Optimal Reconstruction of a Sequence From its Probes

Alan M. Frieze* Franco P. Preparata† Eli Upfal‡

August 5, 1999

Abstract

An important combinatorial problem, motivated by DNA sequencing in molecular biology, is the reconstruction of a sequence over a small finite alphabet from the collection of its probes (the sequence *spectrum*), obtained by sliding a fixed sampling pattern over the sequence. Such construction is required for Sequencing-by-Hybridization (SBH), a novel DNA sequencing technique based on an array (SBH chip) of short nucleotide sequences (*probes*). Once the sequence spectrum is biochemically obtained, a combinatorial method is used to reconstruct the DNA sequence from its 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. We present in this work a novel probe design, crucially based on the use of universal bases (bases that bind to any nucleotide [LB94]) that drastically improves the performance of the SBH process and asymptotically approaches the information-theoretic bound up to a constant factor. Furthermore, the sequencing algorithm we propose is substantially simpler than the Eulerian path method used in previous solutions of this problem.

*Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh PA15213, USA `af1p@andrew.cmu.edu`. Supported in part by NSF grant CCR-9530974.

†Computer Science Department, Brown University, Box 1910, Providence, RI 02912-1910, USA. E-mail: `franco@cs.brown.edu`.

‡Computer Science Department, Brown University, Box 1910, Providence, RI 02912-1910, USA. E-mail: `eli@cs.brown.edu`. Supported in part by NSF grant CCR-9731477.

1 Introduction

The reconstruction of a sequence over a finite alphabet from the set of its subsequences, sampled according to a fixed pattern, is a challenging combinatorial problem, which has received considerable attention in recent years. A pattern can be defined as a binary sequence beginning and ending with a 1, which can be used as a "template" to sample a given sequence, called the target sequence. Specifically, the samples (*probes*) are obtained by sliding the pattern in all positions of complete overlap with the target sequence, and generating from each position the subsequence corresponding to the 1-symbols of the pattern. The resulting collection of probes is called the *spectrum* of the sequence, and the reconstruction task consists of deciding if there is a unique sequence consistent with a spectrum and, if so, to construct it.

Although interesting on a purely information-theoretic level, the motivation for this problem comes from molecular biology, specifically from the sequencing of DNA. In recent times a radically new technique, called *Sequencing by Hybridization*, has been proposed as an alternative to the traditional sequencing by gel electrophoresis [BS91, L+88, D+89]. Sequencing-by-hybridization is based on the use of a chip, fabricated with photolithographic techniques. The active area of the chip is structured as a matrix, each region of which (technically called a *feature*) is assigned to a specific oligonucleotide (or to a set of oligonucleotides), biochemically attached to the chip surface. When a solution of suitably labeled target DNA is applied to the chip, a copy of the target DNA will bind to an oligonucleotide if the latter is Watson-Crick complementary to one of its subsequences. The labeling of the target allows visualization of the binding chip features, thereby yielding the spectrum of the target sequence.

For a fixed cost, expressed by the number of features of the chip (or equivalently by the number k of specified nucleotides of the probes), a challenging combinatorial problem is the design of a most efficient probing scheme, that would yield the maximum length of the sequences for which faithful reconstruction is guaranteed with a given level of confidence.

Pioneering work on this topic [BS91, L+88, D+89] focused on probing schemes based on k-grams (strings of k symbols), which we shall refer to as "classical" probing schemes. To reconstruct the target sequence from its k-grams, original approaches dealt with a subgraph G of the order-k shift-register diagram (De Bruijn graph), so that a consistent reconstruction is identified with a Hamiltonian path in G. Substantial progress was made by Pevzner [P89], who characterized a consistent reconstruction with an Eulerian path in a subgraph G' of the order-$(k-1)$ shift-register diagram, such that an arc from $(k-1)$-gram u to $(k-1)$-gram v exists if and only if u and v are respectively prefix and suffix of a spectrum k-gram. This insight both simplified and characterized the reconstruction problem for k-gram probes.

