# Sequencing-by-Hybridization Revisited: The Analog-Spectrum Proposal

Franco P. Preparata

**Abstract**—All published approaches to DNA sequencing by hybridization (SBH) consist of the biochemical acquisition of the *spectrum* of a target sequence (the set of its subsequences conforming to a given probing pattern) followed by the algorithmic reconstruction of the sequence from its spectrum. In the "standard" or "uniform" approach, the probing pattern is a string of length $L$ and the length of reliably reconstructible sequences is known to be $m_{len} = O(2^L)$. For a fixed microarray area, higher sequencing performance can be achieved by inserting nonprobing gaps ("wild-cards") in the probing pattern. The reconstruction, however, must cope with the emergence of fooling probes due to the gaps and algorithmic failure occurs when the spectrum becomes too densely populated, although we can achieve $m_{comp} = O(4^L)$. Despite the combinatorial success of gapped probing, all current approaches are based on a biochemically unrealistic spectrum-acquisition model (*digital-spectrum*). The reality of hybridization is much more complex. Departing from the conventional model, in this paper, we propose an alternative, called the *analog-spectrum* model, which more closely reflects the biochemical process. This novel modeling reestablishes probe length as the performance-governing factor, adopting "semidegenerate bases" as suitable emulators of currently inadequate universal bases. One important conclusion is that accurate biochemical measurements are pivotal to the success of SBH. The theoretical proposal presented in this paper should be a convincing stimulus for the needed biotechnological work.

**Index Terms**—DNA sequencing, sequencing-by-hybridization, microarrays, gapped probes, thermodynamics of hybridization, analog spectrum, semidegenerate bases.

## 1 Introduction

As is well-known, more than a decade ago, several research groups [10], [12], [6], [14], [20], [5] proposed a radically new alternative to established wet-lab techniques for DNA sequencing. This novel approach (sequencing by hybridization, or SBH) is based on the property of a DNA sequence to hybridize to its Watson/Crick (WC) complement and opens up the possibility of the simultaneous acquisition of all relevant data in a single laboratory experiment.

The basic idea is the deployment of a set (called a *library*) of oligonucleotides on some solid support, called a "microarray" or "chip." The active area of the chip is structured as a matrix, in each region of which (called a *feature*) a very large number of copies of a specific oligonucleotide are implanted.

The chip is immersed under controlled conditions within a solution of a suitably labeled target DNA sequence and a copy of the target DNA will bind (hybridize) to an oligonucleotide if the oligonucleotide is complementary, in the Watson-Crick sense, to one of its subsequences (labeling of the target allows visualization of this event). In DNA sequencing (denoted *de novo* sequencing), the microarray library is complete, i.e., it contains oligonucleotides for all possible choices of the bases.

All published approaches to SBH agree with the following model. The process consists of two cascaded 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 (such a set is called the sequence *spectrum*). The second step, combinatorial in nature, is the algorithmic reconstruction of the sequence from its spectrum. (Biochemical) spectrum acquisition is the extraction of information from the target sequence, (combinatorial) reconstruction uses this information to reproduce the sequence. Sequence reconstruction is effected by symbol-by-symbol extension of a putative sequence from one end to the other. It is customary to measure performance with reference to the ensemble of i.i.d. Bernoulli sequences of given length.

The key component of an SBH-scheme is its *probing pattern*, a $\text{length-}L$ binary string of the form $1(0 \vee 1)^{L-2}1$: A probe (of the unknown target) is a subsequence of the latter obtained by positioning the pattern along the sequence and extracting the symbols sampled by the 1s of the pattern.

The spectrum is therefore the set of *all* such probes. For each such subsequence, the spectrum provides a single bit (1 for presence, 0 for absence). We call such spectra *digital*. This mode of operation applies to all SBH schemes heretofore proposed and analyzed, of which we shall next review two important examples.

In the "standard" or "uniform" approach, the probing pattern is a string. It has been shown [12] that any sequence compatible with the spectrum corresponds to an Eulerian path in the (Eulerian) digraph whose vertices and edges are, respectively, defined by $(L-1)\text{-tuples}$ and $L\text{-tuples}$ of the target. Pevzner also established necessary and sufficient conditions for failure, the most stringent of which gives the

• *The author is with the Computer Science Department, Brown University, 115 Waterman Street, Providence, RI 02912-1910. E-mail: franco@cs.brown.edu.*

*Manuscript received 21 Apr. 2004; revised 18 June 2004; accepted 23 June 2004.*
*For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0049-0404.*

bound $m^* \approx 1.22 \times \epsilon^{1/4} 2^L$ to the length of sequences reconstructible with confidence $1 - \epsilon$. This bound applies to any SBH scheme with probe length $L$ since sequence reconstruction is identified with the traversal of the Eulerian path described above: This will be evidenced by the notation $m_{len}$. Finding

$$m_{len} = c2^L \tag{1}$$

for some $0 < c < 1$ is in sharp contrast with the so-called "information-theory bound" $O(4^L)$ [7].

