JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 10, Numbers 3–4, 2003

Pp. 499–508

Sequencing by Hybridization by Cooperating Direct and Reverse Spectra

SAMUEL A. HEATH, FRANCO P. PREPARATA, and JOEL YOUNG

ABSTRACT

DNA sequencing by hybridization, potentially a powerful alternative to standard wet lab techniques, has received renewed interest after a novel probing scheme has been recently proposed whose performance for the first time asymptotically meets the information theory bound. After settlement of the question of asymptotic performance, there remains the issue of algorithmic fine tunings aimed at improving the performance "constants," with substantial practical implications. In this paper, we show that a probing scheme based on the joint use of direct and reverse spectra (tandem spectra) for a given gapped probing pattern achieves a performance improvement per unit of microarray area of about 5/4 and does not appear to be susceptible to further improvement by increasing the number of cooperating spectra. In other words, tandem-spectrum reconstruction is the best known technique for sequencing by hybridization.

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

1. INTRODUCTION

MORE THAN A DECADE AFTER ITS ORIGINAL CONCEPTION (Bains and Smith, 1988; Lysov *et al.*, 1988; Drmanac *et al.*, 1989; Pevzner, 1989; Pevzner *et al.*, 1991), sequencing by hybridization has not yet become operational because of the inherent limitations of the suggested probing schemes and of significant biochemical difficulties. As is well known, sequencing by hybridization consists of two fundamental steps. The first, biochemical in nature, is the acquisition, by complementary hybridization with a complete library of probes, of all distinct subsequences (of a selected pattern) of a given unknown target sequence; the set of such subsequences is called the sequence *spectrum.* The second step, combinatorial in nature, is the algorithmic reconstruction of the sequence from its spectrum. Most traditional schemes were restricted to probes in the form of strings of k symbols (k-mers), which could achieve no better performance (expressed as the length of reliably reconstructible sequences) than the square root of the information-theoretic bound (Dyer *et al.*, 1994; Southern, 1996). In this paper, we address the combinatorial aspects of the process, but we are perfectly aware that ultimate success of sequencing by hybridization is predicated on concomitant biochemical progress.

Recently, an innovative approach has been proposed (Preparata *et al.*, 1999; Preparata and Upfal, 2000) whose performance is within a factor of 2 of the optimum, the benchmark being, as customary, the ensemble

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

of maximum-entropy random DNA sequences. The central feature of this approach is the selection of the probing pattern. Bringing to full fruition earlier observations on the potential advantage of a gap in the probing pattern (Khrapko *et al.*, 1989; Pevzner *et al.*, 1991; Drmanac and Crkvenjakov, 1992; Pevzner and Lipshutz, 1994), the new method illustrates the power of multigap probing patterns, where natural nucleotides are separated by gaps of "wild-cards" intended to hybridize nonspecifically with any natural base and proposed to be realized with artificial universal bases (Loakes and Brown, 1994). The deployment of universal bases, which are still the object of biochemical investigations, is crucial in achieving array sizes depending only on the number of natural nucleotides and not on the overall length of the probing pattern.

It must be stressed that, since the question of the asymptotic performance has been settled, the focus of research in SBH sequence reconstruction has shifted to the refinement of the algorithmic procedures in order to fine-tune the computational methods. Only small improvements are to be expected, but even small improvements may have very significant impact in life science applications. We begin by reviewing the combinatorics underlying the approach and the known algorithmic techniques.

2. PRELIMINARIES AND REVIEW

A key feature of SBH is the adopted probing pattern, i.e., the structure of the spectrum probes. Notationally we have:

Definition 1. *A* probing pattern *is a binary string (beginning and ending with a* 1*), where a* 1 *denotes the position of a natural base and a* 0 *that of a universal base.*

Definition 2. *For* $s + r = k$*, an* (s, r) probing scheme *has* direct *and* reverse *patterns* $1^s(0^{s-1}1)^r$ *and* $(10^{s-1})^r 1^s$*, respectively.*

For convenience, a probe is viewed as a string of length $\nu = (r + 1)s$ over the extended alphabet $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}, *\}$, where $*$ denotes the wild-card. Conventionally, a probe occurs at position j of a sequence if j is the position of its rightmost symbol. Two strings over $\mathcal{A}$ of the same length *agree* if they coincide in the positions where both have specified symbols (i.e., symbols different from $*$).