However, the effectiveness of the methods was unsatisfactory. In the model where the target sequences are generated by a memoryless source with identical symbol probabilities (all symbols independent and uniformly distributed), it was observed [P+91] (and a tight bound of the same order was established [DFS94, A+96]) that the expected length of unambiguously reconstructible sequences with k-gram probes was $O(2^k)$. By contrast, an information-theoretic argument yields an upper bound $\Theta(4^k)$.

Probe structures alternative to the classical one have also been been proposed. One such design, not analyzed in any detail, introduces one gap of "don't care"" symbols (or *universal bases*), separating a string of specified symbols and a single specified symbol. Today, there is technological justification for truly universal bases, that –if used in short runs – stack correctly without binding.

Under the same statistical model of a memoryless maximum-entropy sequence generator, in this paper we show that the use of "don't care"s is essential to the attainment of asymptotically optimal efficiencies. Specifically. we exhibit a class of novel probe designs, with a well defined periodic pattern of gaps of "don't care"s, which for any k uses 4^k probes to sequence a target sequence of length $\Theta(4^k)$. Our approach does not involve the construction of an Euler path. This apparent paradox (with respect to Pevzner's characterization) is resolved by the observation that our proposed gap structure trivializes the Euler path identification problem, guaranteeing with extremely high probability, in the chosen statistical model, that the Euler path reduces to a simple path in a virtual De Bruijn graph of order-$\Theta(k^2)$. In other words, the full potential of sequencing by hybridization is predicated on the reliable deployment of universal bases.

Although our considerations are applicable to any finite symbol alphabet $\mathcal{A}$, due to the central relevance of our scheme to biological applications, in the rest of the paper we shall normally assume that $|\mathcal{A}| = 4$.

2 Preliminaries and the (s, r)-gapped probes

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 $\mathcal{A} = \{$ A,C,G,T,*$\}$, where A,C,G, and T denote the standard DNA bases and $*$ denotes the "don't care " symbol, implemented using a *universal base* [LB94].

When the SBH chip is brought in contact with a solution of the target DNA string, a probe binds to the target string if and only if there is a substring of the target that is *Watson-Crick complementary* to the probe (where, conventionally, any of the four bases A,C,G,T is Watson-Crick complementary to a universal base. With this convention, a probe is viewed as a string, rather than a subsequence). Biochemical labeling permits the identification of the set of probes (called the string's *spectrum*)

that bind to the target string.

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.

Since the number of features on an SBH chip is limited by the technology, we are interested in the design of a smallest set of probes adequate for sequencing an arbitrary string of a given length.

The following simple observation gives an information-theoretic lower bound for the size of such a set:

Theorem 1 *The number of probes required for unambiguous reconstruction of an arbitrary string of length m is $\Omega(m)$.*

Proof: The spectrum based on t probes is a binary vector with t components. There are 2^t such vectors, and each can define no more than one possible sequence. Thus, $4^m \leq 2^t$, or $t \geq 2m$. $\Box$

This theorem also implies that, in the important case $t = 4^k$, we have $m \leq 4^{k-1/2}$. Past research [P+91, DFS94, A+96] analyzed the performance of SBH chips in the context of random strings of length m, drawn uniformly at random from the set $\mathcal{A}^m$. A similar lower bound holds in that model:

Theorem 2 *For any fixed probability $P > 0$, the number of probes required for unambiguous reconstruction with probability P of a random string of length m is $\Omega(m)$.*

Proof: Since the algorithm must unambiguously reconstruct $P4^m$ sequences, the number of probes t must satisfy $P4^m \leq 2^t$, or $t = \Omega(m)$. $\Box$

In this paper we focus on a special *pattern* of probes which we name (s,r)-*gapped probes* and denote $GP(s,r)$.

Definition 1 *For fixed parameters s and r the set $GP(s,r)$ of $(s,r,)$-gapped probes consists of all probes of the form $X^s(U^{s-1}X)^r$ where X ranges over the 4 standard DNA bases (A,C,G, and T) and U is the universal base.*

Since there are $s + r$ locations with an X symbol in each probe in $GP(r,s)$, the set of probes $GP(s,r)$ consists of exactly 4^{r+s} individual probes.

