Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 10089–10093, November 1991
Genetics

DNA sequencing by hybridization: 100 bases read by a non-gel-based method

ŽAKLINA STREZOSKA, TATJANA PAUNESKU, DANICA RADOSAVLJEVIĆ, IVAN LABAT, RADOJE DRMANAC, AND RADOMIR CRKVENJAKOV*

Institute of Molecular Genetics and Genetic Engineering, P.O. Box 794, 11000 Belgrade, Yugoslavia

Communicated by Paul Doty, August 14, 1991

ABSTRACT **Determination of the sequences of human and other complex genomes requires much faster and less expensive sequencing processes than the methods in use today. Sequencing by hybridization is potentially such a process. In this paper we present hybridization data sufficient to accurately read a known sequence of 100 base pairs. In independent reactions, octamer and nonamer oligonucleotides derived from the sequence hybridized more strongly to this DNA than to controls. The 93 consecutive overlapping probes were derived from a 100-base-pair segment of test DNA and additional probes were generated by incorporation of a noncomplementary base at one of the ends of 12 of the basic probes. These 12 additional probes also had a full-match target in one of the control DNAs. The test and one of five control DNAs spotted on nylon filters were hybridized with 83 octamers and 22 nonamers under low-temperature conditions. A stronger signal in DNA containing a full-match target compared to DNA with only mismatched targets was obtained with all 105 probes. In 3 cases (2.9%), the difference of signals was not significant (<2-fold) due to inefficient hybridization and the consequently higher influence of background. The hybridization pattern obtained enabled us to resequence the 100 base pairs by applying an algorithm that tolerates an error rate much higher than was observed in the experiment. With this result, the technological components of large-scale DNA sequencing using the sequencing by hybridization method are in place.**

Large DNA sequencing projects are seen as vehicles for the advancement of biology. Genome sequences are expected to yield a wealth of information. Improvements in sequencing technology are among the stated goals of the Human Genome Project (1). Various proposals for methods involving gel-based and other approaches have been advanced. Among them are multiplex gel sequencing (2), scanning tunneling microscopy (3), single molecule fluorescence detection (4), laser x-ray diffraction (5), and sequencing by hybridization (SBH) (6, 7).

The basic idea behind SBH is that longer sequences can be obtained by the maximal and unique overlap of their constituent oligomers. For example, the three octamers

ATCAGGTC,
TCAGGTCT, and
CAGGTCTG

uniquely define the decamer ATCAGGTCTG.

No knowledge of the frequency or the position of the oligomers is needed; the knowledge of oligomer sequences and hybridization results suffices. To obtain accurate lists of the constituent oligomer content of DNAs of unknown sequence, two modes of oligomer hybridization have been proposed: (*i*) DNA bound to a surface and oligomers in solution for mapping (8) and sequencing (6, 7) or (*ii*) bound oligomers with free DNA as the probe (6, 9–12). The optimal probe lengths are related to target DNA complexity (insert and vector). Informatical, biochemical, and technological factors determine the optimal lengths of the probes. Our analysis of future SBH projects on the order of 1×10^9 base pairs (bp) based on the assessment of likely trends in development of hybridization technology gave the following results. When DNA is bound, probe lengths should be 6–10 nucleotides (7, 12). When the oligonucleotides are attached to surface, oligomers of 11–15 nucleotides are far more effective (12). The number of oligomers required for complete sequencing is 65,536 for octamers, 4.2×10^6 for 11-mers, etc. In either mode, hybridizations must discriminate between those samples containing duplexes with a perfect match and those having hybrids with the mismatched base pairs to compile accurate lists of constituent oligomers.

The destabilizing effect of a single internal mismatched base pair on oligomer hybrids of more than 11 nucleotides has been demonstrated (13). The least destabilizing situation is a single end mismatch, so the crucial test for sequence-grade oligomer hybridization is the ability to discriminate fully matched from end-mismatched duplexes. Appropriate hybridization conditions applicable to short (6–10 nucleotides) oligomers were found in a model study comprising reactions of 28 probes to two M13 clones bound to a nylon membrane (14). The optimal conditions involve performing the hybridization reaction at the lowest practical temperature to maximize the hybrid formation. Extended washes of the resulting hybrids at the same temperatures maximize discrimination by allowing the higher dissociation rate of mismatched hybrids to take effect. The conditions are applicable for hybrids with all possible sequences of a given length, but the hybrid yield is sequence-dependent. Tetramethylammonium chloride as a component of hybridization buffer has been useful for longer oligomers (15). Under our conditions it proved ineffective in equalizing sequence differences for shorter oligonucleotides, as expected (16). In support of the SBH concept, we report the sequencing of 100 bp of model DNA as proof that hybridization can provide sequencing data in addition to its other known applications. We used the mode in which the DNA to be sequenced was attached to a filter and the sequencing reaction was performed with octamer and nonamer probes; these conditions are likely to be appropriate for immediate applications of SBH.