Definition 3. *Given a sequence* $\mathbf{a}$ *over* $\mathcal{A}$ *the* spectrum *of* $\mathbf{a}$ *is the set of all of its* probes*, i.e., its subsequences conforming to the chosen probing pattern.*

Example 1. For sequence $\mathbf{a} = CGGATACACTTGCAT$ and (direct) pattern 111001001, the spectrum is $ACA**T**A(14)$, $ATA**C**G(12)$, $CAC**G**T(15)$, $CGG**A**C(9)$, $GAT**A**T(11)$, $GGA**C**T(10)$, $TAC**T**G(13)$ (probes listed lexicographically, their positions within parentheses).

The reconstruction of the sequence is algorithmically carried out symbol-by-symbol from one end to the other (Preparata *et al.*, 1999; Preparata and Upfal, 2000; Heath and Preparata, 2001). The reconstruction is successful if the constructed *putative* sequence coincides with the original *target* sequence.

Definition 4. *A spectrum probe is said to be a (current)* feasible extension *if its* $(\nu - 1)$*-prefix coincides with the (current)* $(\nu - 1)$*-suffix of the putative sequence.*

The algorithmic primitive of the reconstruction is the interrogation of the spectrum with a query of the form $u*$, where u is a string of length $(r + 1)s - 1$ over $\mathcal{A}$, which returns all agreeing spectrum probes.

We ignore in this paper the details concerning initiation and termination of the reconstruction process, which have been discussed elsewhere (Preparata *et al.*, 1999), and concentrate instead on the advancing

mechanism (extension of the putative sequence by concatenation of symbols to its right). The algorithm interrogates the spectrum with the query $u*$ defined above, the query returns all feasible extension probes contained in the spectrum. The following construct is the source of reconstruction ambiguities.

Definition 5. *A* fooling probe *is a feasible extension probe for position i which occurs as a subsequence at position $j \neq i$ in the target sequence.*

For convenience of presentation, if we observe the reconstruction algorithm between two consecutive extensions of the putative sequence, we distinguish two modes of operation:

- **Extension mode.** The query $q*$ returns a single match and the sequence is extended by a single symbol.
- **Branching mode.** The query $q*$ returns more than one match (ambiguous branching). The algorithm attempts the extension of all paths issuing from the branching (and of all other paths spawned in turn by them) on the basis of spectrum probes. The breadth-first construction of such tree is pursued up to a maximum depth H (a design parameter), unless at some stage of this construction it is found that all surviving paths have a common prefix, which is then concatenated to the putative sequence.

Failure occurs when, operating in the branching mode, at depth H the common prefix mentioned above is empty.

The rationale of this advancing mechanism is that, whereas the correct path is deterministically extended, the extension of the spurious paths rests on the (probabilistic) presence of fooling probes in the spectrum. The parameter H should be chosen large enough to make the probability of spurious paths vanishingly small. The behavior of the described algorithm has been analyzed in some detail in (Preparata and Upfal, 2000). Since, except for trivial lengths of the target sequence, there is always a nonzero probability of ambiguous reconstruction, performance is naturally measured as the length m of sequences that are reconstructible with a given confidence level, under the standard hypothesis that the target sequence is generated by a maximum-entropy memoryless source (i.i.d. symbols).

3. SEQUENCING WITH MULTIPLE SPECTRA

The use of more than a single spectrum of a given target sequence provides additional information that may conceivably improve the performance of the reconstruction process. We must remark, however, that the use of multiple spectra implies the deployment of the corresponding microarrays and their cost. Therefore, as an appropriate performance measure, we should take the ratio *nucleotides-per-feature*, i.e., the length of a reliably reconstructible sequence divided by the size of the microarray library.

For a given pair (s, r), suppose that we have separately obtained with two distinct microarrays (of identical cost) the direct and reverse spectra $D(a)$ and $R(a)$, respectively, of a given target sequence a of length m.

