JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Number 4, 2004

Pp. 753–765

DNA Sequencing by Hybridization Using Semi-Degenerate Bases

FRANCO P. PREPARATA[1] and JOHN S. OLIVER[2]

ABSTRACT

One way to enhance the performance of hybridization microarrrays for DNA *de novo* sequencing is the use of probing patterns with gaps of unsampled positions. Ideally, such gaps could be realized by the inclusion into microarray oligos (probes) of wild-card compounds, referred to as universal bases (which bind nonspecifically to natural bases). The suggested alternative is to deploy in the gap positions degenerate bases, i.e., uniform mixtures of the four natural bases, with ensuing deterioration of the hybridization signal. In this paper, we show that such signal loss is a minor shortcoming, compared with the fact that degenerate bases cannot be treated as universal. Indeed, the substantial spread of hybridization energies at any microarray feature is such that on overwhelming number of mismatches bind more strongly than legal matches. We observed, however, that much narrower energy spreads are exhibited by pairs of bases in the same strength class (A-T and C-G). We call *semi-degenerate* a gap position realized with bases in the same energy class and show that well-known sequence reconstruction algorithms can be modified to achieve substantial improvements in sequencing effectiveness. For example, with a 4^9-feature microarray and an acceptable weakening of the hybridization signal, one may achieve lengths of about 4,000 bases (compared with < 250 of the standard uniform method). Our approach also incorporates the use of a spectrum expressed in terms of observed feature melting temperatures (analog spectrum), rather than binary decisions made directly at the biochemical level (digital spectrum). While universal bases represent the ultimate goal of sequencing by hybridization, semidegenerate natural bases are the most effective known substitute.

Key words: DNA sequencing, sequencing by hybridization, microarrays, gapped probes, thermodynamic model, semidegenerate bases.

1. INTRODUCTION

NOTWITHSTANDING THE REMARKABLE IMPROVEMENTS of the Sanger electrophoresis-based DNA sequencing method in the past 10 years, it does not appear likely that this approach can meet the projected demands of the medical and biological communities. Thus, the exploration of alternative solutions is a timely endeavor. Sequencing by hybridization (SBH) was proposed well over a decade ago (Bains

[1]Computer Science Department, Brown University, Providence, RI 02912-1910.
[2]Chemistry Department, Brown University, Providence, RI 02912-1910.

and Smith, 1988; Lysov *et al.*, 1988; Drmanac *et al.*, 1989; Pevzner, 1989; Pevzner *et al.*, 1991; Waterman, 1995) as one such alternative technique with a strong potential. As is well known, sequencing by hybridization consists of two fundamental steps, the first biochemical, the second combinatorial. The biochemical step essentially involves hybridization of an unknown target sequence to a microarray, designed to acquire information about the sequence (the so-called sequence *spectrum*, consisting of information about subsequences of the target DNA sequence); the combinatorial step is the algorithmic reconstruction of the target sequence from its spectrum.

Serious difficulties concerning both steps have prevented SBH from becoming operational. A first difficulty was the combinatorial inefficiency of the traditional "string" probing patterns, whose performance is substantially lower than the theoretical information-theory bound (Dyer *et al.*, 1994; Southern, 1996). The objective here is to *increase performance without increasing the cost* (in our case, microarray size). Substantial progress in this direction has been recently made (Hannenhalli *et al.*, 1996; Preparata *et al.*, 1999; Preparata and Upfal, 2000; Frieze and Halldorsson, 2002; Skiena and Snir, 2002); among these performance-enhancing approaches, we restrict our attention to schemes using probing patterns which include unsampled gaps between sampled positions. A formal justification of the advantages provided by sampling gaps is given by Preparata *et al.* (1999) and Preparata and Upfal (2000); suffice it to mention here that, since the length of the probing pattern governs the performance of sequencing, under the constraint of fixed microarray size (i.e., fixed probe library size), larger pattern length can be achieved by the adoption of gaps of unsampled positions. The very fact that the probing pattern includes gaps involves the availability of biochemical "wild cards," i.e., chemical compounds that exhibit hybridization nonspecificity, to be deployed for the realization of gaps. Such components are referred to as *universal bases*, and specific examples thereof have been recently discussed in the literature (Loakes, 2001), essentially as chemical curiosities. However, no decisive progress has yet been made in the realization of a perfect universal base. The presently available products do not appear to be sufficiently well behaved for *de novo* sequencing by SBH.

In this paper, we shall consider approaches to the implementation of gapped patterns that do not use universal bases. The realization of wild-card positions had been considered earlier in the literature (see, e.g., Drmanac *et al.* [1989], Pevzner *et al.* [1991]), and the proposed solution was to deploy in each "don't-care" position a complete equimolar mixture of the four nucleotides (a probe position of this kind is said to contain a *degenerate* base). Unfortunately, such an approach has limited scope, since each degenerate base in the probing pattern causes a reduction of the number of hybridizing strands by a factor of 4, thereby correspondingly weakening the intensity of the hybridization signal. Thus, only small numbers of degenerate bases can be included in the probing pattern, although, with current fluorescence-based technology, one can withstand to some extent this drawback. However, the weakening of the hybridization signal is only the lesser shortcoming of the approach. In fact, we shall show in Section 2 that degenerate bases are not universal (in that they cause a significant performance deterioration with respect to ideal universal bases). Instead, we shall describe in Section 3 a novel approach to the realization of probing gaps by the deployment of nucleotide combinations called here *semi-degenerate* bases, whose behavior is basically universal (except for the expected weakening of the hybridization signal). Our analysis is a feasibility study of an approach whose implementation we are intensely pursuing in the laboratory.

