

ADAPTIVE CONTROL OF HYBRIDIZATION NOISE IN DNA SEQUENCING-BY-HYBRIDIZATION

HON-WAI LEONG

School of Computing, National University of Singapore
leonghw@comp.nus.edu.sg

FRANCO P. PREPARATA*

Computer Science Department, Brown University
franco@cs.brown.edu

WING-KIN SUNG[†] and HUGO WILLY[‡]

School of Computing, National University of Singapore
[†]*ksung@comp.nus.edu.sg*
[‡]*hugowill@comp.nus.edu.sg*

Received 12 October 2003

Revised 16 July 2004

Accepted 20 July 2004

We consider the problem of sequence reconstruction in sequencing-by-hybridization in the presence of spectrum errors. As suggested by intuition, and reported in the literature, false-negatives (i.e., missing spectrum probes) are by far the leading cause of reconstruction failures. In a recent paper we have described an algorithm, called “threshold- θ ”, designed to recover from false negatives. This algorithm is based on overcompensating for missing extensions by allowing larger reconstruction subtrees. We demonstrated, both analytically and with simulations, the increasing effectiveness of the approach as the parameter θ grows, but also pointed out that for larger error rates the size of the extension trees translates into an unacceptable computational burden. To obviate this shortcoming, in this paper we propose an adaptive approach which is both effective and efficient. Effective, because for a fixed value of θ it performs as well as its single-threshold counterpart, efficient because it exhibits substantial speed-ups over it. The idea is that, for moderate error rates a small fraction of the target sequence can be involved in error recovery; thus, expectedly the remainder of the sequence is reconstructible by the standard noiseless algorithm, with the provision to switch to operation with increasingly higher thresholds after detecting failure. This policy generates interesting and complex interplays between fooling probes and false negatives. These phenomena are carefully analyzed for random sequences and the results are found to be in excellent agreement with the simulations. In addition, the experimental algorithmic speed-ups

*115 Waterman Street, Providence, RI 02912-1910, USA. Supported partially by the National Science Foundation under Grant DBI-9983081 and by the Kwan Im Thong Chair at the National University of Singapore.

of the multithreshold approach are explained in terms of the interaction amongst the different threshold regimes.

Keywords: Sequencing-by-hybridization; microarrays; gapped probes; hybridization errors; false negatives; graceful degradation.

1. Introduction

DNA sequencing-by-hybridization (SBH) was proposed over a decade ago¹⁻⁶ as a potentially powerful alternative to current electrophoresis techniques. As is well known, sequencing by hybridization consists of two fundamental steps. The first, biochemical in nature, is the acquisition, by complementary hybridization with a complete library of probes, of all subsequences (of a selected pattern) of a given unknown target sequence; the set of such subsequences is called the sequence *spectrum*. The second step, combinatorial in nature, is the algorithmic reconstruction of the sequence from its spectrum.

However elegant in its conception, the approach is plagued by serious difficulties, both biochemical and combinatorial. While considerable progress has been made on the combinatorial side, through the identification of probing patterns capable to nearly achieve optimal performance (effectiveness),^{7,8} the biochemical aspect remains more problematic due to difficulties in reliably controlling the dynamics of array hybridization. This is further complicated by the fact that the novel information-effective method⁸ contemplates the use of universal bases, ideally exhibiting no hybridization specificity for the four natural bases (wild-card bases). Little is known about chemical compounds with such properties, and what is known fits imperfectly the model of non-specificity.

Normally, the spectrum is assumed to contain *exactly* the subsequences of the target sequence conforming to a chosen pattern (*noiseless spectrum*). However, the serious inadequacy of such an assumption was early recognized and several suggestions have been made^{4,5,9-12} to confront the problem of *noisy* hybridization. Of course, any approach to error control presupposes an error model, i.e., a formalization of the random process producing the hybridization errors (hybridization noise), in the form of *false negatives* (errors of the first kind or misses) and of *false positives* (errors of the second kind or false-hits). Unfortunately, knowledge of the hybridization process is currently inadequate for a precise quantification of the hybridization model; judging from available data for natural bases,^{13,14} it appears likely that a realistic oligonucleotide hybridization model may have to be probe-specific.

In the absence of sufficient data for realistic modeling, researchers have directed their attention to the question of a graceful degradation of the efficiency of the reconstruction process in the presence of noise. Such studies are typically based on the following error process:

Noisy Standard Model

- (1) Any correct spectrum probe can be suppressed with a fixed probability (false negatives);

- (2) any probe at Hamming distance 1 from a correct spectrum probe can be added to the spectrum with a fixed probability (false positives); and
- (3) hybridization noise is expressed in terms of error rates for false negatives and positives.

In this model, Doi and Imai¹⁵ recently investigated whether the method by Preparata *et al.*⁷ and Preparata and Upfal,⁸ which provably achieves asymptotic optimality in the noiseless model, remains viable in the presence of noise. Their study, based on simulations, reached very negative conclusions, exhibiting a dramatic drop in performance.

In a very preliminary version of this paper,¹⁶ we have refuted the conclusions by Doi and Imai¹⁵ by showing that proper adaptation of the reconstruction algorithm achieves an acceptably graceful performance degradation; for completeness, a brief analysis of the poor effectiveness of the approach outlined by Doi and Imai¹⁵ is reported as an appendix. After the original submission by Leong *et al.*,¹⁶ a noteworthy new approach to fault-tolerant SBH has been proposed by Halperin *et al.*,¹⁷ based on voting for the majority-supported choice for sequence extension.

In this paper we present an algorithm that is much more effective and sophisticated than that presented in by Leong *et al.*¹⁶ This fault-tolerant sequence reconstruction algorithm is based on adapting the robustness of the reconstruction to the locally detected severity of the hybridization noise. The performance analysis, albeit simplified, adequately agrees with extensive simulation data. Of course, the preservation of acceptable effectiveness of fault-tolerant reconstruction is necessarily achieved at the price of computational cost, although the proposed (multithreshold) adaptiveness is substantially more efficient than its single-threshold counterpart.

2. Preliminaries

We briefly review the probing scheme and the reconstruction algorithm in the error-free case.

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

Definition 2. An (s, r) *probing scheme*, of weight $k = r + s$, has *direct* pattern $1^s(0^{s-1}1)^r$ and *reverse* pattern $1(0^{s-1}1)^r1^{s-1}$. The *spectrum* of a given target sequence is the collection of all its subsequences conforming to the chosen probing pattern.

For notational convenience a probe is viewed as a string of length $(r + 1)s = \nu$ over the extended alphabet $\mathcal{A} = \{A, C, G, T, *\}$, where $*$ denotes the “wild-card”. A probe occurs at position i of the target if i is the position of its rightmost symbol. Two strings over \mathcal{A} of identical length are said to *agree* if they coincide in the positions where both have symbols different from $*$.