The most naive use of $D(a)$ and $R(a)$ is the separate execution of the two reconstruction processes. Failure occurs only if both reconstructions fail. An analysis of the statistical dependence of the two processes appears very arduous and perhaps not worth the effort in view of the conclusion reached at the end of this discussion. Suffice it to say that extensive simulations suggest that the two processes may, for all practical purposes, be treated as independent and with nearly identical probabilities of failure. Denoting P_1 and P_2 the probabilities of single- and combined-spectrum failures, respectively, equations $P_1 = \epsilon$ and $P_2 = \epsilon$ define the lengths of reliably reconstructible sequences in the two cases, for a fixed parameter ϵ. It has beeen shown in (Preparata and Upfal, 2000) that P_1 is, to a rough approximation, proportional to m^{k+1}; since $P_2 \approx P_1^2$, the performance gain is about $\epsilon^{-\frac{1}{2(k+1)}}$, an insignificant value (for example, for $\epsilon = 0.1$, we obtain a factor < 1.14), for a doubling of the cost.

The just-outlined scheme uses the spectra in a "global" fashion. In the next section, we shall demonstrate that a substantial advantage is to be gained by using the two spectra in a "local" (i.e., *probe-by-probe*) fashion.

4. PROBE-BY-PROBE COOPERATING SPECTRA (TANDEM SPECTRA)

Given spectra $R(a)$ (primary) and $D(a)$ (secondary) of a sequence a for a given (s, r) probing scheme, we propose the following reconstruction algorithm (with the previous terminology):

- In the extension mode only, the primary spectrum $R(a)$ is used; in the branching mode, each reverse probe is accepted only if confirmed by a direct probe ending at the same position (called a "companion probe").
- The algorithm fails when, in the tree of paths issuing from a branching and extended to a predetermined maximum length H, there are at least two paths differing in their initial symbol.

There are essentially two mechanisms which may cause failure, analyzed below:

- **Failure Mode 1.** *There are two paths identical except for their initial symbol* (in the branching position). This failure has been caused by $2k$ fooling probes scattered, with possible overlaps, along the target sequence.

Example 2. For direct probing pattern 111001001, suppose the algorithm detects the following situation:

```
...A  A  C  G  G  T  T  A  C  A  [T]  T  G  A  T  A  T  G  G  A  ...
                                [G]  T  G  A  T  A  T  G  G  A  ...
```

i.e., an ambiguous branching at the position indicated by the symbols within square brackets, where the paths issuing from the branching are identical except in their initial symbol. Assuming that the top path is correct, the spurious path is due to fooling probes 1–10 listed below:

```
... A  A  C  G  G  T  T  A  C  A  [T]  T  G  A  T  A  T  G  G  A ...
                                  [G]  T  G  A  T  A  T  G  G  A ...

 1        C  *  *  T  *  *  C  A  [G]
 2           G  *  *  T  *  *  A  [G]  T
 3              G  *  *  A  *  *  [G]  T  G
 4                       A  *  *  [G]  *  *  A  T  A
 5                                [G]  *  *  A  *  *  T  G  G
 6        C  G  G  *  *  A  *  *  [G]
 7              T  T  A  *  *  [G]  *  *  A
 8                       C  A  [G]  *  *  A  *  *  T
 9                          A  [G]  T  *  *  T  *  *  G
10                             [G]  T  G  *  *  A  *  *  G,
```

of which 1–5 belong to the reverse spectrum and 6–10 to the direct spectrum. On the basis of the spectrum alone, there is no way to decide which of the symbols T and G detected at the branching represents the correct extension.

