Nucleic Acids Research, 2003, Vol. 31, No. 19 ***5667–5675***
DOI: 10.1093/nar/gkg757

Advanced computational techniques for re-sequencing DNA with polymerase signaling assay arrays

Itsik Pe'er[1,2,*], Naama Arbili[2], Yi Liu[3], Colby Enck[3], Craig A. Gelfand[3] and Ron Shamir[2]

[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel, [2]School of Computer Science, Tel Aviv University, Tel Aviv, Israel and [3]Orchid Biosciences Inc., Princeton, NJ, USA

Received May 28, 2003; Revised July 13, 2003; Accepted August 6, 2003

ABSTRACT

Re-sequencing, the identification of the specific variants in a sequence of interest compared with a known genomic sequence, is a ubiquitous task in today's biology. Universal arrays, which interrogate all possible oligonucleotides of a certain length in a target sequence, have been suggested for computationally determining a polynucleotide sequence from its oligonucleotide content. We present here new methods that use such arrays for re-sequencing. Our methods are applied to data obtained by the polymerase signaling assay, which arrays single-based primer extension reactions for either universal or partial arrays of pentanucleotides. The computational analysis uses the spectrum alignment algorithm, which is refined and enhanced here in order to overcome noise incurred by the use of such short primers. We present accurate re-sequencing results for both synthetic and amplified DNA molecules.

INTRODUCTION

Genomic sequence is now abundant: the genomes of more than a hundred species including human have already been sequenced. Despite this profusion of data, sequencing is still a routine task in laboratory work. This demand for sequencing is to a large extent targeted at molecules whose nucleic acid sequences are approximately known in advance. This is the case in validation of sequences, in cDNA sequencing, and in detection and typing of polymorphisms or germline/somatic mutations. All these tasks can be categorized as 're-sequencing' tasks, i.e. the determination of a nucleotide sequence which is known to be a variant of some previously sequenced reference molecule. This promotes re-sequencing as a key endeavor in today's biology (1).

The identification of millions of human polymorphisms (2) and the ongoing mapping of all common human haplotypes (3,4) will lead in the near future to a situation where virtually all common sequence variations have been mapped. Nevertheless, due to the more modern expansion of the human race, much of the observed variation is comprised of rare polymorphisms and familial mutations. To determine the correct alleles of a certain locus borne by a specific individual it is thus insufficient to type only known single nucleotide polymorphisms (SNPs) that are abundant in the population: one would ultimately need to detect sporadic variations as well, and so, for many studies, complete re-sequencing will remain a key task in accurate genetic typing of individuals.

Sequencing by hybridization (SBH) was invented as an alternative to gel-based sequencing (5–7). This method makes use of a universal DNA microarray, which harbors all oligonucleotides of length k (called k-mers or 'words') as probes. These oligonucleotides are assayed with an unknown DNA target fragment, whose sequence we would like to determine. Under ideal conditions, this target molecule would hybridize to all words whose Watson–Crick complements occur somewhere along its sequence. Thus, in principle, one could determine in a single microarray reaction the set of all k-long sub-sequences of the target (this set is called the target's spectrum) and try to infer the sequence from these data.

SBH is greatly impeded by ambiguity in target reconstruction. Depending on k and on the target length, there may be several—or many—alternative sub-sequences, within any tested target, that have the same spectrum and are thus indistinguishable. Hence, spectrum data simply do not contain sufficient information to uniquely resolve targets of reasonable lengths (8,9). Alternative sources of information have been suggested to complement the spectrum data (10–12).

One possible source of complementary information for SBH is the reference sequence. Re-sequencing by hybridization (RSBH) is the task of reconstructing the target sequence using its spectrum and a reference sequence to which the target is similar. We recently developed a computational method, called spectrum alignment, for RSBH (13). Spectrum alignment uses a probabilistic representation for the reference sequence information and the spectrum signals. It computes the most likely target sequence given these data. Here we shall study the adaptation of spectrum alignment to handle real experimental data, and we validate the theoretical method with actual re-sequencing results.

An alternative strategy for array-based re-sequencing uses straight hybridization to specific, partial arrays. This strategy uses the probe array as a massively parallel assay for single

*To whom correspondence should be addressed. Tel: +972 8 9344907; Fax: +972 8 9344487; Email: peer@wicc.weizmann.ac.il nucleotide primer extension genotyping, and seeks a mutation or potential polymorphism at each site along the target. The probes on such an array are a set of *k*-mers that tile the target sequence and its variants. The advantage of this approach is that it does not require the complete set of *k*-mers to be used, allowing the use of fewer and longer oligonucleotides with reasonable cost. The disadvantage is that such arrays can only detect the pre-determined sequence variations for which they were designed. Moreover, a new array needs to be created for each new re-sequencing reference. Such arrays have been designed for re-sequencing, among others, the P53 tumor suppressor gene (14,15), HIV protease (16) the HLA locus (17) and other genomic regions (18). See Tillib and Mirzabekov (19) for a review of such methods.

