Lecture 3: April 11, 2002

*Lecturer: Ron Shamir* *Scribe: Tamir Tuller and Kobi Lindzen*

# 3.1 Gapped Chips for SBH

## 3.1.1 Motivation

As we saw in Pevzner's algorithm for SBH reconstruction [10], if we use a classical SBH chip with capacity of $t$ probes, the length of random sequence (four independent letters, with probability of $\frac{1}{4}$ each) that we are able to reconstruct unambiguously is $O(\sqrt{t})$, which is quiet inefficient. In this lecture we shall study more efficient techniques for SBH.

## 3.1.2 An Upper Bound

In their work Perparate, Frieze and Upfal [1] analyzed (base on information-theoretic argument) the maximal length that can be reconstructed with a general SBH chip, that is, any variant of SBH chip. The error probability is denoted by $p$ and the *Spectrum* being refereed in this lecture is a binary spectrum.

**Theorem 3.1 ([1])** *For any fixed probability $p > 0$, the length of a random string that can be unambiguously reconstructed with probability $p$ by a chip with $t$ probes is $O(t)$.*

**Proof:** For each probe on the chip we have the information whether it appeared in the target (at least once) or not. Hence, there are $2^t$ possible spectrum's vectors and $4^m$ sequences of length $m$ (see Figure 3.1).

In order to have an unambiguous reconstruction, each spectrum's vector can define no more than one possible sequence. Thus,

$$4^m \leq 2^t \Rightarrow m \leq \frac{1}{2}t \Rightarrow m = O(t) \tag{3.1}$$

Since there are vectors that do not match any $m$-long sequence we get $m < \frac{1}{2}t$. In order to reconstruct unambiguously a fraction $p$ of the $4^m$ possible sequences we must have

$$p4^m \leq 2^t \Rightarrow m = O(t) \tag{3.2}$$

■ Pevzner's algorithm uses all $t = 4^k$ probes of length $k$. The expected length of sequences

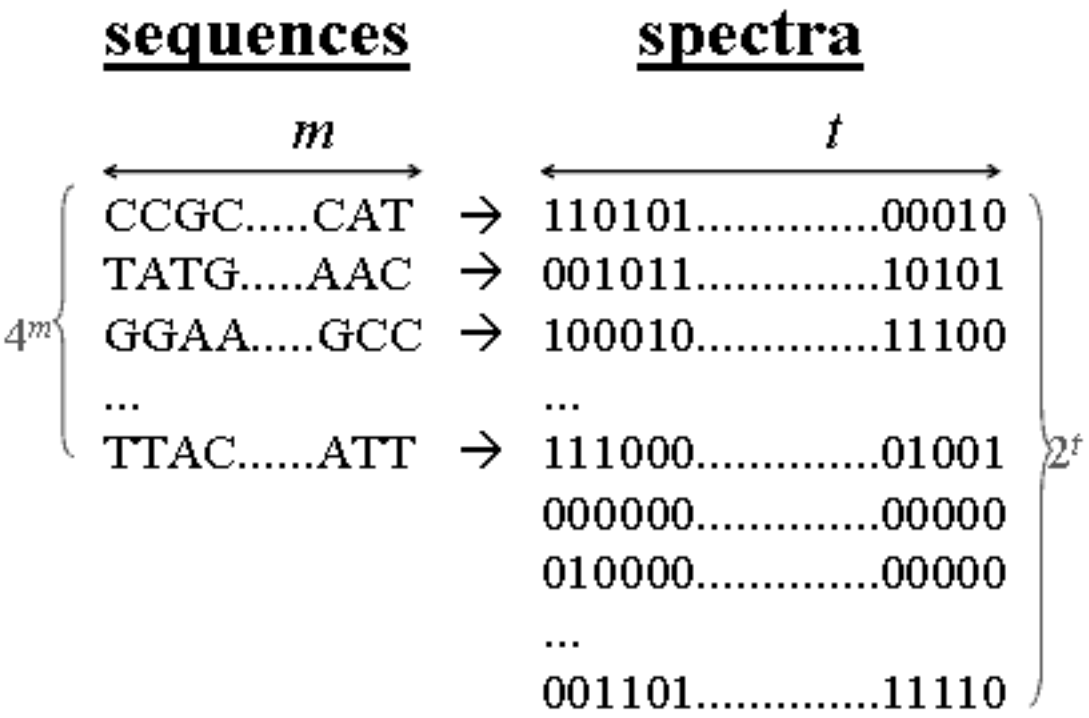


Figure 3.1: Number of possible sequences Versus the number of possible vectors in the spectrum.


Figure 3.2: An example of matching to universal probe, where $s = 3, r = 3$.

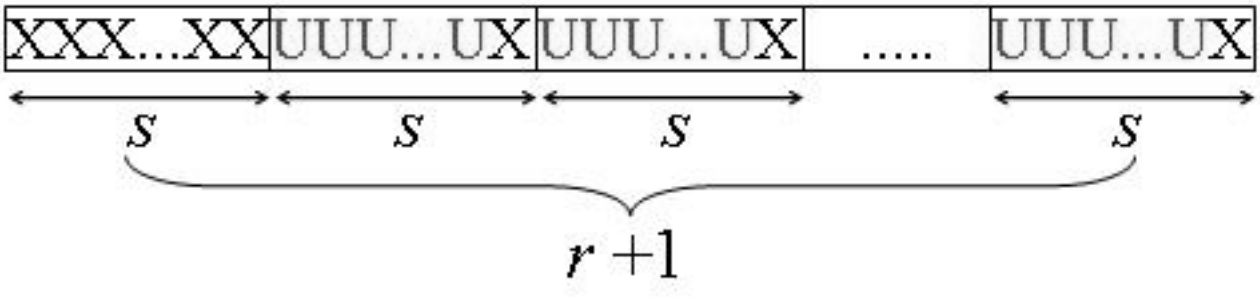


Figure 3.3: The scheme of a probe in PFU's algorithm.

it can unambiguously reconstruct is only

$$m = O(2^k) = O(\sqrt{t}) = o(t) \tag{3.3}$$

We shall now describe an algorithm by Preparate et al. (PFU) [1] which achieves $m = \Theta(t)$.

### 3.1.3 The PFU Algorithm

The key idea is to use *"universal"* (don't care) bases within the probes, that is, bases that bind to any nucleotide (see Figure 3.2). Our probes will include specified bases and unspecified (=universal) bases. For example the probe AC*T*G will be able to hybridize with TGACAA, TGACCA ,TGACGA, TGACTA, TGCCAA, etc. Preparate et al. proposed the following design for the probes (see Figure 3.3). Each probe is of the form $X^s(U^{s-1}X)^r$ where $X \in \{A, C, G, T\}$ is a specified base and U denotes a universal base. For a given $r$ and $s$, the chip contains all probes of this form. According to the scheme we need to determine $r + s$ specified positions (the X's in Figure 3.3) $\Rightarrow$ The total number of probes is $t = 4^{r+s}$.

### 3.1.4 The Reconstruction Algorithm

In order to reconstruct the sequence, based on the universal probes, Preparta et al. gave the following algorithm:

We assume that we are given the $s(r+1)$-prefix of the target string (we will show later how to handle this assumption). Let $Z$ be the putative sequence reconstructed so far (with length greater than $(s(r+1)-1)$). Find set $M_0$ of all the probes in the spectrum such that, the $(s(r+1)-1)$-prefix of each of the probes matches the $(s(r+1)-1)$-suffix of $Z$. Let $B_0$ be set of possible extensions (the base located in position $s(r+1)$ of probes $\in M_0$) - the next symbol in $Z$. In case where the set $B_0$ is empty, no extension exists and the algorithm terminate with failure. In case where the set $|B_0|$ contains only one element, a single extension was found and the corresponding symbol is appended to $Z$. Otherwise, ($|B_0| > 1$), find set $M_1$ of all the probes such that their $(sr-1)$-prefix matches $(sr-1)$-suffix of $Z$ and their $(s+1)$-suffix "agrees" with any probe $\in M_0$. Let $B_1$ be set of possible extensions (the bases in position $sr$ of probes $\in M_1$) $\bigcap B_0$. Once again: In case where the set $|B_1|$ is empty declare failure. If $|B_1| = 1$-extend $Z$ and continue, otherwise ($|B_1| > 1$) continue with $M_2$ according to the general step. Figure 3.4 demonstrates this process.

In each phase we have $s-1$ new specified bases that have to agree with $Z$.

### The Algorithm

Denote by $L$ the sequence length. A formal description of the PFU algorithm follows.

$M_0, M_1, \ldots, M_r \leftarrow \emptyset$.
$B_0, B_1, \ldots, B_r \leftarrow \emptyset$.
`For` $n = s(r+1)+1$ `to` $L - s(r+1)$ `do`

1. $j \leftarrow 0$.
2. $M_0 \leftarrow \forall$ `probes` $P_x \in Spectrum$ `where` $(s(r+1)-1)$`-prefix of` $P_x$`=` $(s(r+1)-1)$`-suffix of` $Z$.
3. $B_0 \leftarrow P_{s(r+1)} \forall P \in M_0$.
4. `If` $|B_0| = 0$
   `Declare "Failure".`
   `End.`
5. `If` $|B_0| = 1$
   $ch \leftarrow B_{0_1}$
   $Z \leftarrow Z + ch$.
   `Next` $n$.
6. `If` $|B_0| > 1$
   `selectCandidate(`$++j$`).`

`Next` $n$

`selectCandidate`$(j)$

1. `If` $j > r$
   `Declare "Failure".`
   `End.`
2. $M_j \leftarrow \forall$ `probes` $P_x \in Spectrum$ `where` $(s(r+1-j)-1)$`-prefix of` $P_x$`=` $(s(r+1-j)-1)$`-suffix of` $Z$ `AND` $\exists$`probe` $Q_y \in M_{j-1}$ `which "agrees" with the` $(sj+1)$`-suffix of` $P_x$.
3. $B_j \leftarrow P_{s(r+1-j)} \ \forall P \in M_j \bigcap B_{j-1}$.
4. `If` $|B_j| = 0$
   `Declare "Failure".`
   `End.`
5. `If` $|B_j| = 1$
   $ch \leftarrow B_{j_1}$.
   $Z \leftarrow Z + ch$.
   `Next n.`
6. `If` $|B_0| > 1$
   $j \leftarrow j + 1$.
   `selectCandidate(`$j$`).`

### 3.1.5 Advantages of Gapped Probes

After we had reconstructed $(n-1)$-prefix of $Z$, we have two options for the $Z_n$ symbol. The case where we have only one candidate is trivial, but when prefix of several probes matches the suffix of $Z$, the question is how to use *other* probes to disambiguate. For every position in the sequence we have several probes ($k$ - in the classical SBH, $r$ - in this case) that can confirm the original symbol. In classical SBH we would like to use the shifted probes to confirm the correct base, but due to the fact that the information within the shifted probes is dependent we cannot use them, as exemplified in Figure 3.5. Using gapped probes, the probability of ambiguous extension in each aligned probe (with respect to a randomly generated sequence), is almost independent of the other probes (see Figure 3.6). This is the feature that gives this method its "power".

