Run Probability of High-Order Seed Patterns and its Applications to Finding Good Transition Seeds

J. L. Yang* and L. X. Zhang

Department of Mathematics, National University of Singapore,
Singapore 117543, Singapore
**E-mail: g0306107@nus.edu.sg*

Transition seeds exhibit a good tradeoff between sensitivity and specificity for homology search in both coding and non-coding regions. But, identifying good transition seeds is extremely hard. We study the hit probability of high-order seed patterns. Based on our theoretical results, we propose an efficient method for ranking transition seeds for seed design.

Keywords: High-order seed pattern, transition Seed, sensitivity analysis, run statistics.

1. Introduction

Biomolecular sequence comparison is one of the most important tasks in bioinformatics in the study of molecular evolution, genomics and molecular medicine. As a result, many sequence comparison programs have been developed to meet the challenge of the rapid increase in size of sequence databases.

The seed alignment is the dominant technique for homology search and genomic sequence alignment. Such a technique was first implemented in BLASTN program.[1] In BLASTN, a local alignment is found by first identifying exact matches of eleven contiguous residues between the two input sequences, called *seed hits*, and then extending them on either side for approximate matches by dynamic programming. The resulting alignments are scored for acceptance. In recent years, more general patterns of conservation have been proposed as seeds for sequence alignment.[5,8,14,20,24] Different seeds are also used as anchor point in whole-genome and multiple sequence alignments.[3,6,11]

Good spaced seeds improve tremendously the sensitivity of seed alignment while keeping speed unchanged.[20] Hence, seed design is an important aspect of seeded alignment. Identifying good seeds relies on efficient computation of seed sensitivity. Dynamic programming[13] and recurrence[10] methods were proposed for computing the sensitivity of the spaced seeds on a simple i.i.d. ungapped alignment model. It has been shown that computing the sensitivity of a spaced seed is NP-hard.[18] Hence, efficient heuristic methods were also developed for identifying good spaced seeds.[7,9,10,15,23,27,29] Algorithms for multi-seed design were also developed.[16,19,21,26,28]

Transition and transversion were first incorporated into seed design in BLASTZ.[24] This leads to the study of the transition seeds that contain fixed match and transition positions. Transition seeds exhibit a good tradeoff between sensitivity and specificity for homology search in both coding and non-coding regions.[27,31] However, identifying good transition seeds is a difficult task. Here we study the run probabilities of high-order seed-like patterns, which include spaced seeds and transition seeds as special cases. We generalize the theoretic study of spaced seeds[30] to high-order seed-like patterns. Using these results, we propose an efficient method for ranking transition seeds for the purpose of seed design.

Due to the space limit, we omit the proofs of the theorems stated in this extended abstract. The reader is referred to the full version of this paper for all proofs.

2. Seeds, Sensitivity and Specificity

2.1. *Spaced seeds*

A **(basic) spaced seed** is defined as a list of indices $\{i_1, i_2, \ldots, i_w\}$ with $i_1 = 0$. It can also be specified by a string $1 *^{i_2-1} 1 *^{i_3-i_2-1} 1 \ldots 1 *^{i_w-i_{w-1}-1} 1$ over alphabet $\{1, *\}$. Two sequences S_1 and S_2 exhibit a seed match at positions x and y if, for $1 \le k \le w$, $S_1[x+i_k] = S_2[y+i_k]$. The number of match positions w is defined to be the weight of the seed; the span of the checked positions, $i_w + 1$, is called the length of the seed.

A **transition spaced seed** is defined as a pair of disjoint lists of indices:

$$M = \{i_1, i_2, \ldots, i_{m'}\}, \quad Z = \{j_1, j_2, \ldots, j_{t'}\}$$

with $i_1 = 0$ or $j_1 = 0$. Two sequences S_1 and S_2 exhibit a match of the transition seed at positions x and y if, for $1 \le k \le m'$, $S_1[x+i_k] = S_2[y+i_k]$ and, for $1 \le k \le t'$, $S_1[x+j_k] = S_2[y+j_k]$, or two residues $S_1[x+j_k]$ and $S_2[y+j_k]$ are both pyrimidines or purines. The positions in M are called match positions; m' is defined to be the match weight of the seed. The positions in Z are called transition positions; t' is defined to be the transition weight of the seed. The length of the seed is defined to be $\max\{i_{m'}, j_{t'}\}+1$. Equivalently, we specify a transition seed of length L_Q by a string of length L_Q over alphabet $\{1, \#, *\}$ in which 1s represent match positions, #s transition positions, and $*$s other so-called 'don't care' positions.