The use of SBH in practice raises two conflicting issues. On the one hand, short oligonucleotide hybridizations are less specific and less accurate than longer ones. On the other hand, the number of probes in a universal array grows exponentially with the probe length, so feasible lengths are severely limited. While universal arrays of 9mers have been constructed (20), for universal arrays to be cost-effective for most applications, they need to be much smaller than these gigantic sets of $4^9 \approx 2.5 \times 10^5$ oligonucleotides. Other formats of arrays have been attempted, using the probes sequentially, rather than on one large array (21), or pooling subsets of the probes (22). The natural, generic strategy would be to use shorter probes, which allow the complete set of oligonucleotides to be cost-effectively used in each array (23). The reader is referred to reviews on SBH formats (24,25).

Small universal arrays must use very short probes, which worsen the hybridization specificity problem. In order to create a practical assay for re-sequencing by universal arrays one needs a probing assay which is as specific as possible even for 5mers or 6mers (23), and computational means to overcome the low specificity and high noise in the assay. In this work, we have used the polymerase signaling assay (PSA), a method for probing a target sequence for the presence of short oligonucleotides using enzymatic discrimination, based on single-nucleotide primer extension, giving better accuracy than hybridization alone (26–28). Even using this method, noise is a major problem. We therefore develop algorithms to increase signal specificity by computational means, and report the results of applying these algorithms on partial and universal PSA arrays.

MATERIALS AND METHODS

Target molecules

Target samples included 10 synthetic double-stranded DNA molecules of length 25–35 bp and 32 PCR amplicons of length 100–140 bp (see Table 1).

PSA

The PSA is an accurate method for parallel re-sequencing examination of polynucleotides up to several thousand base pairs long (26–28). PSA uses a glass slide, onto which probes are spotted in an arrayed fashion. Our plates included 192 spots each, where 16 spots are used as controls, and 176 represent a unique five-base probing sequence, representing 5mers and sequence variations specifically related to the target sequence being tested. Used for analysis of AGT exon 2 and CFTR exon 11, these experiments simplify the approach from the true 'universal array' of 5mers. A complete universal array, which may be used for analysis of any arbitrary sequence, has a unique five-base probe for each of the $4^5 = 1024$ possible pentanucleotide combinations. These larger arrays were constructed by using several sub-arrays. Each probe is a 29mer, comprising a capture motif, which is identical for all probes, and a probe-specific oligonucleotide pentamer. The capture motif, U_{18}, is separated from the attachment moiety by a spacer amidite (Spacer 18; Glen Research). The capture and probe motifs are separated by a segment of six bases of completely random content (i.e. the amidite source for these synthesis steps is made of a mixture of 25% each of A, C, G and T), such that this segment is a structural spacer, and should not play any role in the re-sequencing event. The probe-specific nucleotide combinations are designed to perfectly match every possible 5mer segment along a target.

Target molecules are PCR amplicons, one short (ideally 100–150 bases) amplicon for each of the target sequence, with an attached poly(A) tail, appended to the 5′ end of one of the PCR primers. Any reasonable set of PCR primers is acceptable, so long as the short amplicon can be achieved in good yield. After PCR, the double-stranded amplicons are degraded to a single strand, using a combination of a phosphorothioated PCR primer and an exonuclease, as described previously (14). When applied to the slide at 37°C, these tails bond to the poly(U) segment of the capture probe, and the entire target molecule is thus always in the vicinity of the specific probe. Whenever the target sequence contains a sequence that perfectly matches the specific probe, the complementary sequences can form a transient duplex. PSA also incorporates DNA polymerase and fluorescently labeled terminating nucleotide triphosphates, which allow single-base extension of the specific probes that have been matched. The presence of the enzyme stabilized the transient duplexes within a primed tertiary complex, and thus stabilization and primer extension occur as a linked biochemical process. The fluorescent slide is then scanned to detect the spots that signal a match between their probes and the target sequence. Exact details of this assay are described elsewhere (26–28).

Computational analysis: enhancements to spectrum alignment

Computational analysis is based on the spectrum alignment algorithm (13). We outline this method in brief and then describe the improvements made in this study in order to transform spectrum alignment into a useful, practical method, which could handle the real PSA signals.

The input to spectrum analysis comprises of a reference sequence in the form of a hidden Markov model (HMM) for a polynucleotide sequence, as well as the results of the array experiment. HMMs are a powerful method for describing sequence profiles and variations (29). Intuitively, it can be thought of as a sequence of states that describe match/mismatch, insertion, or deletion in the target with respect to the genomic reference sequence. The HMM registers probabilities for transition between those states, as well as probabilities of producing each nucleotide at each state. Results of the arrayed probing reactions are formulated as a