Notationally, let $a_{(1,m)} = a_1, ..., a_m$ be the target string, and for any $1 \leq i < j \leq m$ let $a_{(i,j)} = a_i, ..., a_j$. Given $a_{(i,j)}$ and $i < h < j$, $a_{(i,h)}$ and $a_{(h,j)}$ are respectively the $(h - i + 1)$-prefix and the $(j - h + 1)$-suffix of $a_{(i,j)}$. Hereafter we assume that the set of probes $GP(s,r)$ was used to obtain a spectrum of the string $a_{(1,m)}$.

3 The sequencing procedure

We describe a simple procedure for sequencing the string a using the spectrum information obtained from the (s,r)-gapped probes. To simplify the presentation we assume that we are given the $s(r+1)$-prefix of the target string. (Section 6 explains how to remove this assumption.)

The procedure produces a *putative* sequence b which represents the reconstruction of the sequence a. It starts with the prefix $b_{(1,s(r+1))} = a_{(1,s(r+1))}$. At each iteration the procedure tries to extend a current *putative* sequence $b_{(1,\ell-1)} = b_1, ..., b_{\ell-1}$, $\ell - 1 \geq s(r+1)$ with a new symbol b_ℓ.

To take full advantage of the $GP(s,r)$ probes, we use each probe in up to r different possible alignments with the current sequence.

The extension is attempted as follows. We find the set M_0 of all probes in the spectrum such that the $(s(r+1)-1)$-prefix of each of the probes matches the $(s(r+1)-1)$-suffix $b_{(\ell-s(r+1)+1,\ell-1)}$ of the putative sequence, with the stated convention about don't care symbols. If M_0 is empty, then no extension exists and the algorithm terminates. Otherwise, if $|M_0| = 1$ a single extension is defined and the corresponding symbol is appended to the putative sequence. The case $|M_0| > 1$ is problematic since it suggests an ambiguous extension. Here we use the power of the $GP(s,r)$ probes, since an ambiguous extension is detected only if confirmed by $r+1$ spectrum probes, as discussed below. If these probes confirm the ambiguous extension, either they occur scattered along the target sequence (and are referred to briefly as "fooling probes") or they originate from a single substring (of adequate length). Intuitively, our approach rests on the facts that $(r+1)$ confirmatory fooling probes are very improbable, and that even more improbable is their arising from a single substring.

When M_0 is not a singleton, 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 $b_{(\ell-sr+1,\ell-1)}$, and their $(s+1)$-suffix agrees [1] 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. More sophisticated algorithms, not discussed in this paper, may explore all branches of an ambiguous extension, in the expectation that after a small number of extensions

[1] Agreement is obviously restricted to the specified positions, appropriately shifted.

only one branch will be supported by the spectrum.

The success of the above algorithm stems from the fact that up to r probes, appropriately aligned along the current sequence, are used to confirm the uniqueness of a one-symbol extension. One could try to extend the "power" of any set of probes by using various alignments with the current string. The advantage of the set $GP(s,r)$ is that the probability of ambiguous extension in each of the alignments, with respect to a randomly generated sequence, is almost independent of the other patterns. This property is central to the analysis presented in the next section.

4 Analysis of the sequencing procedure

We present in this section a relatively simple analysis of the performance of the algorithm described in the previous section when applied to a spectrum obtained using $GP(s,r)$ probes. We will show that the performance of this scheme approaches the information-theoretic lower bound of Theorem 2. To simplify the presentation we assume again that together with the spectrum the algorithm is provided with the $s(r+1)$-prefix of the target sequence. We will show in section 6 that this assumption can be removed without altering the performance of the sequencing scheme.

Theorem 3 *For constants $1 < \gamma = O(\log m)$ and $\beta = o(\log m)$, such that r and s are integers, let:*

$$\begin{aligned} r &= \frac{1}{\gamma}\log_4 m + \beta \\ s &= \log_4 m + 1 + \gamma - r. \end{aligned}$$

Let $\mathcal{E}$ be the event: The algorithm fails to sequence a random string of length m using a $GP(s,r)$ spectrum of the string. Then:

$$Pr(\mathcal{E}) \;\leq\; 4^{-\gamma(1+\beta)}.$$

Proof:

Let $\mathbf{t} = \{t, t_0, t_1, \ldots, t_r\}$, denote a vector of $r+2$ positions in the target string, and let $\mathcal{A}(\mathbf{t})$ denote the event: there are substrings in the target sequence $a_{(1,m)}$ that satisfy the following relations:

$$\begin{array}{lcll} a_{(t_0+1,t_0+s)} & = & a_{(t+1,t+s)} & \mathcal{B}_0(\mathbf{t}) \\ a_{t_0+is} & = & a_{t+is} \qquad 2 \leq i \leq r. & \mathcal{C}_0(\mathbf{t}) \\ a_{t_0+(r+1)s} & \neq & a_{t+(r+1)s} & \mathcal{D}_0(\mathbf{t}) \end{array}$$

For $1 \leq j \leq r$

$$\begin{array}{lcll} a_{(t_j+1,t_j+s-1)} & = & a_{(t+js+1,t+(j+1)s-1)} & \mathcal{B}_j(\mathbf{t}) \\ a_{t_j+is} & = & a_{t_{j-1}+(i+1)s} \qquad 1 \leq i \leq r. & \mathcal{C}_j(\mathbf{t}) \end{array}$$

We focus first on the success of the algorithm in sequencing all but the last $(r+1)s$ symbols of the target sequence.

Claim 1 *The algorithm fails to sequence the $m-(r+1)s$ prefix of the target string if and only if $\exists \mathbf{t}$ such that $\mathcal{A}(\mathbf{t})$ occurs.*

Proof: Assume that the algorithm is trying to extend the current sequence $a_{(1,\ell-1)}$ with the next symbol a_ℓ. Let $t = \ell - s(r+1)$. If $|B_0| > 1$ is not a singleton then there is a probe in the spectrum that matches $a_{(t+1,\ell-1)}$ but its rightmost symbol $b \neq a_\ell$. Denoting by $a_{(t_0+1,t_0+s(r+1))}$ the substring of the target string that binds with that probe, conditions $\mathcal{B}_0$,$\mathcal{C}_0$ and $\mathcal{D}_0$ hold.

If $\cap_{j=0}^{r} B_j$ is not a singleton, then it contains both a_ℓ and b. Thus, for each j there is a probe in the spectrum, and a corresponding substring $a_{(t_j+1,t_j+(r+1)s)}$ in the target sequence, such that the s-prefix of that substring matches $a_{(t+js+1,t+(j+1)s)}$, and the locations t_j+is of the substring, for $2 \leq i \leq r$ match the corresponding locations (with a shift of s positions) of the substring $a_{(t_{j-1}+1,t_{j-1}+(r+1)s)}$ as formulated in conditions $\mathcal{B}_j$ and $\mathcal{C}_j$.

□

Let $\mathcal{T}$ denote the set of all possible vectors $\mathbf{t}$, so that

$$\binom{m-2(r+1)s}{r+2}(r+2)! \leq |\mathcal{T}| \leq \binom{m}{r+2}(r+2)!. \tag{1}$$

Suppose $t_{j_1} < t_{j_2} < \ldots < t_{j_{r+2}}$, where $(j_1, j_2, \ldots, j_{r+2})$ is a permutation of $(-1, 0, 1, \ldots, r)$, $t \equiv t_{-1}$. If $t_u - t_{u-1} \geq s(r+1)$, then all regions of definition of $a_{(t_j+1,t_j+(r+1)s)}$ are disjoint and the $2(r+1)$ $\mathcal{B}$ and $\mathcal{C}$ events are trivially independent.

Suppose now that, for some u, $t_{j_u} - t_{j_u-1} < s(r+1)$. The corresponding $\mathcal{B}$ events are still independent among themselves and of the $\mathcal{C}$ events, since although their regions overlap, they constrain disjoint regions (symbols) of $a_{(t+1,t+(r+1)s)}$. The $\mathcal{C}$ events, however, are independent only if $t_{j_u} \neq t_{j_u-1} \mod s$, since in this case their regions are disjoint; if $t_{j_u} = t_{j_u-1} \mod s$, then their regions overlap and they constrain overlapping regions of $a_{(t+1,t+(r+1)s)}$: in such case, as few as $r+1$ symbols are constrained rather than $2(r+1)$. However, this happens only if t_{j_u} occurs in one of $(r+1)$ evenly spaced positions in the range $[t_{j_u-1}+1, t_{j_u-1}+(r+1)s]$.