A natural approach to achieve higher sequencing performance for a given microarray area is to increase probe length without increasing the number of probing nucleotides. Such a lengthening can only be achieved by inserting nonprobing gaps between probing positions and, in the digital-spectrum model, gapped probing schemes have been shown to achieve performances comparable to the information-theory bound [15], [17]. Gap positions act as "wild-cards" to be ideally realized with "universal bases" (bases with perfect hybridization nonspecificity).

Whatever their physical realization, wild-card gaps give rise to a novel phenomenon: the emergence of fooling probes.

**Definition 1.** *A fooling probe is due to a subsequence conforming to the probing pattern, agreeing with the correct extending subsequence in all but its rightmost symbol, and occurring elsewhere in the target sequence.*

(Note that, in the uniform-array method, multiple query responses are always correct extensions, whereas, in gapped-arrays approaches, they almost never are.)

In sequence reconstruction, all competing alternatives issuing from an ambiguous branching are extended up to some maximum depth $H$ [17]. Success is highly likely because the correct path is deterministically extended, while the extension of the spurious paths rests on the (probabilistic) presence of fooling probes in the spectrum. On the other hand, fooling probes are a serious source of algorithmic inefficiency (as demanded by the construction of spurious paths). When the probability of occurrence of a specific fooling probe attains a value equal to the a priori probability (1/4) of its extension symbol, then we face a runaway branching process with a positive probability that spurious path extension will never terminate [8], [18]. Denoting $F$ the size of the fooling probe set, this occurs when $m\frac{F}{4^L} \approx \frac{1}{4}$ so that we define the bound:

$$m_{comp} = \frac{4^{L-1}}{F} \tag{2}$$

to underscore the fact that performance is limited by computational infeasibility.

We now observe:

1. The performance of uniform-array SBH is governed by $m_{len}$ since, in this case, $F = 0$, i.e., $m_{comp} = \infty$.
2. The performance of gapped-array SBH is governed by $m_{comp}$. Indeed, in the best performing schemes [17], we have $L \approx (k+1)^2/2$ and $F = 4^{L-k}$ ( since the probing pattern has $L - k$ wild-card positions). It follows that

$$m_{comp} = \frac{4^{L-1}}{F} = 4^{k-1} \ll c2^{\frac{k^2+1}{2}} = m_{len}$$

for any realistic value of $k$.

However, the above conclusions are predicated on the validity of the digital-spectrum model. In the next sections, we challenge the adequacy of such a model and propose an alternative, called the *analog-spectrum* model, which more closely reflects the process of hybridization. A consequence of this model and of the adoption of "semidegenerate bases" as suitable substitutes of not yet available universal bases reestablishes probe length as the performance-governing factor.

We conclude this introduction with two significant observations:

1. Accurate biochemical measurements are pivotal to the combinatorial success of SBH. This observation is emblematic of the interaction between the two disciplines, where successful combinatorial/algorithmic research, originally stimulated by biochemistry, shifts the focus to biochemical research for the refinement of the model.
2. The reported analysis is not intended as an experimental validation of SBH; rather, it should be viewed as a persuasive suggestion for the experimental work needed to finally translate SBH research into practical realizations.

## 2 The Inadequacy of the Digital-Spectrum Model

Validation of the standard digital-spectrum model must rest on the biochemistry of spectrum acquisition. As assay temperature rises, copies of an annealed duplex progressively separate (denature) to completion: "conventional" state-transition (annealed/denatured) occurs at the temperature (called melting temperature $T_m$) at which 50 percent of the duplexes have become separated. $T_m$ is an accurate measure of the "binding strength" (internal energy $\Delta G$) of the duplex. The internal energy is a function of the nucleotide sequence, additively and with a property of locality, sufficiently well-modeled by "dimer" parameters [19]. For example:

$$\Delta G(\mathrm{CGTTTGA}) = \mathrm{CG} + \mathrm{GT} + \mathrm{TT} + \mathrm{TT} + \mathrm{TG} + \mathrm{GA} + \Delta \mathrm{G_{init}},$$

where $\Delta G_{init}$, omitted hereafter for brevity, is a special parameter accounting for the end-pairs.

As an underpinning to further considerations, in Table 1, we report the $\Delta G$ dimer match parameters for $T = 65^\circ\mathrm{C}$ [1] in (noncoventional) units of $-10^2$ Kcal/mole.

As this table shows, dimer values have a considerable spread, with first two moments $\mu^{(1)} = 78.9$ and $\sigma^{(1)} = 40.6$.

Energy parameters have also been determined for mismatches [1], [2], [3], [4]. Specifically, such parameters are available for single mismatches in the form of dimers corresponding to configurations of the type

$$\begin{matrix} N_1 & N_2 \\ \overline{N}_1 & N_3, \end{matrix}$$

where $N_1$, $N_2$, and $N_3$ are arbitrary nucleotides and $\overline{N}_1$ is the WC-complement of $N_1$ (mismatches correspond to

1. $T = 65^\circ\mathrm{C}$ is a realistic average melting temperature of the hybridization experiments considered here.