All reconstruction algorithms heretofore proposed construct a *putative* sequence symbol-by-symbol. A reconstruction is successful if the completed putative sequence coincides with the original sequence. We ignore here the details, discussed elsewhere,⁷ of the initiation and termination of the reconstruction process.

Definition 3. A probe is said to be a *feasible extension* if its $(\nu - 1)$ -prefix coincides with the corresponding suffix of the putative sequence.

We now summarize the standard reconstruction algorithm for noiseless spectra,⁸ to provide the background for the necessary algorithmic modifications:

- Given the current putative sequence, the spectrum query returns the set of feasible-extension probes (which is necessarily nonempty in the error-free case if the putative sequence is correct). If only one probe is returned, then we have trivial one-symbol extension of the putative sequence viewed as a graph-theoretic path (algorithm in **extension mode**). Otherwise, we have an ambiguous branching and two or more competing paths are spawned; subsequently the algorithm attempts the breadth-first extension of all paths issuing from the branching (and of all other paths spawned in turn by them) on the basis of spectrum probes (algorithm in **branching mode**). The construction of such tree of paths is pursued up to a maximum depth H (a design parameter), unless at some stage of this construction it is found that all surviving paths have a common prefix. In the latter case, this prefix is concatenated to the putative sequence, and the process is iterated.

The occurrence of ambiguous branchings is due to the following construct:

Definition 4. A *fooling probe* at position i is a feasible extension for position i which occurs as a subsequence at some position $j \neq i$ in the target sequence.

Indeed, fooling probes are the cause of reconstruction failures. Intuitively, for a given probe weight k , as the length m of the target sequence increases, the spectrum becomes more densely populated, and correspondingly the probability of encountering fooling probes increases (we shall later discuss the probability of fooling probes). The rationale for the described reconstruction mechanism in the branching mode is that, whereas the correct path is deterministically extended, the extension of spurious paths rests on the existence in the spectrum of specific fooling probes, and the parameter H should be chosen large enough to make their joint probability vanishingly small.

When the branching-mode extension reaches depth H , we have two distinct failure modes:

- (1) The two extant paths are identical except in their initial symbols (so the same feasible-extension probes extend both paths) (Failure Mode 1); or
- (2) The two paths reproduce two distinct portions of the target sequence (and are deterministically extended). The spurious path contains a segment of length

$\nu - 1$, which agrees, entirely or partially, with an equally positioned segment of the correct path, the disagreements being compensated for by fooling probes (Failure Mode 2).

3. The Modified Reconstruction Algorithm

Intuitively, a false positive is much less detrimental to the reconstruction than a false negative. Indeed, referring to the above description of the algorithm, a false negative irretrievably interrupts the extension process, whereas a false positive simply adds one probe to the spectrum. Such probe will be accessed only if it may act as a feasible extension (in the same way as a fooling probe). In other words, false positives simply increase the pool of fooling probes, and, provided the false-positive rate remain reasonably small, their effect is truly negligible;¹⁵ in the interest of simplicity and clarity in this paper we assume that the false-positive rate is zero. Inclusion of false positives should only minutely complicate the analysis.

Essentially, the success of reconstruction depends upon our ability to recover false negatives. In fact, we shall explain analytically that the poor behavior reported in Doi and Imai¹⁵ is due to inadequate recovery of false negatives.

We therefore propose, and later analyze, a robust reconstruction algorithm, which interacts with a noisy spectrum. We assume therefore that the latter is obtained by eliminating with probability ϵ and independently each probe of the noiseless spectrum. As we shall recognize, the outlined standard algorithm can be viewed as a special case of the new procedure. Informally, we shall introduce two novel features: a modified spectrum interrogation (query), and a path elimination criterion. Specifically:

- **Modified spectrum query**

- (1) The spectrum query always returns four scored extensions (for each of the four DNA bases); the score is 0 if the corresponding probe exists in the spectrum and is 1 otherwise;
- (2) All extant paths have identical length and are extended with additive scoring.

- **Path-termination criterion:** A path is terminated when its score exceeds by a threshold θ (a small integer) the score of the lowest-score extant path.

Clearly, the standard algorithm corresponds to the choice $\theta = 1$. Choices of $\theta > 1$ allow recovery of all false negatives, and the potential for successful reconstruction rests on the fact that lack of fooling probes (for spurious paths) is more likely than occurrence of closely spaced false negatives (on the correct path). It is also intuitive that higher reliability is achieved for higher values of θ .

All other algorithmic details are maintained; the resulting algorithm is referred to as threshold- θ reconstruction algorithm. We shall assume throughout the paper to deal with reverse probing patterns, due to their algorithmic advantages.

4. Failure Analysis of Threshold- θ Algorithms

The events causing the algorithm to fail are collections of false negatives and collections of fooling probes. Whereas the former are by definition independent, fooling probes are not because they may overlap. Thus, the crux of the analysis is the evaluation of the joint probability of collections of fooling probes.

We begin by observing that the probability that a specific probe of weight k (i.e., with k specified symbols) belongs to the spectrum of a sequence of length m is expressed with sufficient accuracy by

$$\alpha = 1 - \left(1 - \frac{1}{4^k}\right)^m \approx 1 - e^{-m/4^k}.$$

Of course, conditioning should be considered because, rigorously, the existence of a given probe alters the prior probability of other probes. This dependence is exclusively due to overlaps. However since overlap spans are enormously smaller than the sequence length, here we make the strongly simplifying assumption of independence. Whereas such an assumption minimally weakens the rigor of the analysis, the excellent agreement with simulation results amply justifies the choice. In the same spirit, the analysis that follows will occasionally sacrifice rigor for the benefit of simplicity based on sound intuition.

As reviewed in Sec. 2, in the noiseless situation the performance of reconstruction algorithms is correctly described in terms of the Failure Modes 1 and 2.

