On the Power of Universal Bases in Sequencing by Hybridization

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

Abstract

Sequencing by hybridization is a novel DNA sequencing technique in which an array (SBH chip) of short sequences of nucleotides (*probes*) is brought in contact with a solution of (replicas of) the target DNA sequence. A biochemical method determines the subset of probes that bind to the target sequence (the *spectrum* of the sequence), and a combinatorial method is used to reconstruct the DNA sequence from the spectrum.

Since technology limits the number of probes on the SBH chip, a challenging combinatorial question is the design of a smallest set of probes that can sequence an arbitrary DNA string of a given length. We show in this work that the use of universal bases (bases that bind to any nucleotide [LB94]) can drastically improve the performance of the SBH process. We present a novel probe design with performance that asymptotically approaches the information-theoretical bound up to a constant factor, and, for any number of probes, is significantly better than previously analyzed probe patterns. Furthermore, the sequencing algorithm we use is substantially simpler than the Eulerian path method used in previous work.

1 Introduction

A central application in molecular biology is the *sequencing of DNA*, i.e., the determination of the sequence of nucleotides of a chosen (fragment of a) DNA molecule. Recently a radically new technique has been proposed as an alternative to the traditional sequencing by gel electrophoresis. Such technique, called *Sequencing by Hybridization*, proposed independently by several different research teams (for example, [BS91, L+88, D+89]), is based on the use of a chip, fabricated on a glass substrate with photolithographic techniques analogous to those employed in the production of integrated circuits. Specifically, 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 (a short sequence of nucleotides), or to a specific set of oligonucleotides. Oligonucleotides are biochemically attached to the chip surface. 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 oligonucleotide is complementary, in the Watson-Crick sense, to one of its subsequences. The labeling of the target allows visualization of the chip features containing binding oligonucleotides, thereby yielding a method for automatically probing the target sequence for specific subsequences.

Although probing for specific subsequences is possible with this technology, its general-purpose application is DNA sequencing. In such application, the objective is the faithful reconstruction of the target sequence using the outcome of the probing process, which is the collection of the oligonucleotides appearing, at least once, as subsequences of the target sequence (and referred to as the *spectrum* of the sequence). For a fixed cost, expressed by the area of the probe chip, a major challenge 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, by Bains and Smith [BS91], Lysov *et al.* [L+88], and Drmanac *et al.* [D+89], focused on "classical" probing schemes, i.e., chips accommodating all 4^k k-symbol oligonucleotide strings (k-mers or "solid" probes with no gaps), the symbols being the well-known DNA bases { A,C,G,T } and k being a technology-dependent integer parameters (currently rather small, but expected to moderately grow). To reconstruct the target sequence from probes that are strings of k symbols, original approaches dealt with a

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

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

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 [L+88, D+89, BS91]. 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, where 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 probe. This insight not only drastically simplified the solution of the reconstruction problem, but it provided a characterization of spectrum-consistency, so that a sequence is unambiguously reconstructible if and only if the corresponding graph G' contains a unique Eulerian path. The analysis of the effectiveness of this novel sequencing method has been statistical (see, e.g., the work of Pevzner [P89, P+91, PL94] and the textbook by Waterman [W95]), based on the model that the target sequence is randomly generated by a memoryless source with identical symbol probabilities. In this model, also adopted in this paper, Pevzner *et al.*[P+91] observed that the expected length of unambiguously reconstructible sequences with solid length-k probes is $O(2^k)$ and a tight bound of the same order has been proven in [DFS94, A+96]. These results were confirmed by extensive simulations. Note, however, that an information-theoretic argument yields an upper bound $O(4^k)$.