TABLE 1
$\Delta G$ Match Dimer Parameters for $T = 65°C$

| | A | C | G | T |
|---|---|---|---|---|
| A | 39 | 82 | 70 | 30 |
| C | 82 | 127 | 140 | 70 |
| G | 69 | 155 | 127 | 82 |
| T | 0 | 69 | 82 | 39 |

$N_3 \neq \overline{N}_2$). The destabilization $\delta(N_1, N_2, N_3, N_4)$ of a mismatch configuration

$$\begin{matrix} N_1 & N_2 \\ & N_3 \quad N_4 \end{matrix}$$

can be easily modeled in terms of available dimer match and mismatch parameters. This modeling enables us to compute (an estimate of) the binding energy of any duplex having at most two nonadjacent mismatches, which is adequate since the case of more than two mismatches is unreliably estimated and uninteresting in the context of SBH. Destabilization $\delta(N_1, N_2, N_3, N_4)$ can be treated as a random variable over the set of mismatches. From the above table, we obtain that the average destabilization value is $\mu_{mis} \approx -245$ and its standard deviation is $\sigma_{mis} \approx 96$.

This sketchy review of the hybridization process is sufficient to seriously invalidate the standard digital-spectrum model, as explained below.

First of all, state-transition temperature has a substantial variance, which rules out single-temperature spectrum-acquisition experiments. In other words, digital-spectrum, if practicable at all, must be *feature-specific*. Although feature-specific detection appears adequate for uniform-array schemes, the conclusion is quite different for gapped-array schemes.

The main issue is that the physical realization of wild-card positions must be realistically modeled. We can envision three alternatives:

1. **Ideal universal bases**. These bases, with perfect hybridization nonspecificity, have been postulated in the combinatorial analysis of gapped schemes, but are not likely to exist. Thus, this alternative is not viable.
2. **Practical universal bases**. These bases are chemical compounds that exhibit imperfect nonspecificity (about two dozen such compounds have been synthesized to-date). To evaluate whether a practical universal base is viable for deployment (referred to here as an *effective* universal base), we model the binding energy of a pair (practical universal base)-(natural base) as a random variable with standard deviation $\sigma_U$ and treat individual pairs as independent. It follows that, if the probe pattern contains $h$ universal bases, the free energies of the set of sequences that may anneal at microarray feature $f$ are distributed around some nominal value with standard deviation $\sqrt{h}\sigma_U$. Obviously, the hybridization-detection energy-threshold $\theta(f)$ for feature $f$ must be set at the smallest value over all members of this set. Note that the condition $\sigma_U = 0$ characterizes an ideal universal base.

   No shortcoming arises if $\sigma_U$ is very small. For larger values of $\sigma_U$, however, (correct) matches may compete with (incorrect) mismatches. This happens because "weak" matches (i.e., matching sequences with an A-T majority in the wild-card positions) may have smaller $\Delta G$ than "strong" mismatches (with a C-G majority). If we model the universal-base fluctuation as a normal random variable with 0 mean and standard deviation $\sigma_U$ and denote $M_0$ the minimum mismatch destabilization for a probe with $h$ universal bases, we obtain the condition

$$\sigma_U \leq \frac{M_0}{6\sqrt{h}}, \tag{3}$$

   which could be taken as a criterion for a synthetic base to be considered "effectively" universal.

   Unfortunately, however, no such acceptable universal base has yet been found, although its discovery and synthesis appear as reasonable biochemical objectives.
3. **Universal-base surrogates**. Universal-base surrogates are mixtures of natural bases intended to emulate the behavior of universal bases. As we shall see, a common shortcoming of such approaches is the weakening of the luminescence signal.

   *Degenerate bases* have been proposed as a solution. A degenerate base is a uniform mixture of the four nucleotides so that the probe at a given microarray feature will consist of a uniform mixture of copies of $4^h$ distinct oligonucleotides, which are identical in their specified positions and exhibit, instead, all possible selections in the wild-card positions. Deploying $h$ such surrogates, the annealing oligonucleotide has a concentration that is only $1/4^h$ of that of the analogous ideal-universal-base oligonucleotide; this fact reduces the hybridization signal and poses a severe limit on the value of $h$, which is not likely to exceed the value 5. Unfortunately, due to the wide spread of internal energies, the deployment of $h$ degenerate bases may cause energy variations with respect to the feature mean value that are comparable to the destabilization produced by one or even two mismatches. The net effect of this interference is a drastic increase of the number of fooling probes, thus seriously affecting the performance of sequence reconstruction. This phenomenon can be carefully analyzed by modeling both the mismatch values and the energies of the $4^h$ assignments of the degenerate bases as normal random variables (see the Appendix for documentation). Suffice it to display here the corresponding histograms (Fig. 1).

