Large Scale Sequencing By Hybridization

Ron Shamir*
School of Computer Science
Tel-Aviv University
rshamir@tau.ac.il

Dekel Tsur
School of Computer Science
Tel-Aviv University
dekelts@tau.ac.il

ABSTRACT

Sequencing by hybridization is a method for reconstructing a DNA sequence based on its k-mer content. This content, called the *spectrum* of the sequence, can be obtained from hybridization with a universal DNA chip. However, even with a sequencing chip containing all 4^9 9-mers and assuming no hybridization errors, only about 400 bases-long sequences can be reconstructed unambiguously.

Drmanac et al. suggested sequencing long DNA targets by obtaining spectra of many short overlapping fragments of the target, inferring their relative positions along the target and computing spectra of subfragments that are short enough to be uniquely recoverable. Drmanac et al. do not treat the realistic case of errors in the hybridization process. In this paper we study the effect of such errors. We show that the probability of ambiguous reconstruction in the presence of (false negative) errors is close to the probability in the errorless case. More precisely, the ratio between these probabilities is $1 + O(p/(1-p)^4 \cdot 1/d)$ where d is the average distance between neighboring subfragments, and p is the probability of a false negative.

We also obtain lower and upper bounds for the probability of unambiguous reconstruction based on errorless spectrum. For realistic chip sizes, these bounds are tighter than those given by Arratia et al. Finally, we report results on simulations with real DNA sequences, showing that even in the presence of 50% false negative errors, a target of cosmid length can be recovered with less than 0.1% miscalled bases.

1. INTRODUCTION

One of the main current endeavors in Life Sciences and Medicine is efficient sequencing of very long DNA molecules. The prevalent sequencing technologies are currently gel-based. Sequencing by hybridization (SBH) [3, 10] was proposed in the late Eighties as an alternative way to DNA sequencing. In this method, the target sequence is hybridized to a universal chip containing all 4^k sequences of length k. For each k-long sequence (or probe) in the chip, if its reverse complement appears in the target, then the two sequences will bind (or hybridize), and this hybridization can be detected. Thus, from the above experiment one can obtain the multi-set of all k-long subsequences of a target sequence (all subsequences referred to in this paper will be contiguous). This multi-set is called the *k-spectrum* of the target, or simply its *spectrum*. We note that in reality, only the *set* of all k-subsequences of the target can be obtained, but like most previous studies, we will assume here that the multi-set is known. Pevzner has shown that reconstructing a sequence from its spectrum is polynomial [13]. Since copies of the universal chip can be economically produced and used to sequence any DNA target, and the computational reconstruction task is efficiently handled, this seems as a promising alternative to standard sequencing techniques.

Unfortunately, often sequence reconstruction is not unique: Other sequences can have the same spectrum as the target's. Therefore, we are interested in telling how likely this event is as a function of the target length. We say that a sequence is *uniquely recoverable* from its spectrum if there is no other sequence with the same spectrum. By selecting a distribution on the sequences of a certain length, one can compute the probability that a random sequence is not uniquely recoverable from its k-spectrum. We refer to this as the *failure probability*. Denote the failure probability for the uniform distribution over N-long sequences by $P(n, k)$, where $n = N - k + 1$ (n is the number of k-tuples in the sequence). An asymptotically exact formula for $P(n, k)$ was given by Arratia et al. [2].

The main shortcoming of classical SBH is ambiguous solutions: The maximum length $n = n(k, \epsilon)$ for which $P(n, k) \leq \epsilon$, is rather small. For example, $n(8, 0.1)$ is about 200 [12]. Several methods for overcoming this limitation were proposed: alternative chip designs [3, 9, 12, 15], interactive protocols [16], using location information [1, 4, 6, 8], and using a known homologous sequence [11]. Currently, SBH is not considered competitive in comparison with standard gel-based sequencing technologies.

Drmanac et al. [7] proposed the following enhancement to SBH (compare Figure 1): Instead of reconstructing a single target sequence from its spectrum, one can obtain the spectra of many short overlapping fragments (clones) of the target. The larger the overlap between two clones, the more similar their spectra would be. Using this similarity

*Supported in part by grants from the Israeli Science Foundation of the Israeli Academy for the Sciences and the Arts, and from the US-Israel Binational Science Foundation.

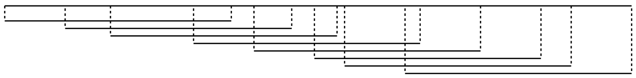

Figure 1: Partitioning a target sequence into IFs. The target (top) is partitioned into IFs by the endpoints of the clones (bottom). There are 8 clones and 15 IFs.

one can infer the position of clones along the target. Moreover, the endpoints of the clones induce a partition of the target into even shorter subfragments, and the spectrum of each of those can be computationally inferred. The DNA stretches between consecutive clone endpoints are called *information subfragments* (IFs). By obtaining clones at high redundancy, the average length of an IF would be short enough to be uniquely recoverable from its spectrum with high probability. When all IFs are uniquely recoverable, so is the complete target. Drmanac et al. suggested sequencing a 10^6 bp long target by obtaining the spectra of 25,000 500 bp-long clones. They provided some simple computational arguments and performed error-free simulations to support their suggestion.

In this paper we shall expand on the analysis of the Drmanac strategy and of Arratia et al. in several ways. First, we improve the estimate of Arratia et al. on $P(n,k)$ for small (and realistic) values of k: We show that if $5(k-1) \leq n \leq 2^{k+1} + 4(k-2)$, then $P(n,k) = \Theta(n^4/4^{2k})$. We then use this result in order to investigate the Drmanac et al. strategy in the presence of hybridization errors: Under some simplifying assumptions which will be described below, we prove that the introduction of false negative errors has a very small effect on the probability of unique recoverability. More precisely, the ratio between the failure probability in the presence of errors, and in the errorless case, is $1 + O(p/(1-p)^4 \cdot 1/d)$, where d is the average IF length, and p is the probability of a false negative error. We also perform simulations with real DNA sequences which show that the technique can reconstruct a target longer than 30kb from spectra containing 50% false negative errors, with an average of less than one miscalled base in 1000 bases.

We need some notation in order to specify our assumptions: Denote the target sequence by A. One first clones many short random subfragments $C_1, C_2, \ldots, C_c$ of A, and obtains the k-spectrum of each clone. We assume that the clones completely cover A. The endpoints of the c clones form a partition of the target A into IFs $J_1, J_2, \ldots, J_l$ where $l \leq 2c-1$. We assume that the positions of the clones along the target have already been inferred from the spectra. In the absence of errors, if a k-tuple P is located in the IF J_i, then P will appear in the k-spectrum of each clone C_j that contains J_i. Using this observation, one can compute the spectrum of each IF.

Short probe hybridizations are error prone. In *false positive* errors, a certain k-tuple appears in the experimental spectrum while in fact it does not appear in the target. The converse occurs in a *false negative* error. False positive errors are quite easy to correct (cf. [5]). Moreover, by increasing the hybridization stringency, the number of false positives can be decreased, at the expense of increasing the false negatives rate, which as we shall show, has little effect on the success of the strategy. Therefore, we will assume that there are no false positive errors.

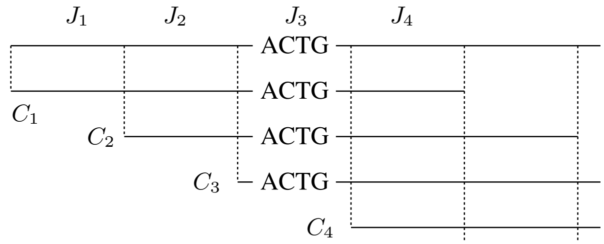

Figure 2: An example of attributing k-tuples to IFs. If there are no errors, then the 4-tuple ACTG appears in clones C_1, C_2, C_3, so we attribute it to IF J_3. If there are false negatives, ACTG can appear, for example, only in C_1, C_2, and therefore it will be mistakenly attributed to IF J_2. As the mistake is only in one direction, we know in this case that the k-tuple ACTG appears in $\cup_{i \geq 2} J_i$.