Under noisy conditions ($\epsilon > 0$), the error-free failure modes remain active, but a more complex failure behavior arises due to the interplay of false-negatives and fooling probes, which we first intuitively motivate. A new mode, referred to as Failure Mode 3 or *correct-path elimination*, emerges and may become a substantial cause of failure for high error rates. The elimination of the correct path is due to the path-rejection rule outlined in Sec. 3 and to the occurrence of closely spaced false-negatives on the correct path. In such an event, the correct path receives a high score and the algorithm may wrongly terminate it; consequently, the algorithm will detect failure shortly thereafter because all spurious paths become extinct. In addition, the co-occurrence of closely spaced false negatives (on the correct path) and of fooling probes (for a spurious path) may prevent the algorithm from discriminating between the two, thus incorrectly emulating failure situations that are standard in error-free operation. In detail, we identify five distinct failure modes, to be now individually analyzed:

- (1) **Failure Mode 1** (conventional). The branching-mode extension tree attains depth H with two identical paths (except for their initial symbol) with identical score 0. As discussed elsewhere,⁸ such a failure occurs when the spectrum contains k specific fooling probes that sample the spurious branching symbols, otherwise agreeing with the correct path. The approximate probability of

occurrence of this event at a specific position in the reconstruction is^{8,18}:

$$(1 - e^{-3m/4^k})\alpha^{k-1} \left(1 + \frac{4^{r+1}}{m}\right)^r \left(1 + \frac{4^s}{m}\right)^{s-1},$$

where the first term is due to the fooling probe causing the branching, the second to the subsequent $k - 1$ fooling probes, and the rightmost two terms account for fooling-probe overlaps. For simplicity, in subsequent discussion we shall approximate this probability with $3\alpha^k$.

- (2) **Failure Mode 2** (conventional). A detailed combinatorial analysis given elsewhere¹⁸ fails to convey an intuitive appreciation of the behavior. We attempt to provide intuitive support with the following analysis of an approximate but quite effective modeling.

Note that the tree of paths issuing from the branching contains the correct path and (at least) one competing (spurious) path. The latter contains an $(\nu - 1)$ -symbol segment (self-sustaining segment), including or following the branching position, which is identical to a segment actually occurring in the sequence, so that its extension is deterministically guaranteed by probes in the spectrum. The self-sustaining segment agrees, entirely or partially, with an equally positioned segment of the correct path, the disagreements being compensated by fooling probes also occurring in the sequence.

Denoting conventionally the branching position as 0, the position-index immediately to the right of the self-sustaining segment is called the segment's *offset* and denoted J . Thus, $J \geq 0$.

The following observation will intuitively support the chosen approximations. The spectrum must contain a set of fooling probes necessary to compensate for the disagreements between the two competing homologous segments (aligned on the two paths). Precisely, no probe is required at a position $0 \leq j < J$ if and only if no disagreement (between the two alternative paths) occurs at positions $\{j, j-1, \dots, j-s+1, j-2s+1, \dots, j-\nu+1\} \cap \{0, 1, \dots, j\}$. Positions $\{j, j-1, \dots, j-s+1, j-2s+1, \dots, j-\nu+1\}$ are referred to as *in-phase* with respect to position j , corresponding to shifting to j the right end of the chosen (reverse) probing pattern; of these, we must exclude those occurring to the left of the branching. Thus, a single disagreement may require the presence of several compensating fooling probes, and since a disagreement is 3 times as likely as an agreement, we may expect that there will be a fooling probe (with its rightmost symbol) at nearly every position in the interval $[1, J - 1]$. Our approximation expresses the probability of this event as $\alpha^{\eta(J-1)}$ (where η slightly less than 1; detailed computer simulations suggest that $\eta \approx 0.9$).

For generic J there are about m^2 ways to select the self-sustaining segment and its correct-path counterpart in the sequence. For $J < \nu - 1$, the two segments agree in $\nu - 1 - J$ positions (with probability $1/4^{\nu-1-J}$), the branching has probability 3α , and each subsequent fooling probe has probability α^η .

We conclude that the probability of Mode 2 failure may be approximated as

$$m^2 \sum_{J \geq 1} \frac{3\alpha}{4^{\nu-1-J}} \alpha^{\eta(J-1)} = \frac{3\alpha m^2}{4^{\nu-2}} \sum_{J \geq 1} (4\alpha^\eta)^{J-1} \leq \frac{3\alpha m^2}{4^{\nu-2}} \cdot \frac{1}{1-4\alpha^\eta}. \quad (1)$$

(The next three failure modes are due to the action of false-negatives.)

- (3) **Failure Mode 3.** This is the only case of termination occurring possibly before the tree construction reaches depth H . The correct path accumulates a score exceeding by θ the score of a spurious path. Here again we have two competing paths: the correct path and a spurious path. The latter, in turn, may be of one of two types, which we call here *Mode-1* or *Random*, with different probabilistic characterizations.

- (a) A Mode-1 path agrees with the correct path in all in-phase positions prior to its termination. Termination beyond position ν would appear as a Mode-1 failure, whence the label of such paths. Only in-phase positions need be considered, because a false negative in off-phase position equally affects both paths. Failure occurs at the i th in-phase position because while the spurious path is supported by i fooling probes (probability $3\alpha^i$), there are θ positions (including the i th one) where the correct path experiences a false-negative, i.e., the event's approximate probability is

$$F_{3,1} = \alpha \sum_{i=\theta}^k 3\alpha^{i-1} \binom{i-1}{\theta-1} \epsilon^\theta = 3\alpha \epsilon^\theta \sum_{j=\theta-1}^{k-1} \alpha^j \binom{j}{\theta-1}.$$

- (b) Random paths. These paths are sustained position by position by fooling probes, the first symbol with probability 3α and each subsequent symbol with probability 4α , while the correct path experiences θ false negatives. It follows that the probability of this event can be expressed as

$$F_{3,2} = \sum_{J=\theta}^{\nu} (1 - \alpha^{s-2+\lfloor J/s \rfloor}) 3\alpha (4\alpha)^{J-1} \epsilon^\theta \binom{J-1}{\theta-1}$$

where the term $(1 - \alpha^{s-2+\lfloor J/s \rfloor})$ excludes Mode-1 paths. We then set

$$F_3 = F_{3,1} + F_{3,2}.$$

- (4) **Emulated failure Mode-1.** This case occurs when the tree-depth reaches value H apparently in Mode 1 and both paths have accumulated identical score $u > 0$ within the probe length. For example, for probe pattern 10001000100010001111, consider the following:

	0	3	7	11	15	19				
...	CCATCAT	T	TAG	GCCA	AGCC	CGAT	GGCA			
	<i>a</i>	<i>t</i>	<i>a</i>	<i>GCC</i>	<i>a</i>	AGCC	CGA	<i>t</i>	GGC	<i>a</i>

A branching occurs at position 0 and the top path is the correct one. Fooling probes are indicated with lower-case and false negatives in bold-face. The