Probe structures alternative to the classical one described above have also been proposed recently [PL94]. One such structure replaces individual nucleotides with subsets, such as {A,T}, {C,G}, {A,G}, or {C,T}. Another structure introduces a gap of "don't care" bases separating a string of specified nucleotides and a single specified nucleotide, but no in-depth analysis has been reported. Originally, it was proposed to realize "don't care"s by a mixture of probes exhibiting in the chosen position all four standard bases. Recently, a much more interesting alternative has been proposed, which uses truly *universal* bases (such as naturally occurring inosine-style bases or synthetic 5-nitroindole [LB94], that –if used in short runs – stack correctly without binding.

In this paper we show that the use of probes with a well defined periodic pattern of gaps is crucial to the attainment of asymptotically optimal efficiencies (i.e., expected sequence length $\Theta(4^k)$). We present a novel probe design that 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 $\Theta(k^2)$-gram De Bruijn graph. Therefore, essential to the attainment of the information-theoretic upper bound is the implementation of gapped probes, i.e., the safe insertion of "universal" (don't care) bases into the oligonucleotide. The full potential of sequencing by hybridization is predicated on the reliable deployment of universal bases.

The analytical results reported here are asymptotic. To establish the validity of our approach for practical chip sizes, we have run extensive simulations for technologically feasible parameters. The simulation results, fully documented in [HPU98], remarkedly match the analysis, and clearly demonstrate the advantage of our probing scheme for any number of probes, and in particular for today's practical range of SBH chips with thousands to (possibly) a few millions probes.

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). Biochemical labeling permits the identification of the complete set of probes (called the string's *spectrum*) that bind to the target string. In this paper, to fairly compare the relative capabilities of different methods, we assume that hybridization is an error-free process, with no missing probes nor false positives.

A *sequencing algorithm* is an algorithm that, given a set of probes and a sequence 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 $t = \Omega(m)$. $\Box$

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.

Definition 2 *Two sequences are said to agree (in a chosen relative alignment) if their symbols are identical in any position in which they are both specified.*

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 basic scheme

We describe a simple procedure, referred to as the "basic scheme", for sequencing the string a using the spectrum information. To simplify the presentation we assume that we are given the $s(r+1)$-prefix of the target string. (We will see later how to remove this assumption.)

By $b_{(1,...)}$ we denote the *putative* sequence constructed by the sequencing algorithm. The procedure 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, each symbol may have to be confirmed by up to $(r + 1)$ probes in different alignments with the current putative 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 current 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. Problematic is the case $|M_0| > 1$, 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 $b_{(\ell-sr+1,a_\ell-1)}$, and their $(s+1)$-suffix agrees, in the sense of Definition 2 and in appropriate shifts, with the probes in M_0. Let B_1 be the set of symbols appearing in the sr-th position of the probes in M_0. If $B_0 \cap B_1$ is a singleton, then we have a unique extension to the string. Otherwise we continue by constructing the set M_2 of the spectrum probes whose $(s(r - 1) - 1)$-prefix matches $b_{(\ell-s(r-1)+1,\ell-1)}$ and $(2s+1)$-suffix agrees with the probes in M_1. From M_2 we construct the corresponding set B_2 of extensions. Again, if $B_0 \cap B_1 \cap B_2$ is a singleton we are done, else we proceed by considering shorter prefixes of lengths $s(r-2), s(r-3), s(r-4), ..., s$ of the spectrum probes. If $|\cap_{j=1}^{i} B_j| = 1$ for some $i \leq r$, then we have an unambiguous extension. Otherwise, in the basic scheme we halt and report the current sequence (a more thorough and better performing technique will be sketched later in this paper).

The success of the above algorithm stems from the fact that up to r additional probes, appropriately aligned along the current sequence, are used to confirm the non-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 basic scheme

We present in this section an analysis of the performance of the simple 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, in addition to the spectrum, the algorithm is provided with the $s(r+1)$-prefix of the target sequence. We will show in the next section that this assumption can be removed without altering the performance of the sequencing scheme.