Table 1. List of targets

Data set	Experiment	Target Type[a]	Locus	From[b]	To[b]	Mutant[c]
1	1	A	AGT[d]	4078	4177	W
	2	A	AGT	4078	4177	ATG281ACG
	3	A	AGT	4078	4177	W
	4	A	AGT	4078	4177	ATG281ACG
	5	A	AGT	4078	4177	W
	6	A	AGT	4078	4177	ATG281ACG
2	1	A	CFTR[e]	107766	107863	GGA542TGA
	2	A	CFTR	107782	107891	GGT551G[G/A]T
	3	A	CFTR	107782	107891	CGA553TGA
	4	A	CFTR	107810	107917	AGG560ACG
	5	S	CFTR	107803	107827	GGA542TGA
	6	S	CFTR	107803	107827	W
	7	S	CFTR	107858	107881	W
3	1	A	CFTR	107766	107863	GGA542TGA
	2	A	CFTR	107782	107891	GGT551G[G/A]T
	3	A	CFTR	107782	107891	CGA553TGA
	4	A	CFTR	107810	107917	AGG560ACG
	5	S	CFTR	107834	107856	W
	6	S	CFTR	107834	107856	GGT551GAT + CGA553TGA
	7	S	CFTR	107858	107881	AGG560ACG
4	1	A	CFTR	107782	107891	GGA542TGA
	2	A	CFTR	107782	107891	GGT551G[G/A]T
	3	A	CFTR	107825	107914	CGA553TGA
	4	A	CFTR	107825	107914	AGG560ACG
	5	A	CFTR	107795	107893	CGA553TGA
	6	A	CFTR	107766	107863	W
	7	A	CFTR	107810	107917	GGT551G[G/A]T
	8	A	CFTR	107766	107863	GGA542TGA
5	1	A	CFTR	107766	107863	GGA542TGA
	2	A	CFTR	107766	107863	W
	3	A	CFTR	107782	107891	GGT551G[G/A]T
	4	A	CFTR	107782	107891	W
	5	A	CFTR	107825	107914	CGA553TGA
	6	A	CFTR	107825	107914	W
	7	A	CFTR	107810	107917	AGG560ACG
	8	A	CFTR	107810	107917	W
6	1	S	Ch 18[f]	44	78	Base 19 A→G
	2	S	Ch 18	44	78	W
	3	A	Ch 18	1	109	Base 62 A→G
	4	A	Ch 18	1	109	W
	5	S	CFTR	107803	107827	W
	6	S	CFTR	107803	107827	GGA542TGA

[a] Synthetic (S)/amplicon (A).

[b] Offset (bp) from translation start site (coding sequences) or from segment start (non-coding).

[c] Either the wild type (W) or a mutant, which is denoted by the original codon, codon number and new codon (coding sequences) or bp number with base change (non-coding). Samples that are heterozygous for a mutation are denoted by, e.g. [A/G].

[d] Genomic sequence at positions 769274–780916 of GI:27477742.

[e] Genomic sequence at positions 42296576–42485274 of GI:22050628.

[f] Genomic sequence at positions 136976–137084 of GI:18677476.

probabilistic spectrum, i.e. a pair $P_1(x)$ and $P_2(x)$ for each k-mer x, where:

$$P_1(x) = \text{Prob(observed signal for probe } x \mid x \text{ matches the target)}$$

and

$$P_0(x) = \text{Prob(observed signal for probe } x \mid x \text{ does not match the target)}$$

Evaluation of $P_1(x)$ and $P_0(x)$ from x's fluorescence signals is discussed below. For probes y that are absent from the array, $P_1(y)$ and $P_0(y)$ are set to reflect occurrence probability of y along putative targets that are randomized variants of the reference sequence.

The enhanced algorithm applies spectrum alignment iteratively. Each iteration re-evaluates the putative target sequence of a specific region along this sequence, assuming correctness of the rest of the reconstructed sequence. Putative incorrect regions are identified by their lower likelihood. This is done in order to correctly interpret probe signals which are positives, but are due to a match that occurred outside the focus region. For re-sequencing a specific, focus region we need to find the most likely chain of HMM states which starts and terminates with the oligonucleotides that match the ends of that region. This is computed by a dynamic program (30).

Table 2. Summary of data sets analyzed

Probe set	Number of probes	Experiments per data set	Data sets	Total number of experiments
Angiotensinogen tiling	176 unique	6	1	6
CFTR tiling	176 (166 unique)	7/8	2–5	30
Universal	1119 (1024 unique)	6	6	6

Per-probe training

$P_1(x)$ is evaluated as follows. When we have sufficient examples of fluorescent signals for the probe x with known positive match to some known target, we can evaluate the mean signal $\mu_1(x)$ for the matched probe, and its standard deviation $\sigma_1(x)$. A goodness-of-fit test does not reject the hypothesis that samples are normally distributed (data not shown). $P_1(x)$ is then set to the P-value for signal $s(x)$ to be drawn from a normal distribution with mean $\mu_1(x)$ and standard deviation $\sigma_1(x)$. Evaluation of $P_0(x)$ is done similarly. (In practice, we assume $\sigma(x) = \sigma_1(x) = \sigma_0(x)$ and evaluate $\sigma(x)$ based on the two sets of samples.)

When we do not have sufficiently many samples of positive/negative matches to the probe x, we enrich the occurrence count of positive/negative matches to the probe x by adding to it the count of another probe y, whose signals are similarly distributed, but whose counts are not sparse. For each candidate probe y we use its computed normal distributions, $N[\mu_1(y), \sigma^2(y)]$ and $N[\mu_0(y), \sigma^2(y)]$, to evaluate the likelihood of the observed (matched and unmatched, respectively) signals for x. The probe $y = y^*$ that maximizes this likelihood is chosen and its counts are added to those of x. The combined count is used to evaluate expectancies μ_1, μ_0 and standard deviation σ for x that together define the normal distributions of its matched/unmatched signals.