   The fact that single mismatches are the source of an overwhelming volume of fooling probes (and so are double mismatches) illustrates the inadequacy of degenerate bases.

   As a practical emulator of a universal base, we have recently proposed the notion of *semidegenerate base* [16]. A semidegenerate base is a uniform mixture of the two nucleotides of similar "strength"

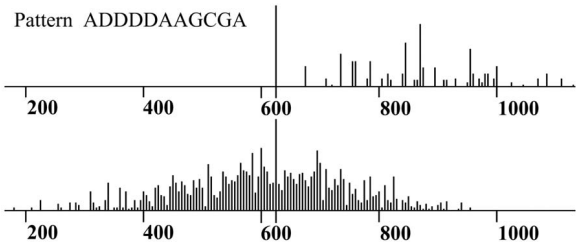


Fig. 1. Histograms of binding energies for a pattern with four degenerate (D) bases (matches above, single mismatches below).

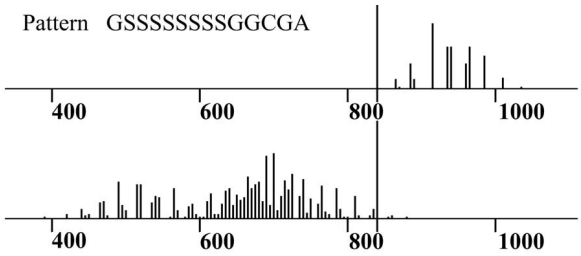


Fig. 2. Distributions of binding energies for a pattern with eight semidegenerate (S) bases (matches above, single mismatches below).

(according to the classification of A, T as "weak" ($W$) bases and of C, G as "strong" ($S$) bases). From Table 1, we obtain standard deviation $\sigma^{(2)} = 10.8$. In this approach, for $h$ wild-card positions, each wild-card position of a given feature will have a unique assignment in the set $\{W, S\}$ and the feature will contain a uniform mixture of copies of the $2^h$ oligonucleotides corresponding to all base selections consistent with the given strength selection. Clearly, each gapped probe (ideally, a single microarray feature) is now collectively represented by $2^h$ microarray features (each associated with a distinct pattern of base-strengths).

Again, on the basis of the data of Table 1, we may now calculate the standard deviation of a string of eight semidegenerate bases, which is only $\sqrt{(8-1)\times 99} = 26.3$, compared with the corresponding value of $\sqrt{(4-1)\times 3147} = 97.16$ for a string of four degenerate bases (see the Appendix). Semidegenerate bases can also be analyzed for interference with mismatches. The histograms of the energies distributions, shown in Fig. 2, clearly illustrate the advantage over degenerate bases.

As a summary of the preceding discussion, for fixed cost and signal reduction (a $4^k$-feature microarray and a $4^h$-fold signal reduction), probe length $k + h$ is achieved either with $k$ definite positions and $h$ degenerate bases or with $k - h$ definite positions and $2h$ semidegenerate bases. This observation seems to suggest that no obvious advantage accrues from the adoption of semidegenerate bases over degenerate bases. However, semidegenerate bases exhibit negligible match/mismatch interference, which makes them viable universal-base surrogates. Therefore, in what follows, the phrase "$h$ degenerate bases" is to be understood as "$h$ equivalent degenerate bases, emulated by $2h$ semidegenerate bases while reducing by $h$ the number of definite bases (to maintain unaltered the microarray size)."

## 3 The Analog-Spectrum Approach

In this section, we let $k$ and $h$, respectively, denote the numbers of definite bases and equivalent degenerate bases in the probing pattern, i.e., $L = k + h$.

We noted in Section 2 that gapped probing patterns, while adequate to overcome the inherent inefficiency of uniform patterns, engender fooling probes, which prove to be the performance-limiting factor.

Therefore, the fact that, in current technology, only small values of $h$ appear feasible may seem to lead to a pessimistic prognosis for SBH. Upon closer examination, however, we realize that the bound $m_{comp}$ (which, by and large, governs gapped-scheme SBH) is essentially due to the stated digital-spectrum acquisition policy. Indeed, in order to avoid the occurrence of false-negatives which are very detrimental to the reconstruction process, for each microarray feature, we are forced to set the detection threshold at the lowest value compatible with the probes that correctly hybridize at the feature. Such a policy, which is at the root of the fooling probe phenomenon, is due to the entirely arbitrary decoupling of biochemistry and combinatorics, which is the basis of all approaches heretofore proposed. As we shall see, such decoupling sacrifices valuable information available to the reconstruction algorithm.