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

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

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:

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

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:

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

For $1 \le j \le r$:

$$\begin{array}{rll} a_{(t_j+1,t_j+s)} = a_{(t+js+1,t+(j+1)s)} & & \mathcal{B}_j(\mathbf{t}) \\ a_{t_j+is} = a_{t_{j-1}+(i+1)s} & 2 \le i \le r. & \mathcal{C}_j(\mathbf{t}) \end{array}$$

We focus first on the success of the algorithm in sequencing all but the last rs symbols of the target sequence.

Claim 1 *The algorithm fails to sequence the $m - sr$ 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 $2 \le i \le r$ 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 $\mathbf{t}$, i.e.:

$$|\mathcal{T}| = \binom{m}{r+2}(r+2)!.$$

For a given vector $\mathbf{t} \in \mathcal{T}$, let $C(\mathbf{t})$ denote the set of components of $\mathbf{t}$ that are within a distance $3rs$ from any other component of $\mathbf{t}$ (in the following definition $t \equiv t_{-1}$):

$$C(\mathbf{t}) = \{j : \exists j' < j \text{ with } |t_{j'} - t_j| \le 3rs\}.$$

Let $\mathcal{T}_i$ denote the set of vectors with $|C(\mathbf{t})| = i$, i.e.:

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

Next we bound 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 $r+1$ events are independent. If $\mathbf{t} \in \mathcal{T}_i$, then all of the $\mathcal{B}$ events are still independent, and all but at most i of the $\mathcal{C}$ events are independent (a $\mathcal{B}$ event involves $s+r-1$ symbols ($s+r$ for $\mathcal{B}_0$), a $\mathcal{C}$ event $r-1$). Thus we prove:

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

and

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

If $\mathbf{t} \in \mathcal{T}_i$ then at least i of $\mathbf{t}$'s components are restricted to the $3rs$-neighborhood of other $r+1$ components. Thus

$$|\mathcal{T}_i| \le |\mathcal{T}| \binom{r+1}{i} \left(\frac{3rs(r+1)}{m}\right)^i$$
$$\le \binom{r+1}{i} m^{r+2} \left(\frac{3rs(r+1)}{m}\right)^i. \qquad (3)$$

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

$$\mathbf{Pr}(\exists \mathbf{t} \notin \mathcal{T}_0 : \; \mathcal{A}(\mathbf{t})) \le$$

$$\sum_{i=1}^{r+1} \binom{r+1}{i} (3rs(r+1))^i m^{r+2-i} 3 \left(\frac{1}{4}\right)^{(r+1)s+r^2-i(r-1)}$$

$$= 3\frac{m^2}{4^{(\gamma+1)r+s}} \sum_{i=1}^{r+1} \binom{r+1}{i} \left(\frac{3rs(r+2)4^{r-1}}{m}\right)^i = o(1).$$

(This bound makes use of the condition $\beta = o(\log m)$.)

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

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

Using (1) we get

$$\mathbf{E}(Z) \leq \binom{m}{r+2} (r+2)! \times 3 \times \left(\frac{1}{4}\right)^{(r+1)s+r^2}$$

$$\leq \frac{3m^2}{4^s} \left(\frac{m}{4^{s+r}}\right)^r \leq \frac{3m^2}{4^s 4^{(\gamma+1)r}} \leq 3 \times 4^{-(\beta\gamma+\gamma+1)}$$

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

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

Finally, if for all $m - rs < t < m$ we do not have the event $\mathcal{B}_0(t) \cap \mathcal{C}_0(t) \cap \mathcal{D}_0(t)$ the last rs symbols are uniquely determined,i.e.:

$$Pr\Big(\bigcup_{j=m-rs}^{m} (\mathcal{B}_0(t) \cap \mathcal{C}_0(t) \cap \mathcal{D}_0(t)) \Big) \leq rs 4^{-(r+s)} = o(1).$$

□

Remark The previous theorem outlines a criterion for the selection of the parameters r and s. For given $\log_4 m$ (assumed integer), in order to reduce the cost of the chip we choose a small value of $\gamma > 1$, say, $\gamma = 2$. To reduce the probability of failure we choose as large a value of β as is compatible with its defining constraint ($o(\log m)$), so that $r = \log_4 m/2 + \beta$ and $s = \log_4 m/2 + 3 - \beta$.