Probe-independent training

For learning the distribution in an unsupervised manner we employ an alternative strategy, which does not build on experience with previous assays, and does not fit the distribution for each probe. Instead, we utilize the distributions of signals with positively and negatively matched probes in the current data set.

Obviously, we do not know in advance for the current data set whether a probe is perfectly matched by the target or not, as the target is yet unknown. However, we can evaluate the probability of that event with respect to a random target that is similar to the reference sequence as much as the true target is assumed to be. Such random targets can be drawn using the HMM, which models the probabilistic space of such sequences. We thus generate $N = 100$ candidate targets, and average the matched/unmatched status of the probe x as follows. We empirically estimate the probability of a perfect match, based on all randomized targets, as the fraction of probes attaining a certain signal among perfectly matched probes:

$$P_1(x) = \frac{\sum\limits_{\text{random target } t} (N^{<}(t, s(x)) + 0.5N^{=}(t, s(x)))}{\sum\limits_{\text{random target } t} N^{<}(t, \infty)}$$

$P_0(x)$ is analogously estimated.

Figure 1. The fluorescence confocal microscope scan of a reacted universal array (data set 6, experiment 6). The array has 992 different unique probes, 32 duplicated probes and 96 positive and negative control probes.

RESULTS

The ultimate goal of this research is to establish the practicality of universal arrays for re-sequencing. The question that we address is whether one can make complete yet reasonably small arrays of probes, by limiting probe length, to accurately re-sequence DNA sequences of practical length. We thus performed a series of blind tests, in which the target sequence was unknown. One set of assays comprised of simple genotyping tests, where the target sequence was either the wild type or a single-nucleotide mutant thereof. Other assays were re-sequencing tests, wherein the target could have been any variant of the known reference sequence.

We first constructed partial, tiling arrays. Some of these arrays consisted of probes that tile variants of exon 2 of angiotensinogen, while others tile exon 11 of CFTR. We then constructed and tested complete, universal arrays. We used arrays of 5mer probes, for which only 1024 different oligonucleotides are needed (see Table 2 for the list of arrays used). Various target molecules were re-sequenced (Table 1). To obtain as much specificity as possible from these short probes, we applied the PSA protocol (see Materials and Methods). The image, a confocal fluorescence scan, of one such universal array is presented in Figure 1.

Arrayed PSA reactions produce data sets of raw fluorescent signals. When reconstructing a target sequence using spectrum alignment, the quantity of interest for each probe is the likelihood of a perfect match. More precisely, given the raw signal $s(x)$ for a probe x, one needs to compute the probabilities $P_1(x)$ = Prob[$s(x) \mid x$ is perfectly matched by the target] and $P_0(x)$ = Prob[$s(x) \mid x$ is not perfectly matched by the target]. Although PSA provides cleaner signals than hybridization, signals are still very noisy. The observed noise might be due either to stochastic effects, causing variation in replicate observations of the same intensity, or to hidden variables that distinguish between signals. As seen below, both factors contribute to the signal distribution, and we can exploit our knowledge of some hidden variables, such as individual probe differences, to improve signal analysis. Overall distributions of signals are presented in Figure 2.

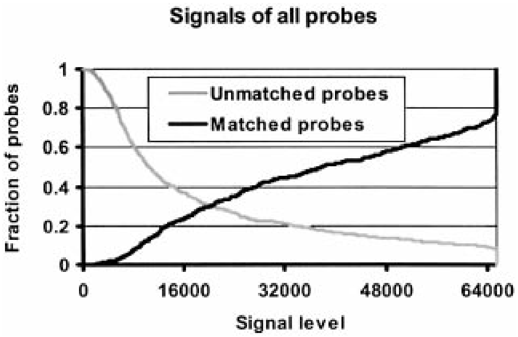

Figure 2. Signal level distributions for matched and unmatched probes. For each level of the fluorescent signal, the black curve displays the fraction of matched probes that produced at least this level of signal. The gray curve displays the fraction of unmatched probes that produced at most this level of signal. Any signal level set as threshold between positive and negative matches will incur a high error rate. Data were collected from data sets 2–6.

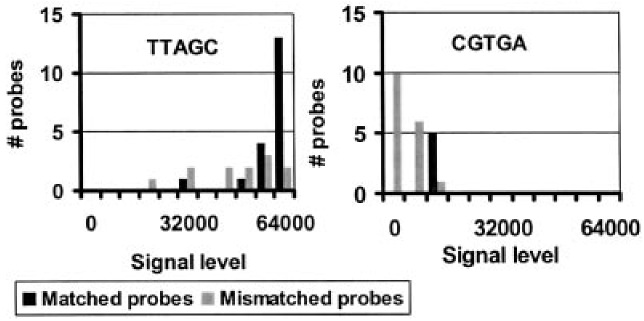

Figure 3. Different probes may have very different distributions. This figure shows signals of two specific probes: TTAGC, whose signals are extremely high, and CGTGA, whose signals are extremely low. For each level of the fluorescent signal, the number of matched (black bars) or unmatched (gray bars) probes that produced this level of signal is displayed. Every threshold rule for calling matched/unmatched by fluorescent signal level would either label all TTAGC probes as matched or all CGTGA probes as unmatched. Nevertheless, analysis of each probe individually separates positive versus negative signals much better. Data were collected from data sets 2–6.