For false negative errors, our probabilistic model assumes that each k-tuple contained in some clone does not appear in its (experimental) spectrum with probability p, independently of the other k-tuples in the clone and of its appearance in other clones. False negative errors have two effects. First, it is possible that some k-tuple P in the target sequence of A will not appear in the spectra of any of the clones that contain P. Therefore, P will be missing from the spectrum which is built for A. As we assume that the clones cover the target with high redundancy, this effect has very low probability. The second type of effect is that the k-tuple P may appear in some, but not all, the spectra of clones that contain it. Thus, when we decide which IF contains P, we may be wrong. This is the effect that we shall study in detail.

For the theoretical analysis we make several simplifying assumptions: First, we assume that the clone positions are not random, but are spread uniformly across the sequence, and we denote by d the distance between the left endpoints of two consecutive clones. Second, for the partition into IFs, we ignore the right endpoints of the clones, and only consider the partition of the sequence A derived from the left endpoints of the clones. (Note that this assumption generates longer and fewer IFs than there really are, so removing this assumption would only improve the results). Technically, to achieve this we assume that the length of the clones is divided by d. This partition forms c IFs $J_1, \ldots, J_c$. When we consider some k-tuple P in the spectrum of A, we attribute P to the fragment J_i where i is the maximum index of a clone in which P appears in its spectrum (for simplicity of representation we assumed here that P appears only in one IF). Since we assumed only false negative errors, the index i is always less than or equal to the correct index i'. So in the case of errors, instead of knowing that P is in fragment $J_{i'}$, we know that P is in the union $\cup_{j \geq i} J_j$. Moreover, the value of $i' - i$ is a random variable with geometric distribution with parameter p. See Figure 2 for an example of the situation described above.

The paper is organized as follows: Section 2 contains problem definitions and preliminaries. In Section 3 we give the upper and lower bounds on the failure probability with no errors. The main result of this paper is given in Section 4. Finally, our simulations are described in Section 5.

2. PRELIMINARIES

For a sequence $A = a_1 \cdots a_r$, denote by $A|_i^l$ the l-subsequence $a_i a_{i+1} \cdots a_{i+l-1}$. For two sequences $A = a_1 \cdots a_r$ and $B = b_1 \cdots b_s$, we denote $A|B$ if the $(s-1)$-suffix of A is equal to the $(s-1)$-prefix of B. If $A|B$, we denote by $A \diamond B$ the sequence $a_1 \cdots a_r b_s$.

The k-spectrum of a sequence A of length N is the multi-set of all the k-subsequences of A and it is denoted by $\mathrm{SP}^k(A)$. The SBH problem is: Given a multi-set M of k-tuples, find a sequences A for which $\mathrm{SP}^k(A) = M$, if there is such a sequence. A sequence A is called *uniquely recoverable w.r.t.* k if there is no sequence $A' \neq A$ such that $\mathrm{SP}^k(A) = \mathrm{SP}^k(A')$.

An alternative way to define the SBH problem is as follows: Given a multi-set $M = \{A_1, \ldots, A_n\}$ of k-tuples, find a permutations $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ for which $A_{\pi(i)}|A_{\pi(i+1)}$ for all $i < n$. For such a permutation π, define the sequence $A^\pi = A_{\pi(1)} \diamond A_{\pi(2)} \diamond \cdots \diamond A_{\pi(n)}$, and note that $\mathrm{SP}^k(A^\pi) = M$. Pevzner [13] gave a formulation of the SBH problem using graphs: For a multi-set M, define the *de-Bruijn* graph G_M whose vertices are all the unique $(k-1)$-tuples that appear in M, i.e. $\cup_{i=1}^n \{ A_i|_1^{k-1}, A_i|_2^{k-1}\}$, and whose edges are $e_i = (A_i|_1^{k-1}, A_i|_2^{k-1})$ for $1 \leq i \leq n$. A permutation π is a solution to the SBH problem iff $P_\pi = [e_{\pi(1)}, e_{\pi(2)}, \ldots, e_{\pi(n)}]$ is an Eulerian path in G_M. Therefore, the SBH problem is polynomial.

The *Positional SBH problem* (PSBH) is defined as follows: Given a multi-set $M = \{A_1, \ldots, A_n\}$ and sets $S(i) \subseteq \{1, \ldots, n\}$ for all $i \leq n$, find a permutation π for which $A_{\pi(i)}|A_{\pi(i+1)}$ for all $i < n$, and $i \in S(\pi(i))$ for all $i \leq n$. Such a permutation π is called a *solution* of the instance M, k, S. The PSBH problem is NP-hard even when $|S(i)| \leq 3$ for all i [4], while it is polynomial when $|S(i)| \leq 2$ for all i [4], or when each $S(i)$ is an interval of length $O(\log n)$ [8].

It is more convenient to denote an instance of PSBH by A, k, S where A is a sequence of length $N = n+k-1$, which naturally defines the multi-set $M = \mathrm{SP}^k(A)$. A solution π of A, k, S is called *trivial* if $A^\pi = A$. A sequence A is called *uniquely recoverable w.r.t.* k, S if there is no nontrivial solution π of the instance A, k, S.

In this paper, we consider two special cases of the Positional SBH problem. Let $I = \{J_1, J_2, \ldots, J_c\}$ be a partition of the interval $[1, n]$ into disjoint intervals (i.e., there are integers $s_1 = 1, s_2, \ldots, s_c, s_{c+1} = n + 1$ such that $J_i = [s_i, s_{i+1} - 1]$). For $i \leq n$, let $I(i)$ be the index such that $i \in J_{I(i)}$. An instance of the *Interval PSBH problem* (IPSBH), denoted by A, k, I, is an instance A, k, S_I of PSBH, where $S_I(i) = J_{I(i)}$ for $i \leq n$. A sequence A is called *uniquely recoverable w.r.t.* k, I if there is no nontrivial solution of the instance A, k, I. The instance A, k, S_I is equivalent to the instance A, k, S'_I, where $S'_I(i) = [s_{I(i)}, n]$. (For proof, suppose that π is a solution of A, k, S'_I. For $i \in [1, s_1 - 1]$, we have $i \in S'_I(\pi(i)) = [s_{I(\pi(i))}, n]$, hence $I(\pi(i)) = 1$ which implies that $\pi(i) \in [1, s_1 - 1]$. Now, for $j \in [s_1, s_2 - 1]$, we have $j \in [s_{I(\pi(j))}, n]$, so $I(\pi(j)) \leq 2$. But since $\pi(i) \in [1, s_1 - 1]$ for all $i \in [1, s_1 - 1]$, it follows that $\pi(i) \geq s_1$. Therefore, $\pi(i) \in [s_1, s_2 - 1]$. By repeating the same argument, we conclude that $\pi(i) \in [s_{I(i)}, s_{I(i)+1} - 1]$ for all i, or in other words, $I(\pi(i)) = I(i)$ for all i. Therefore π is a solution of A, k, S'_I. Conversely, if π is a solution of A, k, S_I then it is also a solution of A, k, S'_I as $S_I(i) \subseteq S'_I(i)$ for all i.)

Let $\Delta = (\Delta_1, \ldots, \Delta_n)$ be a vector of nonnegative integers. An instance of the *Inexact Interval PSBH problem* (IIPSBH), denoted by A, k, I, Δ, is an instance $A, k, S_{I,\Delta}$ of PSBH, where $S_{I,\Delta}(i) = [s_{\max(I(i)-\Delta_i, 1)}, n]$. A sequence A is called *uniquely recoverable w.r.t.* k, I, Δ if there is no nontrivial solution of the instance A, k, I, Δ.

For the rest of this paper we assume that $n = cd$ for some integers c and d. Let I_d be the set of intervals $\{[1, d], [d+1, 2d], \ldots, [(c-1)d+1, cd]\}$ (so $I(i) = \lceil i/d \rceil$). We denote by $P(n, k, d)$ the probability that for a random sequence A of length $n + k - 1$, A is not uniquely recoverable w.r.t. k, I_d. We also denote by $P(n, k, d, p)$ the probability that for a random sequence A of length $n + k - 1$ and for a vector $\Delta = (\Delta_1, \ldots, \Delta_n)$ of independent identically distributed random variables with geometric distribution with parameter p, A is not uniquely recoverable w.r.t. k, I_d, Δ. The main result of this paper is showing that $P(n, k, d, p) = (1 + O(1/d))P(n, k, d)$.