2.2. *Seed sensitivity and specificity*

The sensitivity of a seed is the probability that a biologically meaningful alignment contains a match to the seed. The biological meaningful alignments are usually given through a probabilistic model on nucleotides. Here, we restrict ourselves to the Bernoulli or zero-th order Markov ungapped alignment model. We assume the pair of residues in each position is independently and identically generated from $\{A, G, C, T\} \times \{A, G, C, T\}$. By using 1s and 0s to represent matches and mismatches in the ungapped alignment between two sequences X and Y, a seed match can be viewed as an occurrence of the spaced seed in a binary sequence. Therefore, in the Bernoulli sequence model, the sensitivity of a spaced seed is defined to be the hit probability of a spaced seed pattern in a binary random sequence of a fixed length L (which is 64 by default).

Similar to basic spaced seeds, each match to a transition seed in an ungapped alignment can be viewed as an occurrence of the seed in the corresponding sequence over $\{1, 2, 3\}$, where we use 1s, 2s, 3s to represent matches, mismatches and transversions respectively.

A seed's specificity is defined to be one minus the probability that the seed match occurs in an alignment between two unrelated random sequence by chance. Therefore, the specificity is also a kind of hit probability in a probabilistic alignment model.

3. High-order Seed Patterns and Their Run Probabilities

Motivated by analyzing seed sensitivity and specificity, we study the run probabilities of sequence patterns of a special type in this section.

Let $\Sigma = \{b_1, b_2, \ldots, b_m\}$. An order-$t$ pattern $\mathcal{P}$ consists of a sequence Q of length L_Q on an alphabet $\Sigma' = \{a_1, a_2, \ldots, a_t\}$ and an ordered list of subsets $\{\Sigma_1, \Sigma_2, \ldots, \Sigma_t\}$ such that $Q[1] \neq a_t$, $Q[L_Q] \neq a_t$, and $\Sigma_1 \subset \Sigma_2 \subset \ldots \subset \Sigma_t = \Sigma$. We say the pattern Q to hit a sequence S on Σ at position k if, for $1 \leq i \leq L_Q$, the following condition is satisfied: if $Q[i] = a_j$ for some j, then, $S[i + k - L_Q] \in \Sigma_j$.

Example 3.1. (1)A basic spaced seed π is an order-2 pattern with a sequence over $\{1, *\}$ and the subset list: $\{1\}$, $\{0, 1\}$. (2) A transition seed π_t is an order-3 pattern with a sequence over $\{1, \#, *\}$ and ordered subset list: $\{1\}$, $\{1, 2\}$, $\{1, 2, 3\}$.

We study the hit probability of an order-t pattern in the Bernoulli random sequence on alphabet $\Sigma = \{b_1, b_2, \cdots, b_m\}$, in which a letter b_i is generated with probability p_i at each position and $\sum_{1 \leq i \leq m} p_i = 1$. We use $\mathcal{M}(\Sigma, p_1, p_2, \ldots, p_m)$ to denote this Bernoulli sequence model.

For an order-t pattern Q and a Bernoulli random sequence S, we use q_n to denote the hit probability that the pattern Q hits S before or at position n; we also use f_n to denote the probability that Q first hits S at position n. Let the length of the pattern Q be L_Q. Obviously,

$$f_i = 0, 1 \leq i \leq L_Q - 1,$$

$$f_{L_Q} = p_1^{w_1} (p_1 + p_2)^{w_2} \cdots (\sum_{i=1}^{m-1} p_i)^{w_{m-1}},$$