These distributions, though obviously different, have a broad range of overlap. Consequently, a simple threshold value cannot effectively distinguish between matched probes and unmatched ones. Furthermore, even if we use the probabilities in Figure 2, for most of the signal range, the matched and unmatched probabilities are of the same order of magnitude. Thus, the log-likelihood term $\log[P_1(x)/P_0(x)]$ contributed by most probes is approximately zero, rendering the model statistically weak.

The weak separation of the P_0 and P_1 distributions can have two causes: either the individual per-probe distributions are separated weakly for most probes, or they are separated, and their superposition causes the weak separation. Fortunately, as exemplified by Figure 3, the latter case is in effect. For example, T-rich probes produce very high signals, due to the poly(A) capture probes used in PSA (see Materials and Methods). Therefore, negative signals for such probes would be deemed positive according to the overall signal distribution, which is a mixture of many different per-probe distributions (see Fig. 2). This suggests empirically estimating P_0 and P_1 on a per-probe basis. For each probe x, for each signal level s, we estimate the probability of observing a signal $s(x)$ under the assumption of a perfect match in the target sequence. We assume such signals are normally distributed, with a probe-specific mean and variance, providing the distribution of $P_1(x)$. The distribution of $P_0(x)$ is analogously estimated.

We study and test two scenarios. First, we present a method to estimate P_0 and P_1 from a single array only. Each of these two distributions is assumed to be the same for all probes, for lack of better information. We call this method probe-independent training. Secondly, in cases in which we have several arrays that were assayed using the same protocol, but with different target molecules, better analysis is possible: we estimate individual signal distributions for each probe under the approximate assumption that these arrays are replicates of the same experiment. We call this method per-probe training.

In probe-independent training, in the absence of any prior information on the signal distributions, we use the following approximation: we generate in simulations many random targets, which are variants of the reference sequence, and collect statistics on the signal distributions of matched and unmatched probes. In this manner, we attempt to model the statistical properties of the actual target sequence used in the assay, without having any further information about the actual biochemical outcome of known target variants (see Materials and Methods).

In per-probe training, one has several arrays that were assayed using a similar reference, but with different mutations. This is the case, for example, for each individual data set in Table 2, which used several arrays. This is also the case for all the data sets of the CFTR arrays that together constitute a much richer set. Thus, a number of experiments with extensive perfect match data are available. In order to resolve the target in a specific array, we train each probe using all other arrays with match/mismatch for the current probe. We pool the matched/unmatched signal levels for each probe from all arrays and obtain a richer distribution. When that distribution is not based on sufficiently many probe occurrences, we enrich that distribution by that of another, similar probe (see Materials and Methods). As samples accumulate, richer and richer training sets can be built and exploited this way, and the statistical confidence of any single experiment will increase.

The two training methods present a trade-off. Probe-independent training uses a rich, yet coarse, set of observations, and forms a distribution that may be not representative of the specific probe. The per-probe method uses a finer set of observations, which may be too small a sample, and thus overfit the estimated distribution. We also consider a similar trade-off with respect to the experiments used to learn the per-probe distribution. We compare results of analysis based on learning this distribution from the current data set only, with learning based on all data sets, or on all other data sets except the current one.

Figure 4 presents a comparison of the results obtained by each of the training methods. Per-probe analysis based on all other arrays is superior to probe-independent analysis based on the current data set only. In per-probe methods, there is a trade-off between training which is based only on the same data set and training on all data sets: the more refined, but sparser training per data set makes more false calls at known SNP sites, but reports less spurious false positives due to overfitting.

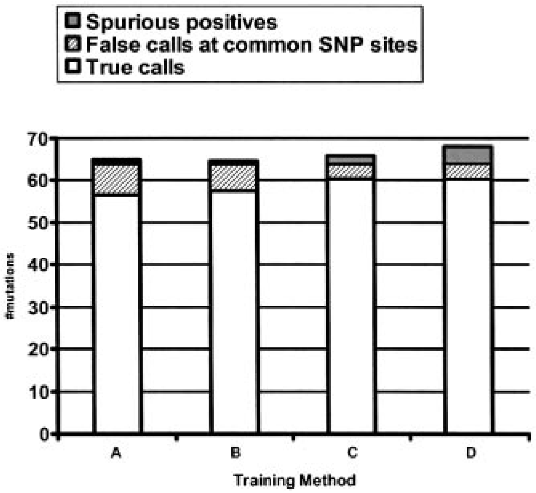

Figure 4. Re-sequencing performance using different training procedures. The training procedures are used for generating probe signal distributions in the spectrum alignment algorithms. Tests were performed on all the CF arrays (data sets 2–5). Bars represent the success rate of genotype calls. For a genomic bi-allelic amplicon target, we count a polymorphism as successfully typed if both predicted alleles match those present in the sample. Half an error is reported for each allele mismatch. Mono-allelic synthetic targets (arrays 5–7 in data sets 2 and 3) were all successfully typed and counted as one success each. (A) Probe-independent training based on the current experiment only (no prior data). (B) Per-probe training, using the current data set for probes with three or more matched and unmatched examples observed. For probes with fewer examples, an enrichment procedure is applied (see Materials and Methods). (C) Per-probe training using all data sets. (D) Per-probe training using all data sets except the data set that contains the target.