We now propose to abandon this standard digital-spectrum model and adopt instead a policy where the biochemical and combinatorial steps interact or, equivalently, the biochemical step provides much richer information to the reconstruction process and lets the latter use it contextually. We call such a policy **analog-spectrum acquisition**, defined as follows:

- The sequence spectrum consists, for each feature undergoing a hybridization transition, of the measurement of the quantity (typically, temperature) at which its state transition conventionally occurs. Formally, if we denote $T(f)$ the measurement at feature $f$:

$$\text{analog spectrum} = \{T(f) | f \in \text{probe library}\}.$$

We must underscore that, although there is no current instrumentation for achieving the desired measurements, its realization (a temperature ramp and a high-precision scanning of the microarray) is well within the state-of-the-art.

A crucial feature of the analog spectrum approach is that, since reconstruction occurs by extension of a putative sequence,

*for any reconstruction path being extended, the algorithm has full knowledge of the* **entire** *sequence suffix, not just the $k-1$ definite positions used in constructing the spectrum query in the digital-spectrum approach.*

This additional information enables the algorithm to specify with some accuracy the temperatures at which each of the four possible state-transitions (corresponding to extension $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$) is expected to occur and therefore classify the responses in a much finer way. In other words, rather than letting the instrument make *hard* decisions as to the presence/absence of a response on a worst-case basis, we require the instrument to provide a finer measurement

and let the algorithm make full use of contextual information to produce the classification.

For a microarray feature $f$, we let $\theta(f)$ denote the nominal transition value generated by the algorithm, either computed or experimentally predetermined; the analogous "measurement" is denoted $T(f)$. In addition, we let $\eta$ denote a tolerance value intended to account for unavoidable experimental errors. The spectrum query will now have three responses, rather than two as in the digital spectrum:

1. $T(f) < \theta(f) - \eta$.
   This event is interpreted as a **mismatch**.
2. $\theta(f) - \eta \leq T(f) \leq \theta(f) + \eta$.
   This event is interpreted as a **match**.
3. $\theta(f) + \eta < T(f)$.
   This event is interpreted as a **fooling probe**.

This 3-way classification, to be contrasted with the 2-way $\{\text{mismatch}, \text{match/fooling} - \text{probe}\}$ classification of the digital-spectrum approach, affords significant algorithmic flexibility and higher performance. In fact, in the most simplistic implementation, path extension could be performed only for matches. A more subtle implementation could quantitatively score both mismatches and fooling probes and terminate spurious paths on the basis of their score. Notice, however, the asymmetry between Cases 1 and 3: Due to the fact that a Case-3 response may actually conceal a correct match (and not so a Case-1 response), a mismatch could be ignored, whereas a fooling probe must be handled as such.

With reference to performance, we show below that the main cause of the improvements is the drastic reduction of the fooling probe set. In order to be able to obtain a quantitative estimate of this reduction, we adopt some simplifications that—although not rigorously justifiable—do not seem to alter the nature of the process. These simplifications are essentially: 1) normal distribution for random variables of known mean and variance and 2) statistical independence when needed. Let $N(x; \mu, \sigma)$ denote the density function of a Gaussian random variable $x$ with mean $\mu$ and variance $\sigma^2$.

Let $G$ be the average of the energy of the probes correctly hybridizing at feature $f$ and let $\sigma' = \sqrt{h}\sigma_{eff}^{(2)}$, where $\sigma_{eff}^{(2)} = 9.94$ is the effective standard deviation of a single semidegenerate base (see the Appendix). The density function of their energies is $N(t; G, \sigma')$. If $T$ denotes the measurement, then, observing that $\eta$ may be assumed to be small with respect to $\sigma'$, the number of (match) fooling probe is estimated as

$$4^h \int_{\theta-\eta}^{\theta+\eta} N(t; G, \sigma')dt \approx 4^h \cdot 2\eta \cdot N(\theta; G, \sigma') \leq 4^h \cdot \frac{2\eta}{\sigma'\sqrt{2\pi}}.$$

The corresponding analysis carried out for mismatches shows that their contribution to the fooling probe pool is negligible. Thus, the size of the fooling probe pool is estimated as

$$F = \frac{1}{\sqrt{2\pi}} \left(\frac{2\eta}{\sigma'}\right) \cdot 4^h.$$

It is apparent that the crucial role is played here by the parameter $2\eta/\sigma'$, which, in some sense, measures how fine the quantization of the analog measurement is, i.e., how rich the information content of the analog spectrum is.

Specifically, referring back to (2), in our case, we obtain $m_{comp} = 4^{L-1}/F = 4^{k-1} \cdot \frac{\sqrt{2\pi}\sigma'}{2\eta}$; correspondingly, referring to (1), we have (for some $c \approx 0.5$)) $m_{len} = c2^L = c2^{k+h}$ so that the target condition $m_{len} \leq m_{comp}$ becomes

$$\frac{2\eta}{\sigma'} \leq \frac{2^{k-h}}{4c}.$$

In current technology, it is reasonable to assume $k = 8, h = 5$, and $c = 0.6$. In addition $\sigma' = \sqrt{2h}\sigma_{eff}^{(2)} = \sqrt{2 \times 5} \times 9.94 = 31.43$. This results in the inequality $\eta \leq 53$, which is apparently well within current experimental capabilities. This finding leads to the following very significant conclusion:

- The adoption of analog-spectrum SBH and the realization of gaps by means of semidegenerate bases (and, possibly, a few effective additional practical universal bases) reestablishes probe length as the parameter controlling the length of reliably reconstructible target sequences.

