Sequencing-by-hybridization at the information-theory bound: an optimal algorithm

Franco P. Preparata* Eli Upfal*

Abstract

In a recent paper [PFU99] we have introduced a novel probing scheme for DNA sequencing by hybridization (SBH). The new *gapped-probe* scheme combines natural and universal bases in a well defined periodic pattern. It was shown in [PFU99] that the performance of the gapped-probe scheme (in terms of the length of a sequence that can be uniquely reconstructed using a given library size of probes) is significantly better than the standard scheme based on oligomer probes.

In this paper we present and analyze a new, more powerful, sequencing algorithm for the gapped-probe scheme. We prove that the new algorithm exploits the full potential of the SBH technology with high-confidence performance, that comes within a small constant factor (about 2) of the information-theory bound. Moreover, this performance is achieved while maintaining running time linear in the target sequence length.

1 Introduction

Sequencing by hybridization [BS91, L+88, D+89, P89, PL94, W95] is a novel DNA sequencing technique in which an array (SBH chip) of short sequences of nucleotides (*probes*) is brought in contact with a solution of (replicas of) the target DNA sequence. A biochemical method determines the subset of probes that bind to the target sequence (the *spectrum* of the sequence), and a combinatorial method is used to reconstruct the DNA sequence from the spectrum. Since technology limits the number of probes on the SBH chip, a challenging combinatorial question is the design of a smallest set of probes that can sequence an arbitrary DNA string of a given length.

Current implementations of SBH use "classical" probing schemes, i.e., chips accommodating all 4^k k-mer oligonucleotide ("solid" probes with no gaps), the symbols being the well-known DNA bases { A,C,G,T } and k being a technology-dependent integer parameter. Pevzner *et al.* [P+91, PL94, W95] observed that the expected length of unambiguously reconstructible sequences with solid length-k probes is $O(2^k)$ and a tight bound of the same order has been proven in [DFS94]. These results were confirmed by extensive simulations. Note, however, that an information-theoretic argument yields an upper bound $O(4^k)$.

In a recent paper [PFU99] we have introduced a novel probing scheme for DNA sequencing-by-hybridization. This method, which uses probing patterns with a well-defined periodic gap structures (and rests on the deployment of universal bases for the realization of the gaps) over-

*Computer Science Department, Brown University, 115 Waterman Street, Providence, RI 02912-1910, USA. E-mail: {franco, eli}@cs.brown.edu.

RECOMB 2000 Tokyo Japan USA
Copyright ACM 2000 1-58113-186-0/00/04. $5 00

comes the well-known shortcomings of traditional SBH based on oligomer probes, which had raised a negative prognosis for the competitiveness of the approach. We had shown that a simple algorithm, which reconstructs the target sequence from its spectrum symbol-by symbol and halts the process (declares failure) when more that one extension is confirmed by a chosen number of probes, dramatically improves over the oligomer method and, with a high level of confidence, can correctly reconstruct sequences whose length m is "asymptotically" optimal (for example, for 8 specified nucleotides and confidence 0.95, the simple algorithm achieves $m \approx 2000$, against the information-theoretic bound of 32768).

The asymptotic result, however, despite its inherent significance for a problem that has been the focus of considerable research interest for a decade, did not fully reveal the potential of the approach. In this paper we present a novel, more powerful algorithm, that provably exploits the potential of the probing scheme. In addition, we present a combinatorially subtle probabilistic analysis, based on the hypothesis of target sequences generated by a maximum-entropy memoryless source, and show that the high-confidence performance comes *within a constant factor* (about 2) of the information-theory bound. Our analysis is, of course, confined to sequences generated by the above random process, as has been the practice in previous analogous analyses. Unfortunately, very little is known about a corresponding probability model for natural sequences, but extensive simulations with sequences of known genomes (*Haemophilus influenzae, Escherichia coli*) show, despite an expected minor degradation due to the constrained randomness of natural DNA, analogous behavior.