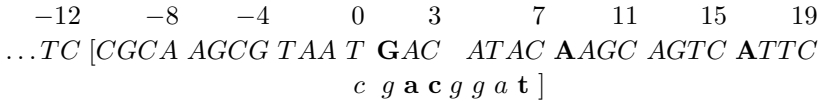
branching disagreement requires fooling probes at 1, 2, 3, 7, 11, 15, and 19. It also happens that the correct and spurious paths experience false negatives at 15 and 11, respectively. Again, only in-phase positions need be considered. Provided the score has been accumulated in the first ν positions, the correct path suffered some number u of false negatives and the spurious path lacked u fooling probes, so that both path have identical scores. For given u , the false-negative and fooling-probe positions can be chosen independently among k (in-phase) positions, so that the corresponding probability is

$$3\alpha^{k-u}\epsilon^u \binom{k}{u} \binom{k}{u}.$$

Combining conventional and emulated Mode-1 failures we obtain:

$$F_1 = 3\alpha^k \sum_{u=0}^{\theta} \left(\frac{\epsilon}{\alpha}\right)^u \binom{k}{u}^2.$$

- (5) **Emulated failure Mode 2.** As in the previous case, identical scores have been accumulated by both paths in the first ν positions, so that the conventional case handled above correspond to score 0 for both paths. For example, again for probe pattern 10001000100010001111 and with the same conventions as above, consider the following situation:



Here, the self-sustaining segment begin at position -11 and ends at 7 (shown within brackets); its initial portion, $[-11, -1]$ coincides with the corresponding portion of the correct path. There are disagreements in position 0 (branching) and in positions 4, 5, and 7, each of which demands a sequence of fooling probes ending within the span of the segment, so that positions 0–7 are each a fooling-probe site. Those in positions 2, 3, and 7 are false negatives; likewise, the correct path experiences false negatives in positions 1, 8, and 16.

In general, the correct path suffers u false negatives and the spurious path lacks u fooling probes. It follows that the corresponding probability is obtained multiplying each term $4\alpha^{\eta J}$ in expression (1) by the term

$$1 + \sum_{u=1}^{\theta} \left(\frac{\epsilon}{\alpha}\right)^u \binom{J}{u} \binom{\nu}{u},$$

since false negatives are to be chosen among ν positions and fooling probes among J positions. Thus, we obtain

$$F_2 = \frac{3\alpha m^2}{4^{\nu-2}} \sum_{j \geq 1} (4\alpha^{\eta})^{j-1} \left(\sum_{u=0}^{\theta} \left(\frac{\epsilon}{\alpha}\right)^u \binom{j}{u} \binom{\nu}{u} \right).$$

Combining the results of the preceding analysis, since the obtained expressions refer to a fixed sequence position, we may conclude that the probability of success of the threshold- θ algorithm is given by

$$e^{-m(F_1+F_3)+F_2}.$$

5. The Adaptive Multithreshold Algorithm

Intuitively, it is clear that the observed high computational cost of the single-threshold algorithm for values $\theta > 1$ is due to the fact that the full robustness of the method is applied uniformly over the entire sequence reconstruction, even where it is not needed. For example, for $m = 10,000$ and $\epsilon = 0.01$ there are about 100 false-negative events, each involving a short stretch (20–30 symbols) where the higher-threshold machinery must be applied; thus, the remaining 7000–8000 symbols are presumably reconstructible with the much more efficient ($\theta = 1$) algorithm. This intuition is the basis of the adaptive multithreshold approach to be described next. All simulation results reported in this section pertain to random DNA sequences with i.i.d. symbols and to a probing scheme using a (4, 4) reverse pattern, i.e., the pattern (refer to Definitions 1 and 2):

1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 1 1 1

We begin with an informal description. The algorithm's normal operating mode is the standard one (noiseless), described by Preparata and Upfal⁸ and succinctly reviewed in Sec. 2. The putative sequence is extended from one end to the other on the basis of the spectrum. A false-negative causes an *interruption* of this process, either because none of the four possible extension probes are immediately found in the spectrum, or because an initiated spurious path becomes extinct. Thus, such an interruption occurring prior to the expected termination of the target sequence (we assume that the parameter m is known with reasonable accuracy) reveals the occurrence of a false negative. On such circumstance, the algorithm switches to threshold $\theta = 2$ and attempts local reconstruction using this threshold. As described earlier, in this new operating mode paths are scored by the number of missing characters introduced by the algorithm and extant paths are eliminated on the basis of their accumulated score. If this action succeeds in reliably extending the putative sequence beyond the detected interruption, then algorithm reverts to $\theta = 1$; otherwise it switches to operation with the next higher threshold $\theta = 3$. This policy is upheld up to some threshold θ_{\max} .

However the interaction of different-threshold subalgorithms is not seamless, and there are several significant details to be addressed:

- (1) *Threshold policy.* In a single threshold algorithm path elimination occurs when the score of a path exceeds by θ the minimum path score, as illustrated in Sec. 3. Such criterion, called *relative scoring*, is meant to guard against an improbable concentration of false-negatives that may locally penalize the correct path, but

Table 1.

length	succ	accesses	succ	ch-1	access-1	ch-2	access-2	ch-3	access-3	J^*
1000	100	118824	100	905	3964	52.4	2553	2.6	2522	3
2000	100	321866	100	1801	8576	148.5	7662	9.5	10597	5
3000	100	619822	100	2648	13609	291	16250	21	23141	7
4000	100	1074961	100	3376	19387	543	32540	41	54353	11
5000	100	1787054	99.3	4194	26041	698	51745	68	113885	11
6000	98.8	2878040	99.7	4771	33768	1063	91525	126	258406	15
7000	99	4602999	98.7	5523	42882	1263	137822	174	470840	15
8000	95.7	7237105	96	6254	53694	1457	205413	249	940193	15
9000	88.7	11033926	91.3	6639	66783	1929	339477	392	1978891	19

is computationally more costly since it uniformly adopts a policy targeted to infrequent events. Thus, an alternative policy is to maintain relative scoring only in connection with θ_{\max} (since there is no further resort), and to adopt instead the following less costly policy for any threshold $1 \leq \theta < \theta_{\max}$ (absolute scoring): branching-tree extension is halted when the score of all extant paths attains the value θ , the rationale being that a potentially eliminated correct path is recoverable at higher thresholds.

- (2) *Reconstruction back-up.* The multithreshold approach introduces the following complicating behavior. Suppose we are operating with threshold 1 and that the algorithm detects a false-negative interruption. The detected false-negative, however, has not necessarily occurred at the position following the interruption. Indeed, a branching may have occurred a few positions prior to the interruption, and the algorithm has undertaken the construction of the trees spawned off the correct path. Therefore, when switching to a higher threshold, reconstruction should back up a number of positions $J^*(m, \epsilon)$. We have chosen to select J^* empirically as the value for which a further increase does not affect the frequency of correct reconstruction in simulations (as an example, refer to the last column of Table 1, pertaining to $\epsilon = 0.01$ for a (4, 4) reverse-probe algorithm). The algorithm will remain in higher-threshold mode until the putative sequence has been extended at least one position beyond the original interruption, a policy which is consistent with the recovery of the false negative.