Failure Mode 1 is well understood and discussed elsewhere (Preparata and Upfal, 2000). Suffice it here to recall the following. Analytical estimates should hold in the ranges of parameters for which the approach is targeted, i.e., for probability of failure not exceeding some value $\epsilon \approx 0.2$, roughly corresponding to m not larger than 4^{k-1}. In this general range of parameters, an ambiguous extension (the branching expressed by the first fooling probe in Example 2) occurs with approximate probability

$$1 - \left(1 - \frac{3}{4} \cdot \left(\frac{1}{4}\right)^{k-1}\right)^m \approx 1 - e^{-\frac{3m}{4^k}}$$

since the corresponding fooling probe occurs somewhere in the sequence, $k - 1$ of its symbols are fully constrained (each with probability $1/4$), and one of them is constrained to 3 out of 4 choices. Analogously,

each of the remaining $2k-1$ fooling probes occurs with probability $\alpha = 1 - e^{-m/4^k}$, so that the overall probability of Mode 1 failure is sufficiently well represented (Preparata and Upfal, 2000) by the product

$$P^{(1)} = m\left(1 - e^{-\frac{3m}{4^k}}\right)\alpha^{2k-1} \tag{1}$$

where the first factor corresponds to the choice of the failure position in the sequence, and we have omitted any correction for probe overlaps, with an insignificant loss of accuracy (refer to the analysis presented in Preparata and Upfal [2000], pp. 624–628).

- **Failure Mode 2.** *There are no two paths as specified for Failure Mode 1.* If H is large enough to make probabilistic extension of a spurious path to length H extremely unlikely, then there must be a $(\nu-1)$-segment of the spurious path containing the branching position or beginning closely thereafter that is identical to an actual segment occurring *elsewhere* in the sequence. Obviously, since the length of such segment is $\nu - 1$, extension of the spurious path is deterministically assured since the spectrum contains all necessary probes, and the reconstruction fails. (Such segments are conveniently referred to as *self-sustaining.*) The self-sustaining segment agrees, entirely or partially, with an aligned segment of the correct path, the disagreements being compensated for by fooling probes also occurring in the sequence.

Example 3. Again, for direct probing pattern 111001001, suppose the algorithm detects the following situation:

...A	C	G	A	G	T	C	(C	T	[G]	A	G	T	G	A	T	A	T	A	T	...
									[T]	A	G	T	A	A)	T	C	T	G	G	...

where the pair [G][T] is the ambiguous branching, the top path represents the correct extension, and in the spurious bottom path, enclosed within parentheses, is the length-8 self-sustaining segment CTTAG-TAA. This segment occurs *elsewhere* in the sequence. Clearly indefinite extension of the spurious path is guaranteed by the spectrum. Segment CTTAGTAA is brought about by an appropriate collection of fooling probes, which compensate for disagreements between the two paths. Below, the disagreements are evidenced within brackets.

									0	1	2	3	4	5	6	7	8	9	...	
...A	C	G	A	G	T	C	(C	T	[G]	A	G	T	[G]	A	T	A	T	A	T	...
									[T]	A	G	T	[A]	A)	T	C	T	G	G	...
1	C	*	*	G	*	*	C	T	T											
2		G	*	*	T	*	*	T	T	A										
3			A	*	*	C	*	*	T	A	G									
4						C	*	*	T	*	*	T	A	A						
5					T	*	*	T	*	*	G	T	A							
6	C	G	A	*	*	C	*	*	T											
7				G	T	C	*	*	T	*	*	T								
8					T	C	C	*	*	A	*	*	A							

Of these, probes 1–5 and 6–8 belong, respectively, to the reverse and direct spectra. Moreover, in the reverse spectrum, the branching disagreement [G-T] is compensated by probes 1–4, and disagreement [G-A] is compensated by 4,5. In the direct spectrum, branching [G-T] is compensated by 6,7, and [G-A] by 8. No other fooling probe is required, because the required extending probes are guaranteed in the spectrum.

We denote conventionally as 0 the branching position. The position immediately to the right of the self-sustaining segment is called the *offset* and denoted J. Thus, $J \geq 0$.

The failure corresponding to $J = 0$ arises from the situation where there are two identical length-$(\nu-1)$ segments occurring at different places in the target. Such an event is constructed by selecting two positions

in the sequence (in $\binom{m}{2} \approx m^2/2$ ways), of which the leftmost (occurring earlier in the reconstruction) identifies the correct path and the rightmost the spurious path. The corresponding probability is therefore

$$\frac{m^2}{2}\frac{1}{4^{\nu-1}}\frac{3}{4} = m^2\frac{3}{4^{\nu-1}\cdot 8} = m^2\pi_0$$

since the $\nu - 1$ symbols of the self-sustaining segment are fully constrained and the branching symbol is selectable in three ways.

