JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 7, Numbers 3/4, 2000
Mary Ann Liebert, Inc.
Pp. 621–630

Sequencing-by-Hybridization at the Information-Theory Bound: An Optimal Algorithm

FRANCO P. PREPARATA and ELI UPFAL

ABSTRACT

In a recent paper (Preparata *et al.*, 1999) we 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 has been shown (Preparata *et al.*, 1999) that the performance of the gapped-probe scheme (in terms of the length of a sequence that can be uniquely reconstructed using a given size library 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.

Key words: DNA sequencing, sequencing by hybridization, gapped probes, probabilistic analysis.

1. INTRODUCTION

SEQUENCING BY HYBRIDIZATION (Bains and Smith, 1998; Lysov *et al.*, 1988; Drmanac *et al.*, 1989; Pevzner, 1989; Pevzner and Lipschutz, 1994; Waterman, 1995) 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. As technology limits the number of probes on the SBH chip, a challenging combinatorial question is the design of the smallest set of probes that can sequence an arbitrary random DNA string of a given length.

Current implementations of SBH use "classical" probing schemes, i.e., chips accommodating all 4^k k-mer oligonucleotides ("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.* (Pevzner *et al.*, 1991; Pevzner and Lipschutz, 1994; Waterman, 1995) 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 Dyer *et al.* (1994). These results were confirmed by extensive simulations. Note, however, that an information-theoretic argument yields an upper bound of $4^{k-\frac{1}{2}}$.

Computer Science Department, Brown University, 115 Waterman Street, Providence, RI 02912-1910.

In a recent paper (Preparata *et al.*, 1999), we have introduced a novel probing scheme for DNA sequencing-by-hybridization. This method, which uses probing patterns with a well-defined periodic gap structure (and rests on the deployment of universal bases for the realization of the gaps) overcomes 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 upon the oligomer method and, with a high level of confidence, can correctly reconstruct random sequences whose length m is "asymptotically" optimal (for example, for 8 specified nucleotides and confidence 0.95, the simple algorithm achieves $m \approx 2{,}000$, against the information-theoretic bound of 32,768).

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 memory-less 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 upon the one of Preparata *et al.* (1999). 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.

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

$$\mathcal{A} \cup \{*\}$$

where $\mathcal{A} = \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$ is the alphabet of the standard DNA bases and $*$ denotes the "don't care " symbol ("blank"), implemented using a *universal base* (Loakes and Brown, 1994).

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* (Preparata *et al.*, 1999) 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, $\overline{u}$ 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:*

$$u = 1, v = \overline{u}^{s-1}u, \qquad \text{or} \qquad v = 1, u = v\overline{v}^{r-1}$$

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

Considering without loss of generality only 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

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

Hereafter, we shall view an (s, r)-probe as an $s(r+1)$-symbol string 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 k-mer scheme is a very special case since it uses $(k, 0)$-probes.

For given s and r, the collection of all the probes that are Watson–Crick complementary to a subsequence of a target sequence a is the $(s, r)-$*spectrum* of a, or, briefly, its *spectrum.* For convenience of discussion, we view a spectrum probe as the *actual* subsequence of a, rather than its WC-complement annealing to it. Therefore, these probes are collected by placing the leftmost position of the probing pattern to correspond to the i-th position of a, for

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

and extracting the sampled subsequence.

In this paper, we focus on a sequence reconstruction process that, based on the spectrum, constructs symbol-by-symbol 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 an arbitrary sequence c (the current putative sequence), c_i denotes its i-th symbol and $c_{(i,j)} = c_i c_{i+1} \ldots c_j$. The fundamental primitive operation of sequence reconstruction is *extension*, i.e., the concatenation of extra symbols (normally one) to the right of the currently constructed prefix of the putative sequence. The following algorithm extends a prefix $b_{(1,\ell)}$ of the putative sequence to its right, possibly to its rightmost end. Obviously $\ell \geq (r+1)s$.

The following 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."