We can now give a more technical description of the multithreshold algorithm, which accepts as input a (putative) sequence P , a maximum operating threshold θ_{\max} , and a recovery back-up J^* .

multithreshold(P, θ_{\max}, J^*)

$\theta_{curr} = 1$

extend P until failure occurs

upon failure

$L^* = |P|$, $L = L^* - J^*$

remove J^* -suffix from P

$\theta_{curr} = 2$

```

while ( $1 < \theta_{curr} \leq \theta_{max}$ )
  while ( $L \leq L^*$ )
    start  $\theta_{curr}$  algorithm at  $L$ 
     $E \leftarrow$  prefix returned by tree extension to depth  $H$ 
    if  $|E| = 0$  then
      if  $\theta_{curr} < \theta_{max}$ 
        then  $\theta_{curr} = \theta_{curr} + 1, L = L^* - J^*$  else return FAIL endif
      else
         $P = PE, L = |P|$ 
        if  $P$  ends at terminating primer then return  $P$ 
        else if  $L > L^*$  then  $\theta_{curr} = 1$  endif
    endif endif

```

Figure 1 jointly display the analytical as well the experimental performance curves for $\theta = 3$ and error-rates $\epsilon = 0.001, 0.005, 0.01, 0.02$, from right to left; the heavier curve to the right gives the analytical performance for the noiseless situation. Each experimental point has been obtained by means of 400 hundred random runs for fixed m and ϵ . This diagram clearly illustrates the expected graceful degradation of performance as the hybridization noise increases.

Finally, we wish to substantiate the advantages offered by multithreshold over single threshold reconstruction. As a starting point of this analysis we report

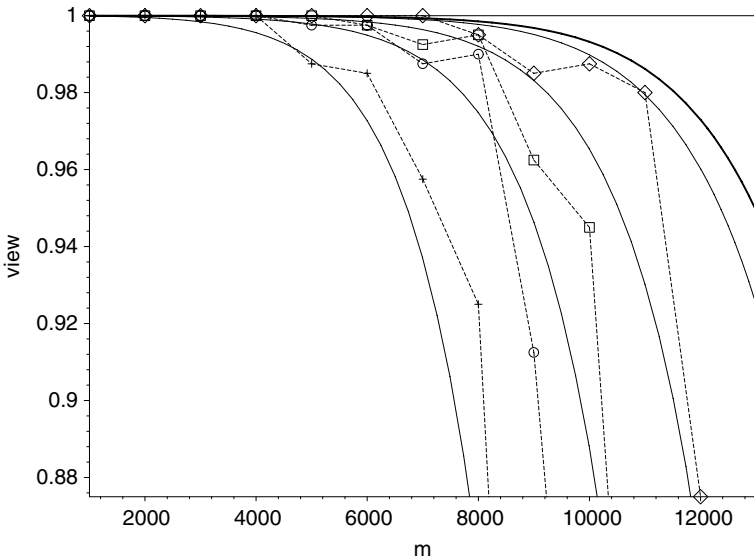


Fig. 1. Analytical and experimental performance curves for $\theta = 3$ and error-rates $\epsilon = 0.001, 0.005, 0.01, 0.02$, from right to left. Heavier curve gives noiseless performance.

and summarize voluminous simulation data.^a For each data point $\{(m, \epsilon) | m = 1000, 2000, \dots, 9000; \epsilon = 0.001, 0.005, 0.01, 0.02\}$ we have carried out 400 simulation runs for multi- and single-threshold reconstructions for $\theta = 3$ and averaged the corresponding performance and spectrum accesses (as a measure of computational load). The most significant of these results are displayed in Figs. 2 and 3 (see also Table 1 for detailed data referring to $\epsilon = 0.01$). Figure 2 confirms, as expected, that the performances of multithreshold-3 and single-threshold-3 reconstructions are almost identical; thus the comparison must hinge on computational advantages. The speed-up afforded by multithreshold is given in Fig. 3.

The latter figure demonstrates a speed-up for all parameter choices, which, by itself, is persuasive experimental evidence of the superiority of multithreshold reconstruction. On the other hand, the reported results present a rather puzzling spread of speed-up values, which calls for some specific analysis. Qualitatively, we observe that the speed-up varies moderately for fixed noise level as a function of m (increasingly for small noise and decreasingly for larger noise), whereas it varies appreciably with the noise level. A detailed, rigorous analysis of the extremely intricate reconstruction process, as driven by random DNA sequences, is of a daunting complexity and hardly justified in consideration of its very narrow scope. Rather, we shall attempt below a grossly simplified analysis that captures, and roughly

^aThe software used to carry out these simulation is maintained by co-author Hugo Willy at the National University of Singapore.

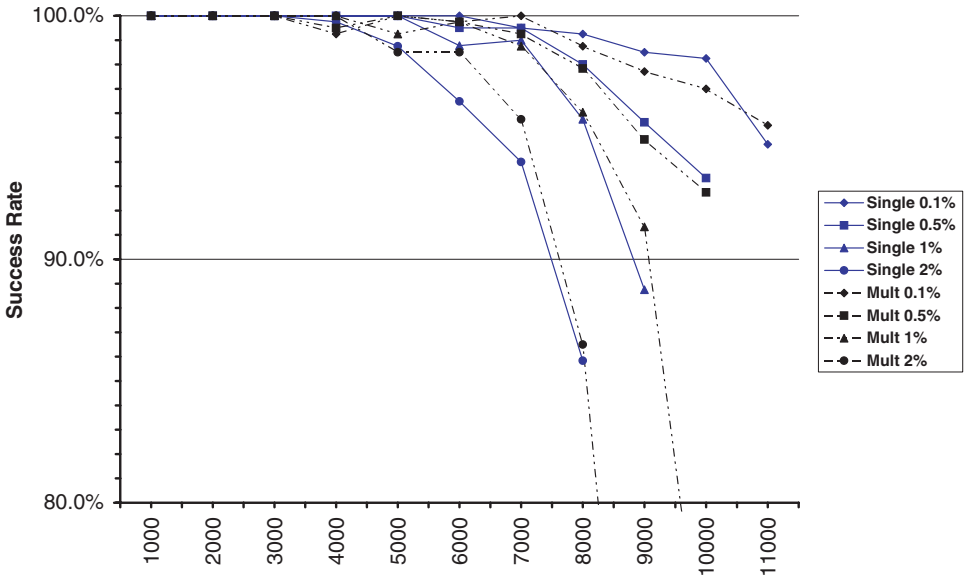


Fig. 2. Experimental performances for single- and multi-threshold algorithms for $\theta = 3$ (solid lines refer to single-threshold reconstructions).