The estimated probabilities serve as input to the spectrum alignment computational engine. Table 3 presents results for blind tests of genotyping and for re-sequencing tests. For angiotensinogen exon 2, targets were either wild type or mutated for a specific polymorphism. The algorithm was not calibrated beforehand with any prior information regarding the identity of this polymorphic site, i.e. the reference sequence model was considered to have an equal likelihood to contain a mutation at any point along the target sequence. The genotype call on this site was correct for six out of six samples, and no spurious calls were made (although permitted by the algorithm). Analysis for arrays in this data set was carried out using probe-independent training. Although each of the 5mer probes may not necessarily give an entirely specific assay signal, their joint analysis using the spectrum alignment algorithm (13) utilizes all the statistical information available to produce a strong, combined signal.

Figure 5 presents results for the CFTR exon 11. For re-sequencing this target (with either partial or universal arrays), we used as reference not only the genomic sequence, but also known mutations from the Human Genome Mutation Database (www.hgmd.org). Altogether, in 30 arrays, we re-sequenced 2.6 kb of DNA. Out of 64 known polymorphisms, 60.5 were correctly typed (see Fig. 4 for details on the counting procedure), and two additional spurious mutations were falsely detected. This true-positive rate of 95% is to be contrasted with the 30% error rate introduced by pentamer biochemistry (Fig. 2). Observe that this analysis was carried out without any attempt to detect heterozygocity. While genotyping does require the detection of heterozygotes (see Discussion), we employed a first, simple approach to test the feasibility of our methodology, which ignored heterozygocity, and therefore technically counted heterozygotes as errors. Out of the 56 homozygotes, we had only one error.

A non-coding region on chromosome 18 was also re-sequenced by universal arrays (data set 6, arrays 1–4). For this target sequence we had no prior knowledge of the mutant sites. For this segment we missed one of the mutations in four re-sequenced targets of total length of 300 bp. Both CFTR targets assayed with universal arrays (data set 6, arrays 5 and 6) were successfully re-sequenced.

Although we account for per-probe signal effects by per-probe training, the major source of remaining error appears to be systematic bias, rather than stochastic effects between replicates: most of the failed genotypes involve the GGT551G[G/A]T mutation. Thus, apparently, averaging many experiments will not be helpful in eliminating such errors, but further understanding and modeling of the causes of such systematic bias may solve the problem.

The spectrum alignment algorithm was implemented on both Windows and Unix platforms. The implementation incorporates a refined analysis of heterozygote samples, although the results presented were analyzed without this feature. The heterozygotes analysis would obviously need to be added for full functionality. In addition, we implemented a visualization tool, called SNP-o-gram, for presentation of re-sequencing results. This Windows application displays the reference and re-sequenced target, along with plots that indicate the likelihood of each base call, similar to standard traces of gel-based sequencing machines. Figure 6 displays the SNP-o-gram of two re-sequenced targets.

DISCUSSION

Re-sequencing genomic information is a major task in today's biology, providing essential data for various research and diagnostics applications. The term 're-sequencing' implies that one has significant information on the reference, thus determination of the target sequence should avoid complete sequence determination *de novo*. A promising strategy for re-sequencing is the use of arrayed short probes. An array containing all possible probe sequences of a particular length can serve as a universal assay for all possible target sequences. In order to be economical, one should minimize probe number, and therefore probe length. However, shorter probes can reduce the accuracy of the assay, so robust assay conditions and analytical processes need to be developed in concert with this simplified array approach.

This study shows the practicality of re-sequencing by an array of probes and detection by polymerase-mediated single-base extension. A combination of the polymerase, which imparts a much higher specificity than DNA hybridization alone, and enhanced computational methods of the resulting data from the DNA array assay, make it possible to handle noisy signal incurred by the use of short probes. Our method is sufficiently versatile to handle different sequences, yet utilize the information from the reference sequence. The analytical method is applied here to actual assay data for the first time.