Let A be a sequence of length $N = n+k-1$. A pair (i, j) is called a *repeat* if $A|_i^{k-1} = A|_j^{k-1}$. A repeat (i, j) is called *rightmost* if $j \neq n+1$ and $(i+1, j+1)$ is not a repeat (i.e., if $a_{i+k-1} \neq a_{j+k-1}$). A repeat (i, j) is called *weakly rightmost* if either $j = dI(i) + 1$, or (i, j) is rightmost. For example, let $k = 4$, $d = n = 11$ and $A =$ AGCTT ACGCT TCTT. The repeats of A are $(2, 8)$, $(3, 9)$, $(9, 12)$, $(3, 12)$, the weakly rightmost repeats are $(3, 9)$, $(9, 12)$, $(3, 12)$, and the single rightmost repeat is $(3, 9)$.

A pair of repeats $((i, j), (i', j'))$ is called *R-pair* if (i, j) is rightmost, and it is called *Rr-pair* if (i, j) is rightmost, and (i', j') is weakly rightmost. The pair is called *interleaved* if $i \leq i' < j < j'$ and $I(i) = I(j' - 1)$.

3. ESTIMATING THE FAILURE PROBABILITY

In this section we show that $P(n, k, d) = \Theta(d^3 n / 4^{2k})$. We note that an asymptotically exact formula for $P(n, k)$ was given in [2], but it does not give a good estimate for small values of n and k.

We begin with showing an upper bound on $P(n, k, d)$. A necessary and sufficient condition for unique recoverability w.r.t. k was given by Arratia et al. [2] (based on result from [14]). In the following theorem, we generalize this result for unique recoverability w.r.t. k, I_d. We also note that our characterization is simpler than the one in [2].

Theorem 1. *A sequence A is not uniquely recoverable w.r.t. k, I_d iff either (1) A contains an interleaved R-pair, or (2) $A|_1^{k-1} = A|_{d+1}^{k-1} = \cdots = A|_{cd+1}^{k-1}$ and there are indices $i_1, i_2, \ldots i_c$ with $(l-1)d + 1 < i_l < ld + 1$, and $A|_{i_1}^{k-1} = A|_{i_2}^{k-1} = \cdots = A|_{i_c}^{k-1} \neq A|_1^{k-1}$.*

Proof. If $((i, j), (i', j'))$ is an interleaved R-pair, then we have $A|_{i-1}^k \mid A|_j^k$ (if $i > 1$), $A|_{j-1}^k \mid A|_i^k$, $A|_{i'-1}^k \mid A|_{j'}^k$, and $A|_{j'-1}^k \mid A|_{i'}^k$. Thus, we define a permutation π as follows:

$$\pi(1), \ldots, \pi(n) = 1, 2, \ldots, i-1, j, j+1, \ldots, j'-1, i', i'+1, \\ \ldots, j-1, i, i+1, \ldots, i'-1, j', j'+1, \ldots, n$$

and it is easy to verify that π is a solution of A, k, I_d. Furthermore, $A|_i^k \neq A|_j^k = A^\pi|_i^k$ (as (i, j) is a rightmost repeat), so $A \neq A^\pi$. Therefore, A is not uniquely recoverable w.r.t. k, I.

Suppose that A satisfies case (2) of the theorem with indices $i_1, i_2, \dots i_c$. We define a permutation π as follows:

$$\pi(1), \dots, \pi(n) = i_1, i_1+1, \dots, d, 1, 2, \dots, i_1-1, i_2, i_2+1, \dots, 2d, d+1, d+2, \dots, i_2-1, i_3, \dots, i_c-1.$$

The conditions of case (2) implies that π is a solution of A, k, I_d and $A \neq A^\pi$.

We now prove the second direction of the theorem. Suppose that A is not uniquely recoverable w.r.t. k, I_d, and let π be a nontrivial solution. To simplify the proof, we use the de-Bruijn graph $G = (V, E)$ of $\mathrm{SP}^k(A)$. Denote $E = \{e_1, \dots, e_n\}$ where $e_i = (A|_i^{k-1}, A|_{i+1}^{k-1})$. For two edges e_i, e_j we write $e_i \equiv e_j$ if e_i and e_j are parallel edges, namely if $A|_i^k = A|_j^k$. Clearly, both $P = [e_1, e_2, \dots, e_n]$ and $P' = [e_{\pi(1)}, e_{\pi(2)}, \dots, e_{\pi(n)}]$ are Eulerian paths in G. We also define subgraphs $G_l = (V, E_l)$ of G, where $E_l = \{e_{(l-1)d+1}, \dots, e_{ld}\}$. Again, both $P_l = [e_{(l-1)d+1}, \dots, e_{ld}]$ and $P'_l = [e_{\pi((l-1)d+1)}, \dots, e_{\pi(ld)}]$ are Eulerian paths in G_l.

The proof is based on comparing the paths P and P'. We will show that if the paths P and P' start from the same vertex (i.e. $A^\pi|_1^{k-1} = A|_1^{k-1}$) then case (1) happens, and otherwise case (2) happens.

We first consider the case when P and P' start from the same vertex. Let i be the minimum index such that $e_i \not\equiv e_{\pi(i)}$ and let $j = \pi(i)$. W.l.o.g. we assume that $e_l = e_{\pi(l)}$ for all $l < j$. Thus $\pi(i) \notin \{1, \dots, i\}$, so $i < j$. From the minimality of i, and the assumption that P and P' start from the same vertex, it follows that the edges e_i and e_j has the same start vertex, so $A|_i^{k-1} = A|_j^{k-1}$. As $e_i \not\equiv e_j$, the edges e_i and e_j has different end vertices, hence $A|_{i+1}^{k-1} \neq A|_{j+1}^{k-1}$. Therefore, (i, j) is a rightmost repeat.

As π is a solution we have that $I(i) = I(j)$. The edges $e_i, e_{i+1}, \dots, e_{j-1}$ form a cycle in $G_{I(i)}$, which we denote by C, and we denote by V_C the vertices of C. As $P'_{I(i)}$ is an Eulerian path, it must pass through all the edges of C, and let l be the minimum index such that $e_{\pi(l)}$ is in C. By definition, the edge $e_{\pi(i)} = e_j$ is not in C (though it can be parallel to an edge from C), and therefore, $l > i$. Clearly, $i \leq \pi(l) \leq j-1$, $\pi(l-1) \geq j$ (as the edge $e_{\pi(l-1)}$ is not in C), and $I(\pi(l-1)) = I(i)$ (as the edge $e_{\pi(l-1)}$ is in $G_{I(i)}$). Since the end vertex of $e_{\pi(l-1)}$ is equal to the start vertex of $e_{\pi(l)}$, we have that $A|_{\pi(l)}^{k-1} = A|_{\pi(l-1)+1}^{k-1}$, namely $(\pi(l), \pi(l-1)+1)$ is a repeat. Therefore, $((i,j), (\pi(l), \pi(l-1)+1))$ is an interleaved R-pair.

We now consider the case when P and P' do not start from the same vertex (P starts from $A|_1^{k-1}$, and P' starts from $A^\pi|_1^{k-1}$). Since P'_1 is an Eulerian path in G_1 which doesn't start from vertex $A|_1^{k-1}$, it follows that the vertex $A|_1^{k-1}$ has an even degree in G_1. Since P_1 is an Eulerian path in G_1 which starts from vertex $A|_1^{k-1}$, it must also end in $A|_1^{k-1}$, hence $A|_1^{k-1} = A|_{d+1}^{k-1}$. With similar arguments, we obtain that vertex $A^\pi|_1^{k-1}$ has even degree in G_1 (as P_1 doesn't start from $A^\pi|_1^{k-1}$), and therefore P'_1 ends in $A^\pi|_1^{k-1}$. Thus, $A^\pi|_1^{k-1} = A^\pi|_{d+1}^{k-1}$.