Algorithm. $sequence(S; b_{(1,\ell)})$

```
1. u ← continue
2. while (u = continue) do
3.     q ← b_(ℓ−s(r+1)+2,ℓ)*
4.     (b, w) ← extend(S; q)
5.     if (w = continue)
6.         then
7.             b_(1,ℓ+|b|) ← b_(1,ℓ) b
8.             ℓ ← ℓ + |b|
9.     u ← w
10. return (b_(1,ℓ), 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 single 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 nonempty short string), $(\epsilon,$ *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} are 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 Preparata *et al.* (1999) 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 (indicating a potential ambiguous extension), we let B_0 denote 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 $b_{(\ell-sr+1,\ell-1)}$ 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_1. 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. Otherwise, we proceed by considering shorter prefixes of lengths $s(r-2), s(r-3), s(r-4), \ldots, s$ of the putative sequence. If $|\cap_{j=1}^{i} B_j| = 1$ for some $i \leq r$, then we have a single-symbol 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 come *nonasymptotically* very close to the information theory bound with high confidence).

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. 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 to be adequately larger than rs).

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. All our considerations rest on the hypothesis that the target sequence is a maximum-entropy random sequence of length m.

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 by additional rs symbols (so that both diverging paths achieve length $rs+1$ beyond the branching).*

Proof. Let $p_{(1,(2r+1)s)}$ denote the segment of the correct (legitimate) path such that the ambiguous extension occurs at position $t = (r+1)s$. Let p_t be the correct extension and $c_t \neq p_t$ be a spurious extension.

Since we have an ambiguous extension at position t, the spectrum contains at least one set of $(r+1)$ fooling probes $q^{(1)}, q^{(2)}, \ldots, q^{(r+1)}$ supporting the (incorrect) extension. These fooling probes collectively specify $(r+1)$ arbitrary symbols $c_t, c_{t+s}, \ldots, c_{t+rs}$, with c_{t+js} possibly different from the correct p_{t+js} for $j > 0$.

For all positions in the range $[t+1, 2t-s] - \mathcal{I}$, where $\mathcal{I} = \{t+is : i = 1, 2, \ldots, r\}$, the probe that extends the correct path in that position (which is guaranteed to exist) also extends the spurious path since it does not overlap with any of the symbols $c_t, c_{t+s}, \ldots, c_{t+rs}$. 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 $H > rs$ and a quantitative criterion will be formulated on the basis of Theorem 1. We assume conventionally as position 1 the position of the ambiguous branching. Beyond position rs, the correct path is deterministically extended, but spurious paths must be supported by fooling probes present in the spectrum.

Whereas in the basic algorithm (Preparata *et al.*, 1999), 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 result. A probe is said to be the probe j for position i of sequence a if it samples symbol a_i with its j-th specified position from the right. Thus, for the (s,r)-scheme, the k probes that sample a_i have their initial symbol in positions $i-(r+1)s+1, i-rs+1, i-(r-1)s+1, \ldots, i-s+1, i-s+2, \ldots, i$ of the target sequence a.