where w_i is the number of occurrences of the letter a_i in the pattern, and

$$q_n = \sum_{1 \leq i \leq n} f_i. \tag{1}$$

Let $\Sigma = \{b_1, b_2, \cdots, b_m\}$. Given an order-$t$ pattern $\mathcal{P}$ with sequence Q and ordered list of subsets of $\Sigma : \{\Sigma_1, \Sigma_2, \ldots, \Sigma_t\}$ and a sequence S on Σ. By encoding the letters in $\Sigma_i - \Sigma_{i-1}$ by a new letter b'_i, we transform the sequence S into a sequence S' on $\Sigma'' = \{b'_1, b'_2, \ldots, b'_t\}$. Let $\mathcal{P}'$ be the order-t pattern with sequence Q' and ordered list of subsets $\{\{b'_1\}, \{b'_1, b'_2\}, \ldots, \{b'_1, b'_2, \ldots, b'_t\}\}$. It is easy to see that the hit probability of $\mathcal{P}$ on sequence S in Bernoulli model $\mathcal{M}(\Sigma, p_1, p_2, \ldots, p_m)$ is equal to the hit probability of $\mathcal{P}'$ on sequence S' in Bernoulli model $\mathcal{M}'(\Sigma'', p'_1, p'_2, \ldots, p'_t)$, where $p'_i = \sum_{j:b_j \in \Sigma_i - \Sigma_{i-1}} p_j$. Therefore, for simplicity, we will focus on order-t patterns with a sequence and an order list of t subsets of an alphabet with size t in the rest of paper.

3.1. *A recurrence formula for hit probability*

Let Q be an order-t pattern and S be a random sequence in Bernoulli model $\mathcal{M}(\Sigma, p_1, p_2, \ldots, p_t)$. We use E_n be the event that Q hits sequence S at position n and $\bar{E}_n$ its complement event. We use $M_Q = \{Q_1, Q_2, \ldots, Q_h\}$ to denote all $h := \prod_{i=2}^{t} i^{w_i}$ distinct sequences obtained from Q by replacing each occurrence of a_i with a letter in $\Sigma = \{b_1, b_2, \ldots, b_t\}$. Taking a transition seed $Q = 1 * \#1$ as an example, we have $M_Q = \{1111, 1211, 1311, 1121, 1221, 1321\}$.

For $1 \leq i \leq h$, we define $E_n^{(i)}$ to be the event that Q_i hits S at position n. Obviously, $E_n^{(i)}$ and $E_n^{(j)}$ are disjoint for $1 \leq i \neq j \leq h$. Define $f_n^{(i)} = P[\bar{E}_1 \bar{E}_2 \cdots \bar{E}_{n-1} E_n^{(i)}]$, the probability that Q first hits S at position n and $S[n - L_Q + 1, n] = Q_i$. Clearly, Q hits S at position n if and only if some Q_i hits S at position n and so $E_n = \bigcup_{1 \leq i \leq h} E_n^{(i)}$. This implies that

$$f_n = \sum_{i=1}^{h} f_n^{(i)}. \tag{2}$$

Let x be a sequence with length $|x|$. For an integer $k \leq |x|$, we use $x\langle k]$ and $x[k\rangle$ to denote the length-k suffix and prefix of x respectively. For any i, j and k, $1 \leq i, j \leq h$, $1 \leq k \leq L_Q$, we define

$$p_k^{(ij)} = \begin{cases} P[\, Q_j\langle L_Q - k] \,] & \text{if } k \leq L_Q - 1 \text{ and } Q_i\langle k] = Q_j[k\rangle \\ 1 & k = L_Q \text{ and } i = j \\ 0 & \text{otherwise} \end{cases}$$

we have

$$(1 - q_n)p_j = \sum_{i=1}^{h} \sum_{k=1}^{L_Q - 1} f_{n+k}^{(i)} p_k^{(ij)} + f_{n+L_Q}^{(j)} \tag{3}$$

for $1 \leq j \leq h$, where p_j is the probability that Q_j occurs at the position L_Q.

Using (1) - (3), one can compute the hit probability of a pattern recursively. It was first proved for the basic spaced seeds.[10]

3.2. *An Inequality on Hit Probability*

In this section, we present an inequality that relating the first hit probability to hit probability at different positions.

Theorem 3.1. *Let Q be an order-t pattern of length L_Q. Then, for any $2L_Q - 1 \leq k \leq n$, $f_k(1 - q_{n-k+L_Q-1}) \leq f_n \leq f_k(1 - q_{n-k})$ in a Bernoulli sequence model.*

3.3. *Asymptotic Analysis of Hit Probability*

Buhler *et al.*[7] proved that for any basic spaced seed Q, there exist two constants α_Q and λ_Q such that $1 - q_n \sim \alpha_Q \lambda_Q^n$. Similar results are established by Solov'ev.[25] Such an approximation also exists for our high-order pattern.