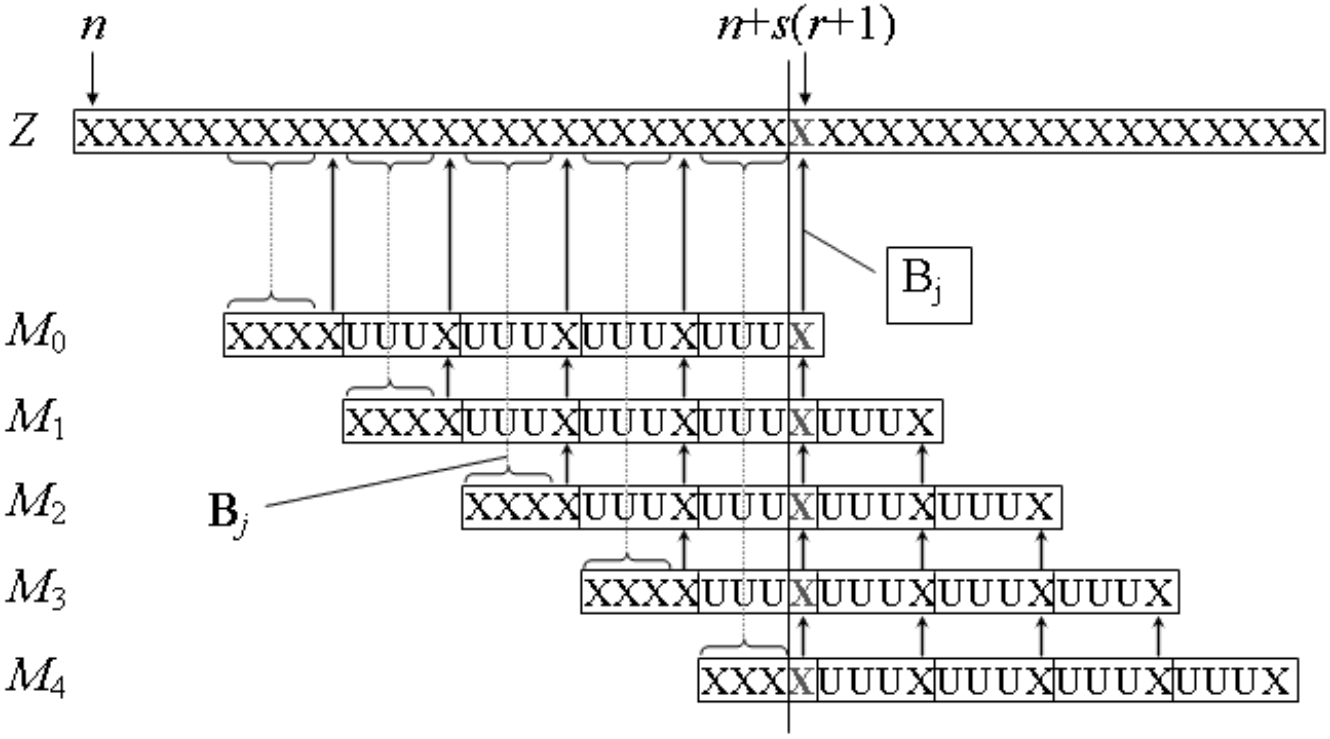


Figure 3.4: An example of the algorithm where $s = 4, r = 4$.

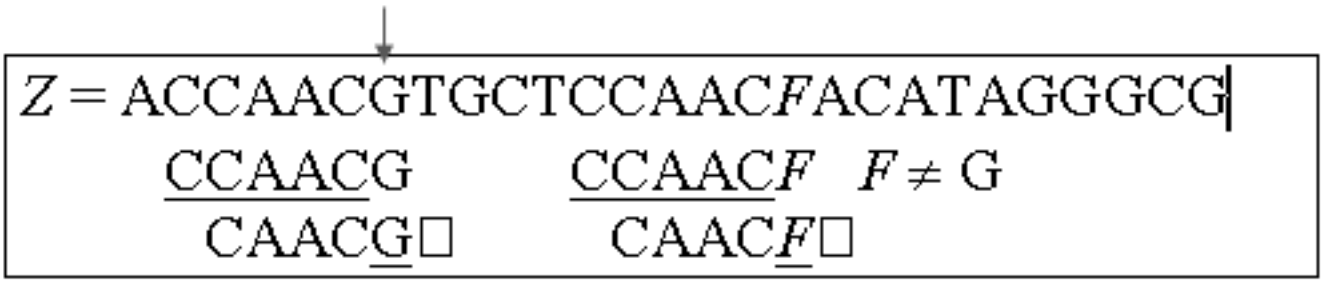


Figure 3.5: An example of the classical probe where $n = 7$. The confirmation does not work because CCAAC**F**_ implies CAAC**F**_ in the spectrum.

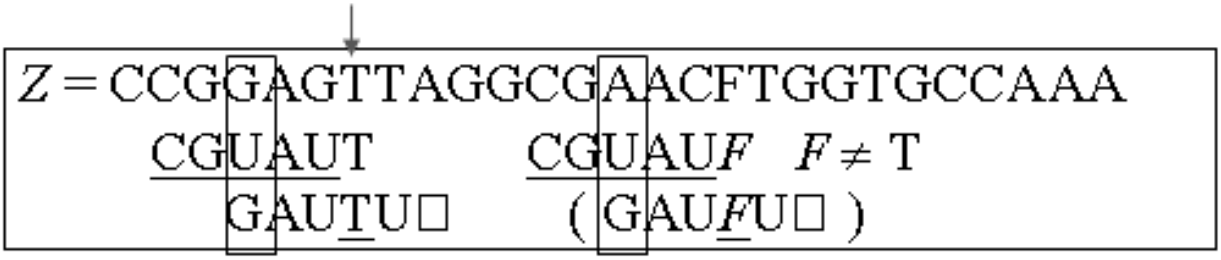


Figure 3.6: An example of an universal probe where $n = 7$. CGUAU**F** does not imply GAU**F**U_. The spectrum contains GAU**T**U_ and not GAU**F**U_, hence, $Z_7 = T$.

### 3.1.6 Removing the Prefix/Postfix Requirements

The PFU algorithm requires the $(s(r+1)-1)$-prefix of $Z$ as an input. We have two ways to resolve this requirement:

1. Biological solution - Attach a synthesized known string of length $s(r+1)$ as a prefix/postfix to the target DNA (this can be achieved as part of the cloning process of the target DNA).

2. Reveal the needed prefix/postfix by standard sequencing machine or classical SBH chip (for short sequences the results are reliable).

### 3.1.7 Universal Chips in Practice

There are several works that claim for molecules that can serve as universal bases (for example [8]), but this issue is still vague. In any case, no universal chip has been realized yet.

### 3.1.8 The Performance of the PFU Algorithm

In order to analyze the PFU algorithm $r$ and $s$ are set as follows: Two constants $\alpha > 1$, $\alpha = O(\log m)$ and $\beta = o(\log m)$ are chosen such that $r > 0$ and $s > 1$ are integers:

1. $r = \frac{1}{\alpha}\log_4 m + \beta (= O(\log m))$.

2. $s = \log_4 m + 1 + \alpha - r (= O(\log m))$.

   As a result of those values we get that the length of a probe is $s(r+1) = O(\log^2 m)$.

Denote by F the event that the PFU Algorithm fails to reconstruct a random target string of length $m$. Perparate et al. prove in their paper [1] the following theorem (we omit the proof here):

**Theorem 3.2** *([1]) If $r, s$ are chosen as above, $Pr(F) \leq 4^{-\alpha(1+\beta)}$.*

As a result of those values, we conclude the following theorem:

**Theorem 3.3** *The length of a random string that the PFU algorithm (using chip with capacity $t$) succeeds to reconstruct with high constant probability is $\Theta(t)$.*

**Proof:** As we saw, the number of probes on the chip is $4^{r+s}$.

$$t = 4^{r+s} = 4^{\log_4 m+1+\alpha} = m4^{1+\alpha} \Rightarrow \tag{3.4}$$

$$m = 4^{-(1+\alpha)}t \Rightarrow \tag{3.5}$$

$$m = \Theta(t) \tag{3.6}$$

■

### 3.1.9 Results

Figure 3.7 demonstrate the universal chip performance for $r + s = 8$. The capacity of this chip ($t$) is $4^{r+s} = 65,536$ probes (equivalent to a standard chip where $k = 8$). The $s(r + 1)$ column shows the probes' length, $Pr(F)$ shows the requested failure rate of the algorithm while $m$ the sequence's length is being determined according to this value. The $\frac{m}{t}$ column presents the chip's performance. As Preparata et al. proved, its upper bound is 50%. The best design we can achieve for failure rate of 1% is with values $r = 6, s = 2$, where the scheme of the probes will be XXUXUXUXUXUXUX. This allows reconstructing up to 2,739 bases. Obviously the results are better than classical SBH.

### 3.1.10 Summary

In their article Preparate et al. [1] proved:

1. Theoretical upper bound on length of random targets that any chip can unambiguously reconstruct.
2. First time that a periodic gapped probe design was analyzed and a reconstruction algorithm was given to it.
3. Provable performance reaches upper bound (up to a constant) - asymptotically optimal design.

| $r$ | $s$ | $s(r+1)$ | $\mathbf{Pr(F)} = 4^{-\alpha(1+\beta)}$ | $m$ | $m/t = 4^{-(\alpha+1)}$ | $\alpha$ | $\beta$ |
|---|---|---|---|---|---|---|---|
| 3 | 5 | 20 | 0.01% | 373 | 0.57% | 2.73 | 1.43 |
| | | | 0.10% | 591 | 0.90% | 2.40 | 1.08 |
| | | | 1.00% | 937 | 1.43% | 2.06 | 0.61 |
| 4 | 4 | 20 | 0.01% | 700 | 1.07% | 2.27 | 1.92 |
| | | | 0.10% | 1028 | 1.57% | 2.00 | 1.50 |
| | | | 1.00% | 1509 | 2.30% | 1.72 | 0.93 |
| 5 | 3 | 18 | 0.01% | 1099 | 1.68% | 1.95 | 2.41 |
| | | | 0.10% | 1527 | 2.33% | 1.71 | 1.91 |
| | | | 1.00% | 2122 | 3.24% | 1.47 | 1.25 |
| 6 | 2 | 14 | 0.01% | 1540 | 2.35% | 1.71 | 2.90 |
| | | | 0.10% | 2054 | 3.13% | 1.50 | 2.33 |
| | | | 1.00% | 2739 | 4.18% | 1.29 | 1.57 |

Figure 3.7: The chip performance for all the possible values for $r + s = 8$. The number of probes is $t = 4^8 = 65,536$.

| $r$ | $s$ | $s(r+1)$ | $\Pr(F) = 4^{-\alpha(1+\beta)}$ | $m$ | $m/t = 4^{-(\alpha+1)}$ | $\alpha$ | $\beta$ |
|---|---|---|---|---|---|---|---|
| 3 | 6 | 24 | 0.01% | 1130 | 0.43% | 2.93 | 1.27 |
| | | | 0.10% | 1791 | 0.68% | 2.60 | 0.92 |
| | | | 1.00% | 2839 | 1.08% | 2.26 | 0.47 |
| 4 | 5 | 25 | 0.01% | 2224 | 0.85% | 2.44 | 1.72 |
| | | | 0.10% | 3264 | 1.25% | 2.16 | 1.30 |
| | | | 1.00% | 4791 | 1.83% | 1.89 | 0.76 |
| 5 | 4 | 24 | 0.01% | 3606 | 1.38% | 2.09 | 2.18 |
| | | | 0.10% | 5010 | 1.91% | 1.85 | 1.69 |
| | | | 1.00% | 6961 | 2.66% | 1.62 | 1.05 |
| 6 | 3 | 21 | 0.01% | 5181 | 1.98% | 1.83 | 2.63 |
| | | | 0.10% | 6909 | 2.64% | 1.62 | 2.07 |
| | | | 1.00% | 9213 | 3.51% | 1.42 | 1.35 |
| 7 | 2 | 16 | 0.01% | 6869 | 2.62% | 1.63 | 3.08 |
| | | | 0.10% | 8871 | 3.38% | 1.44 | 2.45 |
| | | | 1.00% | 11458 | 4.37% | 1.26 | 1.64 |

Figure 3.8: The chip performance for all the possible values for $r + s = 9$. The number of probes is $t = 4^9 = 262,144$.

