilities at position $n \leq 2L-1$, it will take similar time as calculating $\mathbb{HP}_{2L-1}$, so the two bounds are not practical predictor.

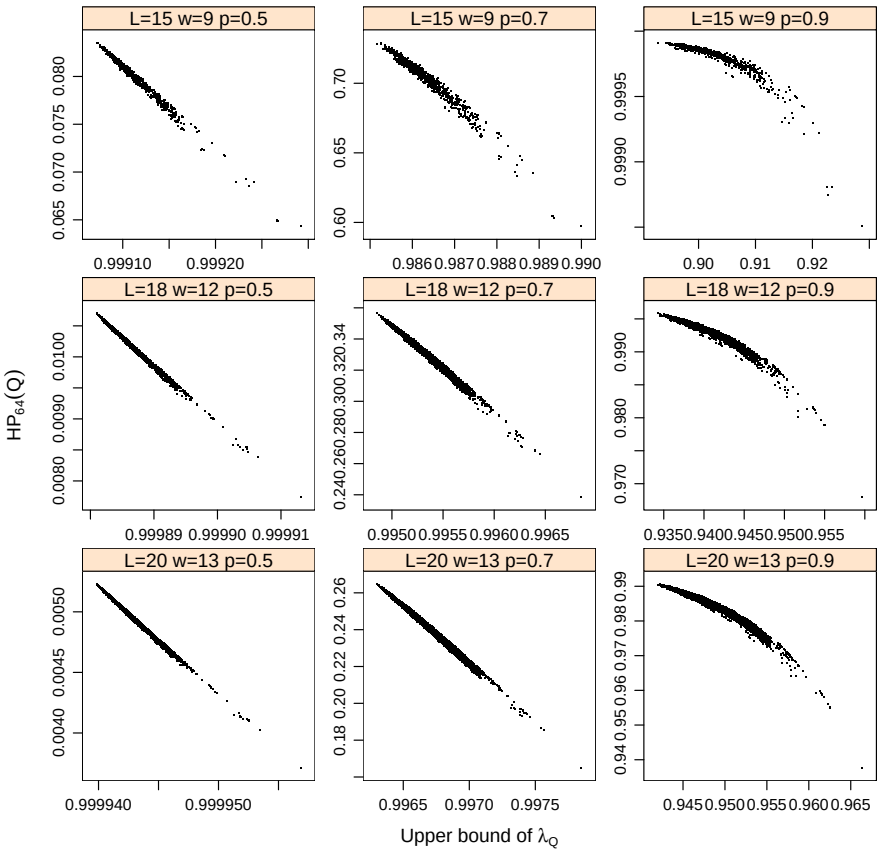

Figure 5.2 *Plots of $\mathbb{HP}_n(\mathcal{Q})$ vs the lower bound of λ_Q for $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ and $\mathcal{Q}_{20,13}$ (rows from top to bottom) for $p = 0.5, 0.7, 0.9$ (columns from left to right).*

5.2 Estimate λ_Q

Applying Theorem 5.1, we have

$$\frac{\overline{\mathbb{HP}}_n}{f_n} = \frac{\overline{\mathbb{HP}}_n}{\overline{\mathbb{HP}}_{n-1} - \overline{\mathbb{HP}}_n} \approx \frac{\beta_Q \lambda_Q^n}{\beta_Q \lambda_Q^{n-1} - \beta_Q \lambda_Q^n} = \frac{\lambda_Q}{1-\lambda_Q},$$

so

$$\log \frac{\overline{\mathbb{HP}}_n}{f_n} \approx \log \frac{\lambda_Q}{1-\lambda_Q}.$$