The preceding discussion indicates that the crucial feature of a vector $\mathbf{t}$ is the number $\nu(\mathbf{t})$ of its independent $\mathcal{C}$ events, which ranges between 1 and $r+1$. Therefore we define the following sets which partition $\mathcal{T}$:

$$\mathcal{T}_i = \{\mathbf{t} \in \mathcal{T} : \ \nu(\mathbf{t}) = i\}.$$

We first bound from above the probability of a given event $\mathcal{A}(\mathbf{t})$. If $\mathbf{t} \in \mathcal{T}_0$ then the $r+1$ probes in the definition of $\mathcal{A}(\mathbf{t})$ are associated with disjoint regions of the

string $a_{(1,m)}$, and thus the $2(r+1)$ $\mathcal{B}$ and $\mathcal{C}$ events are independent. Since a $\mathcal{B}$ event constrains $s-1$ symbols and a $\mathcal{C}$ event constrains r symbols, $\mathbf{t} \in \mathcal{T}_0$ fully constrains $(r+s-1)(r+1)$ symbols of a one more symbol of which is selectable in three ways, so that

$$\mathbf{Pr}(\mathcal{A}(\mathbf{t})) = 3 \times \left(\frac{1}{4}\right)^{(r+1)s+r^2} \qquad \mathbf{t} \in \mathcal{T}_0 \tag{2}$$

If $\mathbf{t} \in \mathcal{T}_i$, then i of the $\mathcal{C}$ events are dependent and

$$\mathbf{Pr}(\mathcal{A}(\mathbf{t})) \leq 3 \times \left(\frac{1}{4}\right)^{(r+1)s+r^2-ir} \qquad \mathbf{t} \in \mathcal{T}_i \tag{3}$$

We now estimate the size of $\mathcal{T}_i$. For $i > 0$, i of $\mathbf{t}$'s components are restricted to $(r+1)$ specified positions within the $(r+1)s$-neighborhood of other $r+2$ components. Thus

$$\begin{aligned} |\mathcal{T}_i| &\leq |\mathcal{T}| \binom{r+1}{i} \left(\frac{(r+1)(r+2)}{m-(r+1)s}\right)^i \\ &\leq \binom{r+1}{i} m^{r+2} \left(\frac{(r+1)(r+2)}{m-(r+1)s}\right)^i. \end{aligned}$$

So,

$$\begin{aligned} \sum_{i=1}^{r+1} |\mathcal{T}_i| &\leq |\mathcal{T}| \sum_{i=1}^{r+1} \binom{r+1}{i} \left(\frac{(r+1)(r+2)}{m-(r+1)s}\right)^i \\ &\leq |\mathcal{T}|(1+o(1)) \frac{(r+1)(r+2)}{m} \\ &= o(|\mathcal{T}|). \end{aligned} \tag{4}$$

We can now bound the probability of an event $\mathcal{A}(\mathbf{t})$ for $\mathbf{t} \in \mathcal{T}_i$, $i \geq 1$:

$$\begin{aligned} \mathbf{Pr}(\exists \mathbf{t} \notin \mathcal{T}_0 : \mathcal{A}(\mathbf{t})) &\leq \sum_{i=1}^{r+1} \binom{r+1}{i} \left(\frac{(r+1)(r+2)}{m-(r+1)s}\right)^i m^{r+2} 3 \left(\frac{1}{4}\right)^{(r+1)s+r^2-ir} \\ &= 3\frac{m^2}{4^{(\gamma+1)r+s}} \sum_{i=1}^{r+1} \binom{r+1}{i} \left(\frac{(r+1)(r+2)4^r}{m}\right)^i \\ &= o(1). \end{aligned}$$

(This bound makes use of the condition $\beta = o(\log m)$ to get $4^r r^2 \ll m$.)