We now wish to evaluate the probability of the following event: Given a segment $a_{(g,g+2(r+1)s-1)}$, along the same sequence a (but not overlapping with $a_{(g,g+2(r+1)s-1)}$) there are fooling probes identical to the correct probes for position $g+(r+1)s$. Specifically we wish to prove the following.

Lemma 2. *For an integer $r' \le r+1$, let σ_j, $j=1,\ldots,r'$, be the j-th fooling probe for position $g+(r+1)s$ of segment $a_{(g,g+2(r+1)s-1)}$. The probability of $\sigma_2,\ldots,\sigma_{r'}$ conditional on σ_1 is at most*

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

Proof. Let t_j be the position of the first symbol of σ_j. The spans of probes are allowed to overlap, but not with $a_{(g,g+2(r+1)s-1)}$. Given σ_i and σ_j, with $i\neq j$ and $t_j-t_i<(r+1)s$, 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 symbol, which implies that two specific symbols of $a_{(t,t+2(r+1)s-1)}$ coincide, thereby constraining one additional symbol. Therefore, in such a case, the total number of constrained symbols is the same as if the probes did not overlap. Thus we shall restrict ourselves to the situations $t_j=t_i+hs$, in which case we must have $j=i+h$, for any other choice will result in a constraint on $a_{(g,g+2(r+1)s-1)}$. In such a case, σ_j, rather than k symbols, constrains just $s-1+h$ symbols of a. If two probes overlap, we say that they share the same *site*.

To describe probe overlap, we imagine a process in which the probes are successively assigned to sites. We begin by assigning σ_1 in an arbitrary position of $a_{(1,m-(r+1)s+1)}$. After $\sigma_1,\ldots,\sigma_{j-1}$ have been assigned, let $u\le j-1$ be the current number of distinct sites: σ_j can be assigned in $u+1$ ways, either isolated or to any of the current sites. Thus the number of assignments is the number of distributions of r' distinct items into up to r' nondistinct cells (i.e., the number of possible equivalence classes of an r'-element set). The process is conveniently described by a rooted tree of $r'+1$ levels from 0 to r', in which a node at level h describes an assignment of $\sigma_1,\ldots,\sigma_h$, and the arcs exiting this node are labeled by the conditional probabilities of the assignment of σ_{h+1}. If σ_{h+1} is isolated (a new site), then the corresponding probability is (unconditionally) $\approx m/4^k$; in all other cases the conditional probability is of the form 4^{-p} where p is the number of *additional* symbols of a constrained by the chosen assignment of σ_{h+1}. Therefore, the sought probability is the sum of the products of the labels over all leafward paths of this tree.

The particular structure of our probing pattern lends itself to a simple upper bound. If all nodes of the tree at levels ≥ 1 had identical sets of exiting arcs of total probability P, then the sought probability would be at most $P^{r'-1}$. We now observe that no two distinct overlap assignments of a probe constrain the same number of symbols of a, so that the sum of the labels of the exiting arcs for every node at levels ≥ 1 is bounded above by

$$\frac{m}{4^k}+\frac{1}{4^{s-1}}\sum_{h=1}^{\infty}\frac{1}{4^h}<\frac{m}{4^k}+\frac{1}{3\cdot 4^{s-1}}.$$

This establishes the lemma. ∎

By an identical argument we can establish the following.

Lemma 3. *For an integer $s' \le s$, let σ_j, $j=r+1,\ldots,r+s'$, be the j-th fooling probe for position $g+(r+1)s$ of segment $a_{(g,g+2(r+1)s-1)}$. The probability of $\sigma_{r+2},\ldots,\sigma_{r+s'}$ conditional on σ_{r+1} is at most*

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

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*