MATERIALS AND METHODS

DNA Samples and Oligonucleotide Probes. The test DNA was the sequenced 922-bp *Eco*RI–*Bgl* II human genomic fragment containing the β-interferon gene (17). The segment between positions 627 and 726 was chosen for resequencing by hybridization; it has 42% G+C content on average. The segment sequence dictated the sequences of 72 octamer and

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: SBH, sequencing by hybridization.

*Present address of the authors: Biological and Medical Research Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439.

21 nonamer probes that occur in the segment in consecutive overlapping frames displaced by one or two bases. Our choice of octamer or nonamer was based on the desire that the probes be contained in the 100,000-probe set proposed for sequencing mammalian genomes (7). Preferentially octamers contain three or more G+C bases and nonamers one or two G+C bases. As a negative control, all 93 probes were also hybridized to another DNA of known sequence chosen not to contain a full match. The following five DNAs served as controls: M13mp18 (18), pBR322 (19), pUC18 (20), pHE4 (21), and pN1 (22). An additional 12 probes were designed by the inclusion of noncomplementary bases on one of the ends of basic probes. Each of these additional probes had a full match in one of the control DNAs.

Table 1 shows the control used and the number and type of relevant targets in both the test and control DNAs for each probe. Probes 69 (pHE4), 81 (pUC18), and 90 (M13mp18), with a full-match target in both the interferon DNA and indicated control DNAs, served to measure the relative amounts of DNA targets available on filters. All probes were made by Genosys, Houston. The probes were not further purified after the deprotection step.

Spotting and Hybridization Conditions. Base-denatured DNA (20 ng of interferon DNA and equimolar amounts of control DNA in 1.5 M NaCl/0.5 M NaOH at 10 ng/μl) were spotted on GeneScreen membranes (New England Nuclear) wetted in the same denaturing solution. Hybridization was according to Drmanac *et al.* (14). Briefly, [γ-^{32}P]ATP end-labeled probes (3.3 pmol; 10 μCi, 3000 Ci/mmol; 1 Ci = 37 GBq; Amersham) without separation of unincorporated radioactivity were hybridized at 10 ng/ml and 12°C in 0.5 M Na_2HPO_4, pH 7.2/7% (wt/vol) sodium lauroyl sarcosine for 3 h. Hybrids were washed in 6× standard saline citrate at 0°C for 40 min and autoradiographed for 4–48 h.

Compilation of Hybridization Data. The intensity of hybridization signals was visually estimated on the basis of several examples measured in a scintillation counter. All relative values for full-match DNA dot (H_{fm}) and end-mismatch DNA dot (H_{emm}) were determined relative to the value of 10 given to the strongest signal. A discrimination factor for a given pair of DNA dots was calculated using equation with the normalization term

$$D = (H_{fm}/H_{emm}) \times (H_{emm,Pm}/H_{fm,Pm}),$$

where the subscript Pm denotes the hybridization results obtained on the same pair of dots using a DNA control probe (i.e., a probe that has identical full matches in the DNAs of both dots). In this way, the signal intensities were corrected for variations in the amount of DNA in the dots.

RESULTS

Sequencing a 100-bp Test Segment. To demonstrate DNA sequencing by octamer and nonamer hybridization, we analyzed a 100-base-long region of a 922-bp human β-interferon gene fragment of known sequence (17). The 93 probes having full matches and 12 probes having end-mismatched targets in the test segment were used for hybridization. Each probe was hybridized to a two-dot filter containing the test DNA and the control DNA. To see if the hybridization reaction is discriminating sufficiently for sequencing purposes, the occurrence of positive hybridization with probes from the first group and its absence with probes of the second group was scored. Positive hybridization is defined as a significantly higher signal from the dot containing the full match. Hybridization results are shown in Fig. 1. Without exception, the DNA with the full match hybridized more strongly than the DNA that did not contain the full match. Thus probes 1–93 hybridized more strongly with interferon DNA; the additional 12 probes hybridized more strongly with the control DNA. The results with the probes having a perfect match in both test and control DNA show that the discrimination is not due to the higher amounts of test DNA compared to control DNA on the filters (Fig. 1*C*). In many cases, there was more DNA in the control dot than in the interferon test dot.