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* with shift s. We now observe that the same $GS(s,r)$ spectrum, used in forward sequencing, can also be used for sequencing in reverse. Indeed, reverse sequencing using a standard pattern $X^s(U^{s-1}X)^r$ with shift 1 is trivially equivalent to forward sequencing using the reverse pattern $(XU^{s-1})^r X^s$ with shift 1. The latter can be readily shown to be equivalent to forward sequencing using the standard pattern $X^{r+1}(U^rX)^{s-1}$ with shift $(r+1)$, to which Theorem 3 fully applies, with the simple modification of interchanging parameters r and $s-1$. We conclude:

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

$$s = 1 + \frac{1}{\gamma} \log_4 m$$
$$r = \log_4 m + 1 + \gamma - s.$$

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 $O(\log m)^4)$ solid (no gaps) probes of length $2\log m$ (standard method). (2) Use $GP(s,r)$ probes for the forward sequencing of the portion of the target from the isolated piece to (almost) the sequence end. (3) Use the same set of $GP(s,r)$ probes for the reverse sequencing of the portion from the isolated piece to the sequence beginning.

A third approach to the construction of a "seed" selects a probe π at random from the spectrum. Of course, such a probe is not a string of specified symbols (it has all the gaps corresponding to the "don't care's" of the probing pattern), so that it must be "filled", i.e., all unspecified positions must be filled consistently with the spectrum. This is done using the initial s-symbol solid segment of π as the guide, namely, accepting as a possible candidate any probe whose $(s-1)$-prefix coincides with the homologous suffix of the initial segment of the

seed, and so on, $s-1$ times, until a set $R(\pi)$ of strings of length $s(r+1)+s-1 = s(r+2)-1$ has been obtained. Presumably, especially if m is very large and s is rather small, the size of $R(\pi)$ may be quite large.

Once the set $R(\pi)$ has been obtained, we begin the forward extension process. In the general case when $|R(\pi)| > 1$, each of its members is successively extended one symbol at a time by the process described earlier. In principle, only a small number (possibly, just one) of the members of $R(\pi)$ are actual substrings of the target sequence (are *legitimate*) and all the others are spurious "paths". We have shown that the expected length of spurious paths is very small, so that the extension process will rapidly eliminate them and concentrate on the legitimate members of $R(\pi)$ (not belonging to spurious paths). Again, this approach involves both forward and reverse reconstruction.

6 Further results

In the absence of ambiguous extensions, the basic scheme is perfectly adequate in reconstructing the target sequence. However, we have already observed that an ambiguous extension spawns a spurious path, for which the spectrum is very unlikely to contain confirmatory evidence. This case is addressed by a more advanced algorithm which does not to halt when encountering an ambiguous extension, but rather extends both the (unknown) legitimate path and the spurious path(s), till either all but the legitimate path cannot be extended, or two branching paths have been both extended beyond a threshold length h. Such policy is based on the expectation that a spurious path will rapidly terminate because found to be non-extensible. This policy is obviously expected to process correctly larger target sequences. Indeed, it can be shown that by choosing an appropriate value of h (and tolerating the ensuing computational overhead) the length of the target sequence which can be reliably reconstructed can be made as close to the information-theoretic upper bound (4^{k-1}) as desired.

Finally, we wish to substantiate our earlier assertion that our approach trivializes the Euler path difficulties. In fact, the probability of a recurrent state is negligibly small for the chosen length m of the target sequence. so that the Euler path with very high probability degenerates to a simple path (the states being the $((r+1)s-1)$-grams of the sequence, linked, where appropriate, through the shift-register relation). It can be shown, that for practical values of the parameter k, the expected number of pairs of recurrent states is less than 1.