There is an additional provision that affords performance improvements (by a factor of about 2 in terms of reliably reconstructible DNA sequences). This novel provision involves the nature of the spectrum acquired by the biochemical step. It is appropriate at this point to review the most important primitive for sequence reconstruction: the spectrum query. The algorithm reconstructs the sequence symbol by symbol, by extending a putative sequence. Through the spectrum query, the algorithm uses the current suffix of the putative sequence to interrogate the spectrum as to the possible extensions; if the query has more than one response, all spurious responses are due to subsequences occurring elsewhere along the target sequences and are conveniently referred to as *fooling probes*. The fooling probe mechanism is a direct product of the model underlying all conventional approaches: hybridization of target DNA with the probes of the microarray occurs at a single temperature, which is intended to be the common temperature at which the target forms duplexes with the microarray probes (*melting* temperature). Therefore, each subsequence agreeing with a probe in its specified positions is assumed to hybridize; correspondingly, the spectrum is described by a binary vector, each component of which identifies a probe (0 for absence, 1 for presence) and appropriately denoted *digital*. In reality, each subsequence of the target sequence hybridizes with a microarray feature at a distinct temperature, due to the specific energetics of the hybridization occurring

TABLE 1. DIMER PARAMETERS FOR WATSON/CRICK-MATCHES

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

at the feature. It is technologically feasible to observe such temperature for each individual feature, by continuously varying the experiment temperature (through a temperature ramp) and by concomitantly detecting the hybridization phase-transitions feature by feature. The melting temperature of each feature could be typically measured with 3–4 significant binary digits (rather than the single bit provided in the digital spectrum approach). It is important to underscore that binding energy and melting temperature are essentially co-gradient, so that we may refer equivalently to either physical quantity.

A spectrum given in terms of melting temperatures, rather than hard binary decisions, is appropriately called *analog*. In the presence of fooling probes, the superiority of the analog-spectrum approach (over the digital-spectrum) is that decision as to the suitability of a given symbol as the current extension is deferred to an algorithmic evaluation, rather than being entrusted to an instrument operating under incomplete information.

2. DEGENERATE BASES ARE NOT UNIVERSAL

In this section, we consider a microarray with 4^k features, each hosting a probe. All probes conform to the same probing pattern, consisting of L nucleotides, of which k are specified (definite) and $h = L - k$ are wild cards. In what follows, we assume for simplicity that the wild-card positions are contiguous, although extensive simulations suggest that this causes no loss of generality.

A wild card is here assumed to be realized as a degenerate base, defined as follows.

Definition 1. *A microarray feature (probe) of length L is said to have a* degenerate base *in position j if, for any choice of the remaining nucleotides, position j contains (an equimolar mixture of)* A,C,G,T.

There are various models of hybridization, with various degrees of realistic validity. The coarsest model (upon which the standard digital-spectrum approach is based) assumes that all Watson/Crick-complementary pairs have identical hybridization strength. A less coarse model assigns distinct hybridization strengths to strong (C-G) and weak (A-T) pairs, while still assuming independence among oligonucleotide positions. An adequately realistic model replaces independence with locality, specifying that the hybridization strength of a complementary pair depends upon the nature of the adjacent pairs on either side and can be quantified in terms of consecutive pairs, called dimers (see, e.g., Breslauer *et al.* [1986]). This model, adopted by SantaLucia and others (see, e.g., SantaLucia [1998]), is characterized by a 4×4 matrix K_2 of parameters, called *nearest-neighbor* (or *dimer*) parameters, that specify how the hybridization strength (internal energy ΔG) of a base-pair is affected by its 5'-neighboring base-pair. The dimer parameter matrix is displayed as Table 1; row and column headings, respectively, specify the 5' and 3' terms of the dimer. The values, expressed as -10^{-2}Kcal/mole,[1] give the internal energy at $T = 65°$C, which is more representative of experimental condition than the normally reported 37°C temperature. Hereafter, we shall frequently use the index correspondence A$\leftrightarrow$ 1, C$\leftrightarrow$ 2, G$\leftrightarrow$ 3, and T$\leftrightarrow$ 4.

This table exhibits a substantial disagreement with the hypothesis underlying the conventional SBH model (digital-spectrum), which implies that all (dimer) parameters have identical values. However, digital-spectrum SBH is reconcilable with nonuniform dimer parameter values under the condition that, for any

[1]For intuitive convenience, we deviate from the usual convention of viewing the binding strength as a negative free energy.

given microarray feature, the binding-energy detection threshold is set at the lowest value compatible with the selections for the bases occurring in the degenerate positions. A consequence of such a policy is that there are 4^h potential fooling probes for any feature if the probing pattern contains h degenerate bases. We may, therefore, define a universal base as follows.

Definition 2. *A base is* universal of degree h *if, deploying h such bases in the probing pattern, there are no more than 4^h fooling probes for any microarray feature.*

An increase of the number of fooling probes above the (minimum) value of 4^h degrades performance. We now show that a degenerate base is not universal except for insignificant values of h.