The discrimination coefficient D was determined for each of the probes. Estimated D values are listed in Table 1. Due to the variation of values near background and their involvement in observed low D values, only D values greater than 2 were considered significant. On this basis, 90 probes out of 93 were included in the significant list. All 12 probes designed to lack a full match had D values greater than 2. The probes 31, 84, and 85 from the basic group had D value less than 2. The D values of the 90 probes ranged from 3 to 40. Since corrections were introduced for the effects of variation in input DNA, the D value is expressed in a way that eliminates the differences in hybridization efficiency among the probes. Thus, probe 16 is considered positive even if its signal with test DNA is weaker than the signal of probe 31 with the control DNA. The complete and accurate sequence of the initial 100 bp was reconstructed from the list of positive probes using an algorithm developed for sequence generation (24). The exclusion of the three false-negative probes (2.9%) from the list did not prevent complete regeneration of the sequence or introduce any errors. Furthermore, the same result was obtained when the 12 control probes were included; i.e., even with 2.9% false-negative and 11% false-positive probes, we still obtained the correct sequence. This confirms the reconstruction efficiency of the SBH algorithm, which was >95% successful in another test using simulated data set [50-kilobase (kb) DNA, 100,000 probes] containing 5% false-positive and 5% false-negative data points (24).

The Factors of Discriminative Hybridization. We have analyzed the hybridization patterns obtained (*i*) to determine the dominant factor that leads to a decrease of discrimination with some of the probes and (*ii*) to estimate the possibility of predicting the hybridization efficiency of untested probes on the basis of sequence. The variations in discrimination are primarily the result of hybridization efficiency, which allows visualization of the signal over the background only if the efficiency is sufficiently large. Another important factor is the kind of mismatched target present in the control DNA. We analyzed the influence of these two factors in more detail.

First, the amount of hybrid obtained in test DNA with a certain probe, H_{fm}, was estimated and given relative values ranging between 0.3 and 10. More than 79% of the probes belong to the group with an H_{fm} value greater than 1 (Table 2). In this group the average discrimination value was 12 and the minimal discrimination value was higher than 3. In 5 out of 25 probes with end-mismatch targets in control DNA, the D value was $\geq$10. A possible explanation can be the large destabilization effect of end mismatches other than G/T or G/A (25) or errors in DNA control sequences.

In the group of probes with a hybridization efficiency H_{fm} of $\leq$1, the average D value is 4.1. The three probes (probes 31, 84, and 85) with low D values belong to this group. The average D value was 3.4 for the probes in the group that had end-mismatched targets in the control DNA. Due to the inefficiency of hybrid formation in this group of probes, the influence of the background over the kind of mismatched target in control DNA becomes dominant. Two possible reasons exist for low hybrid formation: (*i*) prevention of hybrid formation and (*ii*) hybrid instability.

The inability to form a hybrid can be inherent to the probe (for example, self-complementarity), the target may not be available (presence of internal loops), or there could be errors in the sequence of the DNA or the probes. The first reason was excluded, as none of the probes are palindromic. Probe 85 was shown to hybridize with its full match in the two

Table 1. Probes used, their target composition, obtained hybridization values, and calculated free energies