Now, P'_2 is an Eulerian path in G_2 which doesn't start from vertex $A|_{d+1}^{k-1}$ (as $A^\pi|_{d+1}^{k-1} = A^\pi|_1^{k-1} \neq A|_1^{k-1} = A|_{d+1}^{k-1}$) and therefore vertex $A|_{d+1}^{k-1}$ has even degree in G_2. Again, it follows that $A|_{d+1}^{k-1} = A|_{2d+1}^{k-1}$. We also obtain that $A^\pi|_{d+1}^{k-1}$ has even degree in G_2 and $A^\pi|_{d+1}^{k-1} = A^\pi|_{2d+1}^{k-1}$.

By repeating the same arguments, we obtain that $A|_1^{k-1} = A|_{d+1}^{k-1} = \dots = A|_{cd+1}^{k-1}$ and $A^\pi|_1^{k-1} = A^\pi|_{d+1}^{k-1} = \dots = A^\pi|_{cd+1}^{k-1}$. We therefore define $i_l = \pi((l-1)d+1)$ for $l = 1, \dots, c$, and A satisfies the conditions of case (2) with the indices $i_1, \dots, i_c$. $\square$

For bounding the failure probability, we will use a slightly different characterization:

Theorem 2. *A sequence A is not uniquely recoverable w.r.t. k, I_d iff either A contains an interleaved Rr-pair, or case (2) of Theorem 1 happens.*

Proof. It suffices to prove that if A contains an interleaved R-pair, then it also contains an interleaved Rr-pair. Let $((i,j), (i', j'))$ be an interleaved R-pair. If (i', j') is weakly rightmost, then we are done. Otherwise, $j' \neq dI(i')+1 = dI(i)+1$ and (i', j') is not rightmost. Since $I(j'-1) = I(i)$, it follows that $I(j') = I(i)$. As (i', j') is not rightmost and $j' \neq n+1$, we have that $(i'+1, j'+1)$ is a repeat. If $i' < j-1$ then denote $i_2 = i$, $j_2 = j$, $i'_2 = i'+1$, and $j'_2 = j'+1$. If $i = j'-1$ then denote $i_2 = i$, $j_2 = j$, $i'_2 = i$, and $j'_2 = j'+1$ (note that (i'_2, j'_2) is a repeat as (i, j) and $(j, j'+1)$ are repeats). In both cases, $((i_2, j_2), (i'_2, j'_2))$ is an interleaved R-pair. We repeat this process until we reach a pair $((i_r, j_r), (i'_r, j'_r))$ which is an interleaved Rr-pair. $\square$

By Theorem 2, $P(n, k, d)$ is less or equal to the probability that there is an interleaved Rr-pair, plus the probability that case (2) happens. The latter probability is less than $1/4^{(k-1)n/d}$. Denote by $P_{i,i',j,j'}$ the probability that $((i,j), (i', j'))$ is an interleaved Rr-pair.

Lemma 3. *For all $i \leq i' < j < j'$, $P_{i,i',j,j'} \in \{0, 9/4^{2k}\}$ for $j' < dI(i)+1$, and $P_{i,i',j,j'} \in \{0, 12/4^{2k}\}$ for $j' = dI(i)+1$.*

Proof. We only prove the case $j < dI(i)+1$, as the proof of the second case is similar. Denote by $a_1, \dots, a_N$ the letters of the sequence A. By definition,

$$P_{i,i',j,j'} = \Pr\left[\begin{array}{cc} a_{i+t} = a_{j+t} & t = 0, \dots, k-2 \\ a_{i'+t} = a_{j'+t} & t = 0, \dots, k-2 \\ \multicolumn{2}{c}{a_{i+k-1} \neq a_{j+k-1}, a_{i'+k-1} \neq a_{j'+k-1}} \end{array}\right].$$

For the proof of the lemma we build a graph $G_{i,i',j,j'}$. The vertices of $G_{i,i',j,j'}$ are the indices of the letters that appear in the above equalities and inequalities, and the edges correspond to the equalities. Formally, the vertices of $G_{i,i',j,j'}$ are $\{i, \dots, i+k-1\} \cup \{j, \dots, j+k-1\} \cup \{i', \dots, i'+k-1\} \cup \{j', \dots, j'+k-1\}$ and its edges are $\{(i+r, j+r) | 0 \leq r \leq k-2\} \cup \{(i'+r, j'+r) | 0 \leq r \leq k-2\}$. Let $C_1, C_2, \dots, C_c$ be the connected components of $G_{i,i',j,j'}$, and let v_l, e_l denote the number of vertices and edges in C_l, respectively. The pairs (i,j) and (i', j') are repeats iff for each connected component C_l, all the corresponding letters in A are equal. This probability is exactly $\prod_{l=1}^{c} (1/4)^{v_l - 1}$.

We consider three cases. In case 1, we assume that the number of edges in $G_{i,i',j,j'}$ is less than $2(k-1)$, which implies that $\{(i+r, j+r) | 0 \leq r \leq k-2\} \cap \{(i'+r, j'+r) | 0 \leq r \leq k-2\} \neq \phi$, and therefore $i' - i = j' - j = r$ for some $0 \leq r \leq k-2$. A repeat at $(i', j') = (i+r, j+r)$ implies that $a_{i+r+l} = a_{j+r+l}$ for all $l < k-1$, and in particular, for $l = k-1-r$ we get $a_{i+k-1} = a_{j+k-1}$. But a rightmost repeat at (i,j) implies that $a_{i+k-1} \neq a_{j+k-1}$. Thus, $((i,j),(i',j'))$ can not be an interleaved Rr-pair, so $P_{i,i',j,j'} = 0$.

Denote by $G'_{i,i',j,j'}$ the graph obtained from $G_{i,i',j,j'}$ by adding the edges $e_1 = (i+k-1, j+k-1)$ and $e_2 = (i'+k-1, j'+k-1)$ (corresponding to the two inequalities). For case 2, assume that $G'_{i,i',j,j'}$ has no cycles, and therefore $G_{i,i',j,j'}$ has no cycles, so $e_l = v_l - 1$ for any l. Therefore, the probability that (i,j) and (i',j') are repeats is $\prod_{l=1}^{c} 1/4^{e_l} = 1/4^{\sum_{l=1}^{c} e_l} = 1/4^{2(k-1)}$ where the last equality follows from the fact that $G_{i,i',j,j'}$ has $2(k-1)$ edges. Furthermore, as the edges e_1 and e_2 do not create a cycle, it follows that $P_{i,i',j,j'} = (1/4)^{2(k-1)}(3/4)^2 = 9/4^{2k}$.

Now, for case 3, suppose that $G'_{i,i',j,j'}$ contains a cycle. Note that a cycle in $G'_{i,i',j,j'}$ cannot pass through e_2 as the vertex $j'+k-1$ has only one neighbor (the vertex $i'+k-1$). We claim that $G'_{i,i',j,j'}$ contains a cycle that passes through e_1. Let $C = [v_1, v_2, \ldots, v_{l-1}, v_l = v_1]$ be some cycle in $G'_{i,i',j,j'}$. If C passes through e_1 we are done. Otherwise, for any edge $e = (v_i, v_{i+1})$ in C, as $e \neq e_1, e_2$, then $(v_i + 1, v_{i+1} + 1)$ is also an edge in $G'_{i,i',j,j'}$. Therefore, $C' = [v_1+1, v_2+1, \ldots, v_l+1]$ is also a cycle in $G'_{i,i',j,j'}$. We repeat this process until we obtain a cycle that passes through e_1 (and doesn't pass through e_2). Therefore the vertices $i+k-1$ and $j+k-1$ are in the same connected component of $G_{i,i',j,j'}$ which implies that $((i,j),(i',j'))$ can not be an interleaved Rr-pair. Thus, $P_{i,i',j,j'} = 0$. $\square$

Note that the equivalent of Lemma 3 is not actually proved in [2]. Arratia et al. proved the bound on $P_{i,i',j,j'}$ provided that $\max(j'-j, j-i', i'-i) \geq k$, and used computer computations to bound the other cases. Our proof of the first two cases is similar to theirs, while the third case is new.