For $J > 0$, the positions of the two homologous segment are not interchangeable, so that there are about m^2 ways of selecting the event, and its probability can be expressed as $m^2\pi_J$, for some coefficient π_J.

A detailed analysis of the terms π_J for $J > 0$, with appropriate approximations, is developed in an appendix to this paper (refer to Formulae (3) and (4) in the appendix). Using also the value of π_0 defined above, we obtain an estimate for the probability of Mode 2 failure:

$$\begin{aligned} P^{(2)} &= m^2 \sum_{J=0}^{(r+1)s} \pi_J \\ &= \frac{3m^2}{4^{\nu-2}}\left[\frac{1}{32} + \alpha^2\left(\frac{1-(\alpha(1+3\alpha))^s}{1-\alpha(1+3\alpha)} + (\alpha(1+3\alpha))^{s-1}\frac{1-(4\alpha^\mu)^{rs}}{1-4\alpha^\mu}\right)\right] \end{aligned} \tag{2}$$

where parameter μ satisfies the condition $\mu \geq 15/8$. As we shall discuss in the next section, the analytical probability $(1 - P^{(1)} - P^{(2)})$ of successful reconstruction[1] is in satisfactory agreement with extensive experimental results.

5. DISCUSSION

A fair comparison of alternative schemes for sequence reconstruction must be made with reference to a fixed microarray area (in a chosen technology). In other words, if Scheme 1 with area A_1 achieves performance m_1 and Scheme 2 has corresponding parameters A_2 and m_2, the meaning of the statement "Scheme 1 outperforms Scheme 2" is that

$$m_1 > \frac{A_1}{A_2}m_2$$

(i.e., that Scheme 1 has a higher *nucleotide-per-feature* ratio than Scheme 2). Therefore, when comparing single-spectrum operation (Scheme 1) with tandem-spectrum operation (Scheme 2), since $A_2 = 2A_1$, we must compare m_2 against $2m_1$. In Fig. 1, we display the theoretical probabilities of failure for the single-spectrum and tandem-spectra operations for a $(4, 4)$ scheme, together with their corresponding failure frequencies obtained in extensive simulations with random sequences. Thus, data for single-spectrum operation are displayed with the abscissa dilation discussed above.

Referring for simplicity just to Failure Mode 1 (an analogous argument can be made for Mode 2), if we expand the exponential to second-degree terms and ignore probe overlap, we obtain the following approximations for single- and tandem-spectrum operation (with the correct abscissa dilation)

$$p_1 = \frac{3m}{2}\left(\frac{m}{2\cdot 4^k} - \frac{1}{2}\left(\frac{m}{2\cdot 4^k}\right)^2\right)^k$$

[1] More accurately, the probability of correct reconstruction should be expressed as $e^{-(P^{(1)}+P^{(2)})}$, the limiting value of

$$\left(1 - \frac{P^{(1)} + P^{(2)}}{m}\right)^m.$$

However, in the interesting range of values (≈ 1), the two expressions are indistinguishable.