Therefore, the new algorithm improves by a substantial constant factor over the one of [PFU99]. This fact, despite its minor significance in asymptotic analysis, may have enormous practical repercussions. We also note that the superior performance is achieved while maintaining $O(m)$ running time, under the criterion to adopt the smallest feasible k for the given m. In Figure 1 we display the diagrams of the probabilities of success (for random sequences) of the basic and of the advanced algorithms: The success probability is on the vertical axis, while the other two axes display sequence length and the parameter r. To validate the analysis, in the Appendix we display for comparison corresponding analytical and experimental diagrams for $(4,4)$-probes.

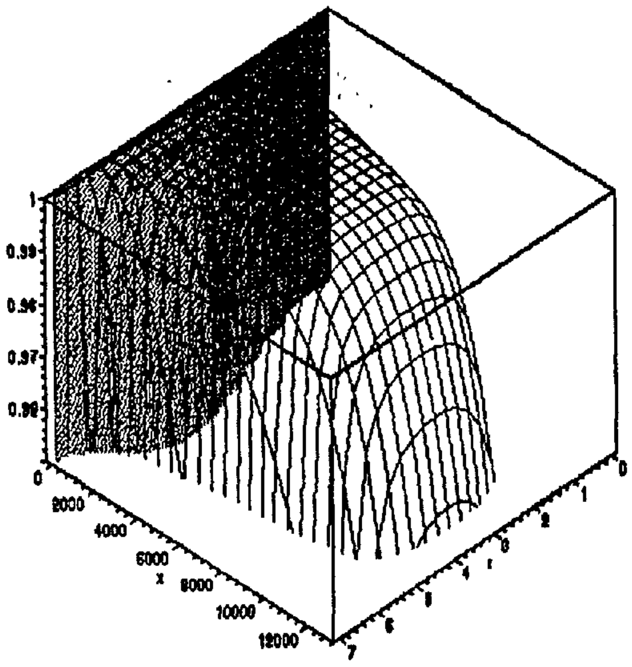

Figure 1: Probability of successful sequence reconstruction for the new algorithm compared to the basic algorithm (shaded graph), as a function of target sequence length (< 13,000) for and all possible choices of (s,r) with $k = 8$.

2 Review of the probing scheme

A *Sequencing by Hybridization (SBH)* chip consists of a fixed number of *features*. Each feature can accommodate one probe. A *probe* is a string of symbols (nucleotides) from the alphabet

$$\mathcal{A} = \{ \text{A,C,G,T,*} \},$$

where A,C,G, and T denote the standard DNA bases and $*$ denotes the "don't care " symbol ("blank"), implemented using a *universal base* [LB94].

The *spectrum* of a target sequence is the set of probes that are Watson/Crick-complementary to a subsequence of the target. A *sequencing algorithm* is an algorithm that, given a set of probes and a spectrum, decides if the spectrum defines a unique DNA sequence, and, if so, reconstructs that sequence.

A *gapped-probe scheme* [PFU99] uses a family of probes with a well defined periodic pattern of gaps ((s,r)-probes). We denote by a^p the p-fold repetition of a string a, and if u is a binary string, $\bar{u}$ is its complementary binary string.

Definition 1 *For integers $r \geq 0$ and $s \geq 1$, a probing pattern is the concatenations $u^s v^r$ of two periodic strings u^s and v^r, where u and v are two binary strings related as follows:*

$$u = 1, v = \bar{u}^{s-1}u, \quad \text{or} \quad v = 1, u = v\bar{v}^{r-1}$$

referred to, respectively, as direct *and* reverse *patterns.*

Considering direct patterns, the corresponding probes have the form $X^s(*^{s-1}X)^r$, for integer parameters s and r, where X ranges over the alphabet and $*$ is blank. For example, a $(4,3)$-probe has the form

$$XXXX***X***X***X.$$

Formally, it is convenient to view an (s,r)-probe as having $s(r+1)$ symbols over the extended alphabet $\mathcal{A} \cup \{*\}$. Of these $s(r+1)$ symbols $r(s-1)$ are blanks, and, since in each probe there are $s+r$ positions with an X symbol, the set of (s,r)-probes has exactly $|\mathcal{A}|^{r+s} = |\mathcal{A}|^k$ members. Note that the classical scheme is a very special case since it uses $(k,0)$-probes.