Hereafter, we adopt the following convention to calculate the binding energy $E(j_1, \ldots, j_h)$ of an h-tuple $j_1, \ldots, j_h$. Letting $K_2(h, k)$ denote the energy of dimer (h, k), we have

$$E(j_1, \ldots, j_h) = \sum_{i=1}^{h} v_i$$

where $v_1 = \sum_{i=1}^{4} K_2(i, j_1)/4$ and $v_i = K_2(j_{i-1}, j_i)$ for $i = 2, \ldots, h$. The variable $E(j_1, \ldots, j_h)$ (binding energy) assumes 4^h values defined here $e_1 \le e_2 \le \ldots \le e_{4^h}$, so that the interval $[e_1, e_{4^h}]$ is the range of its values. It is convenient to define the function $TAIL(e)$ of the continuous variable e over the domain $[e_1, e_{4^h}]$ as follows:

$$TAIL(e) = 4^h - j \quad \text{for} \quad e \in [e_j, e_{j+1}].$$

These notions are illustrated in Fig. 1 for $h = 2$. Note that the shape of $TAIL(e)$ only depends upon the parameter h, the number of degenerate bases; the translation of this curve along the e-axis depends instead upon the binding energy of the k designated bases.

We now consider the situation of mismatch hybridization, i.e., the binding energy of a target subsequence which in the definite-base positions does not match (as Watson/Crick-complement) the bases of the feature.

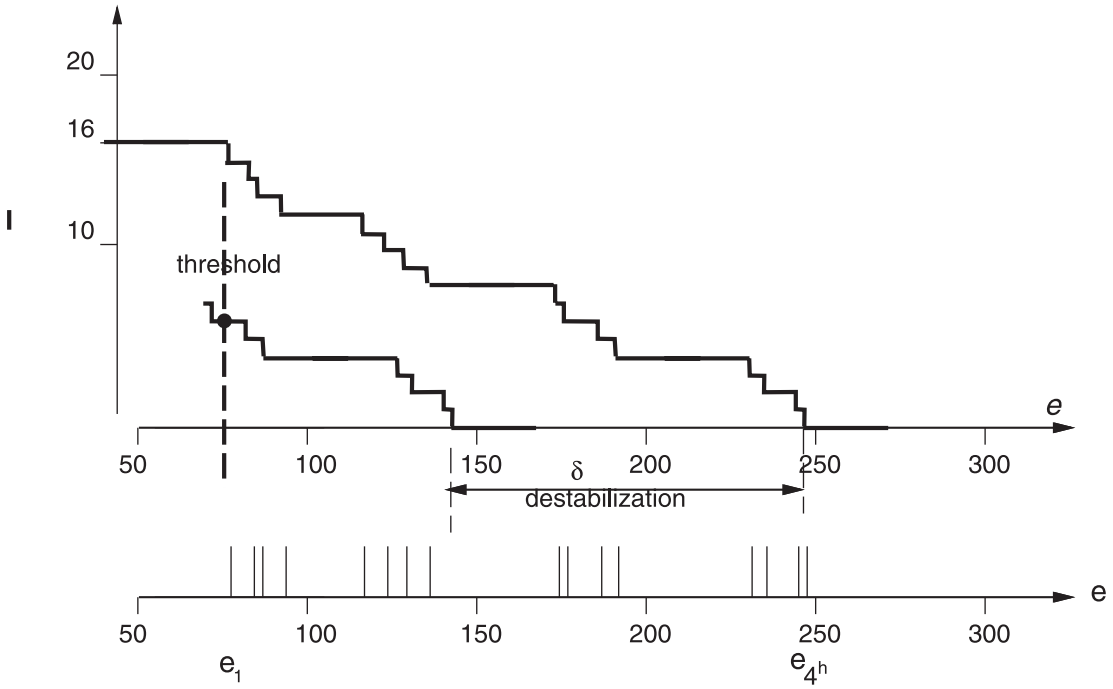

FIG. 1. The distribution of values of the binding energy and its associated TAIL function for $h = 2$. The $4^2 = 16$ abscissae, each carrying a 1-valued vertical bar, are the values of the binding energies of the possible choices of the two bases in {A,C,G,T,}; *TAIL* is the corresponding cumulative distribution.

TABLE 2. DIMER PARAMETERS FOR SINGLE MISMATCHES

	A	*C*	*G*	*T*
A	−62	−36	−7	30
C	−51	−80	−46	70
G	−41	−58	58	82
T	−33	−77	−49	39

$Z = \mathrm{A}$

	A	*C*	*G*	*T*
A	−74	−148	70	−89
C	−64	−93	140	−56
G	−40	−59	127	−63
T	−69	−55	82	−76

$Z = \mathrm{C}$

	A	*C*	*G*	*T*
A	−17	82	−11	−30
C	−7	127	−27	9
G	26	155	65	24
T	−46	69	−38	−49

$Z = \mathrm{G}$

	A	*C*	*G*	*T*
A	39	−63	−69	−95
C	82	−72	14	−34
G	69	−47	21	−63
T	0	−96	−47	−70

$Z = \mathrm{T}$

We shall conservatively contemplate single-base mismatches, although multiple mismatches will further expand the fooling probe set. Single-base mismatch dimer parameters have also been experimentally evaluated by SantaLucia and others (Allawi and SantaLucia, 1997, 1998a, 1998b, 1998c; Peyret *et al.*, 1999) and are reported for convenience below in the form of a three-dimensional matrix K_3. In Table 2 we display the values of the free-energy for each match/mismatch, with the following convention: each two-dimensional subtable corresponds to a nucleotide Z; X is the row heading, and Y the column heading. Each entry denotes the energy of the dimer

$$\begin{array}{cc} \mathrm{X} & \mathrm{Y} \\ & \mathrm{Z} \end{array}$$

where X is matched to its Watson/Crick-complement and (Y, Z) is the mismatch pair. Dimers are presented in the conventional manner, with 5'-ends as row-headings.