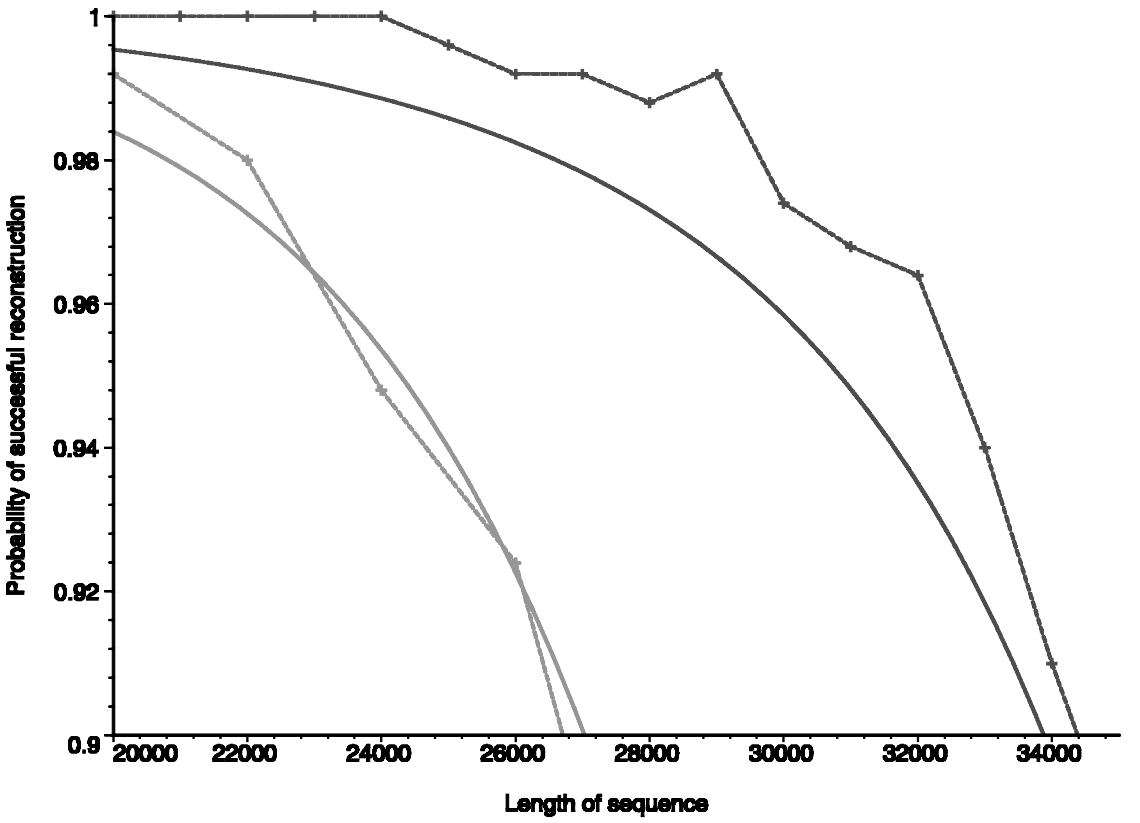

FIG. 1. Analytical and experimental performance diagrams for tandem-spectrum and single-spectrum operations (with abscissa dilation) for a (4,4)-scheme.

and

$$p_2 = 3m\left(\frac{m}{4^k} - \frac{1}{2}\left(\frac{m}{4^k}\right)^2\right)^{2k}.$$

This approximation is rather accurate in the chosen range ($m/4^k < 0.6$). Considering first the single-spectrum case, we introduce the unknown $x = m/(2 \cdot 4^k)$, and, for a chosen confidence level ϵ, we must solve the equation $p_1 = \epsilon$, i.e.,

$$2x - x^2 - \frac{1}{2}\left(\frac{\epsilon}{3x}\right)^{1/k} = 0.$$

Analogously, for the tandem-spectrum case, letting $y = m/4^k$, the equation $p_2 = \epsilon$ becomes

$$2y - y^2 - \left(\frac{\epsilon}{3y}\right)^{1/2k} = 0.$$

To find adequate analytical approximations to the solutions x and y appears very cumbersome. However, the above equations can be numerically solved for a fixed value of ϵ (typically $\epsilon = 0.1$, i.e., 90% confidence) and for different practicable values of the parameter k. From these we may compute the ratio

$$\frac{m_2}{m_1} = \frac{y}{2x}.$$

These values are displayed in Table 1 and provide an analytical confirmation of the previous empirical observation.

TABLE 1. TANDEM-SPECTRUM GAIN AS A FUNCTION OF k

k	5	6	7	8	9	10
x	.1946	.2057	.2146	.2218	.2278	2329
y	.5112	.5443	.5714	.5941	.6134	.6302
m_2/m_1	1.319	1.323	1.331	1.339	1.346	1.352

We also note that as $\epsilon \to 0$, using the Taylor's expansions of the function $f(z) = \sqrt{1-z}$ to first-degree terms, we obtain

$$x \to \left(\frac{\epsilon}{3 \cdot 4^k}\right)^{\frac{1}{k+1}} \quad \text{and} \quad y \to \left(\frac{\epsilon}{3 \cdot 4^k}\right)^{\frac{1}{2k+1}}$$

so that

$$\frac{m_2}{m_1} = \frac{y}{2x} \to \left(\frac{\epsilon}{3 \cdot 2^{3+1/k}}\right)^{-\frac{k}{(2k+1)(k+1)}}$$

which shows that $m2/m1$ is a decreasing function of the parameter ϵ (higher gains for higher levels of confidence).