4. Outperforms previous methods also in practice - with real DNA and for practical values of $s$ and $r$.

5. A year later Preparata and Upfal showed a theoretic design that asymptotically reaches the upper bound (without the constant) [11].

There are two main problems with this work:

1. Handling errors - need to analyze the results when the data on the chip is noisy. A paper regarding this issue was published by Doi and Imai [3].

2. The realization of universal probes.

## 3.2 Handling Long Targets and Error in sequencing by Hybridization

This section summarizes the work of Halperin et al. [5], which includes:

1. A polynomial algorithm that handles errors in classical SBH and reconstructs the correct sequence with probability of $1 - \epsilon$.

2. A novel design of universal probes which gives improved results in the aspect of noise.

3. A polynomial algorithm for an error-prone SBH, using such arrays, with success probability of $1 - \epsilon$ and within $\log^2 m$ factor of the upper bound ($m$ denotes the sequence length).

### 3.2.1 Classical SBH with Noise

#### A Trivial Algorithm

Extend by one base each time according to *one* of the positive probes that were not used till now and agrees with the known prefix. For example, in Figure 3.9 the only way to extend CG is by A, supported by the probe CGA.

#### Problems in Practice

1. False negative error, implying that the "extending" probe might be "absent".

2. False positive errors, that might create ambiguity.

3. Fooling probes - probes that match other locations in the sequence and imply wrong extensions in the current position.

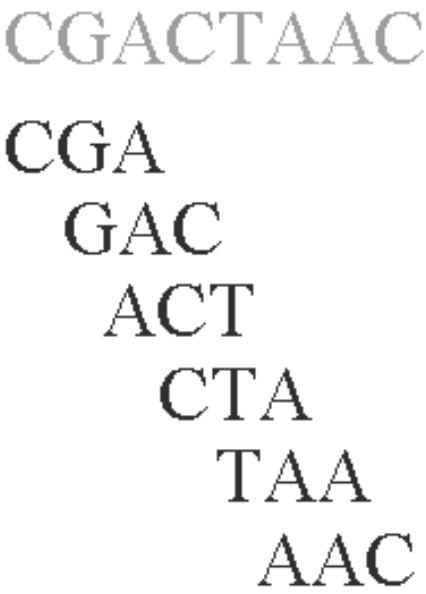


Figure 3.9: CGACTAAC - the unknown sequence. The prefix CG is known as well as the probes ($k = 3$).

Figure 3.10 exemplified the errors types. In order to handle those problems the idea is to look ahead in the sequence and to check *all other* probes which can confirm or deny our candidate. We will check *all* the $k$-length possible extensions and count how many probes support each extension. In the end we will choose the "winning" extension and extend our sequence with the first symbol of that sequence.

**Definition** A *bad path* is a possible extension whose first base is false.

As we can see in Figure 3.10 we have four supporters to the correct path - CGACGTTAC and three supporters to the bad path - CGACAGCTG (two fooling probes and one false positive). In spite the fact that we do have errors we shall prove that the probability that a bad path wins is small. The intuition for that is that "voting process" is less sensitive to errors.

### Lookahead Algorithm

Given a prefix $s_1, \ldots, s_{i-1}$:

1. Enumerate all $4^k$ potential extensions $a = a_1, \ldots, a_k$.
2. Pick an extension $a'$ such that the number of supporting probes $s_1, \ldots, s_{i-1}, a'_1, \ldots, a'_k$ is maximal.

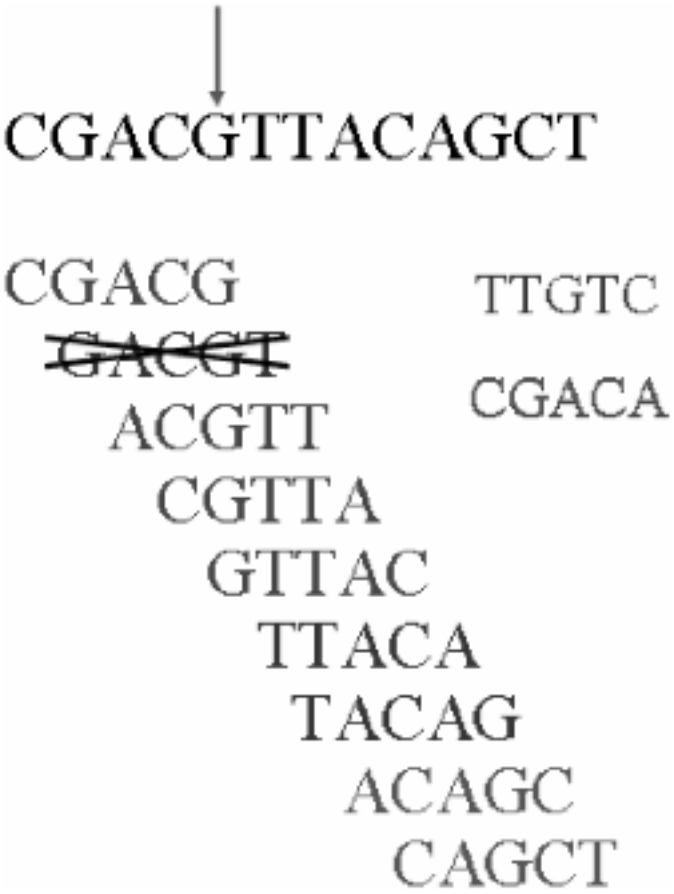


Figure 3.10: CGAC - known prefix. CACGT - false positive. TTGTC,CGACA - false negatives. ACAGC,CAGCT - fooling probes.

3. Set $s_i = a'_1$.

For each reconstructed position ($m$) we need to enumerate $4^k$ possible extensions and check them with a $k$-size window around the current position. Hence, the running time is $O(m4^k k)$.

### The Main Theorem

For analyzing the algorithm we make the following assumptions:

- False negative rate is $q$.
- False positive rate is $p = O(2^{-k})$. In a classical $k$-chip we have $4^k$ points and we showed already that we can unambiguously reconstruct sequences of length $O(2^k)$. Such $p$ implies $p4^k = O(2^k)$ false positive probes (the same order as our reconstructed sequence). If we allow $p > 2^{-k}$ most of the signal we get will be false positive, thus the requirement of such $p$ is reasonable.
- The length of the sequence is $m = O(2^{(1-3q)k})$.
- The probability of errors in each probe are equal and independent.

**Theorem 3.4** *The lookahead algorithm fails to reconstruct the sequence with probability bounded by an arbitrary small constant* $(1 - \epsilon)$.

### Proof sketch

- We expect to see in the spectrum $k$ supporters to the correct symbol. Due to false negative error each one of them has probability of $1-q$ to be seen. Hence, the number of supporting probes for the correct path $X \sim B(k, 1-q)$.
- For a bad path with $j$ fooling probes, the number of supporting probes is $Y = Y1+Y2$, where:

  The number of fooling probes is $Y1 \sim B(j, 1-q)$ - Their probability to appear is also $1-q$ (since they are also "exposed" to the false negative error). The number of false positives is $Y2 \sim B(k-j, p)$.
- The algorithm fails if at some point $X < Y$.
- We can sum over all possible values for $j$

$$\sum_{j=0}^{k} P[X < Y | jfooling] \cdot P[jfooling].$$

- Eventually we get two cases for values of $j$:

  Large $j$: many fooling probes in the bad path - small probability, since it requires resemblance to the current location substring.

  Small $j$: means that $X$ is even smaller and the reason for this is many false negatives for the correct path probes. This event also has a low probability since we have a geometric distribution for it.

### 3.2.2 Gapped Probes Design

Following PFU's algorithm, Halperin et al. suggest to use a universal bases within the probes, but while Preparta et al. suggested a periodic chip this work presents a random design, means, the location of the *don't care* on the probe will be set randomly. Each such setting will be a probe family (the common feature of the family member will be their unspecified bases location), all the families will have the same number of specified (or un specified) bases (see Figure 3.11). The probes are designed with the following features:

- The length of each probe is $ck + 1$ (where $c > 1$).
- Each probe has $k + 1$ specified bases, the rest are universal.
- The last symbol is always a specified base.
- There are $\beta k$ ($\beta > 1$) families of probes.
- For each family: pick randomly set of $k$ positions of specified bases, and implement all $4^{k+1}$ possibilities (the last position is also specified).
- The total number of probes (capacity of the chip) is $t = \beta k 4^{k+1}$.

### 3.2.3 The Reconstruction Algorithm

Again we will count how many probes 'vote' for a possible extension:

- For each specified base $x$, count how many probes support $s_1, \ldots, s_{ck}, x$.
- Set $s_{ck+1}$ to be the base with maximal support.

#### Complexity

For each position ($m$) and for each family ($O(k)$), we need to check for positive signals only for four members (which equal in the $ck$-prefix and differ in their last position) $\Rightarrow O(km)$. Each family chooses different specific locations and this adds disqualification elements. Due to the random selection, the families are independent and that enabled the authors to analyze the algorithm's performance.