No.	Sequence	IF fm	IF emm	Control Name	Control fm	Control emm	H_{fm}	D	$\Delta G°$, kcal/mol
1.	GTCTGAAA	1	0	H	0	1	7	8	8.8
2.	TGTCTGAA	1	0	H	0	1	4	8	8.6
3.	TTGTCTGA	1	0	H	0	1	1.5	8	8.6
4.	CTTGTCTG	1	0	H	0	0	3	15	8.6
5.	TCTTGTCT	1	0	H	0	0	1	3	8.4
6.	ATCTTGTC	1	0	H	0	0	4	15	8.3
7.	AATCTTGTC	1	0	H	0	0	10	40	10.6
8.	GAATCTTG	1	0	H	0	0	10	40	9.0
9.	TGAATCTTG	1	0	H	0	0	2	7	11.1
10.	ATGAATCTT	1	0	H	0	0	3.5	25	10.8
11.	GATGAATC	1	0	H	0	0	3	25	8.4
12.	AGATGAATC	1	0	H	0	0	1.5	20	10.3
13.	TAGATGAAT	1	0	H	0	0	1	13	9.7
14.	CTAGATGA	1	0	H	0	0	7	20	7.5
15.	GCTAGATG	1	0	H	0	0	10	20	9.2
16.	TGCTAGAT	1	0	N	0	0	0.5	3	9.2
17.	GTGCTAGA	1	0	H	0	0	7	30	8.9
18.	AGTGCTAG	1	0	H	0	0	5	20	9.1
19.	CAGTGCTA	1	0	H	0	0	4	8	9.3
20.	CCAGTGCT	1	0	H	0	0	4	15	11.5
21.	GCCAGTGC	1	1	H	0	1	10	7	13.1
22.	AGCCAGTG	1	1	N	0	0	5	10	11.5
23.	CAGCCAGT	1	0	H	0	0	5	8	11.5
24.	CCAGCCAG	1	0	H	0	0	10	15	13.3
25.	TCCAGCCA	1	0	H	0	0	3	5	13.2
26.	TTCCAGCC	1	0	H	0	1	3	5	13.4
27.	ATTCCAGC	1	0	H	0	1	7	10	11.8
28.	CATTCCAG	1	0	H	0	0	3	10	10.4
29.	TCATTCCA	1	0	H	0	0	1	3	10.3
30.	CTCATTCC	1	0	H	0	0	4	8	10.1
31.	TCTCATTC	1	0	U	0	0	0.5	1.7	8.5
32.	GTCTCATT	1	0	H	0	0	5	10	8.3
33.	AGTCTCAT	1	0	H	0	0	1	5	7.9
34.	TAGTCTCA	1	0	H	0	0	0.5	4	7.2
35.	ATAGTCTC	1	0	H	0	0	0.3	3	7.0
36.	AATAGTCTC	1	0	H	0	0	7	15	9.3
37.	CAATAGTC	1	0	H	0	0	3	8	7.7
38.	ACAATAGTC	1	0	H	0	0	6	15	9.3
39.	AACAATAGT	1	0	H	0	0	3	10	9.8
40.	CAACAATAG	1	0	H	0	0	7	20	10.3
41.	TCAACAATA	1	0	H	0	0	7	15	10.2
42.	CTCAACAA	1	0	H	0	0	4	12	9.1
43.	TCTCAACA	1	0	H	0	0	2	4	8.6
44.	TTCTCAAC	1	0	H	0	0	0.5	9	8.8
45.	GTTCTCAA	1	0	H	0	1	0.5	6	8.8
46.	GGTTCTCA	1	0	H	0	1	10	20	9.9
47.	AGGTTCTC	1	0	H	0	0	3	4	9.7
48.	GAGGTTCT	1	0	H	0	0	4	15	9.7
49.	GGAGGTTC	1	0	H	0	0	8	15	11.2
50.	AGGAGGTT	1	0	H	0	0	7	15	11.3
51.	CAGGAGGT	1	0	H	0	0	7	15	11.1