Theorem 3.2. *For an order-t pattern Q, there exist constants α_Q and λ_Q do not depend on n, such that $q_n = 1 - \alpha_Q \lambda_Q^n (1 + o(R^n))$ with $0 < R < 1$ in a Bernoulli model $\mathcal{M}(\Sigma, p_1, p_2, \cdots, p_t)$.*

The single term $1 - \alpha_Q \lambda_Q^n$ gives a very close approximation to q_n even for relative big n. Consider a specific transition seed $1 * 1$ containing no #s. In the Bernoulli sequence model $\mathcal{M}(\{1, 2, 3\}, p = 0.6, q = 0.3, r = 0.1)$, we obtain that $\lambda_Q = 0.7291502607$ and $\alpha_Q = 1.058452825$ using Maple.

In general, it is not easy to compute α_Q and λ_Q for an order-t pattern when t and L_Q are large. However, we will establish good bounds for λ_Q using the average distance between successive non-overlapping hits.

4. The Average Distance Between Successive Non-overlapping Hits

Renewal theory studied recurrent events connected with repeated trials. A recurrent event qualifies for the theory if the number of trials between successive occurrences of the event are jointly independent random variables with identical distribution. An non-overlapping hit of a pattern Q is a recurrent event under the following assumption: If a hit at position i is selected as a non-overlapping hit, then the next non-overlapping hit is the first hit at or after position $i + L_Q$.

The average distance between successive non-overlapping hits μ_Q is a very important parameter in the renewal theory. For the purpose of studying the hit probabilities of a pattern, it is formally rewritten as

$$\mu_Q = \sum_{i=0}^{\infty} i f_i.$$

Since $\sum_{j=L_Q}^{\infty} f_j = 1$ and $1 - q_i = \sum_{j=i+1}^{\infty} f_j$ for all $i \geq L_Q$, μ_Q may also be defined as

$$\mu_Q = L_Q + \sum_{i=L_Q}^{\infty} (1 - q_i).$$

4.1. *Bounding μ_Q*

Let Q be an order-t pattern on alphabet $\Sigma' = \{a_1, a_2, \dots, a_t\}$. For $1 \le k \le t$, we define $\mathcal{RP}(k)$ to be the ordered list of indices i such that $Q[i] = a_k$. For any $0 \le j \le L_Q - 1$, we define

$$\mathcal{RP}(k) + j = \{i + j \mid i \in \mathcal{RP}(k)\}.$$

For $1 \le k, k' \le t$ and $1 \le j \le L_Q - 1$, set

$$O_Q^j(k, k') = |\mathcal{RP}(k) \cap (\mathcal{RP}(k') + j)|,$$

which is the number of common indices in $\mathcal{RP}(k)$ and $\mathcal{RP}(k') + j$. Note that $O_Q^j(k, k') \ne O_Q^j(k', k)$ for different k' and k in general. For $1 \le k \le t$, define

$$O_Q^j(k) = O_Q^j(k, k) + \sum_{k' < k} O_Q^j(k', k) + \sum_{k' < k} O_Q^j(k, k').$$

Theorem 4.1. *With notations above,*

$$\mu_Q \le \sum_{j=0}^{L_Q - 1} \frac{1}{\prod_{k=1}^{t-1} \left(\sum_{i=1}^{k} p_i \right)^{O_Q^j(k)}}. \tag{4}$$

This is a generalization of a result proved for basic spaced seeds in.[13] Applying it to transition seed, we have the following fact.

Theorem 4.2. *For any transition seed Q,*

$$\mu_Q \le \sum_{j=0}^{L_Q - 1} \frac{1}{p^{O_Q^j(1)} (p + q)^{O_Q^j(2)}} \tag{5}$$

in a Bernoulli sequence model $\mathcal{M}(\{1, 2, 3\}, p, q, r)$.

The bound given above is quite tight when the generation probabilities are large. Consider transition seed $Q = 11\#\#1 * \#11$ in the Bernoulli sequence model $\mathcal{M}(\{1, 2, 3\}, p, q, r)$. Figures 1 shows both the exact μ_Q and its upper bound in Theorem 4.2. As shown in the figure, μ_Q and the bound get closer and closer when one of the generation probabilities goes to large.