For given s and r, the collection of all the probes of a target sequence a is called the $(s,r)-$*spectrum* of a, or, briefly, its *spectrum*. These probes are collected by placing the leftmost position of the probing pattern to correspond to the i-th position of a, for

$$i = 1, 2, \dots, |a| - s(r+1) + 1,$$

and extracting the sampled subsequence.

The sequence reconstruction task is a the symbol-by-symbol construction from the spectrum of a putative sequence b, intended to be identical to the target sequence that originated the spectrum. Reconstruction succeeds if and only if sequence b coincides with sequence a.

Given a sequence b (the current putative sequence), b_i denotes its i-th symbol and $b_{(i,j)} = b_i b_{i+1} \dots b_j$. The fundamental primitive operation of sequence reconstruction is *extension*, i.e., the addition of one extra symbol to the current putative sequence. The following algorithm extends a prefix $b_{(1,\ell)}$ of the putative sequence to its right, possibly to its rightmost end. Obviously $\ell \geq (r+1)s$.

Algorithm *sequence*$(S; b_{(1,\ell)})$
The algorithm uses as a subroutine a function *extend*$(S;q)$, for some probe q, which returns a pair (b,w), in which b is a nonempty string (normally, a single symbol), or a set of symbols, or the empty symbol ϵ, and, correspondingly, the parameter w is "continue", or "ambiguous", or "complete".

```
1. $u \leftarrow$ continue
2. while ($u$ = continue) do
3.     $q \leftarrow b_{(\ell-s(r+1)+2,\ell)}*$
4.     $(b,w) \leftarrow extend(S;q)$
5.     if ($w$ = continue)
6.         then
7.             $b_{(1,\ell+|b|)} \leftarrow b_{(1,\ell)}b$
8.             $\ell \leftarrow \ell + |b|$
9.     $u \leftarrow w$
10. return ($b_{(1,\ell)}, w$)
```

The"while"-loop 2-9 normally extends the putative sequence one symbol at a time. In line 3 a query probe is prepared as the $((r+1)s -$

1)-suffix of the current putative sequence extended with a single "blank" (intended to sample the extension symbol). This query is used by the function *extend* (line 4) to interrogate the spectrum (see next section), and will obtain the set of all the probes matching the query in their specified positions. If this probe set is a singleton, then the extension is unique, and function *extend* immediately returns a symbol b, with a certificate $w = continue$. Otherwise it will interrogate the spectrum for additional evidence, and will ultimately return a pair (b, w) of the forms $(b, continue)$ (b a symbol), $(\epsilon, complete)$ (ϵ the empty symbols), or $(B, ambiguous)$ (B a set of symbols, $|B| > 1$). Extension is implemented in line 7. The semantics of the designations { continue, complete, ambiguous} is straightforward. Specifically, "ambiguous" means that the algorithm is unable to return a unique extension, and therefore the process of complete reconstruction fails (only a proper prefix of the target sequence has been produced).

3 An optimal SBH algorithm and its performance analysis

Clearly, the crucial component of the method is the implementation of the function *extend*$(S; q)$. In [PFU99] we proposed an implementation, referred to here as the "basic algorithm", with the following failure mechanism.

When the interrogation of the spectrum returns a set M_0 consisting of more than one probe (i.e., a potential ambiguous extension), let B_0 be the set of the possible extensions. The verification is executed as follows. We construct the set M_1 of all probes in the spectrum such that their common $(sr-1)$-prefix matches $b_{(\ell-sr+1,a_\ell-1)}$, and their $(s+1)$-suffixes agree, in appropriate shifts, with the probes in M_0. Let B_1 be the set of symbols appearing in the sr-th position of the probes in M_0. If $B_0 \cap B_1$ is a singleton, then we have a unique extension to the string. Otherwise we continue by constructing the set M_2 of the spectrum probes whose $(s(r-1)-1)$-prefix matches $b_{(\ell-s(r-1)+1,\ell-1)}$ and $(2s+1)$-suffix agrees with the probes in M_1. From M_2 we construct the corresponding set B_2 of extensions. Again, if $B_0 \cap B_1 \cap B_2$ is a singleton we are done, else we proceed by considering shorter prefixes of lengths $s(r-2), s(r-3), s(r-4), ..., s$ of the spectrum probes. If $|\cap_{j=1}^{i} B_j| = 1$ for some $i \leq r$, then we have an unambiguous extension. Otherwise, in the basic scheme we halt and report the current sequence.