52.	CCAGGAGG	1	0	H	0	0	5	8	12.9
53.	GCCAGGAG	1	0	M	0	0	5	7	12.9
54.	AGCCAGGA	1	0	N	0	0	8	8	12.9
55.	TAGCCAGG	1	0	N	0	1	6	4	12.4
56.	TTAGCCAG	1	0	H	0	1	3	8	11.3
57.	ATTAGCCA	1	0	H	0	0	4	5	11.2
58.	CATTAGCC	1	0	H	0	0	2	4	11.1
59.	ACATTAGC	1	1	H	0	0	2	4	9.4
60.	GACATTAG	1	0	H	0	0	5	10	7.7
61.	AGACATTAG	1	0	H	0	0	10	20	9.6
62.	TAGACATTA	1	0	H	0	0	4	8	8.9
63.	ATAGACATT	1	0	H	0	0	5	8	9.5
64.	GATAGACA	1	0	H	0	0	10	20	7.2
65.	TGATAGAC	1	0	H	0	0	0.5	3	7.2
66.	ATGATAGAC	1	0	H	0	0	10	10	9.0
67.	GATGATAG	1	0	H	0	0	7	12	7.4
68.	TGATGATAG	1	0	H	0	0	5	10	9.4
69.	CTGATGAT	1	0	N	0	0	1	4	8.3
70.	TCTGATGA	1	0	N	0	0	1	3	8.3
71.	ATCTGATG	1	0	H	0	0	5	8	8.3
72.	TATCTGATG	1	0	H	0	0	8	8	9.5
73.	TTATCTGAT	1	0	H	0	0	1	3	9.7
74.	TTTATCTGA	1	0	M	0	0	0.5	4	10.2
75.	GTTTATCTG	1	0	H	0	0	2	4	10.0
76.	GGTTTATC	1	0	H	0	1	5	8	9.4
77.	TGGTTTATC	1	0	H	0	0	2	3	11.5
78.	ATGGTTTAT	1	0	H	0	0	5	5	11.5
79.	GATGGTTT	1	0	H	0	0	7	12	10.3
80.	AGATGGTT	1	2	H	0	0	5	12	9.9
81.	CAGATGGT	1	2	H	0	0	8	25	9.7
82.	TCAGATGG	1	0	H	0	2	5	8	9.9
83.	TTCAGATG	1	0	N	0	1	8	5	8.8
84.	CTTCAGAT	1	0	N	0	0	0.5	1.5	8.6
85.	TCTTCAGA	1	1	U	0	1	0.5	1.3	8.6
86.	GTCTTCAG	1	1	H	0	1	8	15	8.4
87.	TGTCTTCA	1	0	H	0	0	1	3	8.6
88.	CTGTCTTC	1	0	H	0	0	1	5	8.4
89.	ACTGTCTT	1	0	H	0	0	1	3	8.2
90.	GACTGTCT	1	0	H	0	0	10	15	7.7
91.	GGACTGTC	1	0	H	0	0	10	25	9.2
92.	AGGACTGT	1	0	H	0	1	10	30	9.3
93.	CAGGACTG	1	0	H	0	1	10	15	9.8
94.	ATTGTCTG	0	1	M	1	1	4	8	8.5
95.	TAGATGAC	0	1	P	1	0	1	3	7.2
96.	GCCAGCCA	0	1	M	1	2	3	6	14.9
97.	TCCAGCCT	0	1	M	1	1	3	4	13.0
98.	TTCCAGCA	0	1	H	1	0	1.5	3	12.1
99.	GCATTCCA	0	1	M	1	1	4	6	12.0
100.	TGTCTCAT	0	1	U	1	0	3	7	8.1
101.	TTCTCAAG	0	1	M	1	2	2	5	9.1
102.	GTTCTCAG	0	1	H	1	2	10	8	8.4
103.	TATCTGATT	0	1	M	1	0	10	8	9.7
104.	TTTCTCCT	0	3	U	1	1	1	3	10.4
105.	AGGAGAAA	0	3	H	1	0	2	3	10.4