$$3m\left[\left(\frac{m}{4^k}\right)^k\left(1+\frac{4^{r+1}}{3m}\right)^r\left(1+\frac{4^s}{3m}\right)^{s-1}+\left(\frac{m}{2\cdot 4^{(r+1)s}}\left(1+\frac{s4^k}{4^k-m}\right)\right)\right]+\left(\frac{m}{4^{k-1}}+\frac{1}{4^2}\right)^{H-rs-1} \tag{1}$$

Proof. With the previous notation, extension beyond position $rs+1$ occurs supported either by fooling probes (probabilistically) or deterministically, because the target sequence contains a substring of length $(r+1)s-1$ identical to a substring of the spurious path. We assume at first that the latter condition does not hold and consider the probabilistic extension.

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 occurring at the same site. We wish to obtain a conservative, but yet useful, upper bound to the probability of the latter event. We observe that the current fooling probe constrains at least two symbols at the site where it occurs (with probability $1/4^2$), except when at least one of these two events happens:

- F_1: The current probe and the one offset by $-s$ positions co-occur at the same site.
- F_2: The current probe and the one offset by $-2s$ positions co-occur at the same site.

Obviously we have

$$\text{Prob}(F_1)+\text{Prob}(F_2)-\text{Prob}(F_1\cap F_2)\le\text{Prob}(F_1\cup F_2)\le\text{Prob}(F_1)+\text{Prob}(F_2)$$

and, considering the numbers of constrained symbols, both $\text{Prob}(F_1)$ and $\text{Prob}(F_2)$ are about $m/4^{k+s-2}$, and $\text{Prob}(F_1\cap F_2)$ is about $m/4^{k+2s-2}$. We conclude that

$$f_{p+1}<f_p\left(\frac{m}{4^{k-1}}+\left(1-2\frac{m}{4^{k+s-2}}+\frac{m}{4^{k+2s-2}}\right)\frac{1}{4^2}+2\frac{m}{4^{k+s-2}}\right)\approx\left(\frac{m}{4^{k-1}}+\frac{1}{4^2}\right)^p. \tag{2}$$

Next we analyze the case of deterministic extension, supported by a substring u of length $(r+1)s-1$ of the target sequence.

2. Event $\mathcal{E}_2$: "Denoting by $v_1v_2av_3$, with $|v_1v_2|=|v_2av_3|=(r+1)s-1$, the correct segment of the target sequence (a being the current symbol), the target sequence also contains a substring v_2bv_3, with $b\neq a$." In general, the positions of v_2av_3 and v_2bv_3 can be chosen in $\binom{m}{2}\approx m^2/2$ ways. When $|v_1|=|v_3|=0$ (i.e., $|v_2|=(r+1)s-1$) $(r+1)s-1$ symbols of v_2 and v_3 are fully constrained and symbol b can be chosen in 3 out of 4 ways, thereby yielding probability $(m^2/2)(3/4)/4^{(r+1)s-1}=3m^2/2\cdot 4^{(r+1)s}$. For $1\le|v_1|\le s$, there must be one additional fooling probe, supporting extension b, that fully agrees with v_3 (thereby constraining k rather than $k-1$ symbols): the corresponding probability is $(m^2/2)s(m/4^k)(3/4)(1/4^{(r+1)s-1})$; analogously, for $s+1\le|v_1|\le 2s$, there must be two additional fooling probes supporting b and agreeing with v_3; the corresponding probability is bounded by $(m^2/2)s(m/4^k)^2(3/4)(1/4^{(r+1)s-1})$, and so on, for $is+1\le|v_1|\le(i+1)s$, so that the total probability is bounded by

$$\frac{3m^2}{2\cdot 4^{(r+1)s}}+\frac{3m^2}{2\cdot 4^{(r+1)s}}s\left(\frac{m}{4^k}\right)+\frac{3m^2}{2\cdot 4^{(r+1)s}}s\left(\frac{m}{4^k}\right)^2+\ldots<\frac{3m^2}{2\cdot 4^{(r+1)s}}\left(1+\frac{s4^k}{4^k-m}\right) \tag{3}$$

Finally, we consider the case when deterministic extension is supported by fooling probes.

3. Event $\mathcal{E}_3$: "Again denoting u_1au_2 the correct string, with $|u_1|=|u_2|=(r+1)s-1$, a denoting the current symbol, the spectrum provides evidence of a substring u_1bu_2." In this case, u_1bu_2 need not be an