4.1.1. *Bounding λ_Q in terms of μ_Q*

In this subsection, we will present lower and upper bounds for the constant λ_Q appeared in Theorem 3.2 in terms of μ_Q. A similar result was proved for basic spaced seeds in.[30]

Theorem 4.3. *For any t-order pattern Q of length L_Q,*

$$1 - \frac{1}{\mu_Q - L_Q + 1} \le \lambda_Q \le 1 - \frac{1}{\mu_Q}. \tag{6}$$

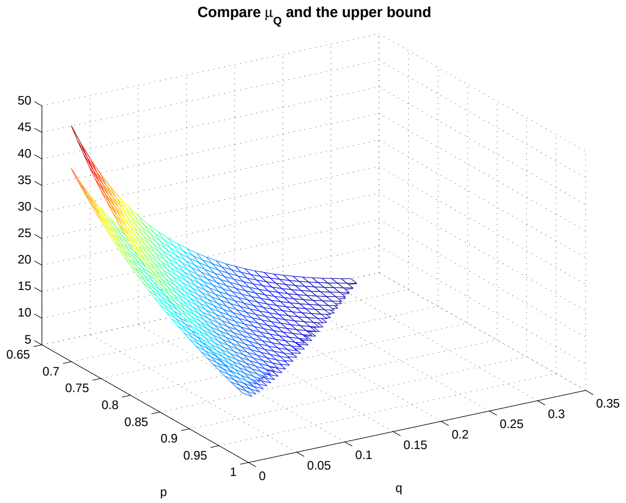

Fig. 1. μ_Q and its upper bound when p from 0.7 to 1 and q from 0 to 0.3 for $Q = 11\#\#1*\#11$.

5. Identifying Good Transition Seeds

Transition seeds exhibit a good tradeoff between sensitivity and specificity for homology search in both coding and non-coding regions.[27,31] However, identifying good transition seeds is a hard task. This is because computing sensitivity is much harder for transition seeds than for basic spaced seeds of the same weight. The weight of a transition seed is defined as its match weight plus the half of its transition weight. By definition, an optimal seed is the seed with the highest sensitivity. In,[17] Kucherov, Noe and Roytberg gave an automata-based method for computing the sensitivity of a basic or transition seed. Such a method takes an exponential number of bit operations in the worst case. Another method for searching good spaced seeds is to use the hill-climbing strategy.[27] Here, based on our theoretical study in the previous sections, we propose an alternative method for the purpose. The efficiency of this method has been demonstrated for basic spaced seed search in[15] and.[29]

Recall that the sensitivity of a spaced seed is defined as the hit probability of the seed in a random sequence of a fixed length L (which is set to 64 traditionally). By Theorem 3.2, the sensitivity of a transition seed Q is closely related to the value of λ_Q. For two transition seeds P and Q, if $\lambda_P < \lambda_Q$, the sensitivity of P is asymptotically larger than Q. Moreover, Theorem 6 indicates that λ_Q can be approximated by a function of μ_Q. Therefore, we propose to identify good transition seeds using the tight bound of μ_Q established in Theorem 4.1. More specifically, we rank the transition seeds Q by the value of $V_Q = \sum_{j=0}^{L_Q-1} p^{-O_Q^j(1)}(p+q)^{-O_Q^j(2)}$. The smaller the value of V_Q is, the higher it is in the ranking list. Given a weight and a Bernoulli model, we identify the ten top transition seeds of the weight and the use the sensitivity in a region of length 64 to select the best one among these ten seeds.

Given a transition seed and a Bernoulli model, the value of V_Q can be simply calculated in a polynomial number of bit operations. Therefore, our heuristic method is much faster than using the sensitivity on a length-64 region to select good transition seeds. In most of cases, our selected seeds are optimal as shown in Table 1 and Table 2.

Table 1. Good transition seeds in Bernoulli model $\mathcal{M}(\{1,2,3\}, 0.7, 0.15, 0.15)$

w	Optimal seeds with w2=2	Sensitivity	Rank	Optimal seeds with w2=4	Sensitivity	Rank
9	111#*1*1**1#*11	0.73745	2	111*#*#1**1#*1#1	0.73806	5
10	111#*1*#*11**1*11	0.60424	8	111#**1#*1*#1*1#1	0.60692	5
11	111*1*1#*1**1#*111	0.47610	1	111#*1*1*#1**1##11	0.48016	1
12	111#*1*1**11*#1*111	0.36368	4	111##*11**1#*1*1#11	0.36692	1
13	111#1**11*1**1#1*111	0.27085	1	111#*1#*11*#*1*1*#111	0.27420	1
14	1111*1*1*#*11*11*#111	0.19760	6	1111#*1*1*#1*#11*#111	0.20077	1
15	1111*#1*1*11**11*1#111	0.14251	3	1111*#11*#1*1#*1*1#111	0.14494	10
16	1111*1*11#*1*11**11#111	0.10165	2	111#11*#*11*1*#11*1#111	0.10360	1
17	11111*#11**11*1*11*1#111	0.07185	3	1111#1*#11*1*1#*11#*1111	0.07333	5