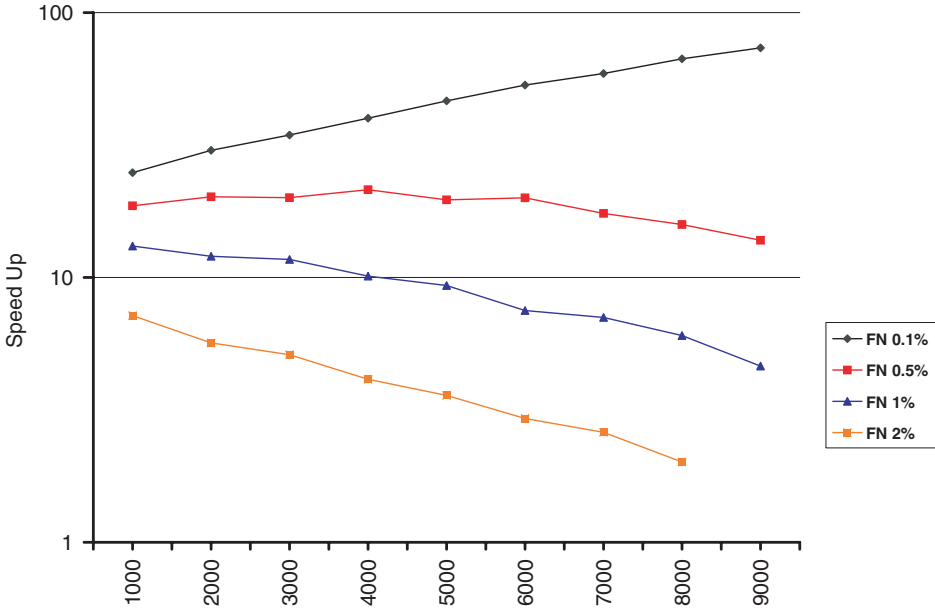


Fig. 3. Multithreshold speed-ups for different noise values.

quantifies, the main features of the process and offers a reasonable explanation of the observed behavior.

The computation time of a successful reconstruction is essentially proportional to the number of spectrum accesses requested by the algorithm. These spectrum accesses are incurred in the path extension process and can be partitioned among the acquired characters of the putative sequence. For each character we estimate the number of accesses performed before the (potentially 3) subtrees initiated by the spurious threads are terminated: we refer to the collection of these subtrees as the *spurious tree*.

As a test case, we shall consider $\epsilon = 0.01$. The relevant data are displayed in Table 1, where the first column gives the target sequence length, the next two columns (success rate as a percentage and number of accesses) pertain to single-threshold reconstruction, and the remaining columns pertain to the multithreshold algorithm in the following order: success rate, number of characters reconstructed and number of accesses under $\theta = 1, 2, 3$, and the value of back-up J^* :

Using this table, we display in Fig. 4 the fractions of sequence characters reconstructed using thresholds 2 and 3 (note that most of characters are obtained using threshold 1, although all characters are processed under $\theta = 1$ before a transition to $\theta = 2$ is made; similarly, all characters reconstructed under $\theta = 3$ had been previously processed under $\theta = 2$). To explain these values, we note that there are on average ϵm FN (false-negatives) events causing threshold transitions: each of these events yields approximately $J^*(m, \epsilon)$ symbols under $\theta > 1$.

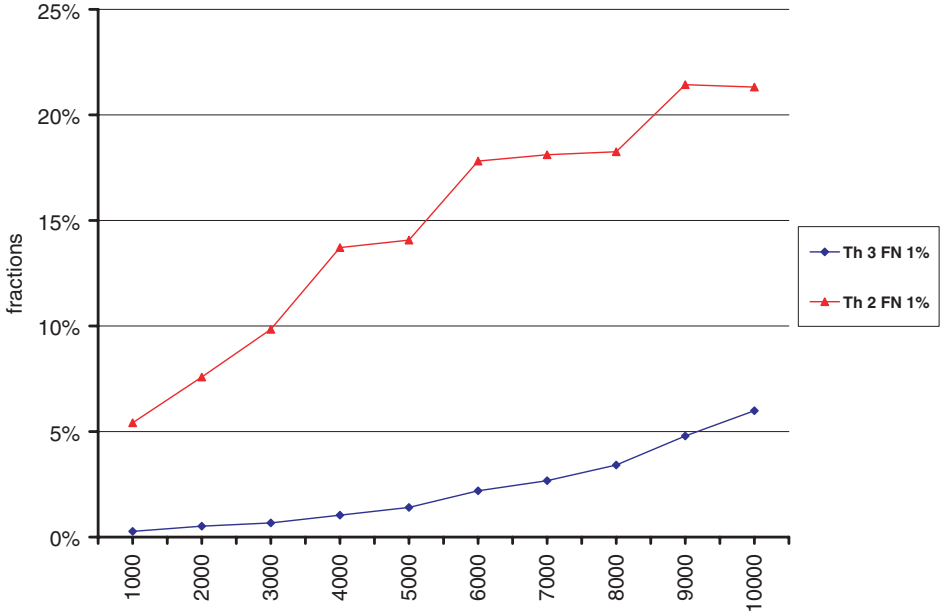


Fig. 4. Fractions of characters reconstructed under $\theta = 2, 3$ for $\epsilon = 0.01$.

Although parameter J^* is empirically determined, from Table 1 we approximate J^* as $J^* \approx c \cdot m$, with $c = 20/9000 \approx 0.0022$. Thus, there are $m\epsilon$ events, each yielding approximately cm characters reconstructed under $\theta > 1$, for a total of $cm^2\epsilon$ characters. A fraction of these are obtained under $\theta = 3$; this happens when an FN is followed by another FN at distance $\leq J^*$, i.e., with conditional probability $J^*\epsilon$, and we conclude that $m(J^*\epsilon)J^* = c^2\epsilon^2m^3$ estimates this number of characters. It can be readily verified that these analytical estimates adequately agree with the experimental results (columns ch-2 and ch-3 of the above table).

We must now address the issue of the computation time by estimating the (average) size of spurious trees. Here we make the strong simplifying assumption that spurious-tree-level extensions are independent events. This assumption ignores the (mild) conditioning among spectrum probes and the detailed structure of the probing pattern, and will result in underestimates of the sizes of large spurious trees.