### 3.1 The Sequence Reconstruction Algorithm

Before addressing the overall sequence reconstruction algorithm, we note that the analog-spectrum policy has beneficial effects on the computational load. In fact, despite the presence of fooling probes (caused by the adoption of gaps in the pattern), each symbol of a path being extended is classified as a match or a fooling probe so that each path is correspondingly scored. Significant score differences can be used for the truncation of paths that are extremely likely to be spurious. This feature drastically lightens the computation of sequence reconstruction.

The gapped-probe sequencing algorithm [17] must be substantially modified to combine it with the traditional Eulerian-path reconstruction algorithm [13]. The resulting algorithm exhibits important features of both its predecessors. The algorithm uses a queue $Q$, designed to store disjoint segments of the sequence to be reconstructed (the symbol $\Leftarrow$ denotes enqueue/dequeue operations). Queue $Q$ is initialized with the length-$(L-1)$ prefix of the target sequence (assumed to be known for simplicity), denoted *seed*. $P$ is a predicate governing the extension. Next, we give a high-level version of the algorithm:

1. $Q \Leftarrow \text{seed}; P := 1$
2. $putative \Leftarrow Q$
3. **while** $(P = 1)$ **do**
4. $\quad R := query(putative)$
5. $\quad\quad \text{Case} |R| = 0 : P := 0$
6. $\quad\quad \text{Case} |R| = 1 : \text{extend } putative$
7. $\quad\quad \text{Case} |R| = 2 : \text{store } completed\ segment$
8. $\quad\quad\quad\quad Q \Leftarrow R(1)$
9. $\quad\quad\quad\quad Q \Leftarrow R(2)$
10. $\quad\quad\quad\quad putative \Leftarrow Q$
11. $\quad\quad \text{Case} |R| = 3, 4 :$ **return** $FAIL$
12. $stitch - Euler - path$

Barring failure (which is correctly detected in Line 11 below), the algorithm first reconstructs a collection of disjoint segments whose union covers the original target

sequence and then rearranges these segments in the correct order in Line 12 (subroutine *stitch-Euler-path*[2]).

The novelty of the procedure resides in the implementation of Line 4: $R := query(putative)$, $R$ being the set of accepted responses. Whereas, in the uniform-array algorithm, this step simply involves four table look-ups, in this case, for each of the four possible $L$-mers, the algorithm must determine the actual threshold either by computation or, more likely, by table look-up of experimental values, and compare it with the corresponding measured value. If all but one of the responses are reliably classified as mismatches, then the singleton set $R$ is returned. Otherwise, we have an ambiguous branching situation and competing paths are extended and scored with the elimination policy outlined above. The path extension process terminates either when all surviving paths have a nonempty common prefix (in which case a singleton set $R$ is returned) or when path-extension reaches some predetermined depth $H$, at which point all distinct initial symbols of the surviving paths are reported for the construction of Eulerian paths.

In Fig. 3, we give the analytical expected performance (based on standard Eulerian graph analysis) of the analog-spectrum method for a $(k = 8)$-microarray with $h = 0, \ldots, 4$. This diagram shows the superiority of the outlined approach over wet-lab methods. Moreover, we note that the addition of each effective universal base nearly doubles the performance without increasing the array size.

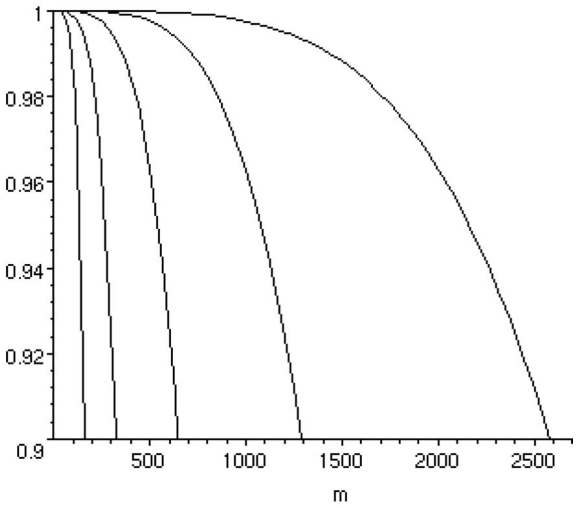


Fig. 3. Expected performance of analog-spectrum approach for $k = 8$ and $h = 0, 1, \ldots, 4$ from left to right.