IF, interferon test DNA; Name, name of control DNA clone (H, pHE4; P, pBR322; M, M13mp18; U, pUC18; N, pN1); fm and emm, the number of full-match and end-mismatched targets, respectively, in either test or control dot for a given probe.

targets with equal efficiency, rendering unlikely the possibility of a loop structure in the test DNA.

An error in the interferon sequence would prevent efficient hybridization with eight successive probes, which was not observed. Finally, seven probes with H_{fm} values of ≤ 1 were resynthesized to check for synthesis errors. An increase in the H_{fm} value of from 0.5 to 3 and D value of from 1 to 8 was obtained with only one of the newly synthesized probes (Fig. 2). The dramatic increase of both H_{fm} and D values is obvious, indicating an error in the synthesis of the first lot. Except for this instance, the rest of the cases cannot be explained by any of the factors that prevent hybrid formation. The low efficiency of hybridization for the second group of probes is more likely to be the result of a sequence-dependent hybrid instability. We calculated the free energy of hybrid dissociation $\Delta G°$ for our probes using values for dimers (26). Although not measured values, the free energy estimates are sufficiently informative for probe comparisons. The range

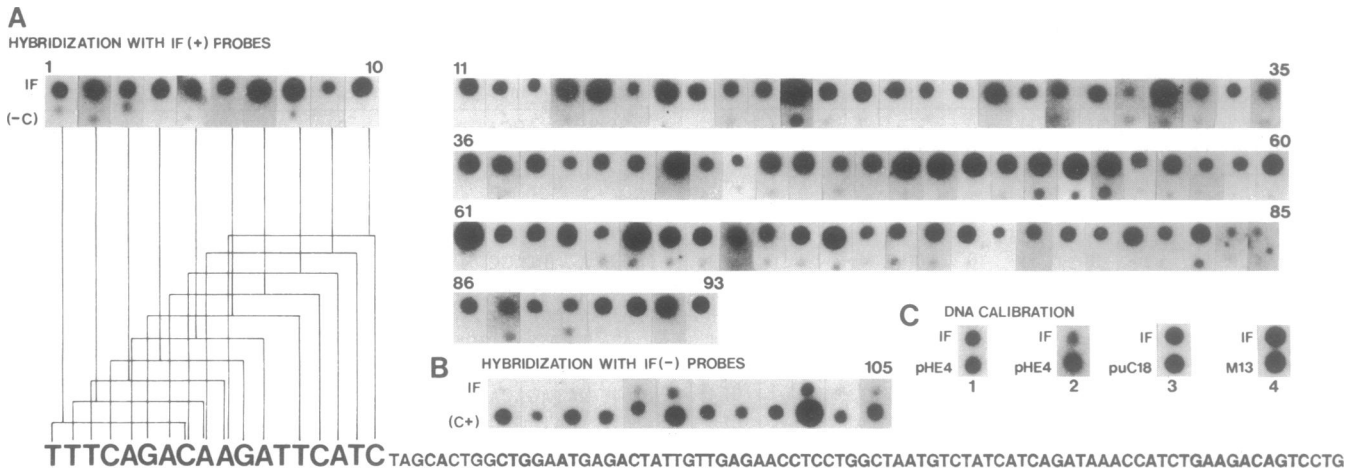

FIG. 1. SBH of 100 bp of the 922-bp human β-interferon gene fragment (IF). (*A*) Hybridization with 93 probes (72 octamers and 21 nonamers) with a full-match in the IF test DNA. IF and control rat globin clones pHE4 and pN1 were PCR-amplified (23), and control M13mp18, pBR322, and pUC18 DNAs were in linearized double-stranded form. The actual 100-bp portion of the interferon gene sequence from positions 627 to 726 is shown with the complementary targets for the first 10 probes indicated. (*B*) Hybridization with 12 probes (11 octamers and 1 nonamer) that have end mismatches in IF fragment. (*C*) Dot DNA calibration. Blots: 1 and 2, IF and pHE4, probe CTGATGAT; 3 IF and pUC18, probe CAGATGGT; 4, IF and M13mp18, probe GACTGTCT. The ratios of DNA amounts in the IF and control dots were 1:1 in blots 1, 3, and 4 and 1:3 in blot 2. Filters with IF and pN1 had 1:2 ratio with probe CTGATGAT (data not shown). These DNA calibration factors were used to correct estimated H_{fm} and D values from *A* and *B*.

calculated for octamers at 12°C was 5–21 kcal/mol (1 cal = 4.184 J), depending on the sequence (Table 1). Of the group of weakly hybridizing octamer probes ($H_{fm} \leq 1$), 84.2% had a $\Delta G°$ of 7–9 kcal/mol (Fig. 3). This indicates the dominant influence of hybrid instability on low hybridization efficiency. Furthermore, the relative proportion of weakly hybridizing probes decreases with an increase in $\Delta G°$. This is also shown by the five cases of weakly hybridizing octamers; probes 6, 35, 37, 60, and 65, which are contained in nonamer probes. The nonamer probes 7, 36, 38, 61, and 66 gave better discrimination results than the octamers they contain.

However, our experiment does not answer the question of the predictive power of $\Delta G°$ with regard to the efficiency of hybrid formation, mainly due to the described way of measuring H_{fm} and H_{emm} leading to the insufficient precision of D values. Examples of probes that have a high $\Delta G°$ but a lower H_{fm} than expected are nonamers 9 and 77. They hybridize more weakly than their constituent octamers 8 and 76, indicating that the possible existence of additional factors influencing hybridization efficiency is not excluded by our experiment.

DISCUSSION

The Reliability of Hybridization Sequencing Data. The overall results of these 105 probes are in agreement with previous data (14) and demonstrate the accurate determination of the presence of full-match targets in test DNA. In addition, the results show that octamers and nonamers can be efficiently used in SBH. There is a low chance (<1 in 105) of obtaining a false-positive signal using simple and identical hybridization conditions for unpurified probes. An unreliable signal can be expected in 2.9% of the probes. This is a high estimate since the frequency of probes with a $\Delta G°$ of <9 kcal/mol is 50% higher in this experiment than expected for a random sequence. These probes do not necessarily give false-negative results, since they can be recognized by the use of calibrating control DNAs with the full matches. The majority of these oligomer sequences will give efficient fingerprints if groups of longer oligomers having them as a core are used as probes. One might also use chemical modifications of probes that enhance hybrid stability (27).

Table 2. Distribution of D values for 105 probes

Mismatched target	Probes, no. $H_{fm} > 1$	Probes, no. $H_{fm} \leq 1$
emm−	13.4 (56.2)	4.1 (17.1)
emm+	8.7 (22.8)	3.4 (3.8)

Values are grouped by both full-match DNA hybridization efficiencies, H_{fm}, and the kind of mismatched target in control DNA (emm−, end-mismatched target absent; emm+, end-mismatched target present). The percentage of probes in each group is shown in parenthesis.

The last factor of importance when sequencing unknown DNAs is the target complexity. This factor was omitted by the design of this experiment. Although most internal and double mismatch targets contribute little to overall hybridization, the presence of 5–10 end-mismatch targets might give a signal comparable to the one obtained from a single complete match (14). In practice, this limits the size of the DNA that can be effectively interrogated with an oligomer of a certain length. Generally, hexamers, heptamers, and octamers can be used on fragments that are 0.5 kb, 2 kb, and 8 kb long, respectively. Two factors can be seen as responsible for the high fidelity of hybridization in the 100-bp sequencing experiment. (*i*) Short oligomer hybridization occurs with "all or none" kinetics (28, 29). (*ii*) The difference in duplex stability between matched and end-mismatched duplex will increase with a decrease in duplex length. However, due to a decrease in binding energy with the decrease in hybrid length, the efficiency of hybrid formation also decreases. Octamers and nonamers perhaps give the best combination of duplex stability and discrimination for sequencing purposes.