actual substring of the target sequence, but may be emulated by fooling probes. We wish to estimate the probability of this event. As in Lemmas 2 and 3, the fooling probes are denoted $\sigma_1, \sigma_2, \ldots, \sigma_k$. Then,

$$\text{Prob}(\sigma_1, \ldots, \sigma_k) = \text{Prob}(\sigma_1, \ldots, \sigma_{r+1})\text{Prob}(\sigma_{r+2}, \ldots, \sigma_k | \sigma_{r+1})$$

since $\sigma_{r+2}, \ldots, \sigma_k$ are independent of $\sigma_1, \ldots, \sigma_r$. We next observe that by Lemma 2

$$\text{Prob}(\sigma_2, \ldots, \sigma_{r+1} | \sigma_1) = \left(\frac{m}{4^k}\right)^r \left(1 + \frac{4^{r+1}}{3m}\right)^r$$

and by Lemma 3

$$\text{Prob}(\sigma_{r+2}, \ldots, \sigma_k | \sigma_{r+1}) = \left(\frac{m}{4^k}\right)^{s-1} \left(1 + \frac{4^s}{3m}\right)^{s-1}.$$

Because $\text{Prob}(\sigma_1) \approx (3m/4^k)$ (since symbol b can be chosen in 3 ways), we conclude that:

$$\begin{aligned}\text{Prob}(u_1 b u_2) &= \frac{3m}{4^k}\left(\frac{m}{4^k}\right)^{k-1}\left(1 + \frac{4^{r+1}}{3m}\right)^r \left(1 + \frac{4^s}{3m}\right)^{s-1} \\ &= 3\left(\frac{m}{4^k}\right)^k \left(1 + \frac{4^{r+1}}{3m}\right)^r \left(1 + \frac{4^s}{3m}\right)^{s-1}. \end{aligned} \tag{4}$$

We can now piece together the preceding intermediate results. The probability of $\mathcal{E}_1$ is conditional on the detection of an ambiguous extension by the basic algorithm, whose probability is trivially majorized by 1. With the only simplifying assumption being that $u_1 a u_2$ and the fooling probes are disjoint, the probabilities of $\mathcal{E}_2$ and $\mathcal{E}_3$ are conditional on the selection of the position of the correct substring. The latter can be chosen in at most m ways, so that multiplying the computed bounds by m achieves an upper bound to the probability of failure. This establishes the theorem. ■

In Figure 1, we display a diagram of a lower bound to the probability of successful sequence reconstruction (the complement to 1 of expression (1)) for $k = 8$ in the range [.9, 1], as a function of the sequence length (in the range [25, 15000]) and of the parameter $r \in [0, 7]$, for $H = 3rs$.

Expression (1) tells us that the corresponding bound (to the probability of $\mathcal{E}_1$) grows dramatically as $m/4^{k-1}$ approaches 15/16. However, choosing $H \geq 3rs$ assures us that the probability of failure is almost insensitive to $\mathcal{E}_1$ for $m < 15 \cdot 4^{k-3}$. Considering now $\mathcal{E}_2$ and $\mathcal{E}_3$, inspection of (1) reveals that the the first term is symmetric in r and $s - 1$, and that the second term is nearly so. A more detailed analysis, treating r as a continuous parameter, yields a minimizing value of r very close to $k/2$ (as Figure 1 illustrates for $k = 8$). Therefore we shall choose

$$r = \left\lfloor \frac{k}{2} \right\rfloor.$$

We now observe that in (1) for $k \geq 6$ the second term is dominant for small and large values of r, but becomes negligible in correspondence of $r \approx k/2$. Therefore, for $r = k/2$, we obtain $(1 + \frac{4^{r+1}}{3m}) \approx (1 + \frac{4^s}{3m}) \approx 1$ and

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

So $\text{Prob}(\textit{failure}) < \epsilon$, for a conveniently small ϵ, leads to

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

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 relatively small values of k (for example, for $\epsilon = 0.05$, $k = 8$, the exponent is $\approx 8 - 1.21$).

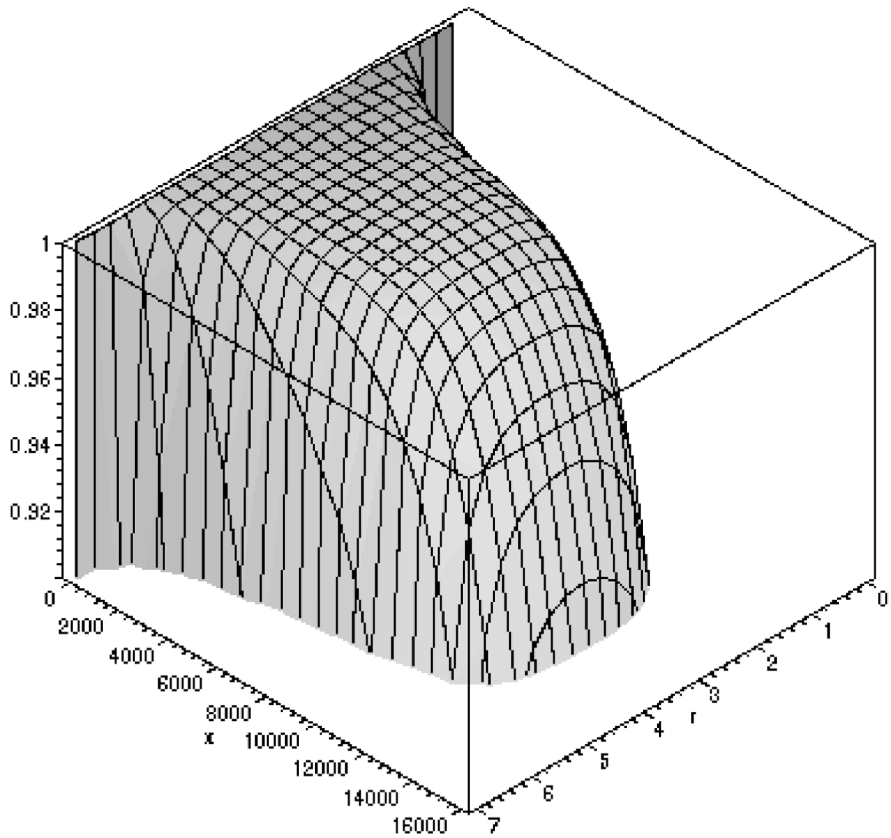