## 4 Conclusion

In this paper, we have proposed the analog-spectrum model of SBH, which is based on the thermodynamics of DNA hybridization. This more realistic modeling tightly couples biochemistry and combinatorics, providing the reconstruction component with physical measurements to be used contextually. Moreover, the analog model enables the analysis of appropriate universal-base surrogates in the form of semidegenerate bases. The latter should be used to achieve significant probe length without increasing the microarray cost. We underscore again that the presented theoretical investigation is intended to motivate serious experimental work, with the potential of transforming SBH into a practical technology.

## Appendix

## Variance of Binding Energy of Oligos

Let random variable $E(j_1, \ldots, j_h)$ denote the binding energy of a string $j_1, \ldots, j_h \in D^h = \{A, C, G, T\}^h$. With our convention to consider only consecutive pairs, if $\mu$ is the average of the dimer energy, then $(h-1)\mu$ is the average of $E(j_1, \ldots, j_h)$. The corresponding variance has the expression:

$$\sigma_h^2(D) = |D|^{-(h-1)} \sum_{j_1,\ldots,j_h \in D} \big(K_2(j_1, j_2) + \ldots K_2(j_{h-1}, j_h) - (h-1)\mu\big)^2.$$

This expression is easily manipulated as follows:

$$\begin{aligned}\sigma_h^2(D) &= |D|^{-(h-1)} \sum_{j_1,\ldots,j_h \in D} \big((K_2(j_1, j_2) - \mu) + \ldots \\ &\quad + (K_2(j_{h-1}, j_h) - \mu)\big)^2 \\ &= |D|^{-(h-1)} \sum_{j_1,\ldots,j_h \in D} \left( \sum_{u=1}^{h-1} (K_2(j_u, j_{u+1}) - \mu)^2 + 2 \right. \\ &\quad \left. \cdot \sum_{r<s \in 1,\ldots,h} (K_2(j_r, j_{r+1}) - \mu)(K_2(j_s, j_{s+1}) - \mu) \right).\end{aligned}$$

Notice that the second sum above is empty when $h = 2$. We now observe that

$$|D|^{-(h-1)} \sum_{j_1,\ldots,j_h \in D} \sum_{j_u, j_{u+1}} (K_2(j_u, j_{u+1}) - \mu)^2 = \sigma^2$$

and that

$$\sum_{j_r, j_{r+1}, j_s, j_{s+1}} (K_2(j_r, j_{r+1}) - \mu)(K_2(j_s, j_{s+1}) - \mu) = 0$$

when $r + 1 \neq s$ since the index sets are disjoint and $\sum_{j_u, j_{u+1}} (K_2(j_u, j_{u+1}) - \mu) = 0$ by definition. Therefore,

$$\sigma_h^2 = (h-1)\sigma^2 + 2|D|^{-(h-1)} \sum_{j_1,\ldots,j_h \in D} \sum_{r=1}^{h-2} (K_2(j_r, j_{r+1}) - \mu)(K_2(j_{r+1}, j_{r+2}) - \mu).$$

Moreover, we have

$$\begin{aligned}&|D|^{-(h-1)} \sum_{j_1,\ldots,j_h \in D} \sum_{r=1}^{h-2} (K_2(j_r, j_{r+1}) - \mu)(K_2(j_{r+1}, j_{r+2}) - \mu) \\ &= 4^{-(h-1)} \cdot 4^{h-3}(h-2) \sum_{i,j,k} (K_2(i, j) - \mu)(K_2(j, k) - \mu).\end{aligned}$$

2. This subroutine is based on standard graph-theoretic notions. Each sequence segment $A$ is represented as a triple $\alpha A \beta$, where $\alpha$ and $\beta$ are, respectively, the $(L-1)$-symbol prefix and suffix of $A$. The *intersection graph* ([20, p. 32]) of Eulerian graph $G$ is a tree if and only if $G$ has a unique Eulerian cycle. A tree has at least two leaves, each of which is associated with a triple of the form $\alpha A \alpha$. *Stitching* is the node-reduction step that replaces $\beta B \alpha, \alpha A \alpha, \alpha C \gamma$ with $\beta BAC\gamma$.

If we now define the correlation term $\rho_2$ as

$$\rho_2 = 4^{-2} \sum_{i,j,k} (K_2(i,j) - \mu)(K_2(j,k) - \mu),$$

we obtain

$$\sigma_h^2 = (h-1)\left(\sigma^2 + 2\frac{h-2}{h-1}\rho_2\right).$$

This relation allows us to interpret the quantity

$$\sigma^2 + 2\frac{h-2}{h-1}\rho_2 \approx \sigma^2 + 2\rho_2 = \left(\sigma_{eff}^{(4)}\right)^2$$

as an "equivalent" variance, as if the dimers where independent random variables. For degenerate bases from Table 1 (see Section 2), we obtain $\sigma^2 = 1646$ and $2\rho_2 = 1501$. The result is that $\sigma_{eff}^{(4)} = \sqrt{1646 + 1501} = 56.01$.