Theoretically, SBH has drawbacks that lead to a limited number of ambiguities in the final generated long sequences

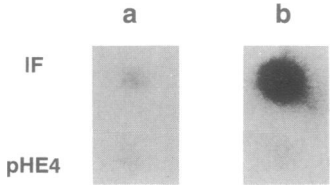

FIG. 2. Hybridization of IF pHE4 filter with two batches of probe TTAGCCAG. Blots: a, lot 532-7; b, lot 622-4.

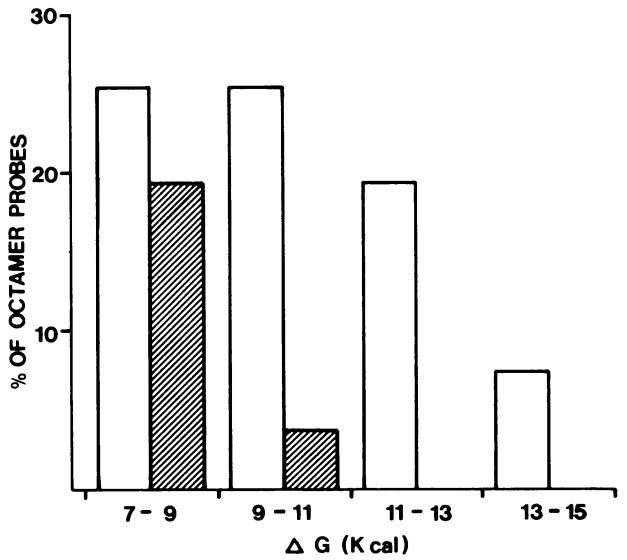

FIG. 3. Correlation of $\Delta G°$ and hybridization efficiency. The relative proportions of octamer probes from each increment (5 kcal/mol) of $\Delta G°$ that belong to the two hybridization efficiency groups are shown (hatched bars, $H_{fm} < 1$; open bars, $H_{fm} > 1$).

(7). Existing complementary methods directed at these ambiguities can be used in the second round of experimentation to complete the affected sequences, mainly in the minisatellite fraction of the genome.

Future Applications. In addition to the applications of oligomer hybridization for physical mapping and complete DNA sequencing, a third application, called partial-SBH, has been proposed by one of us (R.D.) (30). It is based on the idea that the incomplete lists of constituent oligomers will be sufficient to discern biologically relevant information in a comparative analysis of sequences. Partial-SBH should provide direct data for statistically significant inferences of the similarity of unknown DNA fragments to known sequences, to sequences constrained by some imposed rule, or to other unknown sequences. Our expectation is that partial-SBH could localize and define genes on chromosomes with 30–100 times less hybridization data than required for complete sequencing. Partial-SBH could also be used profitably for fingerprinting cDNA libraries and finding and simultaneously following numerous DNA polymorphisms. The data on cDNAs can be used to count and group the expressed genes, assemble complete cDNAs by overlap, compare the patterns of expression of different tissues, etc.

We have demonstrated the utility of the SBH method on dot blots of M13 clones or PCR-amplified inserts (14, 30). A pilot experiment on the hexa- to octamer hybridization of 300 PCR-amplified cDNA clones has shown the feasibility of increasing the number of dots from 2 to 300 (30); dot numbers of 10,000–100,000 per single membrane seem achievable. All techniques and equipment necessary for the large data collection for SBH have been shown to work. After appropriate efforts to achieve a steady throughput, it is conceivable that a single laboratory will be able to collect 100 million hybridization dot data bits over a period of 1–2 years. This is enough to completely sequence a few megabases of DNA or to find most of the genes on a single human chromosome by partial-SBH.

This rate falls short by two orders of magnitude from that needed for a single laboratory to obtain the complete human genome sequence in 1 year. For such sequencing speeds, we have proposed the use of a "sequencing "chip," a microhybridization surface with millions of known positions, each containing a different defined oligomer. The chip will be read by fluorescence microscopy after hybridization with mixtures of fluorescently labeled short pieces resulting from shearing specific large genomic fragments/clones longer than 50 kb (12). Interestingly, the recent work of Fodor *et al.* (31) indicates that the "sequencing chips" of sufficient complexity may soon become a reality.

In summary, an application of nucleic acid hybridization capable of sequencing DNA has been demonstrated, providing a strong incentive for its development into a technology for sequencing human and other genomes.

This work was supported in part by a grant from the Science Fund of Serbia (Yugoslavia), by Department of Energy Grant DE-FG02-88ER60699, and by U.S.-Yugoslav Joint Board Project JF 820.