Corollary 4. *$P(n,k,d) \leq (\frac{3}{8}d^3 + \frac{5}{4}d^2)n\frac{1}{4^{2k}} + 1/4^{(k-1)n/d}$.*

Proof. There are $\frac{n}{d}\binom{d}{4} + \frac{n}{d}\binom{d}{3} = \frac{n}{d}\binom{d+1}{4}$ ways to choose the indices i, i', j, j' with $j' < dI(i) + 1$ (the first term is the number of ways with $i < i'$, and the second term is the number of ways with $i = i'$), and $\frac{n}{d}\binom{d}{3} + \frac{n}{d}\binom{d}{2} = \frac{n}{d}\binom{d+1}{3}$ ways to choose them with $j' = dI(i) + 1$. By Lemma 3, the probability that there is an interleaved Rr-pair is at most

$$\begin{aligned} &\frac{n}{d}\binom{d+1}{4}\frac{9}{4^{2k}} + \frac{n}{d}\binom{d+1}{3}\frac{12}{4^{2k}} \\ &\qquad\leq \frac{9}{4!}n(d^3 - 2d^2)\frac{1}{4^{2k}} + \frac{12}{3!}nd^2\frac{1}{4^{2k}} \\ &\qquad\leq (\frac{3}{8}d^3 + \frac{5}{4}d^2)n\frac{1}{4^{2k}}. \quad \square \end{aligned}$$

We now give a lower bound on $P(n,k,d)$. Denote $D = \lfloor \frac{d}{4} \rfloor$.

Lemma 5. *If $d \geq 4k$ then $P(n,k,d) \geq L(n,k,d)(1 - L(n,k,d)/2)$ where*

$$L(n,k,d) = \frac{n}{d}(D-k+1)^4 \frac{9}{4^{2k}}\left(1 - (D-k+1)^2 \frac{3}{4^k}\right)^2.$$

Proof. For $0 \leq r < n/d$, let $I_{r,1} = [rd+1, rd+D-k+1] \times [rd+d/2+1, rd+3D-k+1]$ and $I_{r,2} = [rd+D+1, rd+2D-k+1] \times [rd+3D+1, rd+4D-k+1]$. Denote by X_r the event that there is an interleaved Rr-repeat $((i,j),(i',j'))$ for some indices i, i', j, j' with $(i,j) \in I_{r,1}$ and $(i',j') \in I_{r,2}$. By Theorem 1, $P(n,k,d) \geq \Pr\left[\bigvee_{r=0}^{n/d-1} X_r\right]$. Clearly, the events $X_0, \ldots, X_{n/d-1}$ are independent and have equal probabilities, so

$$\begin{aligned} \Pr\left[\bigvee_{r=0}^{n/d-1} X_r\right] &= 1 - (1 - \Pr[X_1])^{n/d} \\ &\geq \frac{n}{d}\Pr[X_1](1 - \frac{n}{2d}\Pr[X_1]). \end{aligned}$$

We will now bound $P[X_1]$. Let Z be the event that there is $(i,j) \in I_{1,1}$ such that (i,j) is a rightmost repeat, and let Z' be the event that there is $(i',j') \in I_{1,2}$ which is a rightmost repeat. The events Z and Z' are independent and have equal probabilities. Thus, $\Pr[X_1] = \Pr[Z \wedge Z'] = \Pr[Z]^2$. For a pair $\alpha = (i,j) \in I_{1,1}$, denote by Z_α the event that (i,j) is rightmost repeat, and let $Y_\alpha = Z_\alpha \wedge \bigwedge_{\alpha \neq \beta \in I_{1,1}} \overline{Z_\beta}$. The events $\{Y_\alpha\}_{\alpha \in I_{1,1}}$ are disjoint, so $\Pr[Z] \geq \Pr\left[\bigvee_{\alpha \in I_1} Y_\alpha\right] = \sum_{\alpha \in I_1} \Pr[Y_\alpha]$, and

$$\begin{aligned} \Pr[Y_\alpha] &= \Pr\left[Z_\alpha \wedge \bigwedge_{\alpha \neq \beta \in I_{1,1}} \overline{Z_\beta}\right] \\ &= \Pr[Z_\alpha]\Pr\left[\bigwedge_{\alpha \neq \beta \in I_{1,1}} \overline{Z_\beta} \middle| Z_\alpha\right] \\ &= \Pr[Z_\alpha]\left(1 - \Pr\left[\bigvee_{\alpha \neq \beta \in I_{1,1}} Z_\beta \middle| Z_\alpha\right]\right) \\ &\geq \Pr[Z_\alpha]\left(1 - \sum_{\alpha \neq \beta \in I_{1,1}} \Pr[Z_\beta | Z_\alpha]\right). \end{aligned}$$

As $j - i \geq k$, we have from [2, p. 437] that $\Pr[Z_\beta | Z_\alpha] \in \{0, 3/4^k\}$, for all α and β. Thus, $P[Y_{i,j}] \geq (3/4^k)(1 - (D-k+1)^2 \cdot 3/4^k)$. Therefore, $P(n,k,d) \geq L(n,k,d)(1 - L(n,k,d)/2)$. $\square$

Corollary 6. *If $5(k-1) \leq d \leq 2^{k+1}/c^{1/4} + 4(k-2)$ then $P(n,k,d) = \Omega(d^3 n/4^{2k})$.*

Proof. The bounds on d imply that $D - k + 1 \geq d/4 - d/5 = d/20$ and $L(n,k,d) \leq c(d/4 - k + 2)^4 \cdot (9/4^{2k}) \leq c(2^{k-1}/c^{1/4})^4 \cdot (9/4^{2k}) = 9/16$. By Lemma 5,

$$\begin{aligned} P(n,k,d) &\geq \frac{23}{32}L(n,k,d) \\ &\geq \frac{23}{32}\frac{n}{d}(d/20)^4 \cdot (9/4^{2k}) \cdot (1 - \sqrt{9/16})^2 \\ &= \Omega(d^3 n/4^{2k}). \quad \square \end{aligned}$$

Our results are valid for the classical SBH model with no subintervals, by taking $d = n$. The resulting bounds on $P(n,k)$ improve over Arratia et al. for small (realistic) values of k. See Table 1 for a comparison between the results.

4. ESTIMATING THE FAILURE PROBABILITY IN THE PRESENCE OF ERRORS

Denote by X the event that a random sequence is uniquely recoverable w.r.t. k, I_d but is not uniquely recoverable w.r.t. k, I_d, Δ. In this section we show that $\Pr[X] = O(p/(1-p)^4 \cdot nd^2/4^{2k})$, and therefore $P(n,k,d,p) = P(n,k,d) + \Pr[X] = (1 + O(p/(1-p)^4 \cdot 1/d))P(n,k,d)$.

		Arratia et al.		This paper		
n	k	lower	upper	lower	upper	Simulation
193	8	0	0.5923	0.0051	0.1233	0.0907
791	10	0	0.2648	0.0083	0.1341	0.0996
3175	12	0.0502	0.1500	0.0094	0.1356	0.1009
12195	14	0.0742	0.1000	0.0084	0.1152	0.0875

Table 1: Comparison between the bounds in this paper and in Arratia et al. [2] for $P(n,k)$, and simulation estimates of $P(n,k)$. For the setup of the simulation see Section 5.

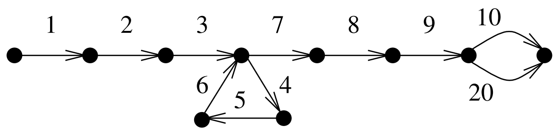

Figure 3: A portion of a de-Bruijn graph showing a example of a bad R-pair (for clarity, not all edges are drawn). The numbers on the edges fit a solution $1,\ldots,n$. Let π be a solution with $\pi(1),\ldots,\pi(7) = 1,2,3,7,8,9,20$. $((4,7),(10,20))$ is an ordered R-pair. Since $\pi(4)=7$, then event $M_{7,4}$ happens, and similarly events $M_{8,5}$, $M_{9,6}$, and $M_{20,7}$ happen. Thus, $((4,7),(10,20))$ is a bad pair.

We denote by $M_{j,i}$ the event that $i \in S_{I_d,\Delta}(j)$, namely, according to the noisy data, the j-th k-tuple can appear at position i in the solution. Note that a permutation π is a solution of A,k,I_d,Δ iff event $M_{\pi(i),i}$ happens for every $i \le n$, and $A|^k_{\pi(i)} \mid A|^k_{\pi(i+1)}$ for every $i \le n-1$. As event $M_{j,i}$ happens iff $I(j)-\Delta_j \le I(i)$, and since the random variable Δ_j has geometric distribution, it follows that $\Pr[M_{j,i}] = p^{I(j)-I(i)}$ for $i \le j$ and $\Pr[M_{j,i}] = 1$ for $i > j$.