The mismatch destabilization of (Y, Z) with respect to $(Y, \overline{Y})$, $\overline{Y}$ being the WC-complement of Y, can be calculated with respect to the configuration

$$\begin{array}{cc} \mathrm{X} & \mathrm{Y} \\ & \mathrm{Z} \quad \mathrm{W} \end{array}$$

where X and W are arbitrary and each is WC-matched. We shall then use the expression

$$mismatch(X, Y, Z, W) = (K_3(X, Y, Z) + K_3(W, Z, Y)) - (K_3(X, Y, \overline{Y}) + K_3(W, \overline{Y}, Y)).$$

We begin by determining for each mismatched pair (Y, Z) the most critical destabilization (smallest absolute value). These results are displayed in Table 3.

TABLE 3. WORST-CASE MISMATCH DESTABILIZATIONS

	A	*C*	*G*	*T*
A	−105	−135	−67	
C	−262	−261		−303
G	−168		−152	−218
T		−211	−105	−174

TABLE 4. MINIMAX DESTABILIZATION VALUES

	A	*C*	*G*	*T*	*Z*
A	−256	−296	−205		−205
C	−417	−463		−398	−398
G	−320		−321	−270	−270
T		−299	−196	−279	−196

TABLE 5. FOOLING PROBES DUE TO MISMATCHES WITH DEGENERATE BASES

h	$\nu(\delta = -196)$	4^h
1	0	4
2	2	16
3	22	64
4	171	256
5	826	1024
6	3722	4096
7	15624	16384

However, in order to realistically model the hybridization process, the values of X, Y, Z, W must be selected so that with almost certainty the computed destabilization will occur for every microarray feature. To achieve this objective, the base Y (belonging to the probe) is held fixed while we consider its three possible mismatches Z, for each Z we must choose the most favorable pair (X, W), i.e., the pair inducing the most favorable (largest in absolute value) destabilization. Finally, for the destabilizations associated with a fixed Y, we shall choose the most unfavorable Z. This will give the most conservative bound, provided that the base Y occurs in a natural-base position of the oligo.

In Table 4, we report the values computed according to the outlined minimax criterion, along with the selection of the most unfavorable Z (last column).

Among the four destabilization values appearing in the right column of Table 4 it is realistic to choose the most unfavorable term, −196, as a criterion for universal bases, since the corresponding base will appear in the near-totality of the microarray features.[2]

We now analyze the interference of mismatched subsequences. Recall that $[e_1, e_{4^h}]$ is the range of binding energies of correctly matched sequences and assume that the detection threshold has been set at e_1 to avoid false negatives (refer again to Fig. 1). Consider all length-L sequences sharing a fixed single mismatch with respect to a chosen feature (with a specific destabilization value δ, a negative number); clearly, there are 4^h such sequences, and the distribution of their binding energies is obtained by a simple translation by δ of the one obtained above. In other words, denoting by $e'_1, \ldots, e'_{4^h}$ the corresponding binding energies, we have $e'_j = e_j - \delta$ for $j = 1, \ldots, 4^h$, and the range of the mismatched sequences is $[e_1 - \delta, e_{4^h} - \delta]$. We observe that all mismatched sequences of this set whose energy exceeds the threshold e_1 act as fooling probes for the feature in question, and their number is given by $TAIL(e_{4^h} - \delta)$. Note that there are $3k$ possible mismatches (three choices for each definite-base position). Thus, the expected number of additional fooling probes due to mismatches is given by $3k \cdot TAIL(e_{4^h} - \delta)$.

To decide whether a degenerate base is universal of degree d, one must evaluate $TAIL(e_{4^h} - \delta)$, for a suitably chosen δ. If $TAIL(e_1 - \delta_{\min}) \gg 0$, the base is not strictly universal of degree d. Below, we report in Table 5 the values of this quantity for $d = 1, 2, \ldots, 7$ for minimax $\delta = -196$, to be compared with the number of possible assignments of the degenerate bases.

We conclude that degenerate bases are universal just for the trivial degree 1.

[2]This destabilization is due to the mismatch $\begin{matrix} \text{T} & \text{T} \\ \text{G} & \text{A} \end{matrix}$.