APPENDIX: AN ESTIMATE OF THE PROBABILITY OF MODE 2 FAILURE

To estimate the probability of Mode 2 failure, we note that each spectrum contains a set of fooling probes necessary to compensate for the branching disagreements (*branching* probes). Precisely, each spectrum contains all the probes that sample the branching position and whose rightmost symbol does not fall to the right of the self-sustaining segment (refer to Example 2, where the segment ends at position 5, the reverse spectrum does not contain the fooling probe ending at position 9, and the direct spectrum does not contain probes ending at 6,7,8). With respect to the branching at position 0, the reverse-spectrum branching probes end at positions

$$S_R = \{1, 2, \ldots, s, 2s, \ldots, (r+1)s\}$$

while those of the direct spectrum end at

$$S_D = \{1, s+1, 2s+1, \ldots, rs+1, rs+2, \ldots, (r+1)s\}.$$

In detail, we have the following.

1. For $J = 1, 2 \ldots, s$, the reverse spectrum requires J fooling probes, and the direct spectrum requires, in addition to the probe ending at 1, a fooling probe in each position of disagreement. Since an agreement has probability $1/4$ and the events are independent, each position contributes a factor

$$\frac{1}{4} + \frac{3}{4} \cdot \alpha = \frac{1+3\alpha}{4}$$

(where we recall that α denotes the probability of a specific fooling probe in a spectrum), so that

$$\pi_J = \frac{1}{4^{\nu-1-J}} 3\alpha^2 \cdot \left(\alpha \frac{1+3\alpha}{4}\right)^{J-1} = \frac{3\alpha^2(\alpha(1+3\alpha))^{J-1}}{4^{\nu-2}}.$$

We conclude that

$$\sum_{J=1}^{s} \pi_J = \frac{1}{4^{\nu-2}} 3\alpha^2 \frac{1-(\alpha(1+3\alpha))^s}{1-\alpha(1+3\alpha)}. \tag{3}$$

2. For $s+1 \le J \le (r+1)s$, the analysis becomes quite cumbersome, because in these cases the requirement of fooling probes ending at a given position $j \le J$ depends not only upon the agreement/disagreement at j, but also upon the agreement/disagreement at sets of positions $i < j$.

Here we make the simplifying assumption (formally not justifiable, but yielding results in good agreement with experimental data) that the total number of fooling probes required by positions $s+1, s+2, \ldots, J$ is given by $\mu(J-s)$, where μ is a lower bound to the average number of fooling probes required by any position in this interval considered in isolation (note that this corresponds to assuming statistical independence). The quantity μ is provided by the following proposition.

Proposition 1. *The average number of fooling probes required at any position $j \in [s+1, (r+1)s]$ is at least*

$$\mu = \frac{15}{8}.$$

Proof. A position $j \in [s+1, (r+1)s]$ is classified on the basis of its membership in the sets S_R and S_D defined above. Recall that for nontrivial gapped probing we must have $s \ge 1$.

(a) $j \in S_D \cap \overline{S_R}$. In this case, position j certainly requires a fooling probe in the direct spectrum because $j \in S_D$. In addition, it may require a fooling probe in the reverse spectrum if any position $j-i$ is the site of a disagreement, where either $i = 0$ or $i \in S_R$ with $i < j$. Since $j \ge s+1$, we immediately recognize that $i = 0, 1, \ldots, s$ satisfy the condition $i < j$, so that the additional fooling probe is required with probability at least $1 - 4^{-(s+1)} \ge 15/16$. We conclude that the average number of fooling probes is at least

$$1 \cdot \frac{1}{16} + 2 \cdot \frac{15}{16} = \frac{31}{16} > \frac{15}{8}.$$

(b) $j \in S_R \cap \overline{S_D}$. By an analogous argument, we recognize that, beside the mandatory probe in the reverse spectrum, an additional probe in the direct spectrum is required for $i = 0, s$, so that the average number of fooling probes is again at least $31/16 > 15/8$.