We now present, and discuss in detail, a more sophisticated technique, referred to as the "advanced algorithm", which we show to fully exploit the power of the probing scheme (i.e., to achieve *non-asymptotically* the information theory bound).

Advanced algorithm

The next-symbol extension is first attempted using the basic algorithm. Upon detection of an ambiguous branching (i.e., the event causing failure of the basic algorithm), the advanced algorithm attempts the extension (based on the spectrum), up to some maximum length H (a design parameter) beyond the branching, of all paths issuing from such branching, and of those spawned by them, in a breadth-first fashion. Beyond the ambiguous branching each path is extended on the basis of a *single* probe: the absence of any such extending probe causes termination of the path. This construction stops either if there remains only one (the correct) path, or upon reaching the threshold H otherwise. In either case, the algorithm extends the putative sequence with the longest common prefix of all surviving paths, and fails only when such prefix is empty. (We show in the next section that the threshold H must be chosen adequately larger than $rs + 1$).

To analyze the performance of the outlined advanced algorithm, we note that the success of our approach (for both the basic and the advanced algorithms) is based on the fact that the probability of the simultaneous occurrence of a large number of fooling probes is adequately small.

We begin by showing the following property of paths beyond an ambiguous branching.

Lemma 1 *After an ambiguous branching with two or more paths, only one of which is legitimate, both the legitimate path and the spurious paths are deterministically extended rs times (so that both diverging paths achieve length $rs+1$ beyond the branching).*

Proof: Let $p_{(1,t)}$ denote the segment of the correct (legitimate) path such that the ambiguous extension occurs at position $t = (r+1)s$. Also, let w denote the probing pattern and let $w^{(i)} = w \cap p_{(i,i+t-1)}$, i.e., the probe corresponding to (its leftmost symbol in) position i of segment $p_{(1,t)}$. Note that $w^{(i)}$ is a string of $t-1$ symbols with "don't care" $*$ in the positions where the probing pattern has universal bases. Since we have an ambiguous extension at position t, the spectrum contains at least one complete set of $(r+1)$ fooling probes $q^{(1)}, q^{(2)}, \ldots, q^{(r+1)}$ supporting the (incorrect) extension symbol $a_1 \neq p_t$. These fooling probes are $q^{(1)} = w^{(1)}a_1$ with $a_1 \neq p_t$, and $q^{(i)} = w^{(s+1)}_{(1,s-1)}q^{(i-1)}_{(s,t-s)}(*^{s-1}a_i)$, with arbitrary a_i. For all positions in the range $[t+1, 2t-s] - \mathcal{I}$, where $\mathcal{I} = t + is, i = 1, 2, \ldots, r$, the (existing) probe that extends the correct path also extends the spurious path since it does not overlap with any of the symbols $a_1, a_2, \ldots, a_r$. Extension in position $t + is \in \mathcal{I}$, $i = 1, 2, \ldots, r$, of the spurious path is provided by fooling probe $q^{(i)}$.

□

This result shows that we must select $H > rs + 1$ and a quantitative criterion will be formulated on the basis of Theorem 1. Assuming conventionally as position 1 the position of the ambiguous branching, beyond position $rs+1$ the correct path is deterministically extended, but spurious paths must be supported by fooling probes present in the spectrum.

Whereas in the basic algorithm [PFU99], which halts upon detection of an ambiguous branching, there is a *single* event that characterizes the algorithm's failure (the presence in the spectrum of $r+1$ fooling probes supporting a spurious extension), we shall see that the advanced algorithm being analyzed has a more complex failure mechanism.