TABLE 6. FOOLING PROBES DUE TO MISMATCHES WITH SEMIDEGENERATE BASES

h	$\nu(\delta = -196)$	2^h
1	0	2
2	0	4
3	0	8
4	0	16
5	0	32
6	0	64
7	0	128
8	2	256
9	10	512
10	92	1024

3. SEMIDEGENERATE BASES ARE (NEARLY) UNIVERSAL

The nonuniversality of degenerate bases is caused by the spread of values of binding energy associated with a degenerate-base position in the probing pattern. In fact, for fixed X and for Y in the degenerate-base position, this spread is expressed as

$$[\min_Y K2(X, Y), \max_Y K2(X, Y)].$$

By inspection of Table 1 we obtain that such intervals are [30,82], [70,140], [69,155], and [0,82] for X =A,C,G, and T, respectively, whose average width is 72. If, on the other hand, rather than individual nucleotides, we consider the subsets A,T (weak bases) and C,G (strong bases), again from Table 1 we obtain the collection of intervals [30,39], [70,82], [70,82], [127,140], [69,82], [127,155], [0,39],[69,82], whose average width is only 17.4. This observation suggests the proposal of replacing degenerate bases (as mixtures of the four natural bases) with mixtures involving the two subsets separately, which we define as follows.

Definition 3. *An oligonucleotide array feature is said to contain a* semidegenerate base *in position j if, for any selection of the bases in the other positions, position j contains only bases of the same-strength set.*

As in Section 2, we now consider a string of h semidegenerate bases and compute the binding energies $e_1 \leq \ldots \leq e_{2^h}$ corresponding to all possible selections of the bases using the same calculation scheme. Again, $[e_1, e_{2^h}]$ is the range of energy values. We obtain for h semidegenerate bases Table 6 (analogous to Table 5), reported below for minimax $\delta = -196$, again compared with the number of possible assignments.

We conclude that semidegenerate bases are strictly universal for $h \leq 7$ and nearly universal for larger h.

4. ANALYTICAL COMPARISON

A better insight into the superiority of semidegenerate bases over degenerate bases can be obtained with an effective analytical approximation.

Since the size of the sample space is 4^h (i.e., sufficiently large for expected values of h), we may assume that the total hybridization energy pertaining to the degenerate bases is a normal random variable.

We shall consider strings of h bases, either all degenerate or all semidegenerate, and calculate and compare the first and second moments of their binding energies, conventionally defined as in Section 2. Let $\mu = 78.9$ and $\sigma = 40.6$, respectively, denote the average and the standard deviation of the dimer binding energy. For simplicity, in the following comparison, we neglect the binding energy pertaining to the leftmost base, but consider only those due to the subsequent $h - 1$ dimers.

We begin by considering degenerate bases. The expected value of the energy of an h-base string is obviously $(h-1)\mu$, since expectations are additive. Of crucial interest to us is the spread around the average; however, variances are not additive since the dimer energies pertaining to a string of bases form a Markov chain, and independence is not assured. In Appendix 1, we derive the fact that the variance of the energy of a string of h degenerate bases is

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

where ρ_2 is a correlation term defined as

$$\rho_2 = 4^{-3}\sum_{i,j,k}(K_2(i,j)-\mu)(K_2(j,k)-\mu).$$

Relation (1) allows us to interpret the quantity

$$\sigma^2 + 2\frac{h-2}{h-1}\rho_2 \approx \sigma^2 + 2\rho_2 = \sigma_{4,eff}^2$$

as an "equivalent" standard deviation, as if the dimers were independent random variables. For degenerate bases from Table 1 (see above), we obtain $\sigma^2 = 1{,}646$ and $2\rho_2 = 1{,}501$. The result is that $\sigma_{4,eff} = \sqrt{1646+1501} \approx 56.1$.

Semidegenerate bases present a much narrower energy spread. In this case, the analysis is more subtle, since each of the 2^h sets of analogous-strength h-tuples must be analyzed separately. This analysis, developed in the appendix, yields $\sigma^2 = 117$ and $2\rho = -18.06$ and, correspondingly, $\sigma_{2,eff} = 9.94$, thus confirming the original expectation.

We may now model the distribution of the binding energies as a normal distribution $N(x; (h-1)\mu, (h-1)\sigma^{(h)})$, where μ and $\sigma^{(h)}$ are the appropriate mean and standard deviation. This modeling increases its adequacy as h grows. For example, Fig. 2 displays the actual distribution and its normal approximation for six degenerate bases.