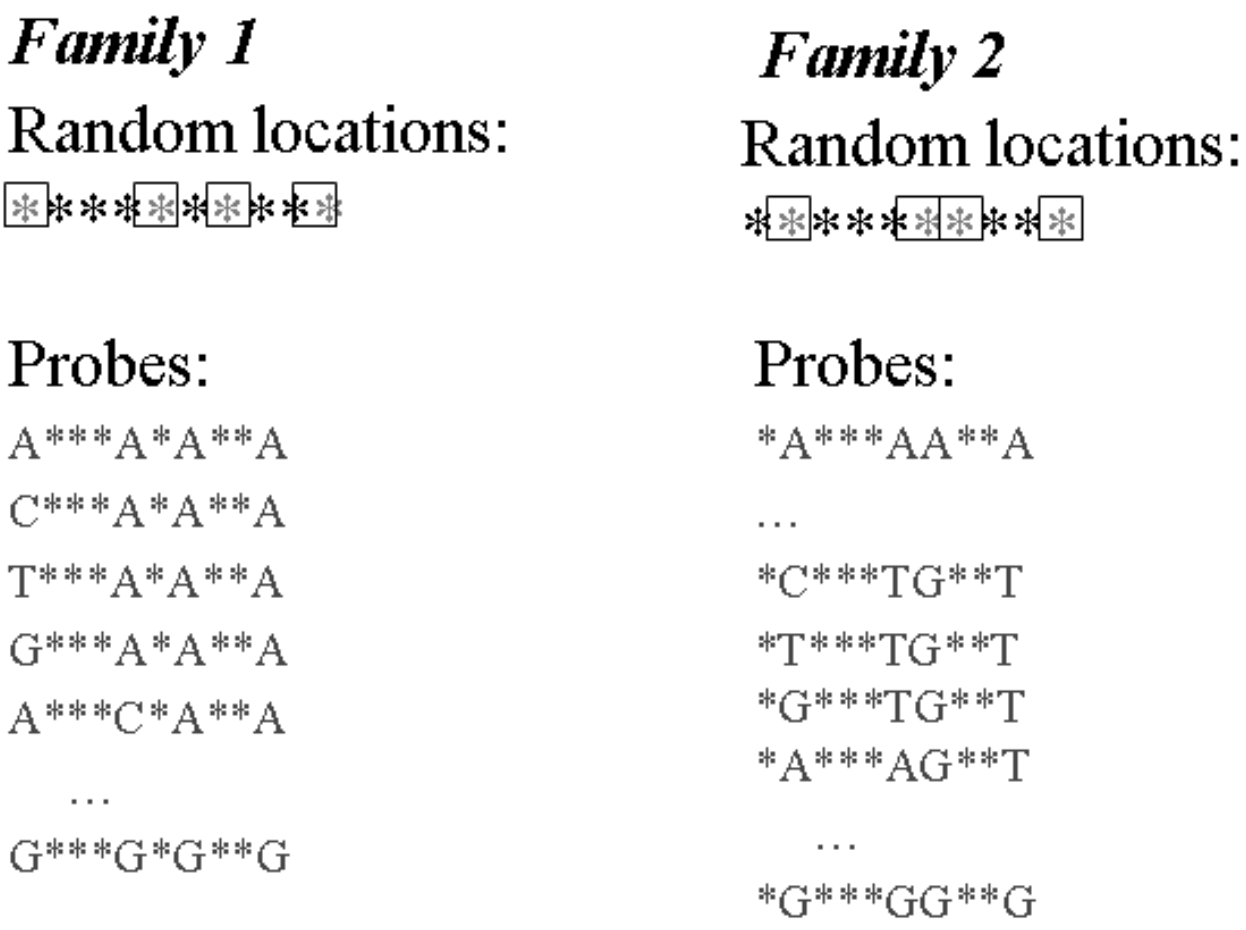


Figure 3.11: Two random families. Each family has $4^4 = 256$ members.

**Gapped SBH - Error-Free Case**

The authors prove in their work [5], the following theorem:

**Theorem 3.5 ([5])** *If there are no hybridization errors, the sequence length is* $m = O(4^k/k)$ *and the number of probes is* $n = \Theta(k4^k)$*, then the algorithm fails with probability bounded by an arbitrary small constant.*

Since the number of probes is $O(k4^k)$, the length of the reconstructed sequence is optimal up to $\log^2$ factor (which is worse than PFU's algorithm).

**Gapped SBH - Error-Prone Case**

The main result of the authors is:

**Theorem 3.6 ([5])** *In the error-prone case, where the length of the sequence is* $m < (\frac{(1-q)4^k}{5k})$*, the number of probes is* $n > \frac{5k4^k}{1-q}$ *and the false positive rate is* $p < \frac{(1-q)}{430}$*, then the algorithm fails with probability bounded by an arbitrary small constant.*

Here we also have $\log^2$ factor distance from the theoretical upper bound with no noise (which holds also in the case of noise).

### 3.2.4 Results

The authors compared their algorithm results on *real DNA* with the results of heuristic implementation of the PFU algorithm with noise by Doi and Imai [3]. Figure 3.12 demonstrates the success rate in reconstructing the sequence of the algorithms as function of the sequence length. A sequence with 1500 bases can be reconstructed perfectly with the algorithm of Halperin et al., while with Preparata et al. algorithm, it can be reconstructed only with 80% success rate.

In Figure 3.13 the chip capacity is $4^{10}$ probes and it presents the possible length of the reconstructed sequence as a function of the error rate when we demand 90% reconstruction success rate. In both figures $p = q$.

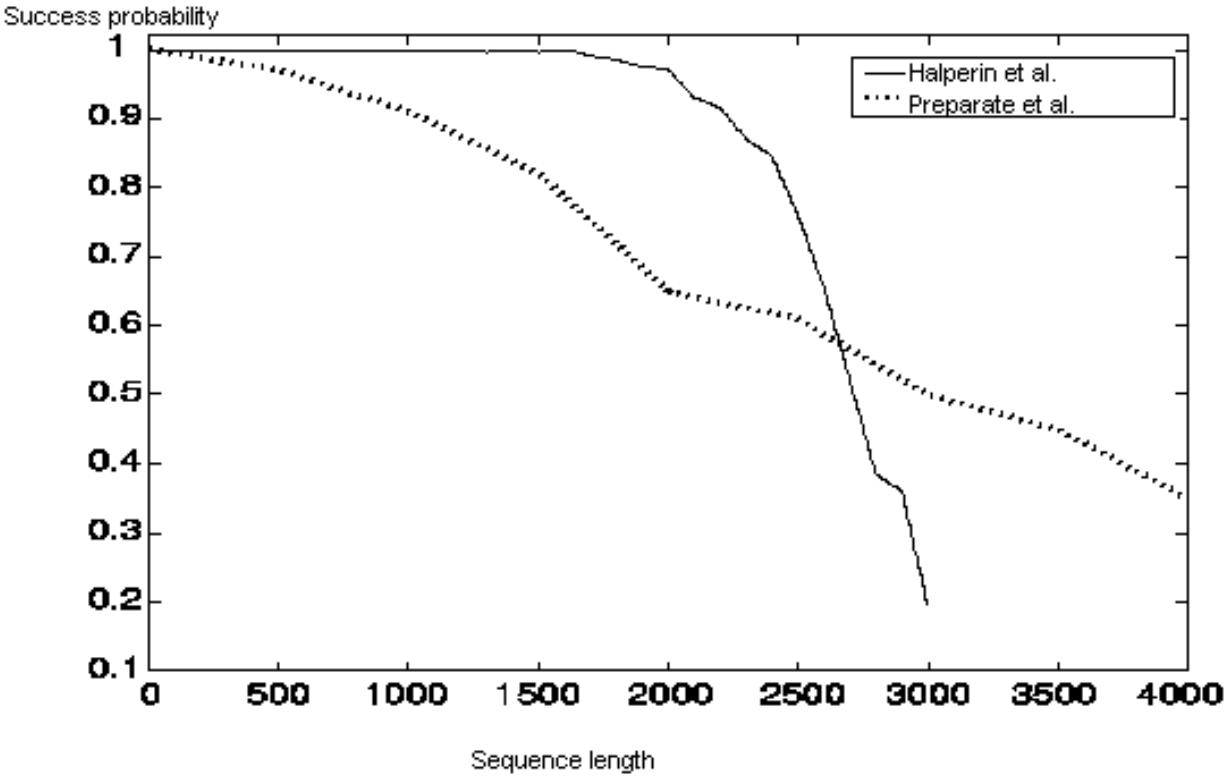


Figure 3.12: Source: [5]. Comparing the gapped design. Chip capacity = $4^8$ probes. In Halperin et al. $p = q = 0.005$, in Preparata et al. $p = q = 0.001$.

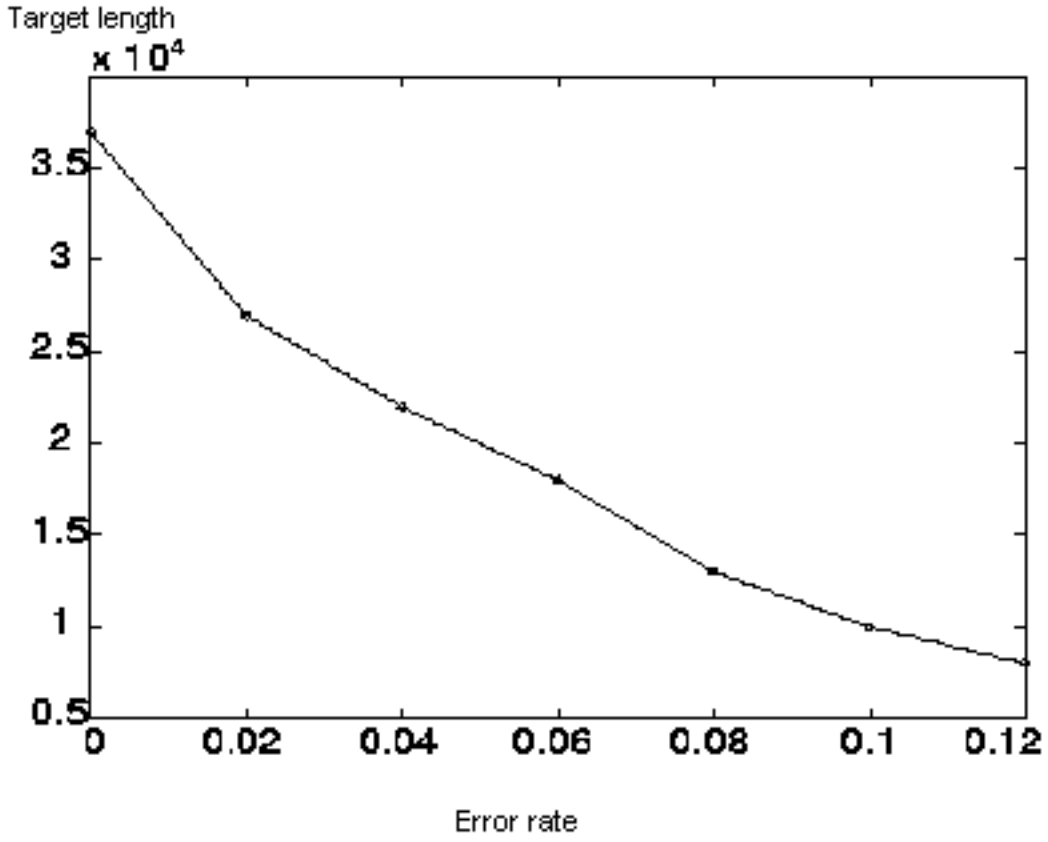


Figure 3.13: Source: [5]. Effect of both False positives and negatives. Chip capacity = $4^{10}$ probes. Assume $p = q$. The results are for 90% reconstruction success.