A pair of repeats $((i,j),(i',j'))$ is called *ordered* if $i < j < j'$, $i' \notin [j,j']$, and $I(i') \ge I(i)$ (note that i' can be either bigger or smaller than j'). Similarly to Lemma 3, the probability that $((i,j),(i'j'))$ is an ordered R-pair is either 0 or $12/4^{2k}$. An ordered R-pair is called *bad* if event $M_{i',j'-j+i}$ happens, event $M_{l,l-j+i}$ happens for every $j \le l \le j'-1$, and either (1) $I(i) < I(j'-1)$, or (2) $j'-1 = dI(i)$ and $j' < i'$. The role of a bad pair is similar to the role of an interleaved Rr-pair in Section 3: We shall show that if event X happens then there is a bad pair (an example is given in Figure 3), so an upper bound on the probability of a bad pair gives us an upper bound on $\Pr[X]$. A bad pair that satisfies condition (1) is called *of type 1,* and otherwise it is called *of type 2.* As the above $M_{\cdot,\cdot}$ events are independent, we get that the probability of a bad pair $((i,j),(i',j'))$ is either 0 or

$$\frac{12}{4^{2k}}\Pr\left[M_{i',j'-j+i} \wedge \bigwedge_{l=j}^{j'-1} M_{l,l-j+i}\right] =$$

$$\frac{12}{4^{2k}}\Pr\left[M_{i',j'-j+i}\right]\prod_{l=j}^{j'-1}\Pr\left[M_{l,l-j+i}\right] = \frac{12}{4^{2k}}p^{Q_{i,i',j,j'}}$$

where $Q_{i,i',j,j'} = \max(I(i')-I(j'-j+i),0)+\sum_{l=j}^{j'-1} I(l) - I(l-j+i)$.

For brevity, we will use the term solution when referring to a solution of A,k,I_d,Δ. Given a solution π, an index l is called a *jump point* if $\pi(l) \ne \pi(l-1)+1$. If $\pi(l) > \pi(l-1)+1$ then the index l is called a *forward jump* and if $\pi(l) < \pi(l-1)+1$ then it is called a *backward jump.* Clearly, any nontrivial solution contains at least one forward jump and at least one backward jump. For a nontrivial solution π, denote by i_1^π the minimum forward jump in π , and by i_0^π the minimum backward jump (clearly, $i_1^\pi < i_0^\pi$).

Claim 7. *If i is a jump point in a solution π, then $(\pi(i-1)+1,\pi(i))$ is a repeat.*

Proof. Since π is a solution, we have $A|^k_{\pi(i-1)} \mid A|^k_{\pi(i)}$ and therefore $A|^{k-1}_{\pi(i-1)+1} = A|^{k-1}_{\pi(i)}$. □

We define a complete ordering on the nontrivial solutions as follows: For two nontrivial solutions π and π', $\pi > \pi'$ if the series of jump points of π, sorted in increasing order, is lexicographically larger than the corresponding series of π'. The maximum nontrivial solution has following property, which we will use later:

Lemma 8. *If π is the maximum nontrivial solution, $i < i_0^\pi$ is a forward jump, and $I(\pi(i-1)+1) = I(i_1^\pi)$ then $(\pi(i-1)+1,\pi(i))$ is a rightmost repeat.*

Proof. Denote $j = \pi^{-1}(\pi(i-1)+1)$. As i_0^π is the first backward jump, it follows that $j \ge i_0^\pi$ and therefore $j > i$. By Claim 7, $(\pi(i-1)+1,\pi(i))$ is a repeat. By contradiction, suppose that it is not a rightmost repeat, so $A|^k_{\pi(i-1)+1} = A|^k_{\pi(i)}$, namely $A|^k_{\pi(j)} = A|^k_{\pi(i)}$.

Define a permutation π' as follows: $\pi'(i) = \pi(j)$, $\pi'(j) = \pi(i)$, and $\pi'(l) = \pi(l)$ for any $l \ne i,j$. Since π is a solution and $A|^k_{\pi(i)} = A|^k_{\pi(j)}$, it follows that $A|^k_{\pi'(i)} \mid A|^k_{\pi'(i+1)}$ for all $i \le n-1$. For any $l \ne i,j$, event $M_{\pi'(l),l}$ happens as event $M_{\pi(l),l}$ happens. Furthermore, $I(\pi'(j)) - \Delta_{\pi'(j)} = I(\pi(i)) - \Delta_{\pi(i)} \le I(i) \le I(j)$ and $I(\pi'(i)) - \Delta_{\pi'(i)} \le I(\pi'(i)) = I(\pi(j)) = I(\pi(i-1)+1) = I(i_1^\pi) \le I(i)$ where the last inequality follows from the fact that $i_1^\pi \le i$. Therefore, π' is a solution, and it is nontrivial as $A^{\pi'} = A^\pi$.

Now, i is a jump point in π, but not in π' (as $\pi'(i) = \pi(j) = \pi(i-1)+1 = \pi'(i-1)+1$). The only possible jump points which exists in π' but not in π are $i+1$, j, and $j+1$, all of which are greater than i. It follows that $\pi' > \pi$, contradicting the maximality of π. Therefore, $(\pi(i-1)+1,\pi(i))$ is a rightmost repeat. □

Theorem 9. *If event X happens then there is a bad pair.*

Proof. Suppose that event X happens, namely A is uniquely recoverable w.r.t. k,I_d but A is not uniquely recoverable w.r.t. k,I_d,Δ. Let π be the maximum nontrivial solution. Denote by i_2^π the maximum forward jump point for which $i_2^\pi < i_0^\pi$ and $I(\pi(i_2^\pi-1)+1) = I(i_1^\pi)$ (such an index exists as $\pi(i_1^\pi-1)+1 = i_1^\pi$, so $I(\pi(i_1^\pi-1)+1) = I(i_1^\pi)$). Denote by i_3^π the minimum jump point which is greater than i_2^π.

Denote $i = \pi(i_2^\pi-1)+1$, $j = \pi(i_2^\pi)$, $j' = \pi(i_3^\pi-1)+1$ and $i' = \pi(i_3^\pi)$. We will show that $((i,j),(i',j'))$ is a bad pair. It is easy to verify that $i < j < j'$, $i' \notin [j,j']$, and $i_1^\pi < i'$. Since $i_1^\pi < i'$, then $I(i) = I(i_1^\pi) \le I(i')$. By Lemma 8, (i,j) is a rightmost repeat. Furthermore, by Claim 7, (i',j') is a repeat. Therefore $((i,j),(i',j'))$ is an ordered R-pair. As π

is a solution, event $M_{l,\pi^{-1}(l)}$ happens for each l. From the definition of i_0^π it follows that $l \le \pi(l-1)+1$ for all $l < i_0^\pi$, and in particular, $i_2^\pi \le \pi(i_2^\pi - 1) + 1 = i$. Thus, for any $l \in [j, j'-1]$ we have $\pi^{-1}(l) = l - j + i_2^\pi \le l - j + i$, and since event $M_{l.\pi^{-1}(l)}$ happens, it follows that event $M_{l.l-j+i}$ happens. Similarly, $\pi^{-1}(i') = i_3^\pi = i_2^\pi + j' - j \le j' - j + i$, hence event $M_{i',j'-j+i}$ happens.

If $I(i) < I(j'-1)$ then $((i,j),(i',j'))$ is a bad pair of type 1. For the rest of the proof assume that $I(i) = I(j'-1)$. We claim that i_3^π is a forward jump.