A tree edge is labelled 1 if it is due to a fooling probe (of probability α) and 0 (of probability $\beta = 1 - \alpha$) otherwise. We denote S_j the (average) size of a tree each path of which has at most j 0s (j is referred to as the *weight* of the tree). Note that under threshold θ we must consider spurious trees of weight $\theta - 1$.

From a node of the correct path we make three queries and observe their responses; we let D_j denote the total number of queries for threshold j . We then

have

$$D_j = 4 + 3\alpha S_j + 3(1 - \alpha)S_{j-1},$$

where the first term is due to the four spectrum interrogations, the second and the third to extension of subtrees of weight j and $j - 1$. We now estimate S_j . We make four queries, so that

$$S_j = 4 + 4\alpha S_j + 4(1 - \alpha)S_{j-1}, \quad \text{or} \quad S_j(1 - 4\alpha) - 4(1 - \alpha)S_{j-1} - 4 = 0,$$

with the initial condition $S_0 = 4 + 4\alpha S_0$ or $S_0 = 4/(1 - 4\alpha)$. The general solution is

$$S_j = \frac{4}{3} \left[\left(\frac{4 - 4\alpha}{1 - 4\alpha} \right)^{j+1} - 1 \right],$$

so that, with straightforward algebra,

$$D_j = \left(\frac{4 - 4\alpha}{1 - 4\alpha} \right)^{j+1}.$$

In this simplified framework, the access costs with thresholds 1, 2, and 3 are estimated as $mS_0/m = S_0, S_1,$ and $S_2,$ respectively (these values are plotted in Fig. 5).

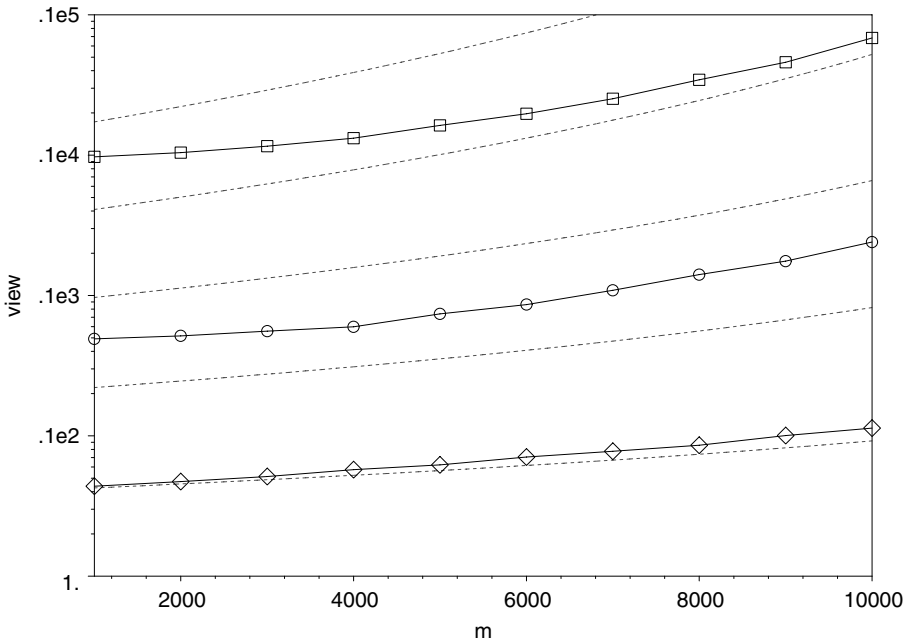


Fig. 5. Experimental per-character costs (solid curves, $\theta = 1, 2, 3$) and plots of S_0, \dots, S_4 (dotted curves, bottom to top).

While there is excellent agreement for $\theta = 1$, the discrepancy for $\theta = 2$ is attributable to the expected underestimate of S_1 . The much stronger discrepancy for $\theta = 3$ is to be attributed to the relative-scoring criterion enforced when using the largest θ . In fact, due to the action of the false negatives on the correct path, most of the reconstruction is likely to involve spurious trees of weights 3 and 4 rather than 2 (compare with the plots of S_3 and S_4 in Fig. 5). This mechanism also explains the modest speed-ups observed for noise $\epsilon = 0.02$, where the threshold-4 and -5 reconstructions presumably account for the largest fraction of the costs of both single-threshold and multithreshold algorithms.

Acknowledgment

The authors wish to thank two referees, whose suggestions greatly contributed to the quality of this paper.

Appendix A. Analysis of the Doi–Imai Approach

The discouraging performance of the Doi–Imai approach¹⁵ is fully explainable by analyzing their sequence reconstruction algorithms. We consider their modification of the reconstruction algorithm by Preparata and Upfal.⁸

Doi and Imai propose the following algorithmic modification. Here C denotes the set of probes returned as a response to the spectrum query. Each of the three cases has a probability that is easily computed accounting for the false negatives and fooling probes associated with it.

- (1) $|C| = 1$. *Extend putative sequence as in the noiseless case (concatenate symbol and proceed)*. The algorithm fails if there is a false negative for the correct response (an unrecovered false negative). Since a single fooling probe exists, the corresponding probability of failure is

$$\phi_1 = \epsilon \cdot 3\alpha(1 - \alpha)^2.$$

- (2) $|C| > 1$. *Extend in the branching mode as in the noiseless case*. The algorithm fails either if there is a false negative for the correct response (probability $\epsilon(3\alpha^2(1 - \alpha) + \alpha^3)$) or, in the absence of the false negative, if the reconstruction fails in the usual extension process. The latter event happens if a correct-path false-negative occurs in concomitance with an in-phase fooling probe for the spurious path. Thus,

$$\begin{aligned} \phi_2 &= \epsilon(3\alpha^2(1 - \alpha) + \alpha^3) + (1 - \epsilon)3\alpha \left(\epsilon \frac{1 - ((1 - \epsilon)\alpha)^{k-1}}{1 - (1 - \epsilon)\alpha} + ((1 - \epsilon)\alpha)^{k-1} \right) \\ &\approx \epsilon(3\alpha^2(1 - \alpha) + \alpha^3) + (1 - \epsilon)3\alpha\epsilon/(1 - \alpha). \end{aligned}$$

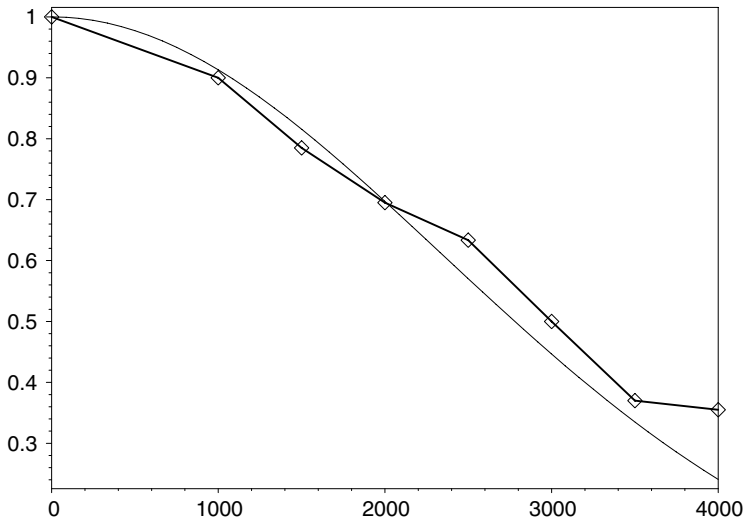


Fig. 6. Analytical diagram of the probability of successful reconstruction for (4,4)-gapped probes and $\epsilon = 0.001$, along with the corresponding experimental result of Ref. 15.