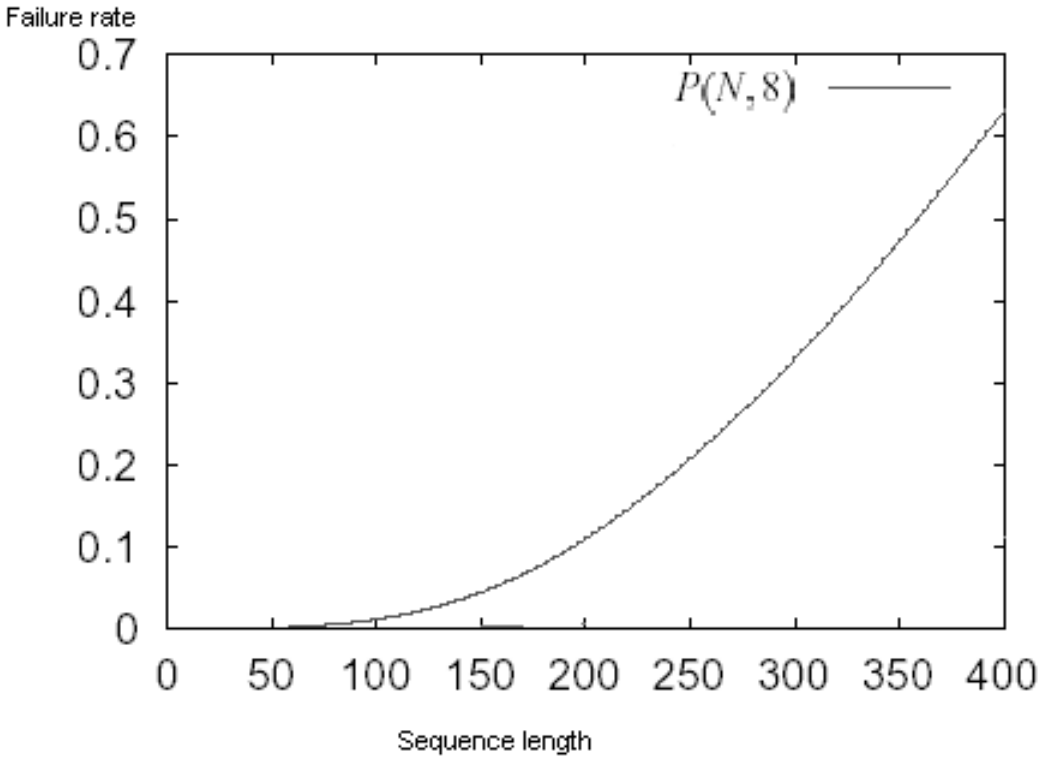


Figure 3.14: The ambiguity probability in Pevzer's algorithm as a function of the reconstructed sequence length.

## 3.3 Large Scale Sequencing By Hybridization

This section presets a work by Tsur et al. [16] that analyzes the performances of shotgun SBH, in the presence of errors.

### 3.3.1 Shotgun *SBH*

As we have seen in the previous sections when using SBH, only relatively short sequences can be reconstruct unambiguously (see Figure 3.14). Dramanac et al. [13] suggested sequencing long DNA target by obtaining the spectra of many short overlapping fragments of the target, inferring their relative positions along the target and computing spectra of sub-fragments that are short enough to be uniquely recoverable.

Shotgun *SBH* consists of the following steps [13]:

1. Fragment the target $S$ into overlapping *clones* and obtain the spectrum of each clone (without trying to reconstruct the clone).

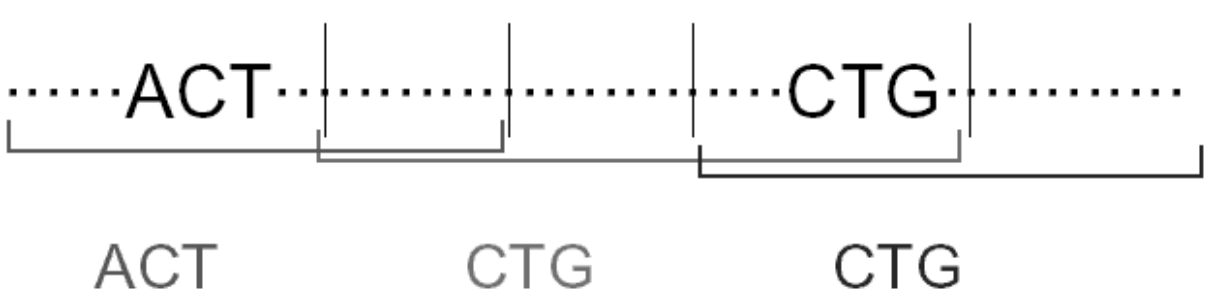


Figure 3.15: Once we know the clones' map, we observe that the $k$-mer CTG is common to the middle and the right-most clones. As a result of this we know that the CTG $k$-mer belong to the fourth IF.

2. Find the correct clone map. See, e.g., [9] for possible mapping.

3. The clones' endpoints form a partition of the sequence $S$ into subsequences called *information fragments (IF)*. For each *IF*, find its spectrum according to the overlaps $k$-mers in each consecutive clone (see Figure 3.15).

4. Reconstruct the sequence of each *IF*.

5. Combine the sequences of each *IF*.

Shotgun *SBH* was shown to work well in the absence of errors [4]. In the following section a rigorous analysis which considers the impact of errors will be described.

### 3.3.2 Hybridization Errors

Since the hybridization experiments are error-prone we are expected for false negative errors, that is, $k$-mers that appear in the clone but does not appear in its measured spectrum (see Figure 3.16). Those errors will affect the output of step 3 in the previous section.

#### Assumptions

The analysis makes the following assumptions:

- Clones positions are known .
- The *IF*s have equal size $d$.
- Each $k$-mer of the target appears in at least one clone spectrum.

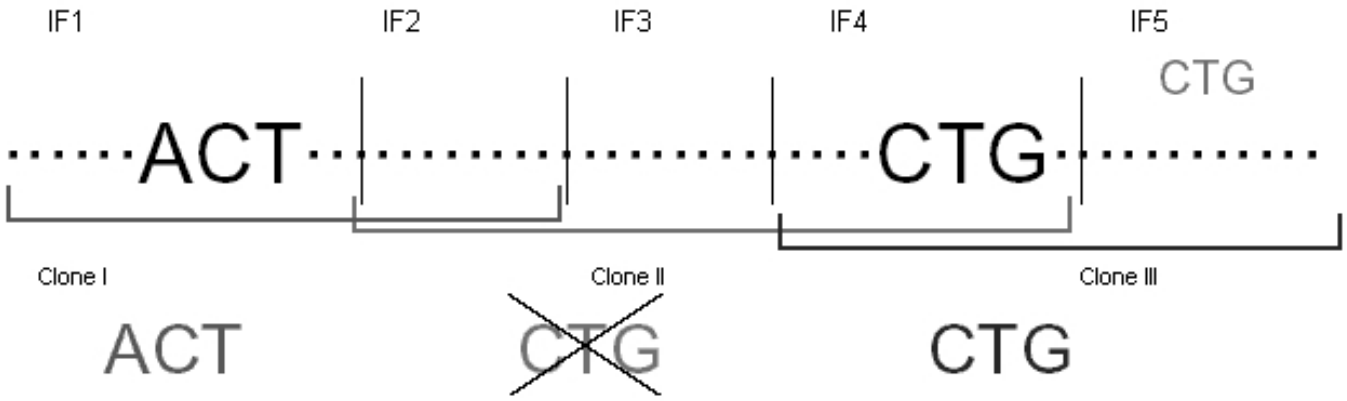


Figure 3.16: The $k$-mer CTG did not appear in the spectrum of clone II due to a false negative error, as a result of this, we will assume that the CTG $k$-mer belong to IF5 instead of IF4.

- Random sequence,that is, the distribution of each nucleotide in the sequence is uniform (with $p = 0.25$) and independent.
- False negative probability $p$ is independently for each $k$-mer and for each clone (see Figure 3.17).

Since false positive errors are quite easy to correct (by increasing the hybridization stringency) [2], we deal here only with false negative errors.

**Main Results**

Let $N$ be the sequence length, $k$ be the probe length, $d$ be the length of the $IF$s (function of the clones quantity) and $p$ be the false negative probability. Denote by $P(N, k, d, p)$ the failure probability, due to the ambiguity within the IF's spectrum.

**Theorem 3.7**

$$P(N, k, d, p) \leq (1 + \frac{c_p}{d})P(N, k, d, 0) \tag{3.7}$$

*where* $c_p \ll d$

According to this theorem we observe that the false negative errors hardly effect the method's success rate. Another result that was presented in the paper was a tighter bound to the probability of reconstruction failure in the case of no errors.

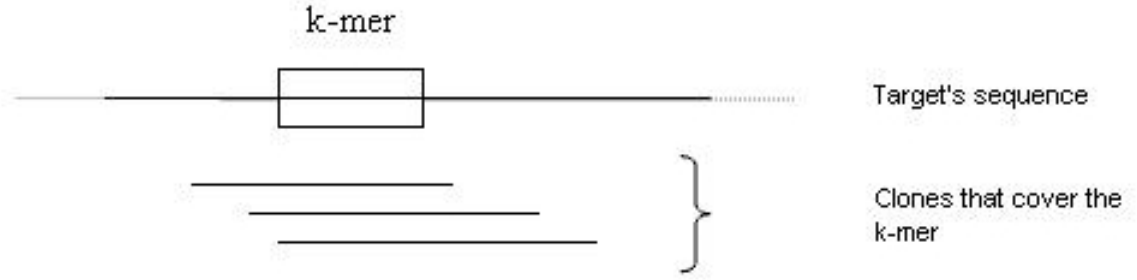


Figure 3.17: The $k$-mer has probability of $1-q$ to appear in the spectrum of each clone that covers it independently ($p$ is the false negative error rate).

| | | Arratia et al. | | Tsur et al. | | |
|---|---|---|---|---|---|---|
| $n$ | $k$ | lower | upper | lower | upper | Simulation |
| 193 | 8 | 0 | 0.5923 | 0.0051 | 0.1233 | 0.0907 |
| 791 | 10 | 0 | 0.2648 | 0.0083 | 0.1341 | 0.0996 |
| 3175 | 12 | 0.0502 | 0.1500 | 0.0094 | 0.1356 | 0.1009 |
| 12195 | 14 | 0.0742 | 0.1000 | 0.0084 | 0.1152 | 0.0875 |

Table 3.1: Source: [16]. A Comparison between the theoretical bounds of Arratia et al. [12] and Tsuer et al. [16] on the failure probability (due to ambiguity) in the errorless case and the simulation results [16], $k$ is the length of the $k$-mer, $n$ is the number of $k$-mers.