Table 2. Good transition seeds in Bernoulli model $\mathcal{M}(\{1,2,3\}, 0.8, 0.1, 0.1)$

w	Optimal seeds with w2=2	Sensitivity	Rank	Optimal seeds with w2=4	Sensitivity	Rank
9	111#*1*1**1#*11	0.97266	2	111*#*#1**1#*1#1	0.97026	6
10	111*1**1*#1#*111	0.93711	8	11#1**1*1#**1#*#11	0.93405	3
11	111*#*11**1**1*1#11	0.88361	1	111#*1*#*1*#1**1#11	0.88046	1
12	111#*1*1**11*#1*1111	0.81402	4	111#*1*1*#1**1#*#111	0.81037	1
13	111#1**11*1**1#1*111	0.73263	5	111#*1#*11*#*1*1*#111	0.73019	5
14	1111*1*#*11**1*11*#111	0.64523	10	1111*#1**1#*1*1*1#*#111	0.64336	2
15	1111*#1*1*1*11**1#*1111	0.55886	9	111#1*1*1#*1*#1**11*#111	0.55729	6
16	1111*11*#1**11#*1*1*1111	0.47593	7	1111#*1#1**1*#11**1*1#111	0.47507	7
17	11111*1*1#*1*11**11*#1111	0.39955	1	1111#1*1*1#*11**1#1*#1111	0.39915	1

In these two tables, we list the ranks of the optimal transition seeds of weight nine to seventeen and transition weight two or four in model $\mathcal{M}(\{1,2,3\}, 0.7, 0.15, 0.15)$ and $\mathcal{M}(\{1,2,3\}, 0.8, 0.1, 0.1)$, respectively. In all the cases considered, the optimal transition seeds are among the top ten transition seeds selected according to V_Q.

In addition, the best transition seeds reported in Table 1 are identical to those reported in[17] for weight from nine to twelve. Here, we also list good transition seeds for weight from thirteen to seventeen.

6. Conclusion

We have studied the run probabilities of a high-order pattern in the Bernoulli sequence model. Both basic spaced and transition seeds are just order-2 and order-3 patterns respectively. We first establish a recurrence formula for computing the hit probability of a high-order pattern; then, we analyze asymptotically the hit probability. We establish a relationship between the hit probability and the average distance between two non-overlapping hits. For future work, one interesting problem is how to generalize the study to higher-order Markov sequence models.

By applying the theoretical results mentioned above, we present an efficient algorithm for identifying good transition seeds. This algorithm can also be adopted to identify multiple transition seeds.

Finally, we list good transition seeds for six different Bernoulli models. The insight gained from our theoretical study and the list of good transition seeds form a useful resource in guiding the selection of seeds in the developing practical applications.

Acknowledgments

This work was partially supported by a NUS ARF grant 146-000-068-112 and a grant from National Natural Science Foundation of China (NSF30528029).