- (3) $|C| = 0$. Proceed as if $|C| = 4$ and extend in the branching mode as in the noiseless case. This is the only case (of probability $\epsilon(1 - \alpha)^3$) where false-negative recovery may occur. There is failure only when path extension fails, so that, arguing as above

$$\phi_3 \approx \epsilon(1 - \alpha)^3 \alpha \epsilon / (1 - \alpha).$$

Since the described failure events are disjoint, an upper bound to the total probability of correct reconstruction is expressed as

$$e^{-m(\phi_1 + \phi_2 + \phi_3)}.$$

This function of m is plotted in Fig. 6 for (4,4)-gapped probes and $\epsilon = 0.001$, along with the corresponding experimental graph reported by Doi and Imai,¹⁵ thereby illustrating the validity of our analysis.

References

1. Smith GC, Bains W, A novel method for DNA sequence determination, *J Theor Biol* **135**:303–307, 1988.
2. Lysov YP, Florentiev VL, Khorlin AA, Khrapko KR, Shih VV, Mirzabekov AD, Sequencing by hybridization via oligonucleotides. A novel method, *Dokl Acad Sci USSR* **303**:1508–1511, 1988.
3. Drmanac R, Labat I, Bruckner I, Crkvenjakov R, Sequencing of megabase plus DNA by hybridization, *Genomics* **4**:114–128, 1989.
4. Pevzner PA, l-tuple DNA sequencing: Computer analysis, *J Biomol Struct Dynamics* **7**(1):63–73, 1989.

5. Pevzner PA, Lysov YP, Khrapko KR, Belyavsky AV, Florentiev VL, Mirzabekov AD, Improved chips for sequencing by hybridization, *J Biomol Struct Dynamics* **9**(2):399–410, 1991.
6. Waterman MS, *Introduction to Computational Biology*, Chapman and Hall, 1995.
7. Preparata FP, Frieze AM, Upfal E, On the power of universal bases in sequencing by hybridization, *Third Annual International Conference on Computational Molecular Biology*, April 11–14, Lyon, France, pp. 295–301, 1999.
8. Preparata FP, Upfal E, Sequencing-by-hybridization at the information-theory bound: An optimal algorithm, *J Comput Biol* **7**(3/4):621–630, 2000.
9. Drmanac R, Labat I, Crkvenjakov R, An algorithm for the DNA sequence generation from k -tuple word contents of the minimal number of random fragments, *J Biomol Struct Dynamics* **8**:1085–1102, 1991.
10. Lipshutz RJ, Likelihood DNA sequencing by hybridization, *J Biomol Struct Dynamics* **11**:637–653, 1993.
11. Pevzner PA, Lipshutz RJ, Towards DNA-sequencing by hybridization, *19th Symp on Mathem Found of Comp Sci LNCS-841*, pp. 143–258, 1994.
12. Blazewicz J, Kaczmarek J, Kasprzak K, Markezicz WT, Weglarz J, DNA sequencing with positive and negative errors, *CABIOS* **13**:151–158, 1997.
13. SantaLucia JJ, A unified view of polymer, dumbbells, and oligonucleotide DNA nearest-neighbor thermodynamics, *Proc Natl Acad Sci USA* **95**:1460–1465, 1998.
14. Le Novere N, MELTING, computing the melting temperature of nucleic acid duplex, *Bioinformatics* **17**(12):1226–1227, 2001.
15. Doi K, Imai H, Sequencing by hybridization in the presence of hybridization errors, *Genome Inf* **11**:53–62, 2000.
16. Leong H-W, Preparata FP, Sung W-K, Willy H, On the control of hybridization noise in DNA sequencing-by-hybridization, *WABI 2002, Rome, Italy LNCS 2452*, 392–403, 2002.
17. Halperin E, Halperin S, Hartman T, Shamir R, Handling long target and errors in sequencing by hybridization, *J RECOMB 2002*, Washington, April 2002, 176–185.
18. Heath SA and Preparata FP, Enhanced sequence reconstruction with DNA microarray application. *COCOON 2001*, 64–74, 2001.



Hon-Wai Leong is an Associate Professor in the Department of Computer Science, National University of Singapore. He received the B.Sc. (Hon) degree from the University of Malaya and the Ph.D. degree from the University of Illinois at Urbana-Champaign. His research interests encompassed the design of practical algorithms for optimization problems from many application areas including computer-aided design of integrated circuits, transportation logistics, networking and multimedia systems, and, most recently, computational biology. He is a member of ACM and IEEE and a senior member of the Singapore Computer Society.



Franco P. Preparata is the An Wang Professor of Computer Science at Brown University. Previously he was a Professor of Electrical Engineering and Computer Science at the University of Illinois at Urbana-Champaign. Over the years he has carried out research in a number of algorithmic domains, with special emphasis on computational geometry and parallel computation. Currently, computational biology is a main focus of his research interests. He is a Fellow of the IEEE and of the ACM.



Wing-Kin Sung received both the B.Sc. and the Ph.D. degree in the Department of Computer Science from the University of Hong Kong. Then, he worked as a Post-Doctoral Fellow in Yale University and worked as a Senior Technology Officer in the University of Hong Kong. Currently, he is an assistant professor in the Department of Computer Science, National University of Singapore, Singapore. He also works as a Senior Group Leader in the Department of Information and Mathematical Science, Genome Institute of Singapore, Singapore.



Hugo Willy received his B.Comp. degree from National University of Singapore (NUS), in 2003. He is currently pursuing his study as a Ph.D. candidate under the joint research programme between the School Of Computing, National University of Singapore and the Institute of Infocomm Research, Singapore. His research interest includes DNA Sequencing-by-Hybridization, RNA structure prediction and Protein-Protein Interaction.

Copyright of Journal of Bioinformatics & Computational Biology is the property of World Scientific Publishing Company and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.