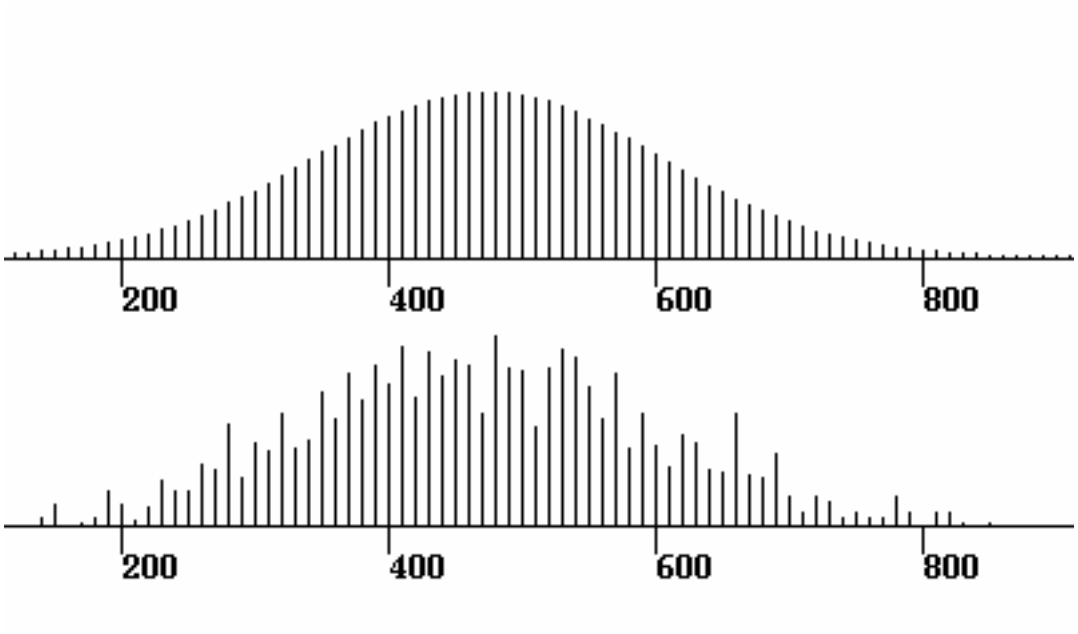

FIG. 2. Experimental distribution (**below**) and its normal approximation for 6 degenerate bases.

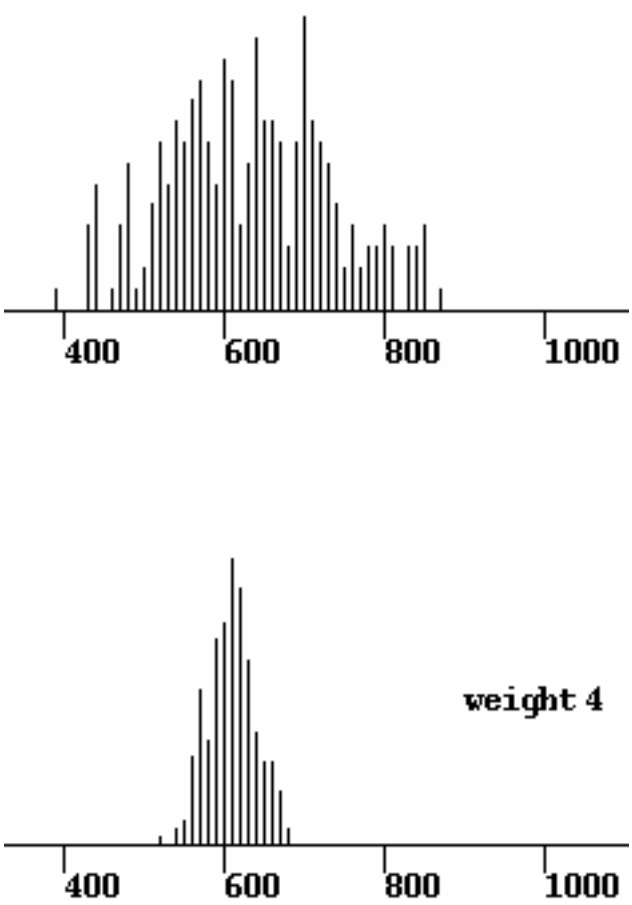

FIG. 3. Distributions of binding energies for the 256 oligos corresponding to either 4 degenerate (**above**) or 8 semidegenerate bases (**below**).

Corresponding diagrams for degenerate and semidegenerate bases are displayed in Fig. 3, under the condition of equivalent weakening of the observation signal, i.e., four degenerate bases are contrasted with eight semidegenerate bases (aligned for convenience with identical mean values). The diagrams effectively display the remarkably narrower energy spread induced by semidegenerate bases.

In this analytical model, we can now compare degenerate and semidegenerate bases from the point of view of universality. We wish to evaluate the number of "fooling probes" introduced by a single mismatch in either case, which, as we have seen in Section 3, is a reasonable measure of deviation from universality. Let n_0 be the conventional number of fooling probes associated with ideal universal bases; i.e., $n_0 = 4^h$ and $n_0 = 2^h$ for h bases in the two cases, respectively. Again, let δ be the value of the destabilization energy common to both cases. Denoting by σ' the standard deviation of the energy distribution, we may take the value $\mu - 3\sigma'$ as defining the left extreme of the energy interval. Referring to the discussion in Section 2, we must evaluate

$$F = \int_{\mu-3\sigma'}^{\infty} N(x; \mu - \delta, \sigma')dx = \int_{\delta/\sigma'-3}^{\infty} N(y; 0, 1)dy = \frac{1 - \mathrm{erf}(\delta/\sigma' - 3)}{2}.$$

The quantity σ' is evaluated at $\sqrt{4} \cdot 56 = 112$ for four degenerate bases and at $\sqrt{8} \cdot 9.94 = 28.1$ for eight semidegenerate bases (two situations corresponding to the same strength of hybridization signal, under which comparison is meaningful); in these two cases, for the average destabilization $\delta = -174$, the quantity F is evaluated, respectively, at 0.97 and $2 \cdot 10^{-7}$, in substantial agreement with the calculations reported in Tables 5 and 6, and providing a confirmation of the superiority of semidegenerate bases.

5. ALGORITHMIC CONSIDERATIONS

Starting from a standard k-base probing pattern, we can realize a semidegenerate-base probing pattern of identical microarray cost by successively replacing a single natural base with two semidegenerate bases. Indeed, such a replacement policy maintains unaltered the number of microarray features (4^k). However, each such substitution weakens the energy of the detection signal by a factor of 4.

Suppose, therefore, that we have replaced h natural bases to obtain a semidegenerate-base array with probes of length $k + h$. The reconstruction algorithm is, of course, based on the (left-to-right) extension of a putative sequence: this requires that the end-bases of the adopted probing pattern be natural bases. The positions of the remaining $k - 2 - h$ natural bases will be chosen to reduce the computational load, as discussed below.