We report successful re-sequencing of 100 bp fragments using pentanucleotide probes. This suits several key applica-

Table 3. Genotyping results

Data set	Experiment	Re-sequencing call[a]	Correct	Log-likelihoods	
			genotypes	Wild type	Mutant
1	1	W	1	–205.847	–221.579
	2	ATG281ACG	1	–206.34	–204.01
	3	W	1	–206.37	–220.155
	4	ATG281ACG	1	–206.953	–198.819
	5	W	1	–205.631	–220.109
	6	ATG281ACG	1	–207.039	–198.845
2	1	GGA542TGA	1	–181.304	–175.85
	2	W	1/2	–204.646	–199.807
	3	CGA553TGA	1	–193.649	–192.389
	4	AGG560ACG	1	–213.685	–210.591
	5	GGA542TGA	1	–240.386	–236.305
	6	W	1	–216.133	–232.219
	7	W	1	–153.507	–171.118
3	1	GGA542TGA	1	–183.781	–177.255
	2	W	1/2	–219.014	–218.487
	3	CGA553TGA	1	–202.959	–197.73
	4	AGG560ACG	1	–208.909	–200.895
	5	W	1	–203.153	–224.416
	6	GGT551GAT + CGA553TGA	2	–186.667	–156.294
	7	AGG560ACG	1	–155.797	–141.592
4	1	W	0	–258.733	–261.182
	2	W	1/2	–198.028	–195.828
	3	CGA553TGA	1	–197.348	–193.998
	4	AGG560ACG	1	–203.601	–200.474
	5	CGA553TGA	1	–194.639	–192.439
	6	W	1	–178.632	–191.412
	7	W	1/2	–205.818	–246.755
	8	GGA542TGA	1	–243.99	–238.935
5	1	GGA542TGA	1	–248.925	–239.479
	2	W	1	–211.087	–222.238
	3	W	1/2	–246.786	–255.151
	4	W	1	–236.699	–248.77
	5	CGA553TGA	1	–212.363	–209.091
	6	W	1	–208.375	–208.806
	7	AGG560ACG	1	–221.538	–220.552
	8	CGA553CAA + AGA555AGC	0	–258.038	–255.874
6	1	Base 19 A→G	1	–1190.56	–1181.98
	2	W	1	–885.905	–906.477
	3	W	0	–907.967	–912.53
	4	W	1	–883.603	–899.166
	5	W	1	–766.15	–781.939
	6	GGA542TGA	1	–691.528	–686.091

[a] Most likely sequence haplotype. See Table 1 for details.

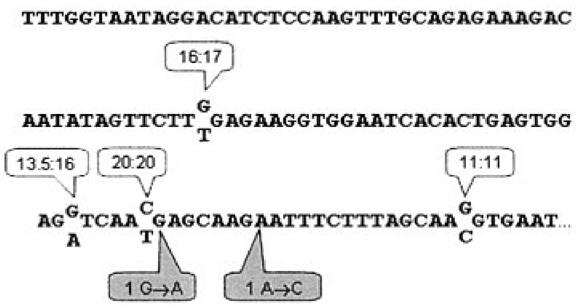
TTTGGTAATAGGACATCTCCAAGTTTGCAGAGAAAGAC
16:17
G
AATATAGTTCTT GAGAAGGTGGAATCACACTGAGTGG
T
13.5:16
20:20
11:11
AG G TCAA C GAGCAAGAATTTCTTTAGCAA G GTGAAT...
A
T
C
1 G→A
1 A→C

Figure 5. Summary of re-sequencing results for CFTR. The wild-type reference sequence is displayed along with call-outs for statistics on the typing of sites with potential mutations found at specific nucleotides. In total, 60.5 out of 64 mutations were correctly typed in common SNP sites (white call-outs). Two mutations were called in spurious sites (gray call-outs).

tions for re-sequencing, in which the sequence of a target exon, for example, may differ from its reference at many polymorphic or mutable sites. Such applications include genetic, diagnostic tests for highly polymorphic genes like CFTR that has over a thousand known mutations, many of them treatable upon proper diagnosis. An additional application involves detecting somatic mutations in onco-related genes. Accurate typing of pathogens can be also be achieved, by re-sequencing genes that are common to all candidate pathogens (e.g. 16S RNA).

This work uses mere 5mer probes, and achieves 100 bp of sequence. Practical prospects for universal arrays in large-scale re-sequencing probably depend on the ability to sequence fragments longer by an order of magnitude. One possibility would be to scale the probe length to include all-8mers, or even all-9mers, arrays that are feasible with some current industrial technologies. Indeed, our simulation studies (13) indicate that the feasible target length for re-sequencing approximately doubles when increasing by one the probe length in a universal array, even without taking into account the potential increase in accuracy due to longer probes. This increased accuracy is expected to enable improvement of the overall fidelity of the re-sequencing process.