1. Department of Energy, Office of Health and Environmental Research (1990) *Fed. Regis.* **55,** 10456.
2. Church, G. M. & Kieffer-Higgins, S. (1988) *Science* **240,** 185–188.
3. Beebe, T. P., Jr., Wilson, T. E., Ogletree, D. F., Katz, J. E., Balhorn, R., Salmeron, M. B. & Siekhaus, W. J. (1989) *Science* **243,** 370–372.
4. Jett, J. H., Keller, R. A., Martin, J. C., Marrone, B. L., Moyzis, R. K., Ratliff, R. L., Seitzinger, N. K., Shera, E. B. & Stewart, C. C. (1989) *J. Biomol. Struct. Dyn.* **7,** 301–309.
5. National Institutes of Health–Department of Energy (1990) *Human Genome News* **2,** No. 2, p. 4.
6. Drmanac, R. & Crkvenjakov, R. (1987) Yugoslav Patent Appl. 570.
7. Drmanac, R., Labat, I., Brukner, I. & Crkvenjakov, R. (1989) *Genomics* **4,** 114–128.
8. Poustka, A., Pohl, T., Barlow, D. P., Zehetner, G., Craig, A., Michiels, F., Ehrich, E., Frischauf, A. M. & Lehrach, H. (1986) *Cold Spring Harbor Symp. Quant. Biol.* **51,** 131–139.
9. Southern, E. (1988) United Kingdom Patent Appl. GB 8810400.
10. Bains, W. & Smith, G. C. (1988) *J. Theor. Biol.* **135,** 303–307.
11. Khrapko, K. R., Lysov, Yu. P., Khorlyn, A. A., Shick, V. V., Florentiev, V. L. & Mirzabekov, A. D. (1989) *FEBS Lett.* **256,** 118–122.
12. Drmanac, R., Labat, I., Strezoska, Ž., Paunesku, T., Radosavljević, D., Drmanac, S. & Crkvenjakov, R. (1991) in *Electrophoreses, Supercomputers and the Human Genome*, eds. Cantor, C. R. & Lim, H. A. (World Sci., Singapore), pp. 47–59.
13. Wallace, R. B., Shaffer, J., Murphy, R. F., Bonner, J., Hirose, T. & Itakura, K. (1979) *Nucleic Acids Res.* **6,** 3543–3557.
14. Drmanac, R., Strezoska, Ž., Labat, I., Drmanac, S. & Crkvenjakov, R. (1990) *DNA Cell Biol.* **9,** 527–534.
15. Wood, W. I., Gitcher, J., Lasky, L. A. & Lawn, R. M. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1585–1588.
16. Jacobs, K. A., Rudersdorf, R., Neil, S. D., Dougherty, J. P., Brown, E. L. & Fritsch, E. F. (1988) *Nucleic Acids Res.* **16,** 4637–4650.
17. Ohno, S. & Taniguchi, T. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 5305–5309.
18. Messing, J. (1983) *Methods Enzymol.* **101,** 20–78.
19. Bolivar, F., Rodriguez, R. L., Greene, P. L., Betlach, M. C., Heyneker, H. W., Boyer, H. W. & Falkow, S. (1977) *Gene* **2,** 95–107.
20. Yanisch-Perron, C., Vieira, J. & Messing, J. (1985) *Gene* **33,** 103–119.
21. Stevanovic, M. & Crkvenjakov, R. (1989) *Nucleic Acids Res.* **17,** 4878.
22. Radosavljević, D. & Crkvenjakov, R. (1989) *Nucleic Acids Res.* **17,** 4368.
23. Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A. & Arnheim, N. (1985) *Science* **230,** 1350–1354.
24. Drmanac, R., Labat, I. & Crkvenjakov, R. (1991) *J. Biomol. Struct. Dyn.* **5,** 1085–1102.
25. Ikuta, S., Takagi, K., Wallace, R. B. & Itakura, K. (1987) *Nucleic Acids Res.* **15,** 797–811.
26. Breslauer, K. J., Frank, R., Blocker, H. & Marky, L. A. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 3746–3750.
27. Asseline, U., Delarue, M., Lancelot, G., Toulme, F., Thuong, N. T., Montenay-Garestier, T. & Helene, C. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 3297–3301.
28. Craig, M. E., Crothers, D. M. & Doty, P. (1971) *J. Mol. Biol.* **62,** 383–407.
29. Porschke, D. & Eigen, M. (1971) *J. Mol. Biol.* **62,** 361–381.
30. Drmanac, R., Lennon, G., Drmanac, S., Labat, I., Crkvenjakov, R. & Lehrach, H. (1991) in *Electrophoreses, Supercomputers and the Human Genome*, eds. Cantor, C. R. & Lim, H. A. (World Sci., Singapore), pp. 60–74.
31. Fodor, S. P. A., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. & Solas, D. (1991) *Science* **251,** 767–773.