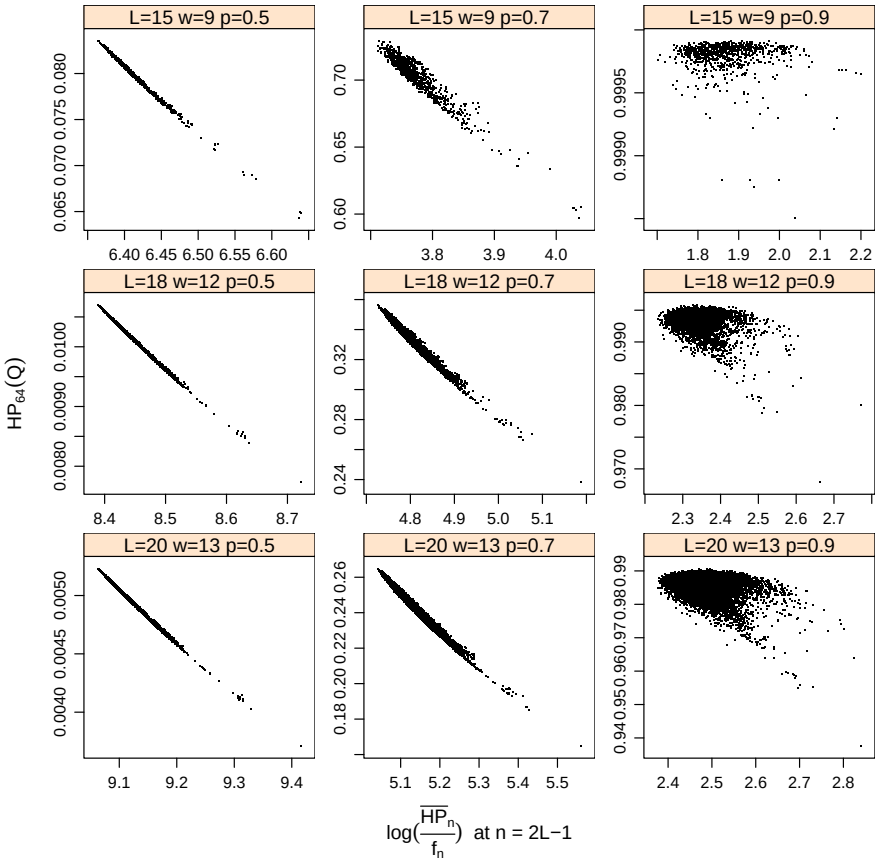

Figure 5.3 *Plots of* $\mathbb{HP}_n(\mathcal{Q})$ *vs* $\log\left(\frac{\overline{\mathbb{HP}}_{2L-1}}{f_{2L-1}}\right)$ *for* $\mathcal{Q}_{15,9}, \mathcal{Q}_{18,12}$ *and* $\mathcal{Q}_{20,13}$ *(rows from top to bottom) for* $p = 0.5, 0.7, 0.9$ *(columns from left to right).*

Easy to see that this is an increasing function of λ_Q. Because λ_Q is negatively correlated with $\mathbb{HP}_n$, it is negatively correlated with $\mathbb{HP}_n$.

Figure 5.3 (on page 60) shows us the correlation of $\mathbb{HP}_n$ with the approximating value of λ, $\log(\overline{\mathbb{HP}}_{2L-1} / f_{2L-1})$. Apparently, there is a good correlation when p is small, but it becomes worse when p increases.

REFERENCE

Altschul, S. F., Gish, W., Myers, E. W. and Lipman, D. Basic local alignment search tool. *J. Mol. Biol.*, **215**:403–410, 1990.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., and Zhang, Z., Miller, W. and Lipman, D. Gapped blastand psi-blast: a new generation of protein database search program. *Nucleic Acids Res.*, **25**:3389–3402, 1997.

Brejovà, B., Brown, D. and Vinař, T. Optimal spaced seeds for hidden Markov models, with application to homologous coding regions. *Predeedings of the 14th Annual Symposium on Combinatorial Pattern Matching (CPM)*, pp. 42–45, 2003.

Buhler, J. Efficient large-scale sequence comparison by locality-sensitive hashing. *Bioinformatics*, **17**:419–428, 2001.

Buhler, J., Keich, U. and Sun Y. Designing seeds for similarity search in genomic

DNA. *Proc. Seventh Annual International Conference on Computational Molecular Biology (RECOMB03)*, pages 67–75, Berlin, Germany, 2003.

Choi, K. P., and Zhang, L. Sensitivity analysis and efficient method for identifying optimal spaced seeds. *Journal of Computer and System Sciences*, **68**:22–40, 2004.

Choi, K. P., Zeng, F., and Zhang, L. Good spaced seeds for homology search. *Bioinformatics*, **20**:1053–1059, 2004.

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., and Salzberg, S. L. Alignment of whole genomes. *Nucleic Acids Research*, **27**:2369–2376, 1999.

Galambos, J., Simonelli, I. *Bonferroni-type inequalities with applications*. Springer, New York, 1996.

Gish, W. WU-Blast 2.0. Website: http://labst.wustl.edu, 2001.

Hardison, R. C., Oeltjen, J. and Miller, W. Long Human-mouse sequence alignments reveal novel regulatory element: A reason to sequence the mouse genome. *Genome Research*, **7**:966–969, 1997.