We begin with a technical lemma. With reference to a segment $a_{(t+1,t+2(r+1)s-1)}$ of the target sequence, define probe t_j, $j = 0, \ldots, r$, as a subsequence such that for $i = 2, \ldots, r+1$

$$a_{(t_j+1,t_j+s)} = a_{(t+js+1,t+(j+1)s)},$$

$$a_{t_j+is} = a_{t+(j+i)s}.$$

The span of a probe is the interval between its first and last designated symbol.

Lemma 2 *The probability*

$$\mathrm{Prob}((t_1, \ldots, t_r)|t_0)$$

of $\mathbf{t} = (t_1, \ldots, t_r)$ *occurring, conditional on* t_0*, in a target sequence of length* m *is bounded above by*

$$\left(\frac{m}{4^k} + \frac{1}{3 \cdot 4^{s-1}}\right)^r = \left(\frac{m}{4^k}\right)^r \left(1 + \frac{4^{r+1}}{3m}\right)^r$$

Proof: Given two distinct probes t_i and t_j, $t_i < t_j$, whose spans are not disjoint (i.e., $t_j - t_i < (r+1)s$), we note that only for $t_j = t_i \bmod s$ they intersect in more than one symbol. In all other cases their intersection is exactly one symbols, but since they constrain different symbols of the correct segment, it is as if their spans were disjoint. When $t_j = t_i + hs, h = 1, \ldots, r$ probe t_j constrains $s - 1 + h$ rather than k symbols. In such case we say that the two probes technically *overlap*.

To describe probe overlap, with each vector $\mathbf{t}$ we associate a vector $\sigma(\mathbf{t}) = (\sigma_1, \ldots, \sigma_r)$ over the integer labels $\{0, 1, \ldots, r\}$, where $\sigma_i = \sigma_j$ if t_i and t_j overlap and the leftmost occurrences of each value form the sequence $0, 1, 2, \ldots$. The probability of vector $\mathbf{t}$ is determined by the number of its "sites" (distinct values of the components of $\sigma(\mathbf{t})$) and by the amounts of overlap between consecutive probes occurring at the same site. Specifically, if p_{j-1} is the probability of the j-prefix of $\sigma(\mathbf{t})$, then $p_j = p_{j-1}q_j$, and q_j is the total probability of the following set of events: either t_j defines a new site (with probability $\approx m/4^k$) or t_j overlaps with a previously defined site (with probability $1/4^{s-1+h_i}$, where $h_i = j - i$, $i < j$, and σ_i is the rightmost probe at that site). It follows that q_j is at most $m/4^k + (1/4^{s-1}) \sum_{i=1}^{j} 1/4^{h_i} < m/4^k + 1/3.4^{s-1}$. By a straightforward induction the lemma follows.

□

By the same argument, we establish that defining as t_j, $j = r+1, \ldots, k-1$, the subsequence

$$a_{(t_j+1,t_j+s)} = a_{(t+rs+j,t+(r+1)s+j-1)},$$

$$a_{t_j+is} = a_{t+j+(r+i)s}i,$$

for $i = 2, \ldots, r+1$, we obtain

Corollary 1 *The probability*

$$\mathrm{Prob}((t_{r+1}, \ldots, t_{k-1})|t_r)$$

of $\mathbf{t} = (t_{r+1}, \ldots, t_{k-1})$ *occurring, conditional on* t_r, *in a target sequence of length* m *is bounded above by*

$$\left(\frac{m}{4^k} + \frac{1}{3 \cdot 4^r}\right)^{s-1} = \left(\frac{m}{4^k}\right)^{s-1} \left(1 + \frac{4^s}{3m}\right)^{s-1}$$

We now prove the main result of this paper.

Theorem 1 *The probability that the advanced algorithm fails to reconstruct a (maximum-entropy) random DNA m-mer is bounded above by*

$$3m \left(\left(\tfrac{m}{4^k}\right)^k \left(1 + \tfrac{4^{r+1}}{3m}\right)^r \left(1 + \tfrac{4^s}{3m}\right)^{s-1} + \tfrac{4^k}{4^k - m} \tfrac{m}{4^{(r+1)s}} \right) \qquad (1)$$

Proof: With the previous notation, extension beyond position $rs+1$ occurs supported either by fooling probes (probabilistically) or by a segment of the target sequence (deterministically). We consider the first case, denoted here Event $\mathcal{E}_1$.