**Theorem 3.8**

$$P(N,k,d,0) = \Theta(\frac{d^3N}{4^{2k}}) \tag{3.8}$$

In the following theorem, the authors showed that the influence of the noise on the method's performance:

**Theorem 3.9**

$$P(N,k,d,p) - P(N,k,d,0) = O(\frac{d^2N}{4^{2k}}) \tag{3.9}$$

Table 3.1 shows a comparison of the failure probability bounds between Tsur et al. [16] and Arratia et al. [12]. It also shows the simulation result from [16], in the errorless case. As we can see, for $k = 14$, the upper bound in [12] is tighter.

**Shotgun SBH Performance**

This section contain simulations results for the error-prone case of generated data under the assumptions used for the theoretical analysis.

| $n$ | $k$ | $d$ | $P(n,k,d,0)$ (%) | $P(n,k,d,0.5)$ (%) |
|---|---|---|---|---|
| 7200 | 8 | 30 | 1.61 | 2.69 |
| 7200 | 8 | 40 | 3.67 | 5.20 |
| 7200 | 8 | 50 | 7.86 | 9.63 |
| 7200 | 8 | 60 | 12.85 | 15.45 |
| 7200 | 8 | 72 | 21.28 | 24.03 |
| 7200 | 8 | 80 | 27.08 | 30.36 |
| 7200 | 8 | 90 | 36.27 | 39.61 |
| 7200 | 8 | 100 | 46.12 | 49.46 |

Table 3.2: Source: [16]. The Impact of $d$: Comparison between the probability of ambiguous reconstruction in the errorless case ($p = 0$) and when the false negative error rate is $p = 0.5$. For changing $d$ (the distance between the neighboring sub-fragments). $P(n, k, d, p)$ is the probability of ambiguous reconstruction, $k$ is the length of the $k$-mers and $n$ is the length of the sequence.

Table 3.2 shows the impact of $d$ on the failure probability. The intuition is that as $d$ grows, the ambiguity is probable. We can see that in spite of the high probability of false negative error ($p = 0.5$), the failure percentage is not dramatically affected.

Table 3.3 shows the ratio between $P(n, k, d, 0)$ and $P(n, k, d, 0.5)$. In order to show it, the $P(n, k, d, 0)$ was held close to 10% (the other parameters ,$n$ and $d$, were adapted to reach this failure probability). One of the assumptions was that the IF's size ($d$) is uniform while in practice this is not the case. Table 3.4 shows the performances when the $IF$ length are poisson distributed with expectation $d$. Still, we can see (as was concluded by the paper [16]) that there is not much difference between $p = 0$ and $p = 0.5$. Table 3.5 shows the performances in the case of real DNA sequences. We can see that the results are not as good as with theoretical assumptions (100% failure with $n = 30000$), but still, the error rate in the reconstructed sequence is less than pro-mil.

Those simulations show that the error have a little effect on the performances even if the assumptions of the previous section are relaxed.

| $n$ | $k$ | $d$ | $P(n,k,d,0)$ | $P(n,k,d,0.5)$ | $\frac{P(n,k,d,0.5)}{P(n,k,d,0)}$ |
|---|---|---|---|---|---|
| 18880 | 8 | 40 | 9.85 | 13.53 | 1.374 |
| 9550 | 8 | 50 | 10.25 | 12.60 | 1.229 |
| 5520 | 8 | 60 | 9.94 | 11.90 | 1.197 |
| 3500 | 8 | 70 | 9.79 | 11.14 | 1.138 |
| 2320 | 8 | 80 | 9.56 | 10.74 | 1.123 |
| 1620 | 8 | 90 | 9.03 | 10.06 | 1.114 |
| 1200 | 8 | 100 | 8.96 | 9.64 | 1.076 |
| 880 | 8 | 110 | 8.90 | 9.50 | 1.067 |

Table 3.3: Source: [16]. The Impact of errors: Comparison between the probability of ambiguous reconstruction in the errorless case ($p = 0$) and when the false negative error is $p = 0.5$. $P(n, k, d, 0)$ is held close to 10%. For changing $d$ (the distance between the neighboring sub-fragments) and $n$ (the length of the sequence). $P(n, k, d, p)$ is the probability of ambiguous reconstruction, $k$ is the length of the $k$-mers.

| | | | Prob(fail) | | E(# errors) | |
|---|---|---|---|---|---|---|
| $n$ | $k$ | $d$ | $p$=0 | $p$=0.5 | $p$=0 | $p$=0.5 |
| 5000 | 9 | 40 | 3.8 | 4.6 | 1.11 | 1.39 |
| 10000 | 9 | 40 | 9.8 | 10.8 | 3.38 | 3.79 |
| 20000 | 9 | 40 | 15.8 | 19.2 | 5.50 | 6.93 |
| 30000 | 9 | 40 | 22.8 | 27.7 | 8.51 | 10.82 |
| 40000 | 9 | 40 | 31.6 | 36.0 | 13.67 | 16.11 |

Table 3.4: Source: [16]. Variable Size IFs (poisson distribution with expectation $d$): Comparison between the probability of ambiguous reconstruction in the errorless case ($p = 0$) and when the false negative error is $p = 0.5$. $d$ is the distance between the neighboring sub-fragments. $P(n, k, d, p)$ is the probability of ambiguous reconstruction, $k$ is the length of the $k$-mers and $n$ is the length of the sequence. The E() columns present the number of errors in the reconstructed sequence

| | | | Prob(fail) | | E(# errors) | |
|---|---|---|---|---|---|---|
| $n$ | $k$ | $d$ | $p$=0 | $p$=0.5 | $p$=0 | $p$=0.5 |
| 5000 | 9 | 40 | 40 | 40 | 5.7 | 6.9 |
| 10000 | 9 | 40 | 50 | 50 | 9.0 | 9.4 |
| 20000 | 9 | 40 | 80 | 80 | 13.4 | 15.0 |
| 30000 | 9 | 40 | 80 | 100 | 26.2 | 31.3 |

Table 3.5: Source: [16]. Real DNA Sequences: Comparison between the probability of ambiguous reconstruction in the errorless case ($p = 0$) and when the false negative error is $p = 0.5$. For changing $n$ (the length of the sequence) $P(n, k, d, p)$ is the probability of ambiguous reconstruction, $k$ is the length of the $k$-mers and $d$ is the distance between the neighboring sub-fragments.

## 3.4 Sequencing by Hybridization to a Universal Micro-array

This work by Pe'er et al. [7] presents an algorithm for reconstructing a target sequence using the SBH technology integrated with known similar reference sequence. The goal is to get the best reconstruction, given these two sources of information.

### 3.4.1 Motivation

Some motivation for using the method:

- Checking for specific mutation, e.g. Tay-Sachs, CF, Fragile-X, Gaucher. To date, those tests are performed separately for each mutation with different technologies. Since we know the sequence of the gene (and sometimes even where the mutation might appears) we would like to use this information for finding the correct path of the SBH output.
- Identifying unknown micro-organism can be done by the sequence of a conserved known gene.
- Identifying a specific strain by more divergent sequences.
- Finding the gene's sequence in one species, given this gene's sequence in another species.
- HLA-typing (graft recognition).
- Detecting somatic mutations.

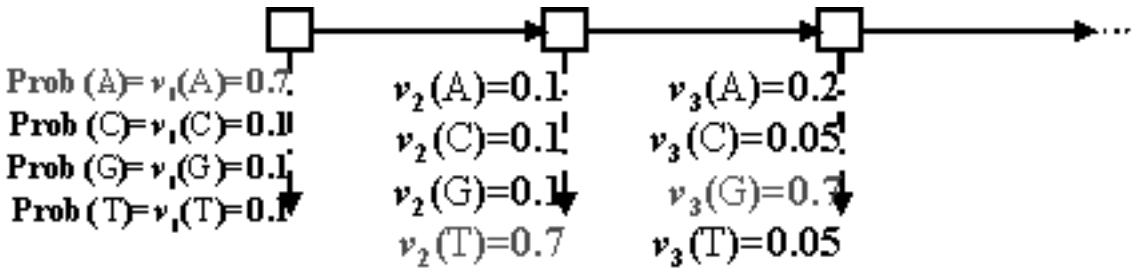


Figure 3.18: A simple example HMM for the reference sequence ATG: The nucleotide T has probability of 0.7 to appear in the second position. In this models there are no deletions or insertions.

All those challenges present one common problem: we need to reconstruct a target sequence given a known similar reference sequence. Current technologies are either target-specific, or ignore existing information on target (redo sequencing the target).

### 3.4.2 The Scheme

The scheme of finding the target sequence that fits best, both to the reference sequence and the hybridization spectrum, includes the following steps:

1. Create a model that presents the reference sequence data.

2. Convert the SBH data into a graph.

3. Using those two presentation we reconstruct the target sequence.

#### HMM Model of the Reference Sequence

The simplest HMM model (PSM), can be described as chain of states, one state for each nucleotide in the reference sequence, with nucleotide frequencies or probabilities $v_k(A)$, $v_k(C)$, $v_k(G)$, $v_k(T)$ per each state $q_k$ (see Figure 3.18). We build this model according to the known distribution of the gene among the population.

A more complicated HMM model can also include delete and insertion states (see Figures 3.19). Here we will present only the simple model, the interested reader is referred to the paper [7]. More details on HMMs can be found in [14, 15].

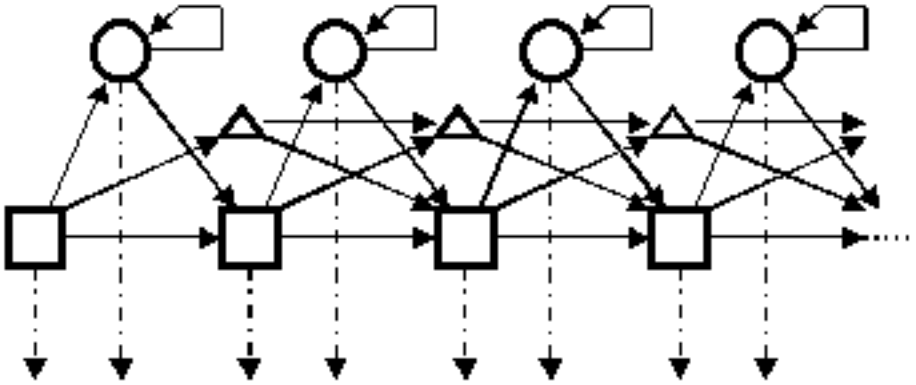

Figure 3.19: An HMM model allowing deletions and insertions. The triangles are deletion states, the circles are insertion states and the squares are match states. The solid arrow connect two states. The dashed arrow specifies emission.

### The Hybridization Graph

Let *Signal(x)* be the intensity of each $k$-mer $x$ upon the chip. For each $k$-mer $x$ we will calculate two probabilities:

$$P_1(x) = Prob(signal(x)|x \in target) \tag{3.10}$$

and the probability to get $signal(x)$ given that $x$ is not occurs along target sequence :

$$P_0(x) = Prob(signal(x)|x \notin target) \tag{3.11}$$

For the hybridization spectrum we build the following directed graph, $G(V, E)$ (see Figure 3.20):

- The set of vertices $V$ contains all the $(k-1)$-mers in the spectrum.
- The set of edges $E$ contains all $k$-mers observed in the hybridization. We build edges between max overlap $(k-1)$-mers, each edge is annotated with its $P_1(x)$ and $P_0(x)$ probabilities.