Keith, U., Li, M., Ma,, B., and Tromp, J. On spaced seeds. unpublished.

Kong, Y. Fast method to find optimal spaced seeds for homology search. in press, 2004.

Huang, X. and Miller, W. A time-effect, linear-space local similarity algorithm. *Adv. Appl. Math.*, **12**:337–357, 1991.

Li, W. H., Gu, Z., Wang, H. and Nekrutenko, A. Evolutionary analysis of the Human genome. *Nature*, **409**:847–849, 2001.

Lipman, D. J., and Pearson, W. R. Rapid and sensitive protein similarity searches. *Science*, **227**:1435–1441, 1985.

Ma, B., Tromp, J. and Li, M. PatternHunter—faster and more sensitive homology search. *Bioinformatics*, **18**:440–445, 2002.

Nicodéme, P., Salvy, B. and Flajoret, P. Motif statistics. *Bioinformatics*, **18**, 2002.

Needleman, S. B., and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**:443–453, 1970.

Preparata, F. P., Zhang, L. and Choi, K. P. Quick, practical selection of effective seeds for homology search. Manualscript.

Rabin, M. O., and Karp, R. Efficient randomized pattern-matching algorithm. *IBM Journal of Reserach Development*, **31**:249–260, 1987.

Scherer, S. W., Cheung, J., MacDonald, J. R., Osborne, L. R., Nakabayashi, K., Herwick, J. A., Carson, A. R., Parker-Katiraee, L., Skaug, J., Khaja, R. *et al.* Human chromosome 7: DNA sequence and biology. *Science*, **300**:767–772, 2003.

Smith, T. F., and Waterman, M. S. Indentification of common molecular subsequences. *J. Mol. Biol.*, **147**:195–197, 1981.

Ureta-Vidal, A., Ettwiller, L. and Birney, E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Res. Genet.*, **13**:251–262, 2003.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexanderson, M. and An, P. Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**:520–562, 2002.

Yeh, R. F., Lim, L. P. and Burge, C.B. Computational inference of homologous gene structrues in human genome. *Genome Research*, **11**:803–816, 2001.

APPENDIX **A**

Derivation of Equation (3.10)

First, we know that

$$\Sigma_3 = \sum_{i\neq j\neq k} \mathbb{P}(A_i A_j A_k) = \sum_{i=1}^{n-L} \sum_{j=1}^{n-L-i} \sum_{l=L}^{n-L-i-j} \mathbb{P}(A_l A_{l+i} A_{l+i+j})$$

Because the probability $\mathbb{P}(A_l A_{l+i} A_{l+i+j})$ is independent of l, we get

$$\Sigma_3 = \sum_{i=1}^{n-L} \sum_{j=1}^{n-L-i} (n-L-i-j+1)\mathbb{P}(A_L A_{L+i} A_{L+i+j}). \tag{A.1}$$

Now we have the following four cases:

(1) $i+j \leq L-1$. In this case, we have the probability

$$\mathbb{P}(A_L A_{L+i} A_{L+i+j}) = p^{3w-\theta(i)-\theta(j)-\theta(i+j)+\theta(i,j)},$$

this simply comes from counting the number of 1's in $Q \bigcup (Q \gg i) \bigcup (Q \gg i+j)$.

The part of summation corresponding to this case in equation (A.1) will be

$$\Sigma_3^{(1)} = \sum_{i=1}^{L-1} \sum_{j=1}^{L-i-1} (n-i-j-L+1)\, p^{3w-\theta(i)-\theta(j)-\theta(i+j)+\theta(i,j)} \tag{A.2}$$

(2) $i+j \geq L$, but $0 \leq i, j \leq L-1$. In this case

$$\mathbb{P}(A_L A_{L+i} A_{L+i+j}) = p^{3w-\theta(i)-\theta(j)}.$$

The part of summation corresponding to this case in equation (A.1) will be

$$\Sigma_3^{(2)} = \sum_{i=1}^{L-1} \sum_{j=L-i}^{L-1} (n-i-j-L+1) p^{3w-\theta(i)-\theta(j)} \tag{A.3}$$

(3) one of $i, j \leq L-1$, the other $\geq L$. In this case, it is easy to know that both the two cases whether $i \geq L$ or $j \geq L$ have the same probability. So we just calculate the case for $i \leq L-1$, $j \geq L$ and then double it. In this case

$$\mathbb{P}(A_L A_{L+i} A_{L+i+j}) = p^{3w-\theta(i)},$$

so the part of summation corresponding to this case in equation (A.1) will be

$$\begin{aligned} \Sigma_3^{(3)} &= 2 \sum_{i=1}^{L-1} \sum_{j=L}^{n-L-i} p^{3w-\theta(i)} \\ &= \sum_{i=1}^{L-1} (n-2L-i+1)(n-2L-i+2) p^{3w-\theta(i)} \end{aligned} \tag{A.4}$$