It is also significant to compare the probabilities that an ambiguous extension is due either to $(r+1)$ fooling probes scattered along the sequence or to a single substring of minimal length that contains them all, since their relative values is the cornerstone of our approach. These two probabilities are, respectively,

$$\binom{m}{r+2}(r+2)!\frac{3}{4}\frac{1}{4^{(k-1)(r+1)}} \quad \text{and} \quad \binom{m}{2}2\frac{3}{4}\frac{1}{4^{(r+1)s-1}}.$$

The first of these expressions has been previously computed (refer to the analysis of set $\mathcal{T}_0$ in the proof of Theorem 3), while the second one is based on the fact that the two configurations coincide in their first $(r+1)s-1$ symbols and differ in their last one. These two probabilities become identical for $r = 0$ (since, in this case, $s = k$), i.e, for ungapped probes. This illustrates in the clearest way the unique role of gaps (universal bases), in achieving the full potential of sequencing by hybridization.

References

[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.

[HPU98] B. Hudson, F.P. Preparata, and E. Upfal, An experimental study of SBH with gapped probes. Technical Report, Dept. of Comp. Sci., Brown University (in preparation), 1999.

[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.

7 Appendix

To experimentally validate the approach, we have recently undertaken a thorough simulation program, currently under way. Our current plan is to assess the cost/effectiveness (in terms of running time vs. length of correctly reconstructed sequence) of several algorithms of increasing complexity. The first coded algorithm is our basic scheme, described in Section 3.

The simulation has been conducted as follows. For a fixed value of k (i.e., for a chip of cost 4^k), we select all possible values of the parameter r, i.e., $r = 0, 1, \ldots, k-2$ (note that the designs $\mathrm{GP}(k, 0)$ and $\mathrm{GP}(1, k-1)$ coincide). For each such selection, increasing values of the length m are adopted. For each value of m a random-number generator is used to generate a sufficiently large sample of target sequences $a_{(1,m)}$. For each such sequence a separate routine produces the spectrum, which then forms the input to the reconstruction algorithm. Once the reconstruction is completed, it is compared with the original sequence and a statistic of failures is compiled.

The results of a sample run are displayed in Figure 1, for $k = 9$ and various values of r. Each plotted point corresponds to a sample of size 250. The leftmost curve corresponds to the classical ungapped probes. Note that for a confidence level 95% the classical approach yields $m \approx 100$, whereas the best result of our method (for $r = 5$) is $m =\approx 8800$.

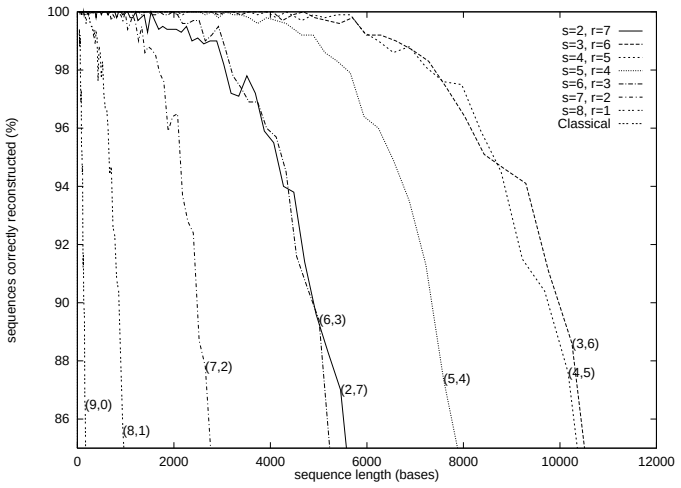

Figure 1: Frequency of successful reconstruction as a function of sequence length for artificially generated random data, $k = 9$ and all possible choices of (s, r).