We now discuss the overall simulation. A probing pattern and the target sequence length m are first selected. Next, we generate an adequate number of sample random sequences of length m (with independent identically distributed symbols). For each such sequence we generate its spectrum as follows: we slide a moving window of width $h+k$ along the target, and for each position, using the parameter tables introduced earlier, we compute the binding energy of the intercepted string and the binding energies of each of the $3(k - h)$ single mismatches in the natural positions; each of these $3(k - h) + 1$ energy values pertains to a specific microarray feature J and is used to update the feature maximum energy $e(J)$.

The spectrum thus computed and the probing pattern (and, if necessary, a putative-sequence primer) are supplied to the reconstruction algorithm, whose spectrum query now works as follows.

- The $(k - h - 1)$-suffix of the putative sequence is completed with the four possible extensions. For each such extension, we compute the binding energy e^* and the microarray feature j^*: the extension is considered present if and only if $e(j^*) \geq e^*$.

This realistic analog-spectrum policy introduces the presence of fooling probes. In case of ambiguous responses, the arising "paths" are extended as supported by the spectrum up to a maximum depth H (a design parameter): failure occurs if and only if the ambiguity is not resolved at depth H.

A simple analysis reveals that, if the competing-path extension reaches depth H, it is by far much more probable that failure-causing fooling probes arise from a single site rather than from multiple sites along the target. On the other hand, to alleviate the impact of fooling probes on the computational efficiency of the algorithm, the probing pattern should be designed with the criterion to decrease the likelihood that consecutive fooling probes arise from the same site of the target sequence. These observations have the following consequences.

1. The probing pattern should have very low off-peak autocorrelation. For example, pattern 100010010001 (where 1 and 0 respectively stand for natural and semidegenerate bases) satifies this criterion.
2. Failure of the length-$(k+h)$ semidegenerate-base scheme is reduced to that of the same-length standard scheme, thereby increasing the achievable reconstruction length from $C2^k$ to $C2^{k+h}$, for the same microarray area and for 4^h-fold weakening of the detectable signal. For example, for a 4^9-feature microarray (the state of the art), with a 4^4-fold signal loss, one may reliably achieve sequencing lengths of about 4,000 bases using exclusively natural-base oligos for the postulated ensemble of random sequences.

APPENDIX: VARIANCE OF BINDING ENERGY OF OLIGOS

We shall consider random variable $E(j_1, \ldots, j_h)$ defined in Section 2. The domain of membership of the h-tuples is the set $D = \{\text{A,C,G,T}\}^h$ of the 4^h DNA h-tuples.

Recall that μ is the average of the dimer energy. With our convention to consider only consecutive pairs, $(h - 1)\mu$ is the average of $E(j_1, \ldots, j_h)$. The corresponding variance has the expression

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

This expression is easily manipulated as follows:

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

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

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

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

$$= 4^{-(h-1)} \cdot 4^{h-3}(h-2) \sum_{i,j,k} (K_2(i, j) - \mu)(K_2(j, k) - \mu).$$

If we now define the correlation term ρ_2 as

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

we obtain

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

The analysis is somewhat more complicated for semidegenerate bases, since here we have to separately consider sets of h-tuples with identical strength pattern, i.e., such that in each position a base may vary only in its strength set (A,T or C,G). Therefore, we shall consider the following table of mean values $\mu(i, j)$ and standard deviations $\sigma(i, j)$.

	W	S
W	27	75.75
S	75.75	137.25

$\mu(i, j)$

,

	W	S
W	16.01	6.26
S	6.26	11.54

$\sigma(i, j)$

There are 2^h strength patterns described by an integer p, each identifying a set of 2^h h-tuples, and the analysis must be confined to one such set D_p at a time and then averaged over the ensemble of patterns. Let $s(j) \in 0, 1$ be the strength of a base $j \in 0, \ldots, 3$. The variance will have the expression

$$var = 2^{-h} \sum_{p=0}^{2^h-1} 2^{-h} \sum_{j_1,\ldots,j_h \in D_p} \left[\sum_{r=1}^{h-1} (K_2(j_r, j_{r+1}) - \mu(s(j_r), s(j_{r+1}))^2 + 2 \sum_{r=1}^{h-2} (K_2(j_r, j_{r+1}) - \mu(s(j_r), s(j_{r+1})))(K_2(j_{r+1}, J_{r+2}) - \mu(s(j_{r+1}), s(j_{r+2}))) \right]$$