(4) $i, j \geq L$. In this case

$$\mathbb{P}(A_L A_{L+i} A_{L+i+j}) = p^{3w}.$$

The part of summation corresponding to this case in equation (A.1) will be

$$\begin{aligned} \Sigma_3^{(4)} &= \sum_{i=L}^{n-L} \sum_{j=L}^{n-L-i} (n-i-j-L+1) p^{3w} \\ &= \frac{1}{6}(n-3L+1)(n-3L+2)(n-3L+3) p^{3w} \end{aligned} \tag{A.5}$$

Now we just add the summations of (A.2)–(A.5) to get equation (3.10).

APPENDIX **B**

Proof of Lemma 4.2

(1) To determine the number of seeds with $b(1 \leq b \leq L-w)$ blocks of $*$'s, it is equivalent to determine which 1's are followed by $*$ and which $*$'s are followed by 1.

To determine the position of 1's followed by a$*$ among all the w 1's, we first exclude the last 1 because it cannot followed by anything. Then among the remaining $w-1$ 1's, we simply choose b and get the number $\binom{w-1}{b}$.

Then to determine the number of $*$'s followed by 1, we follow the same argument above and get the number is $\binom{L-w-1}{b-1}$.

For case (i) L even and w odd, or L odd, and (ii) $w+d$ even, there are no symmetric seeds. Therefore, simply multiply the above number and divided by 2 (take half of the seeds with heavy tail).

Otherwise, there exist symmetric seeds. To count the number of these seeds, we can only consider one half of the spaced seed. It is easy to enumerate all the possible cases to get the number of variation is

$$\binom{\lfloor \frac{w+1}{2} \rfloor - 1}{\lfloor \frac{d}{2} \rfloor} \binom{\lfloor \frac{L-w}{2} \rfloor - 1}{\lceil \frac{d-2}{2} \rceil}$$

just using the same trick as used above.

(2) When a seed has the weight difference $\Delta w (0 \leq \Delta w \leq L-2)$, obviously, if the first condition

$$\text{if } L \text{ even and } \begin{cases} w \text{ odd}, \Delta w \text{ even, or} \\ w \text{ even}, \Delta w \text{ odd} \end{cases}$$

occurs, then there is no seeds satisfying this condition.

Otherwise, when $\Delta w \neq 0$, then there will exist symmetric seeds. Then we just multiply the count of choose 1's from the left half and the count of choosing 1's from the right half to get the result.

When $\Delta w = 0$, there will exist symmetric seeds. In this case both the weight of the two halves are $\lceil \frac{w-1}{2} \rceil$, we just choose these 1's in the first half, and then square it to get all the possible candidate seeds. To get the exact number of seeds with heavy tail, we should discard one half excluding those symmetric seeds. The number of symmetric seeds is just as the number of choose $\lceil \frac{w-1}{2} \rceil$ 1's from the left half. Thus, we get the result.

(3) If k is odd, or L is even and w is odd, then the symmetric seeds will not occur. In this case the number of spaced seed with $h + t = k$ is one half of $k-1$ times (for all possible h and t values) of the combination number $\binom{L-k-2}{w-k}$. Otherwise, there will be symmetric seeds, then before we take one half of the total seeds,

we must add back the symmetric ones, which have the number

$$\begin{pmatrix} \lfloor \frac{L}{2} \rfloor - \frac{k}{2} - 1 \\ \lfloor \frac{w}{2} \rfloor - \frac{k}{2} \end{pmatrix}$$

(4) When k is not 0, there does not exist symmetric seeds, so we just add all the spaced seeds with $h = 1, 2, \ldots, \frac{w+k}{2}$ and then divide by 2. But we know that the numbers of candidate seeds with $h = i, t = i + k$ and those of $h = i + k, t = i$ are the same. So instead divided by 2, we can simply take the seeds with $h < t$. When k is 0 and L even, w odd, we have $h = t$ and there is also no symmetric seeds. So we just needs count all the candidate seeds with $h = t = 1, 2, \ldots, \lfloor \frac{w}{2} \rfloor$ and then divide by 2. In the other case, $h = t$ and there exist symmetric seeds. We also add back the symmetric seeds and then take one half