Let $I(\mathbf{t})$ be a binary variable such that $I(\mathbf{t}) = 1$ if and only if event $\mathcal{A}(\mathbf{t})$ occurs, and let $Z = \sum_{\mathbf{t}\in\mathcal{T}_0} I(\mathbf{t})$. Then

$$\mathbf{Pr}(\exists \mathbf{t} \in \mathcal{T}_0 : \ \mathcal{A}(\mathbf{t})) \ \leq \ \mathbf{E}[Z].$$

Using (1), (2) and (4) we get

$$\begin{aligned}
\mathbf{E}(Z) &= (1+o(1))\binom{m}{r+2}(r+2)! \times 3 \times \left(\frac{1}{4}\right)^{(r+1)s+r^2} \\
&= (1+o(1))\frac{3m^2}{4^s}\left(\frac{m}{4^{s+r}}\right)^r \\
&= (1+o(1))\frac{3m^2}{4^{s+(1+\gamma)r}} \\
&= (1+o(1))3 \times 4^{-(1+\gamma(1+\beta))}.
\end{aligned}$$

Thus, the probability that the algorithm fails to sequence all but the last rs symbols of the sequence is bounded from above by

$$\begin{aligned}
&\mathbf{Pr}(\exists \mathbf{t} \notin \mathcal{T}_0 : \ \mathcal{A}(\mathbf{t})) \ + \ \mathbf{Pr}(\exists \mathbf{t} \in \mathcal{T}_0 : \ \mathcal{A}(\mathbf{t})) \\
&\leq o(1) + (1+o(1))3 \times 4^{-(\beta\gamma+\gamma+1)} \leq 4^{-\gamma(\beta+1)}.
\end{aligned}$$

Finally, if for all $m - 2(r+1)s < t \leq m - (r+1)s$ we do not have the event $\mathcal{B}_0(t) \cap \mathcal{C}_0(t) \cap \mathcal{D}_0(t)$ the last $(r+1)s$ symbols are uniquely determined. But

$$Pr\Big(\bigcup_{t=m-2(r+1)s+1}^{m-(r+1)s} \big(\mathcal{B}_0(t) \cap \mathcal{C}_0(t) \cap \mathcal{D}_0(t)\big) \ \leq \ 3rs4^{-(r+s)} \ = \ o(1).$$

□

At this point one might wish to try to prove that the limiting distribution of Z is Poisson with mean $3 \times 4^{-(\beta\gamma+\gamma+1)}$. We did try to use the Stein-Chen method, see for example [AGG89]. It seems however that Z does not satisfy the requisite conditions and we leave it as an open problem to determine the limiting distribution.

The procedure described and analyzed above, which involves $(r+1)$ fooling probes shifted at regular intervals of s positions, will be briefly referred to as *forward sequencing*. We shall now show that the $GS(s,r)$ spectrum, used in forward sequencing, can also be used for sequencing in reverse.

Let α denote a string over the alphabet $\{X, U\}$. By $\mathrm{FS}_u(\alpha)$ we denote the sequence reconstruction process based on probes of type α, whose confirmatory probes are shifted forward at regular intervals of u positions. By $\mathrm{RS}_u(\alpha)$ we denote the analogous notion for sequencing in reverse. Two sequencing processes are equivalent ($\equiv$) if their respective events of the type $\mathcal{A}(\mathbf{t})$, defined in the proof of Theorem 3, are characterized by the same parameters and occur with the same probabilities. Starting from the standard pattern $X^s(U^{s-1}X)^r$, we shall establish:

1. $\mathrm{RS}_1(X^s(U^{s-1}X)^r) \equiv \mathrm{FS}_1((XU^{s-1})^r X^s)$.

2. $\mathrm{FS}_1((XU^{s-1})^r X^s) \equiv \mathrm{FS}_{r+1}(X^{r+1}(U^r X)^{s-1})$.

Statement 1 is immediate, since it simply corresponds to exchanging right-to-left shifts with left-to-right shifts. Statement 2 is established as follows. We represent a probing pattern by a $0-1$ polynomial in the indeterminate x, where a term x^j corresponds to an X-symbol in the $(j+1)$-st position (from the left). [Thus, $(XU^{s-1})^r X^s$ corresponds to the polynomial $p(x) = \sum_{j=0}^{r-1} x^{js} + x^{rs} \sum_{i=0}^{s-1} x^i$.] If we now subject the pattern to a "shuffle" rearrangement, denoted σ, of its positions:

$$\begin{aligned} \sigma(i) &= i(r+1) \mod ((r+1)s-1), & 0 \le i \le (r+1)s-2 \\ \sigma((r+1)s-1) &= (r+1)s-1. \end{aligned}$$

we transform $p(x)$ ($\mod x^{(r+1)s-1}$) to

$$\sum_{j=0}^{r-1} (x^{r+1})^{js} + (x^{r+1})^{rs} \sum_{i=0}^{s-2} (x^{r+1})^i = \sum_{j=0}^{r-1} x^j + x^r \sum_{i=0}^{s-2} x^{(r+1)i}.$$