Our goal now is to find the target sequence which corresponds to paths in both graphs. In order to achieve it the authors develop scoring method which combine the two information sources.

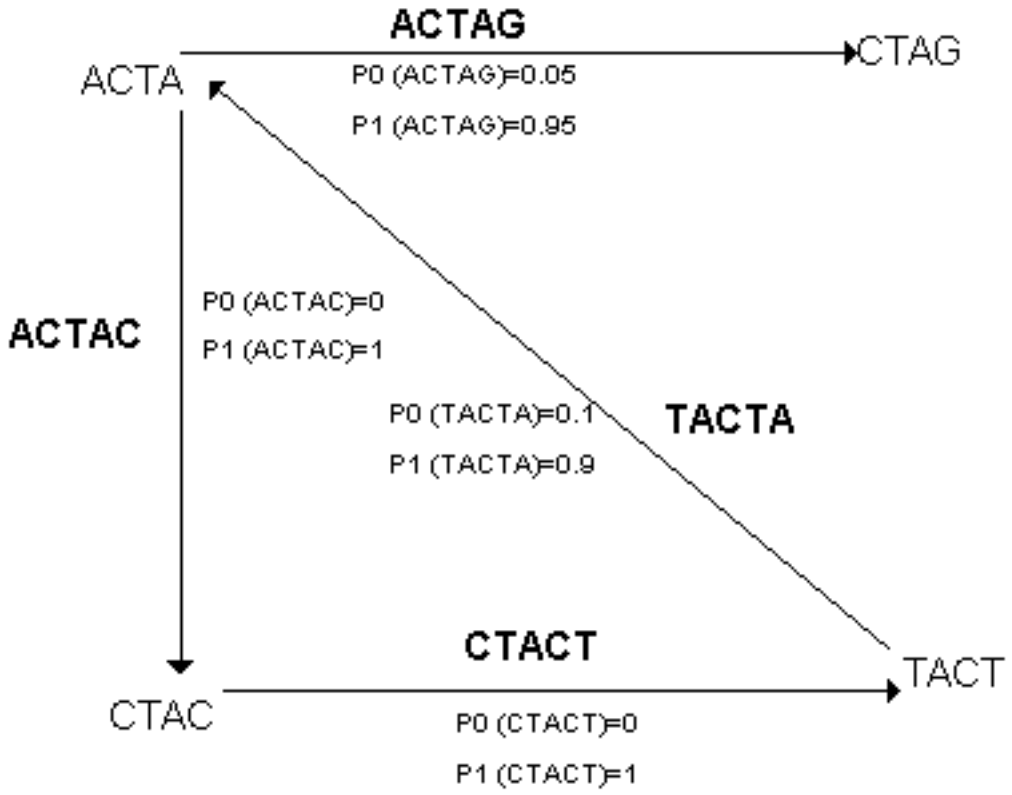


Figure 3.20: Example of the hybridization spectrum graph, the observed $k$-mer are written on the edge with their probabilities. Each vertex presents $k-1$-mer.

**Scoring according to the A-priori Distribution**

The HMM score of the target sequence $T = t_1t_2t_3...t_L$ is the probability of observing sequence, given the HMM model. For the gapless HMM model we get (see Figure 3.18):

$$HMMScoreL_L(T) = \log P(T = t_1t_2t_3...t_L|HMM) = \sum_i log(v_i(t_i)) \tag{3.12}$$

For efficient computation of this score we will use the following identity:

$$HMMScore_L = HMMScore_{L-1} + log(v_L(t_L)) \tag{3.13}$$

The score can be generalized to general HMM. The reader is referred to [7] for details.

**Scoring according to the Spectrum**

The spectrum score of the target sequence is the probability of observing spectrum, given the target sequence:

$$SpecScore_L(T) = \log P(spectrum|T = t_1t_2t_3...t_L) = log\{\prod_{x\in T} P_1(x) \prod_{x\notin T} P_0(x)\} \tag{3.14}$$

Let $W(x) = log(P_1(x)/P_0(x))$. Then:

$$SpecScore_L(T) \approx \sum_x logP_0(x) + \sum_j W(t_{j-k+1}t_{j-k+2}t_{j-k+3}...t_j) \tag{3.15}$$

This score is weight of the path in $G$ which corresponds to the target sequence. Again efficient computation of the score can be achieved using the following identity:

$$SpecScore_L \approx SpecScore_L - 1 + W(\text{current } k\text{-mer}) \tag{3.16}$$

**The Total Likelihood Score**

The total likelihood score is the sum of the spectrum and the HMM scores :

$$TotScore_L(T) = SpecScore_L(T) + HMMScore_L(T) \tag{3.17}$$

We search for $T = t_1t_2t_3...t_L$ which maximizes the total score. We have seen we can solve this problem using dynamic programming. For $0 < j < L$ and a $(k-1)$-mer $x$ let $S_j(x) \equiv max\{TotScore_j(T)|$ sequence $T$ ending with $x\}$. $S_j(x)$ can be calculated using the recursion:

$$S_j(x = x_1x_2\ldots x_{k-1}) = \max_{x_0}\{S_{j-1}(x_0x_1\ldots x_{k-2}) + W(x_0x_1\ldots x_{k-1}) + log(v_j(x_{k-1}))\}$$

$S_{opt} = \max_x(S_L(x))$. $S_{opt}$ gives us the best score for the joint path (with length L) in both graphs. In order to get the corresponds path in each iteration we will keep the character that brought us to this iteration and in the end we will be able to trace back the path (a common technique with all dynamic programming variants).

**Complexity**

It is easy to see that running time and the space requirement are $O(LK)$, as the size of the table in our dynamic programming algorithm. The space requirement can be reduce by using a divide and conquer approach [6],when using this approach we can get time complexity of $O(LKlogL)$ and space requirement of $O(K + L)$.

The typical length of the HMM, $L$, is $10^3 \leq L \leq 10^4$, the probabilistic spectrum is of length $K = O(4^k)$, typically $8 \leq k \leq 10$.

### 3.4.3 Simulation Studies

The algorithm was implemented and tested on simulated data. As a reference, a prefixes of real genomic DNA sequences were used. Each simulated run randomly generates:

1. A target sequence according to a prescribed probability $q$ of substitution.
2. A probabilistic 8 - spectrum according to a prescribed hybridization error .

For simplicity, insertions and deletion were not allowed, substitution were equiprobable, and all the probabilistic parameters were constant, i.e., position/$k$-mer independent. For each parameter set, several simulated data were generated and the algorithm was applied to each.

Two figures of merit were checked:

1. Full success rate - The fraction of runs for which the target sequence was perfectly reconstructed.

2. The percentage miss-called bases - The fractions of base-calling errors made by the algorithm. The simulation results can be seen on Figures 3.21, 3.22, 3.23 and 3.24.

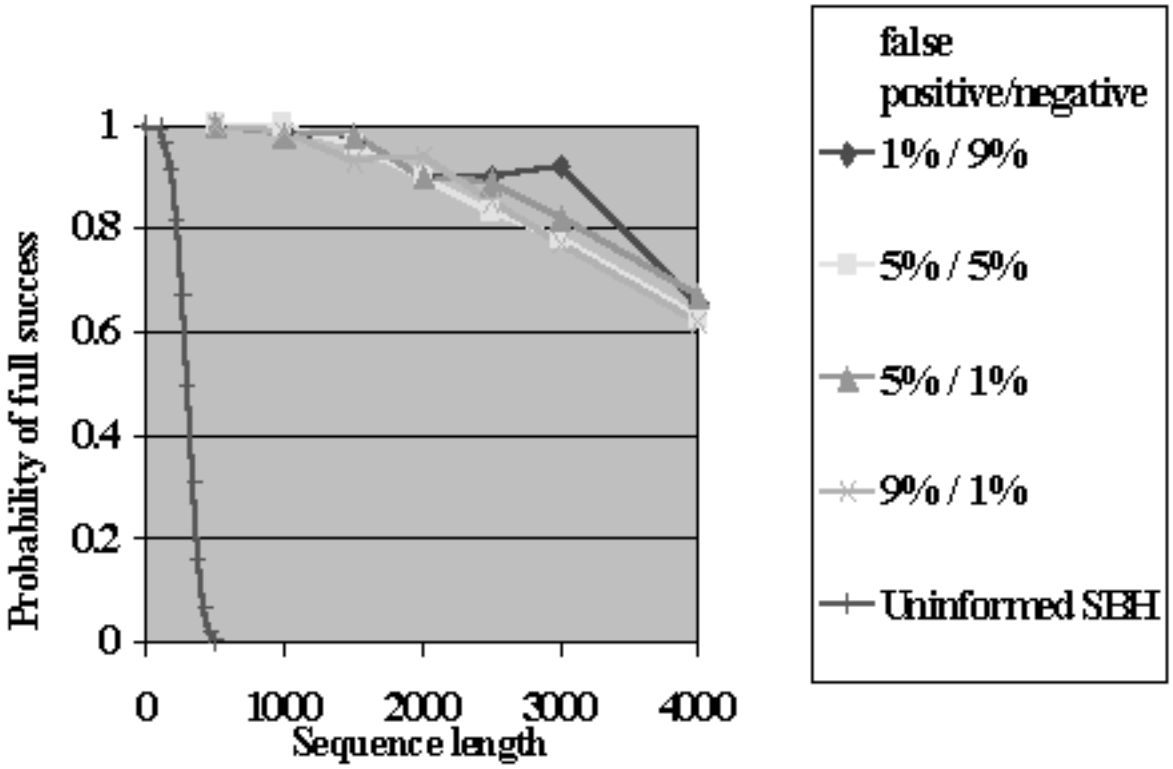


Figure 3.21: Source: [7].Success percentage versus the hybridization error level : The probability of full success in reconstructing the target sequence, as a function of the sequence length, different graphs for different hybridization error levels.