Conversely, suppose that i_3^π is backward jump, hence $i_3^\pi = i_0^\pi$. Let i_4^π be the maximum forward jump point for which $i_4^\pi \le i_2^\pi$ and $\pi(i_4^\pi - 1) < i'$ (there is such a jump point as $\pi(i_1^\pi - 1) < i'$). It is easy to verify that $\pi(i_4^\pi - 1) + 1 \le i' < \pi(i_4^\pi) < j'$ and $I(\pi(i_4^\pi - 1) + 1) = I(i_1^\pi)$. By Lemma 8, $(\pi(i_4^\pi - 1) + 1, \pi(i_4^\pi))$ is a rightmost repeat. Hence, $((\pi(i_4^\pi - 1) + 1, \pi(i_4^\pi)), (i', j'))$ is an interleaved R-pair, and by Theorem 1, A is not uniquely recoverable w.r.t. k, I_d, a contradiction.

Therefore, i_3^π is a forward jump, so $j' < i'$ and $i_3^\pi < i_0^\pi$. Furthermore, by the maximality of i_2^π it follows that $I(j') > I(i_1^\pi) = I(i)$. Since $I(j'-1) = I(i)$, we conclude that $j' - 1 = dI(i)$. Thus, $((i,j),(i',j'))$ is a bad pair of type 2. □

By Theorem 9, $\Pr[X]$ is less or equal to the probability that there is a bad pair. Let P_r denote the probability that there is a bad pair $((i,j),(i',j'))$ with $I(i) = r$. It is easy to verify that $P_1 \ge P_2 \ge \cdots \ge P_{n/d-1}$ and $P_{n/d} = 0$, so the probability that there is a bad pair is at most $(n/d - 1)P_1$. We now bound P_1. We consider five cases, where in the first four cases we consider bad pairs of type 1, and in the fifth case we consider bad pairs of type 2.

Case 1: $I(j) < I(j'-1)$ and $j' < i'$. Denote $q = \lceil (j-i)/d \rceil - 1$, $x = (j-i) - qd$, $r = I(j'-1) - I(j) - 1$, $y = (j'-1) - d(I(j'-1)-1)$, and $s = \lceil (i'-j')/d \rceil - 1$. Note that $q, r, s \ge 0$ and $1 \le x, y \le d$.

Claim 10. *In case 1,* $Q_{i,i',j,j'} \ge s + r + q + \min(x,y)$.

Proof. Recall that $Q_{i,i',j,j'} = I(i') - I(j'-j+i) + \sum_{l=j}^{j'-1} I(l) - I(l-j+i)$. The claim is proven by observing the following:

1. As $I(l) = \lceil l/d \rceil$, we get that $I(i') \ge I(j') + I(i'-j') - 1 = I(j') + s \ge I(j'-j+i) + s$. Hence $I(i') - I(j'-j+i) \ge s$.
2. If $r > 0$, for any $0 \le t < r$, the index $l_t = d(I(j)+t)+1$ satisfies $I(l_t) - I(l_t - j + i) \ge 1$ (as $I(l_t) = I(j)+t+1$ and $I(l_t - j + i) \le I(l_t - 1) = I(j) + t$). Therefore, the indices $l_0, l_1, \ldots, l_{r-1}$ contribute at least r to the value of $Q_{i,i',j,j'}$.
3. $I(j) \ge I(i) + I(j-i) - 1 = I(i) + q$, so $I(j) - I(i) \ge q$
4. For any $l \in [d(I(j'-1)-1)+1, j'-1]$, if $l - d(I(j'-1)-1) \le x$ then $I(l) - I(l-j+i) \ge 1$. Thus, the contribution of the indices in $[d(I(j'-1)-1)+1, j'-1]$ to $Q_{i,i',j,j'}$ is at least $\min(x,y)$. □

We note that the bound in Claim 10 is very crude, but it is suffice for our needs. For fixed values of q, x, r, y, and s, there are at most d ways to choose a value for i (as $I(i) = 1$), and at most d ways to choose a value for i' (the values of j and j' are fixed after choosing a value for i). Therefore, the contribution of case 1 to P_1 is at most

$$\sum_{x=1}^{d}\sum_{y=1}^{d}\sum_{q\ge0}\sum_{r\ge0}\sum_{s\ge0} d^2 \frac{12}{4^{2k}} p^{Q_{i,i',j,j'}}$$

$$\le \frac{12}{4^{2k}} d^2 \sum_{x=1}^{d}\sum_{y=1}^{d}\sum_{q\ge0}\sum_{r\ge0}\sum_{s\ge0} p^{s+r+q+\min(x,y)}$$

$$\le \frac{12}{4^{2k}} d^2 \sum_{x=1}^{d}\sum_{y=1}^{d} \frac{1}{(1-p)^3} p^{\min(x,y)}$$

$$\le \frac{12}{4^{2k}} d^2 \frac{1}{(1-p)^3} 2 \sum_{y=1}^{d}\sum_{x=y}^{d} p^y \le \frac{12}{4^{2k}} \frac{2p}{(1-p)^4} d^3.$$

Since all the cases are similar, we omit the details in calculating the other cases.

Case 2: $I(j) < I(j'-1)$ and $i' < j'$ (so $i' < j$). Define q, x, r, and y as in case 1. Here $Q_{i,i',j,j'} \ge r + q + \min(x,y)$. For fixed values of q, x, r, and y, there are at most d ways to choose a value for i, and $(q+2)d$ ways to choose a value for i' (as $I(i) \le I(i') \le I(j)$ and $I(j) - I(i) \le I(j-i) = q+1$). The contribution of case 2 to P_1 is at most $(12/4^{2k}) \sum_{x=1}^{d}\sum_{y=1}^{d}\sum_{q\ge0}\sum_{r\ge0}(q+1)d^2 p^{Q_{i,i',j,j'}} \le (12/4^{2k})(2p(2-p)/(1-p)^4)d^3$.

Case 3: $I(j) = I(j'-1)$ and $j' < i'$. Let $z = j - d(I(j)-1) - 1$. We have that $j - i > z$ because otherwise, $I(i) = I(j) = I(j'-1)$ contradicting the assumption that $((i,j),(i',j'))$ is a bad pair of type 1. Denote $q = \lceil (j-i-z)/d \rceil - 1$, $x = (j-i-z) - qd$, $y = j' - j$, and $s = \lceil (i'-j')/d \rceil - 1$. Then, $Q_{i,i',j,j'} \ge s + q + \min(x,y)$. For fixed values of q, x, y, and s, there are at most d ways to choose a value for i, and d ways to choose a value for i'. Therefore, the contribution of case 3 to P_1 is at most $(12/4^{2k}) \sum_{x=1}^{d}\sum_{y=1}^{d}\sum_{q\ge0}\sum_{s\ge0} d^2 p^{Q_{i,i',j,j'}} \le (12/4^{2k})(2p/(1-p)^3)d^3$.

Case 4: $I(j) = I(j'-1)$ and $i' < j'$. With z defined as in case 3, we have again that $j - i > z$. Define q, x, r, and y as in case 3. Here $Q_{i,i',j,j'} \ge q + \min(x,y)$. For fixed values of q, x, and y, there are at most d ways to choose a value for i, and $(q+2)d$ ways to choose a value for i'. Therefore, the contribution of case 4 to P_1 is at most $(12/4^{2k}) \sum_{x=1}^{d}\sum_{y=1}^{d}\sum_{q\ge0}(q+2)d^2 p^{Q_{i,i',j,j'}} \le (12/4^{2k})(2p(2-p)/(1-p)^3)d^3$.

Case 5: $j' - 1 = dI(i)$ and $j' < i'$. Denote $s = \lceil (i'-j')/d \rceil - 1$. Here $Q_{i,i',j,j'} \ge s + 1$. For a fixed value of s, there are at most $\binom{d}{2}$ ways to choose values for i and j, and d ways to choose a value for i'. Therefore, the contribution of case 5 to P_1 is at most $(12/4^{2k}) \sum_{s\ge0} \frac{d^3}{2} p^{Q_{i,i',j,j'}} \le (12/4^{2k})(p/2(1-p))d^3$.