FIG. 1. A lower bound to the probability of successful sequence reconstruction for the new algorithm, as a function of target sequence length ($x < 16{,}000$) for all possible choices of (s, r) with $k = 8$.

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 bound on the expected execution time of the algorithm on a randomly generated target sequence. The time performance is expressed in terms of the number of accesses to the spectrum, each assumed doable in $O(1)$ time.

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 $\eta > 0$, and $r = k/2$.

Theorem 2. *If $T(m)$ is the run-time of the algorithm on a random target sequence of length m, then $E[T(m)] = O(m)$.*

Proof. As discussed is Section 3, the sequencing algorithms works in two modes. In the first mode, the algorithm is working with one putative sequence, and each one-symbol extension is confirmed by up to $(r + 1)$ spectrum probes. When the algorithm fails to confirm a unique extension, it switches to the second mode, in which all possible paths extensions are explored up to maximum length H. In the second mode, each one symbol extension is confirmed by only one spectrum probe.

We first need to bound the work done in the first mode, and the probability of an "ambiguous branching," which occurs when more than one extension (the correct one and at least a spurious one) is confirmed by $(r + 1)$ spectrum probes, causing the algorithm to switch to the second mode of "path extension" in an attempt to resolve the ambiguity. The probability of these events is readily supplied by the following corollary to Lemma 2.

Claim 1. *The probability that h probes failed to confirm a unique extension at a specific position of the target sequence is bounded above by*

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

In particular, the probability of an ambiguous branching at a specific position of the target sequence is bounded above by

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

Proof. The only difference from the situation of Lemma 2 is that now each of the fooling probes $\sigma_2, \ldots, \sigma_{h+1}$ constrains at most $k-1$ rather than k symbols (all but the rightmost one). In addition, the probability of the first probe (σ_1 in Lemma 2) is $(3/4)(m/4^{k-1})$, because its last symbol is selectable in 3, not 4, ways. ■

Following an ambiguous branching, the algorithm switches to a second mode in which one-symbol extensions are verified with only one probe. A spurious path is either terminated or is not; in the latter case, it is either extended or it may spawn up to three additional paths. As we did for Event $\mathcal{E}_1$ in Theorem 1, we argue that the expected number of paths extended from a given symbol in that case is very conservatively bounded above by

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

On the other hand, the expected number of branches from the correct path (at a specific position in the single-probe path-extension mode) is bounded by $3\frac{m}{4^k}$, since it depends upon a single fooling probe with 3 choices for its last symbol. Thus, in both cases, the expected number of branches spawned at a given symbol is bounded away from 1.

Lemma 4. *If $A(m)$ denotes the total work on one symbol extensions in the first mode of the algorithm, then $E[A(m)] = O(m)$.*

Proof. The algorithm performs up to m one-symbol extensions in the first mode. By Claim 1, the probability that at least $h+1 \le r+1$ probe accesses are performed at a specific position is given by

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