1. Event $\mathcal{E}_1$. A spurious path, starting at position 1 (deterministically extended up to position $rs+1$ by Lemma 1) is extended up to position H. Extension between positions $rs+2$ and H must be supported by fooling probes. Let f_p be the probability of extension up to position $rs+p$. Clearly, $f_1 = 1$. Extension to position $rs+p+1$ occurs either if the current fooling probe is isolated and therefore constrains all but its last symbol (with probability $m/4^{k-1}$), or if it overlaps with a subset of the preceding $(r+1)s-1$ fooling probes. Arguing as in Lemma 2, we only need consider the closest among the overlapping probes: therefore, arguing in terms of constrained symbols, we conclude that

$$f_{p+1} < f_p \left(\frac{m}{4^{k-1}} + \frac{4}{3}\left(\frac{1}{4^r} + \frac{1}{4^{s-1}}\right)\right)$$

$$= \left(\frac{m}{4^{k-1}}\left(1 + \frac{4^{s-1} + 4^r}{3m}\right)\right)^p$$

It is immediate that the above quantity vanishes exponentially with p, so that, assuming that an appropriate (small) value of p is adopted, Event $\mathcal{E}_1$ will be neglected henceforth.

When the path extension is deterministically supported by a sequence segment, the latter either does not contain (Event $\mathcal{E}_2$) or does contain (Event $\mathcal{E}_3$) the ambiguous-branching position. We now examine these two cases.

1. Event $\mathcal{E}_2$. In this case, the spectrum provides evidence of two segments au_2 and bu_2 with $|u_2| = (r+1)s - 1$, $|a| = |b| = 1$ and $a \neq b$: Extension of both paths proceeds deterministically and the algorithm fails. The target sequence contains the (correct) segment u_1au_2 (with $|u_1| = (r+1)s - 1$) while u_1bu_2 is (normally) emulated by fooling probes. With the only simplifying assumption that u_1au_2 and the fooling probes are disjoint, we remark: The position of u_1au_2 can be chosen in (approximately) m ways, symbol b can be chosen in 3

ways, and the probability that u_1bu_2 be emulated by fooling probes is $\mathrm{Prob}(t_0, t_1, \ldots, t_{k-1})$. Using Lemma 2, Corollary 1, and the fact that the probability of probe t_0 is $(3m/4^k)$, we obtain that the probability of event $\mathcal{E}_1$ is bounded above by

$$m\frac{3m}{4^k}\left(\frac{m}{4^k}\right)^r\left(1+\frac{4^{r+1}}{3m}\right)^r\left(\frac{m}{4^k}\right)^{s-1}\left(1+\frac{4^s}{3m}\right)^{s-1}$$

$$= 3m\left(\frac{m}{4^k}\right)^k\left(1+\frac{4^{r+1}}{3m}\right)^r\left(1+\frac{4^s}{3m}\right)^{s-1}$$

2. Event $\mathcal{E}_3$. The target sequence contains an actual branching point, i.e., it contains the (correct) string $u_1v_1au_2v_2$ (with $|u_1v_1| = |u_2v_2| = (r+1)s-1, |a| = 1$, $|v_1au_2| = (r+1)s$, and $0 \leq |v_1| < (r+1)s$) and a (fooling) string v_1bu_2 with $b \neq a$. In addition, depending upon the length $|u_1|$ there are at most $(k-1)$ fooling probes emulating the subsequence u_1v_1b. The probability of the occurrence of v_1bu_2 is approximately $m/4^{(r+1)s}$, and the probability of the emulated subsequence u_1v_1b is easily shown to be bounded above by $4^k/(4^k - m)$. It follows that the probability of event $\mathcal{E}_2$ is at most

$$3m\frac{4^k}{4^k - m}\frac{m}{4^{(r+1)s}}$$

Since only Events $\mathcal{E}_2$ and $\mathcal{E}_3$ are significant for the failure of the algorithm, the theorem is proved. □