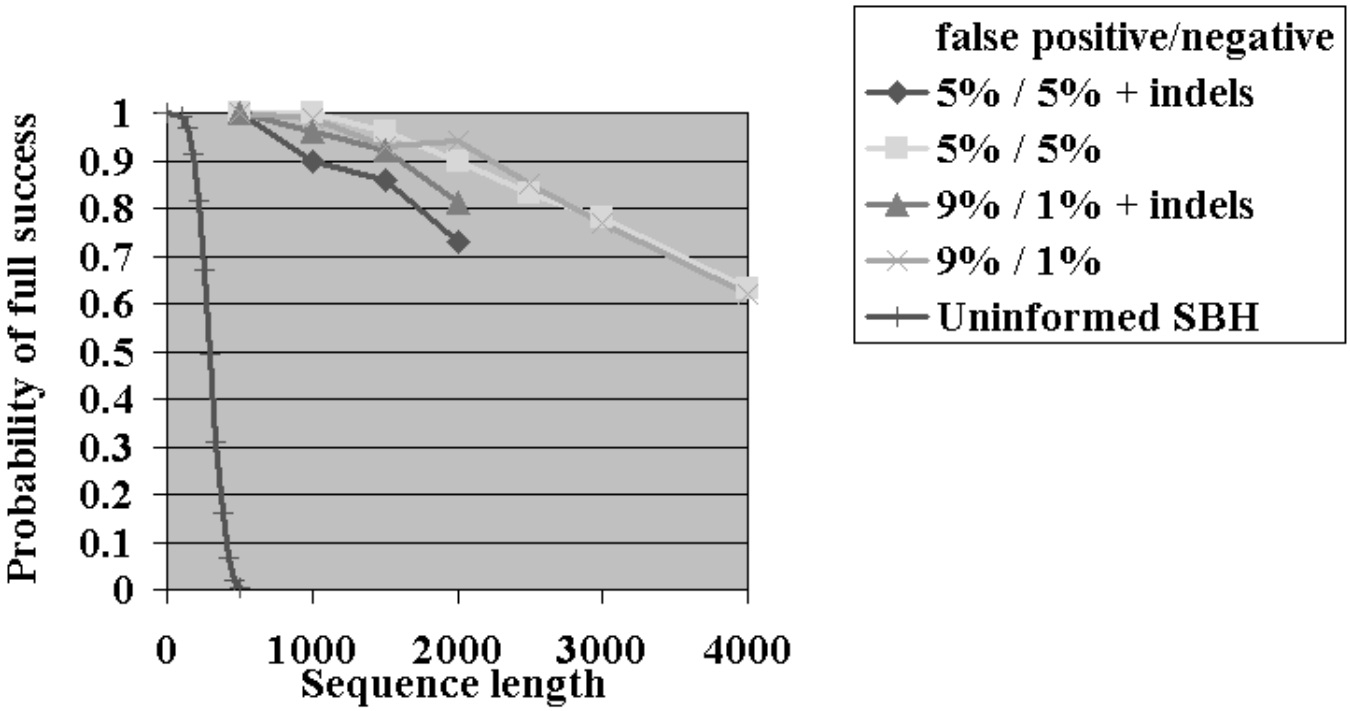


Figure 3.22: Source: [7]. Success percentage versus the insertion/deletion level: The probability of full success in reconstructing the target sequence, as a function of the sequence length, different graphs for different indels and similarity levels between the target and the reference sequences.

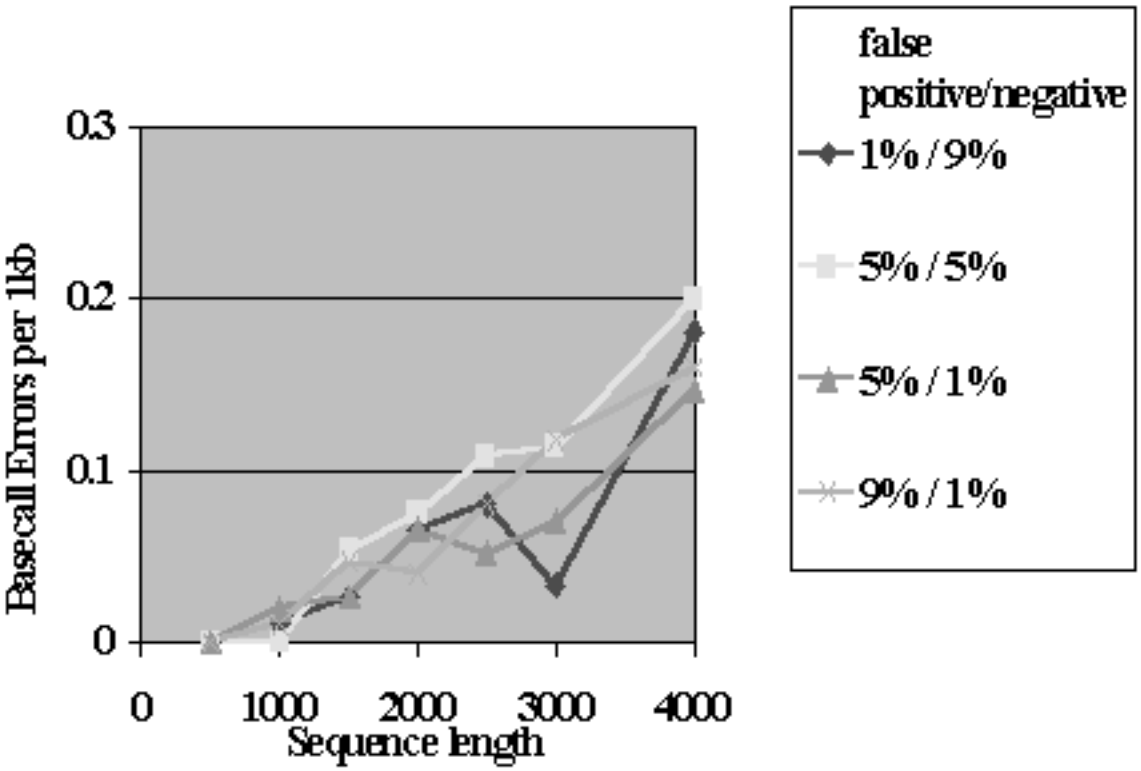


Figure 3.23: Source: [7]. Percentage of miss-called bases as a function of the sequence length versus the hybridization error level.

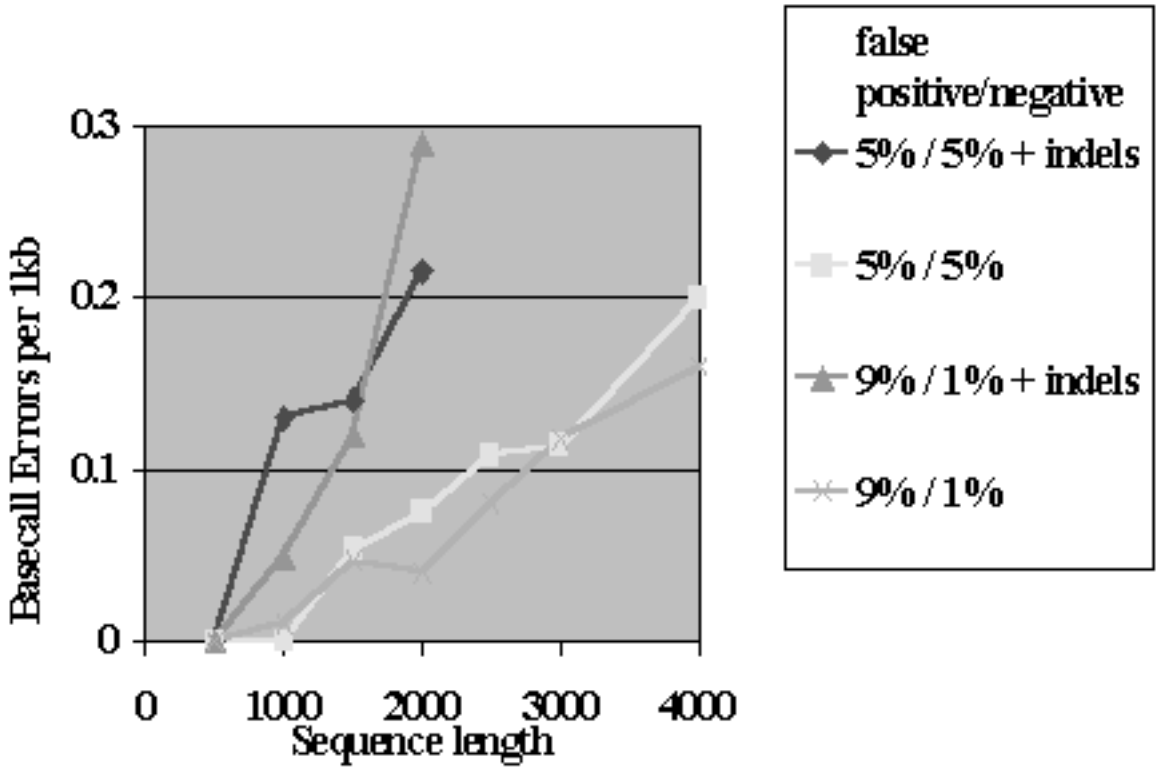


Figure 3.24: Source: [7]. Percentage of miss-called bases as a function of the sequence length versus the insertion/deletion level.

# Bibliography


[1] F. P. Preparata A. M. Frieze and E. Upfal. On the power of universal bases in sequencing by hybridization. In *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB'99)*, pages 295–301, 1999.

[2] J. Blazewicz, P .Formanowicz, M. Kasprzak, W.T. Markiewicz, and J.Weglarz. DNA sequencing with positive and negative errors. *Journal of Computational Biology*, 6(1):113 – 123, 1999.

[3] K. Doi and H. Imai. Sequencing by hybridization in the presence of hybridization errors. In *Proceedings of the Workshop on Genome Informatics (GIW '00)*, volume 11, pages 53–62, 2000.

[4] R. Dramanac and R.Crkvenjakov. Dna sequencing by hybridization. *Yugusalv Patent Application*, 570, 1987.

[5] T. Hrtman E. Halperin, S. Halperin and R. Shamir. Handling long targets and errors in sequencing by hybridization. 2002.

[6] D.S. Hirschberg. Algorithms for the longest common subsequence problem. *J.ACM*, 24:664–675, 1977.

[7] N. Arbili I. Pe'er and R.Shamir. Sequencing by hybridization to a universal microarray. Technical report, Tel Aviv University, 2002.

[8] D. Loakes and D.M. Brown. 5-nitroindole as a universal base analogue. *Nucleic Acid Research*, 22:4039–4043, 1994.

[9] G. Mayraz and R. Shamir. Construction of physical maps from olignucleotide fingerprint data . *proc. RECOMB*, 99:268–277, 1999.

[10] P. A. Pevzner. l-tuple DNA sequencing: computer analysis. *J. Biomol. Struct. Dyn.*, 7:63–73, 1989.

[11] F. P. Preparata and E. Upfal. Sequencing-by-hybridization at the information-theory bound: an optimal algorithm. *Journal of Computational Biology.*, 7:611–630, 2000.

[12] G. Reinert R. Arratia, D. Martin and M.S.Waterman. Poisson process approximation for sequence repeats and sequencing by hybridization. *J. of Computational Biology*, 3(3):425 – 463, 1996.

[13] L. Brukner R. Drmanac, I. Labat and R.Crkvenjakov. Sequencing of megabase pluse dna by hybridization : Theory and methods. *Genomics*, 4:114–128, 1989.

[14] A. Krough R. Durbin, S. Eddy and G. Mitchison. *Biological sequence analysis : Probabilistic Models of Proteins and Nucleic Acids* . Cambridge University press, 1998.

[15] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceeding of the IEEE*, 77(2):257 – 286, 1989.

[16] R. Shamir and D. Tsur. Large scale sequencing by hybridization. In *Proceeding 5th International Conference on Computational Molecular Biology (RECOMB'01)*, pages 267–279, 2001.