Combining all cases, we obtain that $P_1 = O(p/(1-p)^4 \cdot d^3/4^{2k})$. Therefore, we proved the following theorem:

Theorem 11. *$Pr[X] = O(p/(1-p)^4 \cdot nd^2/4^{2k})$.*

Combining Theorem 11 and Corollary 6 gives the following theorem:

Theorem 12. *If $5(k-1) \le d \le 2^{k+1}/c^{1/4} + 4(k-2)$ then $P(n,k,d,p) = (1 + O(p/(1-p)^4 \cdot 1/d))P(n,k,d)$.*

n	k	d	p_0	$p_{0.5}$	$p_{0.5}/p_0$
18880	8	40	0.0985	0.1353	1.374
5520	8	60	0.0994	0.1190	1.197
2320	8	80	0.0956	0.1074	1.123
1200	8	100	0.0896	0.0964	1.076

Table 2: An experimental estimation of $P(n, k, d)$ and $P(n, k, d, 0.5)$. p_0 and $p_{0.5}$ are the estimates of $P(n, k, d)$ and $P(n, k, d, 0.5)$, respectively. For each value of d, we chose a value for n so that p_0 will be approximately 0.1.

n	k	$\bar{d}$	P_0	$P_{0.5}$	E_0	$E_{0.5}$
5000	9	40	0.038	0.046	0.00022	0.00028
10000	9	40	0.098	0.108	0.00034	0.00038
20000	9	40	0.158	0.192	0.00028	0.00035
30000	9	40	0.228	0.277	0.00028	0.00036

Table 3: Results of simulations with random sequences and uniformly distributed clone positions. P_t estimates the fraction of times the reconstructed sequence differs from the target sequence for false-negative probability t. E_t estimates the fraction of incorrectly reconstructed positions in the sequence.

n	k	$\bar{d}$	P_0	$P_{0.5}$	E_0	$E_{0.5}$
5000	9	40	0.4	0.4	0.00114	0.00138
10000	9	40	0.5	0.5	0.00090	0.00094
20000	9	40	0.8	0.8	0.00067	0.00075
30000	9	40	0.8	1.0	0.00087	0.00104

Table 4: Results of simulations with Human DNA sequences and uniformly distributed clone positions.

5. EXPERIMENTAL RESULTS

To complement our theoretical results, we performed simulations with random and real DNA sequences. In the first set of simulations, we randomly generated a sequence, partitioned it into d-long IFs, and computed the error-prone assignment of k-mers into IFs according to our probabilistic model, assuming 50% false negative errors. A simple backtracking algorithm was then used to compute the set of all distinct sequences that are consistent with the data. The same process was performed with noiseless data. (Formally, if the target sequence is A, we computed the set B_0 of all the sequences A' such that $A' = A^\pi$ for a solution π of A, k, I_d, and the set $B_{0.5}$ of all the sequences A' such that $A' = A^\pi$ for a solution π of A, k, I_d, Δ.) The program was run 10,000 times for each combination of n, k and d. Denote by p_t the fraction of runs in which the sequence was not uniquely recoverable in case of false negative probability t ($|B_t| > 1$). Clearly, p_0 and $p_{0.5}$ are estimates of $P(n, k, d)$ and $P(n, k, d, 0.5)$. The results are given in Table 2, and they show that indeed $p_{0.5}/p_0 = 1 + O(1/d)$ as stated in Theorem 12.

Another set of experiments tested the power of the strategy in a more realistic scenario. Here we no longer assumed that the IFs are of equal size. Instead, we randomly chose clone positions so that they are uniformly distributed across the target sequence, and the average distance between adjacent left endpoints of clones is $\bar{d}$. Here we considered both endpoints of the clones for partitioning the sequence into IFs. Note that we assume that the order of the clones and the positions of the endpoints are known. 1000 random target sequences were generated for each value of n, and for each one, the first solution generated by the backtracking algorithm (representing an arbitrary solution) was compared with the target sequence. Two statistics were computed: The fraction of the runs in which the two sequences differed, and the average rate of mismatches between them. (Technically, the algorithm chose one sequence A_0 from B_0, and one sequence $A_{0.5}$ from $B_{0.5}$. We measured P_0, the fraction of runs in which $A_0 \neq A$, and $P_{0.5}$, the fraction of runs in which $A_{0.5} \neq A$. We also computed E_t, the average over all runs of the fraction of positions in which A_t and A differ, for $t = 0, 0.5$.) Table 3 contains the results. While the odds of completely correct reconstruction decrease with target size, the average number of mismatch errors in the reconstruction was very low: between 2 and 4 in 10,000 bp.

Table 4 shows results of the same simulation using real (coding an non-coding) Human DNA sequences. For each target length, 10 disjoint sequences were used. As expected, due to the non-randomness of real DNA, the results worsen. In fact, with sequences of length 30,000 bp and error-prone data, none of the reconstructions was perfectly correct. However, even in that situation, the average number of miscalled base errors was only about one in 1000 bp.

In closing, we note that further simulations making even weaker assumptions can be performed: The assumption that clone order and endpoint positions are known can be relaxed, and the knowledge of the noisy multi-spectrum can be replaced by noisy spectrum without multiplicities. False positives can also be incorporated. We intend to pursue the above in the future. A key difficulty in our analysis and simulations, is the assumption of independence of overlapping clone spectra. Though such dependencies definitely exist, it is currently unclear how to model them adequately.

6. REFERENCES

[1] L. M. Adleman. Location sensitive sequencing of DNA. Technical report, University of Southern California, 1998.

[2] R. Arratia, D. Martin, G. Reinert, and M.S. Waterman. Poisson process approximation for sequence repeats, and sequencing by hybridization. *J. of Computational Biology*, 3(3):425–463, 1996.

[3] W. Bains and G. C. Smith. A novel method for nucleic acid sequence determination. *J. Theor. Biology*, 135:303–307, 1988.

[4] A. Ben-Dor, I. Pe'er, R. Shamir, and R. Sharan. On the complexity of positional sequencing by hybridization. In *Proc. Tenth Annual Symposium on Combinatorial Pattern Matching (CPM 99)*, pages 88–100.

[5] J. Blazewicz, P. Formanowicz, M. Kasprzak, W.T. Markiewicz, and J. Weglarz. DNA sequencing with

positive and negative errors. *J. of Computational Biology*, 6(1):113–123, 1999.

[6] S. D. Broude, T. Sano, C. S. Smith, and C. R. Cantor. Enhanced DNA sequencing by hybridization. *Proc. Nat. Acad. Sci. USA*, 91:3072–3076, 1994.

[7] R. Drmanac, I. Labat, I. Brukner, and R. Crkvenjakov. Sequencing of megabase plus DNA by hybridization: theory of the method. *Genomics*, 4:114–128, 1989.

[8] S. Hannenhalli, P. Pevzner, H. Lewis, and S. Skiena. Positional sequencing by hybridization. *Computer Applications in the Biosciences*, 12:19–24, 1996.

[9] K. R. Khrapko, Yu. P. Lysov, A. A. Khorlyn, V. V. Shick, V. L. Florentiev, and A. D. Mirzabekov. An oligonucleotide hybridization approach to DNA sequencing. *FEBS letters*, 256:118–122, 1989.

[10] Y. Lysov, V. Floretiev, A. Khorlyn, K. Khrapko, V. Shick, and A. Mirzabekov. DNA sequencing by hybridization with oligonucleotides. *Dokl. Acad. Sci. USSR*, 303:1508–1511, 1988.

[11] I. Pe'er and R. Shamir. Spectrum alignment: Efficient resequencing by hybridization. In *Proc. Eighth International Conference on Intelligent Systems in Molecular Biology (ISMB 00)*, pages 260–268.

[12] P. Pevzner, Yu. P. Lysov, K. R. Khrapko, A. V. Belyavsky, V. L. Florentiev, and A. D. Mirzabekov. Improved chips for sequencing by hybridization. *J. Biomolecular Structure and dynamics*, 9:399–410, 1991.

[13] P.A. Pevzner. l-tuple DNA sequencing: Computer analysis. *J. Biomolecular Structure and dynamics*, 7:63–73, 1989.

[14] P.A. Pevzner. DNA physical mapping and alternating Eulerian cycles in colored graphs. *Algorithmica*, 13:77–105, 1995.

[15] F. Preparata, A. Frieze, and E. Upfal. Optimal reconstruction of a sequence from its probes. *J. of Computational Biology*, 6:361–368, 1999.

[16] S. S. Skiena and G. Sundaram. Reconstructing strings from substrings. *J. of Computational Biology*, 2:333–353, 1995.