Remark. For a reasonably small value of p, choosing $H = rs + 1 + p$ guarantees that Event $\mathcal{E}_1$ can be neglected. Referring to Expression (1), the second term is dominant for small and large values of r, but it becomes negligible for the most efficient choices of r, i.e., for $r \approx s - 1$. Therefore, for $r \approx s - 1 \approx k/2$, we obtain $(1 + \frac{4^{r+1}}{3m}) \approx (1 + \frac{4^s}{3m}) \approx 1$ and

$$\mathrm{Prob}(failure) \approx 3m\left(\frac{m}{4^k}\right)^k$$

so $\mathrm{Prob}(failure) < \epsilon$, for a conveniently small ϵ, leads to

$$m < 4^{k-1-\frac{1}{k+1}\log_2\sqrt{\frac{4}{3\epsilon}}}$$

i.e., for any fixed confidence value, the length of the unambiguously reconstructible sequence is within a small constant factor of the information-theoretic bound $4^{k-\frac{1}{2}}$ for very small values of k (for example, for $\epsilon = 0.05, k = 9$, the exponent is $\approx 9 - 1.23$).

4 Running time of the algorithm

Since the algorithm performs a type of "bounded breadth-first-search" of all possible sequence reconstructions from the given spectrum, it is important to verify that the running time of the algorithm is not significantly degraded by this search. In this section we give a high-confidence bound on the execution time. The time performance is expressed in terms of number of accesses to the spectrum, each assumed doable in $O(1)$ average time by standard hashing techniques.

In our analysis, we assume that the algorithm operates at its best performance for a given confidence level, i.e., that m and k are related by $m = 4^{k-1-\eta}$, for some $\eta > 0$.

Theorem 2 *The total number of sequence positions (one-base extensions) associated with ambiguous branchings is w.h.p.* $o(m/\log m)$.

Sketch of proof: We bound the number of ambiguous branchings on the target sequence. Arguing as in Lemma 2, we conclude that their expected number is

$$\nu = \frac{3m}{4}\left(\frac{m}{4^{k-1}}\right)^{r+1}\left(1+\frac{4^{r+1}}{m}\right)^r$$

Since, $m = 4^{k-1-\eta}$, and $r \approx (1/4)\log_2 m$, we conclude that $\nu = m^{1-\frac{\eta}{2}}$, which is strictly sublinear in m. Paths issuing from an ambiguous branching are explored only up to length $H = O(rs) = O((\log m)^2)$. The probability of a branch on any path issuing from an ambiguous branching is bounded by

$$\nu H \frac{3H}{4}\left(\frac{m}{4^{k-1}}\right)^{r+1}\left(1+\frac{4^r}{m}\right)^r = o(1)$$

Thus, we can prove that the total number of accesses associated with the ambiguous branchings is w.h.p. $\nu H = O(m^{1-\frac{\eta}{2}} \log^2 m) = o(m/\log m)$. □

Theorem 3 *The running time of the advanced algorithm is w.h.p. $O(m)$.*

Sketch of proof: The maximum work at a one base extension is $O(r) = O(\log m)$. Since there are $o(m/\log m)$ extensions associated with spurious paths we can restrict our discussion to the work on the remaining $m' = \Theta(m)$ ordinary extensions of the target sequence.

Arguing as in Lemma 2, the probability that at least $h \leq r$ accesses are performed at a specific position is given by

$$Z_h = \frac{3}{4}\left(\frac{m}{4^{k-1}}\right)^h \left(1+\frac{4^{r+1}}{3m}\right)^{h-1}$$

and the expected total amount of work done at ordinary positions is bounded above as follows:

$$m' \sum_{h=1}^{r} Z_h = \frac{3m'}{4}\left(1+\frac{4^{r+1}}{3m}\right)^{r-1} \sum_{h=1}^{r}\left(\frac{m}{4^{k-1}}\right)^h \leq$$

$$\frac{3m'}{4}\left(1+\frac{4^{r+1}}{3m}\right)^{r-1} \sum_{h=1}^{r}\left(\frac{1}{4^{\eta}}\right)^h = O(m') = O(m).$$

To obtain a high probability bound we observe that $Z = \sum_{h=1}^{r} Z_h$ counts the sum of $m(r+1)$ random binary random variables, and that the $r+1$ variables associated with location t are independent of variables associated with locations that are at least $s(r+1)$ positions away from location t.