References

1. S.F. Altschul *et al.*, Basic local alignment search tool. *J. Mol. Biology* **215** (1990), pp. 403-410.
2. N. Balakrishnan and M.V. Koutras, *Runs and Scans with Applications.* John Wiley & Sons, U.S.A. (2002).
3. S. Batzoglou, L. Pachter, J.P. Mesirov, B. Berger, E.S. Lander, Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Research* **10** (2000), pp. 950-958.
4. B. Brejovà, D. Brown, and T. Vinař, Optimal spaced seeds for homologous coding regions. *J. Bioinf. and Comp. Biol.* **1** (2004), pp. 595-610.
5. B. Brejovà, D. Brown, and T. Vinař, Vector seeds: an extension to spaced seeds allows substantial improvements in sensitivity and specificity. *Journal of Computer and System Sciences* **70(3)** (2005), pp. 364-380.
6. M. Brudno, M.A. Chapman, B. Gottgens, S. Batzoglou, B. Morgenstern, Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics* (2003), pp. 4-66.
7. J. Buhler, U. Keich, and Y. Sun, Designing seeds for similarity search in genomic DNA. In *Proc. of RECOMB'03* (2003), pp. 67-75.
8. A. Califano and I. Rigoutsos, FLASH: fast look-up algorithm for string homology, in *Proc. of ISMB'93* (1993), pp. 56-64.
9. K.P. Choi, F. Zeng, and L. Zhang, Good spaced seeds for homology search. *Bioinformatics* **20** (2004), pp. 1053-1059.
10. K.P. Choi, and L. Zhang, Sensitivity analysis and efficient method for identifying optimal spaced seeds. *J. Comput. System Sci.* **68** (2004), pp. 22-40.
11. A.E. Darling *et al.*, Procrastination leads to efficient filtration for local multiple alignment. In *Proc. of WABI'06* (2006), pp. 126-137.
12. W. Feller, *An Introduction to Probability Theory and its Applications. vol. 1. 3rd edition*, John Wiley and Sons, New York (1968).
13. U. Keich, M. Li, B. Ma, and J. Tromp, On spaced seeds for similarity search. *Discrete Appl. Math.* **3** (2004), pp. 253-263.
14. W.J. Kent, BLAT-the BLAST-like alignment tool. *Genome Res.* **12(4)** (2002), pp. 656-664.

15. Y. Kong, Generalized correlation functions and their applications in selection of optimal multiple spaced seeds for homology search. *J. Comp. Biol.* **14** (2007), pp. 238-254.
16. G. Kucherov, L. Noe and M. Roytberg, Multiseed lossless filtration. *IEEE Trans. on Comput. Biol. and Bioinfor.*, **2** (2005), pp. 51-61.
17. G. Kucherov, L. Noe and M. Roytberg, A unifying frame work for seed sensitivity and its application to subset seeds. *INRIA Tech. Report:* N^o *5374* (2004).
18. M. Li and B. Ma, On the complexity of computing the sensitivity of spaced seeds. *J. of Comput. and Sys. Sci.***73(7)**(2007), pp. 1024-1034.
19. M. Li, B. Ma, D. Kisman, and J. Tromp, PatternHunterII: highly sensitivity and fast homogy search. *J. Bioinf. and Comp. Biol.* (2004), pp. 417-440.
20. B. Ma, J. Tromp, and M. Li, PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18** (2002), pp. 440-445.
21. D. Mak, Y. Gelfand, and G. Benson, Indel seeds for homology search. *Bioinformatics* **22** (2006), pp. e341-e349.
22. L. Noe and G. Kucherov, YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Research* **33** (2005), pp. 540-543.
23. F.P. Preparata, L. Zhang, and K.P. Choi, Quick, practical selection of effective seeds for homology search. *Journal of Comput. Biol.* **12** (2005), pp. 1137-1152.
24. S. Schwartz, J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. Hardison, D. Haussler, and W. Miller, Human-mouse alignments with BLASTZ. *Genome Research* **13** (2003), pp. 103-107.
25. A.D. Solov'ev, A combinatorial identity and its application to the problem concerning the first occurences of a rare event. *Theory of Probab. Appl.* **11** (1966), pp. 276-282.
26. Y. Sun, and J. Buhler, Designing multiple simultaneous seeds for DNA similarity search. *Journal of Computational Biology* **12** (2005), pp. 847-861.
27. Y. Sun, and J. Buhler, Choosing the best heuristic for seeded alignment of DNA sequences. *BMC Bioinformatics* **7: 133** (2006).
28. J.B. Xu, D.G. Brown, M. Li, and B. Ma, Optimizing multiple spaced seeds for homology search. In *Proc. of CPM'04* (2004), pp. 47-58.
29. I-H. Yang, S-H. Wang, Y-H. Chen, P-H. Huang, L. Ye, X.Q. Huang, K-M. Chao, Efficient methods for generating optimal single and multiple spaced seeds. In *Proc of BIBE'04* (2004), pp. 411-418.
30. L. Zhang, Superiority of spaced seeds for homology search. *IEEE Trans. Comput. Biology and Bioinformatics* **4** (2007).
31. L. Zhou, L. Florea, Designing sensitive and specific spaced seeds for cross-species mRNA-to-genome alignment. *J. Comput. Biol.* **14** (2007), pp. 113-130.