(c) $j \in \overline{S_R \cup S_D}$. In this case, two fooling probes are required either if j itself is the site of a disagreement (with probability $3/4$) or if there are disagreements at $j-i$ ($i \in S_R$, with probability $\ge 1 - 1/4^s$) and at $j-h$ ($h \in S_D$, with probability $\ge 1 - 1/4 = 3/4$). We conclude that the probability of exactly two fooling probes is at least

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

and the probability of exactly one fooling probe is at least

$$\frac{1}{4}\left(\frac{1}{4}\left(1 - \frac{1}{4^s}\right) + \frac{1}{4^s}\left(1 - \frac{1}{4}\right)\right)$$

so that the average number is at least

$$2\left(\frac{3}{4} + \frac{1}{4}\left(1 - \frac{1}{4}\right)\left(1 - \frac{1}{4^s}\right)\right) + 1 \cdot \left(\frac{1}{4}\left(\frac{1}{4}\left(1 - \frac{1}{4^s}\right) + \frac{1}{4^s}\left(1 - \frac{1}{4}\right)\right)\right) > \frac{15}{8}.$$

(d) $j \in S_R \cap S_D$ ($J = (r+1)s$). In this case, $\mu = 2 > 15/8$, trivially. ■

It follows that

$$\begin{aligned}\pi_J &= \frac{1}{4^{\nu-1-J}} 3\alpha^2 \frac{(\alpha(1+3\alpha))^{s-1}}{4^{s-1}} \alpha^{\mu(J-s)} \\ &= \frac{1}{4^{\nu-2}} 3\alpha^2 (\alpha(1+3\alpha))^{s-1} (4\alpha^{\mu})^{(J-s)}\end{aligned}$$

and

$$\sum_{J=s+1}^{(r+1)s} \pi_J = \frac{1}{4^{\nu-2}} 3\alpha^2 (\alpha(1+3\alpha))^{s-1} \frac{1-(4\alpha^{\mu})^{rs}}{1-4\alpha^{\mu}}. \quad (4)$$

It is easy to verify that for $J > (r+1)s$, the probability of failure is totally negligible.

ACKNOWLEDGMENTS

Thanks are due to the referees, whose comments significantly improved the paper. Coauthor Joel Young, a member of the U.S. Air Force, adds the following statement: "The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government."

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. Theor. Biol.* 135, 303–307.

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

Drmanac, R., and Crkvenjakov, R. 1992. Sequencing by hybridization (SBH), with oligonucleotide probes as an integral approach for the analysis of complex genomes. *Int. J. Genome Res.* 1, 59–79.

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.

Heath, S.A., and Preparata, F.P. 2001. Enhanced sequence reconstruction with DNA microarray application. *COCOON 2001*, Guilin, China, 64–74.

Khrapko, K.R., Lysov, Y.P., Khorlin, A.A., Shik, V.V., Florentiev, V.L., and Mirzabekov, A.D. 1989. An oligonucleotide approach to DNA sequencing. *FEBS Letters* 256, 118–122.

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., Shik, 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. Biomol. Struct. Dyn.* 7(1), 63–73.

Pevzner, P.A., and Lipshutz, R.J. 1994. Towards DNA-sequencing by hybridization. *19th Symp. on Math. Found. of Comp. Sci.* LNCS-841, 143–258.

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. Biomol. Struct. Dyn.* 9(2), 399–410.

Preparata, F.P., Frieze, A.M., and Upfal, E. 1999. On the power of universal bases in sequencing by hybridization. *Proc. 3rd Ann. Int. Conf. Computational Molecular Biology*, 295–301.

Preparata, F.P., and Upfal, E. 2000. Sequencing-by-hybridization at the information-theory bound: An optimal algorithm. *J. Comp. Biol.* 7(3/4), 621–630.

Southern, E.M. 1996. DNA chips: Analysing sequence by hybridization to oligonucleotide on a large scale. *Trends Genet.* 12(3), 110–115.

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