A parallel analysis can be carried out for strings of semidegenerate bases. In this case, we obtain ([16]) $\sigma^2 = 117$, $2\rho = -18.06$, and

$$\sigma_{eff}^{(2)} = 9.94.$$

## ACKNOWLEDGMENTS

The author acknowledges with gratitude technical discussions with biochemist J. S. Oliver. The author was partially supported by the US National Science Foundation under Grant DBI-9983081

## REFERENCES

[1] H.T. Allawi and J. SantaLucia Jr., "Thermodynamics and NMR of Internal G-T Mismatches in DNA," *Biochemistry,* vol. 36, pp. 10581-10594, 1997.
[2] H.T. Allawi and J. SantaLucia Jr., "Nearest-Neighbor Thermodynamic Parameters for Internal G-A Mismatches in DNA," *Biochemistry,* vol. 37, pp. 2170-2179, 1998.
[3] H.T. Allawi and J. SantaLucia Jr., "Thermodynamics of Internal C-T Mismatches in DNA," *Nucleic Acid Research,* vol. 26, pp. 2694-2701, 1998.
[4] H.T. Allawi and J. SantaLucia Jr., "Nearest-Neighbor Thermodynamics of Internal A-C Mismatches in DNA," *Biochemistry,* vol. 37, pp. 9435-9444, 1998.
[5] W. Bains and G.C. Smith, "A Novel Method for DNA Sequence Determination," *J. Theoretical Biology,* vol. 135, pp. 303-307, 1988.
[6] R. Drmanac, I. Labat, I. Bruckner, and R. Crkvenjakov, "Sequencing of Megabase Plus DNA by Hybridization," *Genomics,* vol. 4, pp. 114-128, 1989.
[7] M.E. Dyer, A.M. Frieze, and S. Suen, "The Probability of Unique Solutions of Sequencing by Hybridization," *J. Computational Biology,* vol. 1, pp. 105-110, 1994.
[8] W. Feller, *An Introduction to Probability Theory and Its Applications.* New York: J. Wiley and Sons, 1960.
[9] D. Loakes, "The Application of Universal DNA Base Analogues," *Nucleic Acids Research,* vol. 29, pp. 2437-2447, 2001.
[10] Y.P. Lysov, V.L. Florentiev, A.A. Khorlin, K.R. Khrapko, V.V. Shih, and A.D. Mirzabekov, "Sequencing by Hybridization via Oligonucleotides: A Novel Method," *Dokl. Acad. Sci. USSR,* vol. 303, pp. 1508-1511, 1988.
[11] N. Peyret, P.A. Seneviratne, H.T. Allawi, and J. SantaLucia Jr., "Nearest-Neighbor Thermodynamics and NMR of DNA Sequences with Internal A-A, C-C, G-G, and T-T Mismatches," *Biochemistry,* vol. 38, pp. 3468-3477, 1999.
[12] P.A. Pevzner, "l-Tuple DNA Sequencing: Computer Analysis," *J. Biomoleculular Structure & Dynamics,* vol. 7, no. 1, pp. 63-73, 1989.
[13] P.A. Pevzner, *Computational Molecular Biology: An Algorithmic Approach.* MIT Press, 2000.
[14] P.A. Pevzner, Y.P. Lysov, K.R. Khrapko, A.V. Belyavsky, V.L. Florentiev, and A.D. Mirzabekov, "Improved Chips for Sequencing by Hybridization," *J. Biomoleculular Structure & Dynamics,* vol. 9, no. 2, pp. 399-410, 1991.
[15] F.P. Preparata, A.M. Frieze, and E. Upfal, "On the Power of Universal Bases in Sequencing by Hybridization," *Proc. Third Ann. Int'l Conf' Computational Molecular Biology,* pp. 295-301, Apr. 1999.
[16] F.P. Preparata and J.S. Oliver, "DNA Sequencing-by-Hybridization Using Semidegenerate Bases," *J. Computational Biology,* to appear.
[17] F.P. Preparata and E. Upfal, "Sequencing-by-Hybridization at the Information-Theory Bound: An Optimal Algorithm," *J. Computational Biology,* vol. 7, no. 3/4, pp. 621-630, 2000.
[18] F.P. Preparata, E. Upfal, and S.A. Heath, "Sequence Reconstruction from Nucleic Acid Micro-Array Data," *Analytic Techniques for DNA Sequencing,* B. Nunnally ed., M. Dekker, 2003.
[19] J.J. SantaLucia, "A Unified View of Polymer, Dumbells, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics," *Proc. Nat'l Academy of Science,* vol. 95, pp. 1460-1465, 1998.
[20] M.S. Waterman, *Introduction to Computational Biology.* Chapman and Hall, 1995.

**Franco P. Preparata** is the An Wang Professor of Computer Science at Brown University and the Kwan im Thong Visiting Professor at the National University of Singapore. 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 the ACM.