Thus, the expected work done at a specific symbol of the target sequence is bounded by $\sum_{h=1}^{r+1} Z_h$, and

$$E[A(m)] \le m\sum_{h=1}^{r+1} Z_h = \frac{3m}{4}\left(1+\frac{4^{r+1}}{3m}\right)^{r-1}\sum_{h=1}^{r}\left(\frac{m}{4^{k-1}}\right)^h$$
$$\le \frac{3m}{4}\left(1+\frac{4^{r+1}}{3m}\right)^{r-1}\sum_{h=1}^{r}\left(\frac{1}{4^{\eta}}\right)^h = O(m).$$

■

Next we bound the total amount of work the algorithm performs in its second mode.

The expected number of ambiguous branchings on the target sequence (the expected number of times the algorithm switches to the second mode) is bounded by

$$\nu = \frac{3m}{4}\left(\frac{m}{4^{k-1}}\right)^{r+1}\left(1+\frac{4^{r+1}}{m}\right)^{r} \le m^{1-\frac{\eta}{2}}$$

Each time the algorithm switches to its second mode, it explores up to H symbols on the correct target sequence and at least $rs + 1$ symbols on the spurious path. We can analyze the total work done in the second mode as a collection of branching processes. The branches explored by the algorithm correspond to branching processes with roots at each symbol of the target sequence explored by the algorithm in the second mode, plus the first $sr + 1$ symbols of the confirmed spurious path.

There are a total of $v(H + rs + 1) = O(m^{1-\frac{\eta}{2}} \log^2 m) = o(m/\log m)$ such branching processes. The expected number of offsprings of each node in the branching process is bounded away from 1. Thus, the expected size of each branching process is $O(1)$. The work associated with each node of the tree is $O(1)$, since each branch is confirmed by only one spectrum access. Thus, the total expected work of the algorithm in the second mode 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 $O(m \log^2 m)$.

ACKNOWLEDGMENT

This work was partially supported by the National Science Foundation under grant DBI 9983081.

REFERENCES

Bains, W., and Smith, G.C. 1988. A novel method for DNA sequence determination. *J. Theoret. Biol.* 135, 303–307.

Dyer, M.E., Frieze, A.M., and Suen, S. 1994. The probability of unique solutions of sequencing by hybridization. *J. Comp. Biol.* 1, 105–110.

Drmanac, R., Labat, I., Bruckner, I., and Crkvenjakov, R. 1989. Sequencing of megabase plus DNA by hybridization. *Genomics* 4, 114–128.

Loakes, D., and Brown, D.M. 1994. 5-Nitroindole as a universal base analogue. *Nucl. Acids Res.* 22, 20, 4039–4043.

Lysov, Y.P., Florentiev, V.L., Khorlin, A.A., Khrapko, K.R., Shih, V.V., and Mirzabekov, A.D. 1988. Sequencing by hybridization via oligonucleotides. A novel method. *Dokl. Acad. Sci. USSR* 303, 1508–1511.

Pevzner, P.A. 1989. l-tuple DNA sequencing: Computer analysis. *J. Biomolec. Struct. Dynamics* 7, 1, 63–73.

Pevzner, P.A., Lysov, Y.P., Khrapko, K.R., Belyavsky, A.V., Florentiev, V.L., and Mirzabekov, A.D. 1991. Improved chips for sequencing by hybridization. *J. Biomolec. Struct. Dynamics* 9, 2, 399–410.

Pevzner, P.A., and Lipshutz, R.J. 1994. Towards DNA-sequencing by hybridization. *Nineteenth Symp. on Math. Found. of Comp. Sci.* LNCS-841, 143–258.

Preparata, F.P., Frieze, A.M., and Upfal, E. 1999. On the power of universal bases in sequencing by hybridization. *Third Annual International Conference on Computational Molecular Biology.* April 11–14, Lyon, France, 295–301.

Waterman, M.S. 1995. *Introduction to Computational Biology.* Chapman and Hall, London.

Address correspondence to:
Franco P. Preparata
Computer Science Department
Brown University
115 Waterman Street
Providence, RI 02912-1910

E-mail: franco@cs.brown.edu