The corresponding probe pattern is $X^{r+1}(U^r X)^{s-1}$, appearing in Statement 2. In addition, a 1-position right-shift of the pattern $(XU^{s-1})^r X^s$ corresponds to an $(r+1)$-position right-shift of the pattern $X^{r+1}(U^r X)^{s-1}$. Since only a rearrangement of positions has been executed, the two processes are equivalent.

We now observe that $X^{r+1}(U^r X)^{s-1}$ is a standard probing pattern used in a forward sequencing process. Thus, Theorem 3 fully applies, with the simple modification of interchanging parameters r and $s-1$, and we conclude:

Theorem 4 *For constant $1 < \gamma = O(\log m)$ and $\beta = o(\log m)$, such that r and s are positive integers, let:*

$$\begin{aligned} s &= 1 + \frac{1}{\gamma} \log_4 m + \beta \\ r &= \log_4 m + 1 + \gamma - s. \end{aligned}$$

The algorithm fails to sequence in reverse *a random string of length m using the $GP(s,r)$ spectrum of the string with probability at most $4^{-\gamma(1+\beta)}$.*

5 Removing the prefix requirements

The sequencing procedure outlined above requires a "seed" of length $s(r+1) = O((\log m)^2)$ symbols to "bootstrap" the process. We offer three solutions, two biochemical and one algorithmic, to remove this requirement. The two biochemical methods are more practical.

If the SBH process is used to sequence one string of length m, the simplest solution is to synthesize a short "primer' (a string of length $O((\log m)^2)$) and attach it to the beginning of the string, thus providing the required prefix of the target string.

In most applications, however, one needs to sequence a string that is substantially longer than can be handled by SBH chips, even using our novel scheme. The standard solution is to fragment the target sequence by means of restriction enzymes to produce a collection of overlapping substrings of sizes that can be handled by the SBH method. Once each of the substrings is sequenced, standard techniques [W95] reconstruct the entire string. Since the substrings overlap, it is not necessary to sequence the beginning and the end of each substring. We still, however, need to provide the algorithm with a seed sequence of length $O((\log m)^2)$ for each substring of length m. This could be achieved by the following three steps: (1) Isolate a short, $O((\log m)^2)$, piece of the target sequence and sequence it using $O(4 \log \log m)$ solid (no gaps) probes (traditional method). (2) Use $GP(s,r)$ probes for the forward sequencing of the portion of the target from the isolated piece to (almost) the end of the sequence. (3) Use the same set of $GP(s,r)$ probes for the reverse sequencing of the portion from the isolated piece to the beginning of the sequence.

Finally, we mention a purely combinatorial/algorithmic approach to remove the prefix requirement. A probe is selected at random from the spectrum and its unspecified positions (corresponding to the "don't care" gaps) are "filled" consistently with the spectrum. This results in a number of strings of length $s(r+1)+s-1 = s(r+2)-1$, a subset of which correspond to actual substrings of the target sequence. Only these legitimate substrings are expected to be extensible by forward sequencing. Reverse sequencing of the terms that have been successfully extended in the forward direction, will complete the process.

Acknowledgement: We thank Iddo Lev for carefully reading a draft of the paper and providing several helpful remarks.

References

[AGG89] R. Arratia, L. Goldstein and L. Gordon, Two moments suffice for Poisson approximation: The Chen-Stein method. *The Annals of Probability* (1989) 17, 9-25.

[A+96] R. Arratia, D. Martin, G. Reinert and M.S. Waterman, Poisson process approximation for sequence repeats, and sequencing by hybridization, *Journal of Computational Biology* (1996) 3, 425-463.

[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.

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