Thus, we can partition the sum Z into $s(r+1)^2$ sums, such that the binary variables in each sum are independent. Using the Chernoff bound we show that with high probability the sum Z is $O(m)$. □

We close this section by observing, that when we consider the actual running time of the algorithm for a fixed k and $m \leq 4^{k-1-\eta}$, the work due to the processing of the ambiguous branching becomes the dominant factor for large values of m, so that for $m \in [4^{k-1-\eta}/2, 4^{k-1-\eta}]$ the number of accesses is proportional to $O(m \log^2 m)$.

References

[BS91] W. Bains and G.C. Smith, A novel method for DNA sequence determination. *Jour. of Theoretical Biology*(1988), 135, 303-307.

[DFS94] M.E.Dyer, A.M.Frieze, and S.Suen, The probability of unique solutions of sequencing by hybridization. *Journal of Computational Biology*, 1 (1994) 105-110.

[D+89] R. Drmanac, I. Labat, I. Bruckner, and R. Crkvenjakov, Sequencing of megabase plus DNA by hybridization. *Genomics*,(1989),4, 114-128.

[LB94] D. Loakes and D.M. Brown, 5-Nitroindole as a universal base analogue. *Nucleic Acids Research*,(1994), 22, 20,4039-4043.

[L+88] Yu.P. Lysov, V.L. Florentiev, A.A. Khorlin, K.R. Khrapko, V.V. Shih, and A.D. Mirzabekov, Sequencing by hybridization via oligonucleotides. A novel method. *Dokl. Acad. Sci. USSR*,(1988) 303, 1508-1511.

[P89] P.A.Pevzner, l-tuple DNA sequencing: computer analysis. *Journ. Biomolecul. Struct. & Dynamics* (1989) 7, 1, 63-73.

[P+91] P.A.Pevzner, Yu.P. Lysov, K.R. Khrapko, A.V. Belyavsky, V.L. Florentiev, and A.D. Mirzabekov, Improved chips for sequencing by hybridization. *Journ. Biomolecul. Struct. & Dynamics* (1991) 9, 2, 399-410.

[PL94] P.A.Pevzner and R.J. Lipshutz, Towards DNA-sequencing by hybridization. *19th Symp. on Mathem. Found. of Comp. Sci.*, (1994), LNCS-841, 143-258.

[PFU99] F.P. Preparata, A.M. Frieze, E. Upfal. On the Power of Universal Bases in Sequencing by Hybridization. *Third Annual International Conference on Computational Molecular Biology.* April 11 - 14, 1999, Lyon, France, pp. 295–301.

[W95] M.S. Waterman, *Introduction to Computational Biology.* Chapman and Hall, 1995.

A Simulation results

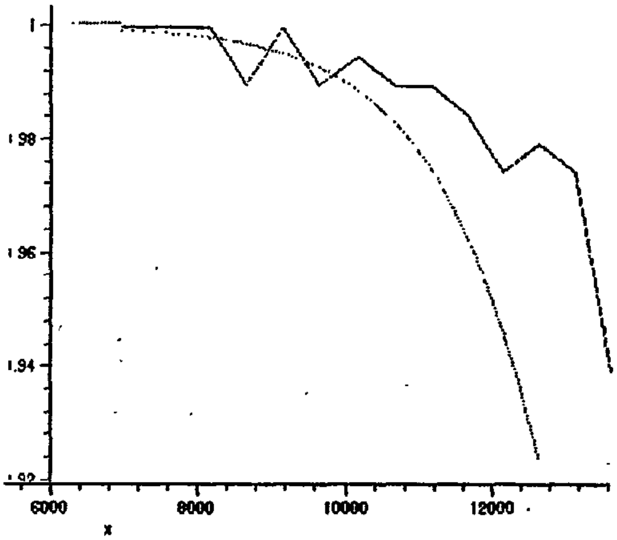

Figure 2: Diagrams of the frequency of correct reconstruction and of a lower bound to the probability of success for (4, 4)-probes as a function of the sequence length