Exchanging the order of summation, we obtain

$$4^{-h} \sum_{p=0}^{2^h-1} \sum_{j_1,\ldots,j_h \in D_p} \sum_{r=1}^{h-1} (K_2(j_r, j_{r+1}) - \mu(s(j_r), s(j_{r+1}))^2$$

$$= 4^{-h} \sum_{r=1}^{h-1} \sum_{p=0}^{2^h-1} 2^{h-2} \sum_{j_r, j_{r+1}} (K_2(j_r, j_{r+1}) - \mu(s(j_r), s(j_{r+1})))^2 = (h-1) \sum_{i,j=0}^{1} \frac{var(i, j)}{4}$$

and

$$4^{-h} \sum_{p=0}^{2^h-1} \sum_{j_1,\ldots,j_h \in D_p} \sum_{r=1}^{h-2} (K_2(j_r, j_{r+1}) - \mu(s(j_r), s(j_{r+1}))) (K_2(j_{r+1}, J_{r+2}) - \mu(s(j_{r+1}), s(j_{r+2})))$$

$$= 4^{-h} \sum_{r=1}^{h-2} \sum_{p=0}^{2^h-1} 2^h \sum_{i,j,k} \frac{(K_2(i, j) - \mu(s(i), s(j))) (K_2(j, k) - \mu(s(j), s(k)))}{8}$$

$$= (h-2)\rho_2'$$

where we have defined the correlation term ρ_2' as

$$\rho_2' = \sum_{i,j,k} \frac{(K_2(i, j) - \mu(s(i), s(j))) (K_2(j, k) - \mu(s(j), s(k)))}{64}.$$

Using the above table for the averages, we evaluate $\rho_2' = -9.03$.

ACKNOWLEDGMENTS

F.P.P. was supported in part by NSF Grant DBI-9983081, and J.S.O by Grant DBI-9983081 as well as NIH Grant HG2181-01A1.

REFERENCES

Allawi, H.T., and SantaLucia, J., Jr. 1997. Thermodynamics and NMR of internal G-T mismatches in DNA. *Biochemistry* 36, 10581–10594.

Allawi, H.T., and SantaLucia, J., Jr. 1998a. Nearest-neighbor thermodynamic parameters for internal G-A mismatches in DNA. *Biochemistry* 37, 2170–2179.

Allawi, H.T., and SantaLucia, J., Jr. 1998b. Thermodynamics of internal C-T mismatches in DNA. *Nucl. Acid Res.* 26, 2694–2701.

Allawi, H.T., and SantaLucia, J., Jr. 1998c. Nearest-neighbor thermodynamics of internal A-C mismatches in DNA. *Biochemistry* 37, 9435–9444.

Breslauer, K.J., Frank, R., Blocker, H., and Marky, W. 1986. Predicting DNA duplex stability from the base sequence. *Proc. Nat. Acad. Sci. USA* 83, 3746–3750.

Bains, W., and Smith, G.C. 1988. A novel method for DNA sequence determination. *J. Theoret. Biol.* 135, 303–307.

Dyer, M.E., Frieze, A.M., and Suen, S. 1994. The probability of unique solutions of sequencing by hybridization. *J. Comp. Biol.* 1, 105–110.

Drmanac, R., Labat, I., Bruckner, I., and Crkvenjakov, R. 1989. Sequencing of megabase plus DNA by hybridization. *Genomics* 4, 114–128.

Frieze, A.M., and Halldorsson, B.V. 2002. Optimal sequencing by hybridization in rounds. *J. Comp. Biol.* 9, 355–369.

Hannenhalli, S., Feldman, W., Lewis, H.F., Skiena, S., and Pevzner, P.A. 1996. Positional sequencing by hybridization. *Computer Applications in the Biosciences* 12(1), 19–24.

Loakes, D. 2001. The application of universal DNA base analogues. *Nucl. Acids Res.* 29, 2437–2447.

Lysov, Y.P., Florentiev, V.L., Khorlin, A.A., Khrapko, K.R., Shih, V.V., and Mirzabekov, A.D. 1988. Sequencing by hybridization via oligonucleotides. A novel method. *Dokl. Acad. Sci. USSR* 303, 1508–1511.

Peyret, N., Seneviratne, P.A., Allawi, H.T., and SantaLucia, J., Jr. 1999. Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A-A, C-C, G-G, and T-T mismatches. *Biochemistry* 38, 3468–3477.

Pevzner, P.A. 1989. l-tuple DNA sequencing: Computer analysis. *J. Biomol. Struct. Dyn.* 7(1), 63–73.

Pevzner, P.A., Lysov, Y.P., Khrapko, K.R., Belyavsky, A.V., Florentiev, V.L., and Mirzabekov, A.D. 1991. Improved chips for sequencing by hybridization. *J. Biomol. Struct. Dyn.* 9(2), 399–410.

Preparata, F.P., Frieze, A.M., and Upfal, E. 1999. On the power of universal bases in sequencing by hybridization. *3rd Int. Conf. Computational Molecular Biology*, 295–301.

Preparata, F.P., and Upfal, E. 2000. Sequencing-by-hybridization at the information-theory bound: An optimal algorithm. *J. Comp. Biol.* 7(3/4), 621–630.

SantaLucia, J.J. 1998. A unified view of polymer, dumbells, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* 95, 1460–1465.

Skiena, S., and Snir, S. 2002. Restricting SBH ambiguity via restriction enzymes. *WABI*, 404–417.

Southern, E.M. 1996. DNA chips: Analysing sequence by hybridization to oligonucleotide on a large scale. *Trends Genet.* 12(3), 110–115.

Waterman, M.S. 1995. *Introduction to Computational Biology*, Chapman and Hall, London.

Address correspondence to:
Franco P. Preparata
Computer Science Department
Brown University
115 Waterman Street
Providence, RI 02912-1910

E-mail: franco@cs.brown.edu

John S. Oliver
Chemistry Department
Brown University
324 Brook Street
Providence, RI 02912-1910

and

Genespectrum, Inc.
49 Pavilion Ave.
Providence, RI 02905

E-mail: John_Oliver@brown.edu