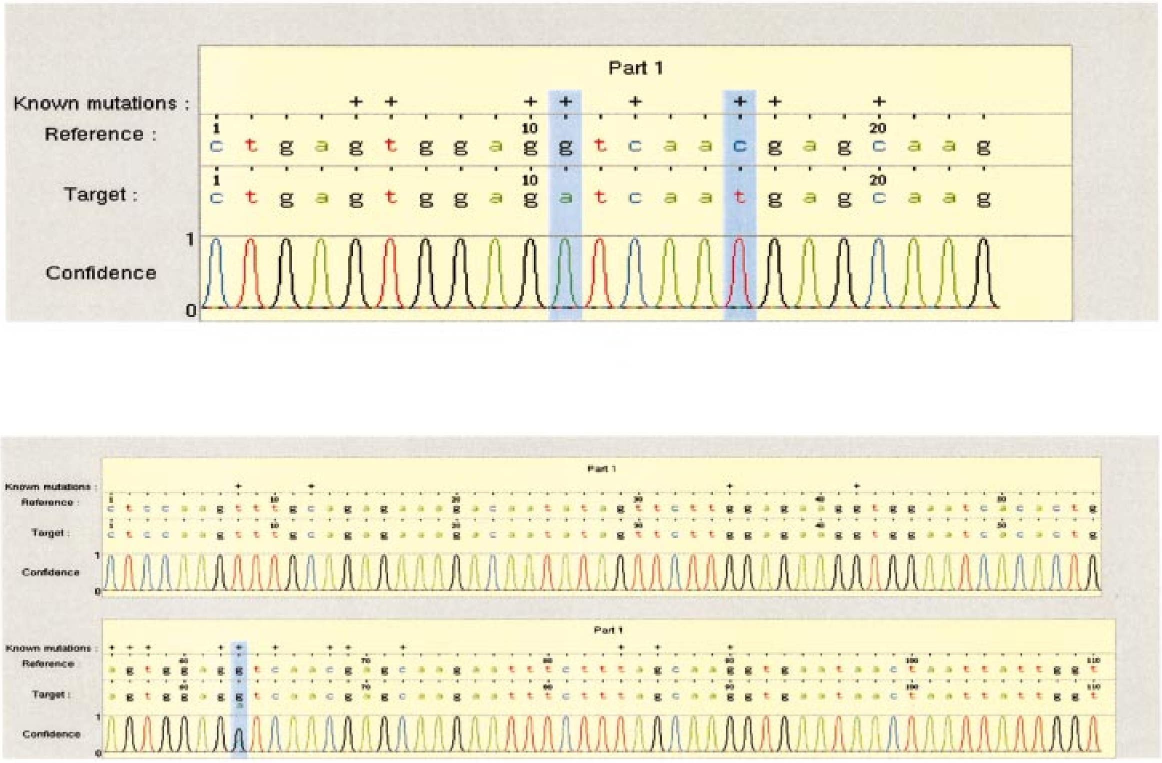

Figure 6. Visualization of re-sequencing results by SNP-o-gram. (Top) A synthetic short target, with two known mutations (array 6, data set 3). (Bottom) A genomic target which is heterozygous for a single known mutation (array 2, data set 4).

In these early studies, it was important for the biochemical process to maximize signal intensity over background. The potential use of longer probes could also go together with more stringent hybridization/extension conditions that would have the potential of reducing spurious biochemical outcomes. Newer, more intense and more sensitive detection molecules and scanning technologies would help access these likely weaker signals, and could push sensitivity well beyond the simple method of incorporation of singly labeled fluorescent nucleotides. Any of these alternatives could be used to increase accuracy, and would enable improvement of the overall fidelity of the re-sequencing process.

A different alternative is to use the 5mer re-sequencing technique to explore the small number of differences that might be the more likely goal in some re-sequencing studies. In this case, detection of small variations with respect to the reference sequence becomes far more important.

The important issue of detecting heterozygotes in this same manner was not addressed in this study. It is possible to achieve these goals within the spectrum alignment framework, in the following manner. In theory, one needs to find a pair of sequences, corresponding to a pair of paths in the spectrum alignment graph, that maximize the likelihood of the signals under the assumption of the two corresponding haplotypes. This likelihood is an expression which sums up individual edge contributions, very similarly to the standard homozygous score. In practice, the two haplotypes are expected to be quite similar to each other. Therefore, the two corresponding paths are intertwined, and often overlap in many edges. The resolution of one haplotype can be performed as in the homozygous case. We propose to continue and resolve regions where the two paths are distinct in a segment by segment fashion. When examining such a segment, one can look for potential heterozygocity by using the distilled spectrum machinery to filter out the spectrum of the first haplotype (13). As the current data set focuses on homozygote targets (37 out of 42), we could not adequately test this approach in this study, and this will be done in a future study.

The computational developments presented here are driven by the nature of errors incurred by the biochemical assay. Nevertheless, these methods are general, and apply also to data obtained by other kinds of assays. In particular, as probing technology improves in quality, cost and throughput, re-sequencing of longer DNA segments is expected to be practical soon. Computational methods such as the ones presented in this work may provide the means to handle such assays in the future.

ACKNOWLEDGEMENTS

The research was supported by NIST Advanced Technology Program grant number 70NANB8H4001. R.S. was supported in part by the Israel Science Foundation founded by the Israel Academy of Sciences and Humanities (grant 309/02). I.P. was supported by a fellowship from Israeli Ministry of Science.

REFERENCES

1. Yan,H., Kinzler,K.W. and Vogelstein,B. (2000) Genetic testing—present and future. *Science*, **289**, 1890–1892.
2. The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
3. National Institute of Health (2002) Large scale genotyping for the haplotype map of the human genome. Request for application HG-02-005.
4. Couzin,J. (2002) Human genome. HapMap launched with pledges of $100 million. *Science*, **298**, 941–942.
5. Southern,E. (1988) Patent GB8810400.
6. Drmanac,R. and Crkvenjakov,R. (1987) Patent YU 1987000000570.
7. Maceviks,S.C. (1989) Patent PS US8904741.
8. Pevzner,P.A. (1989) 1-Tuple DNA sequencing: computer analysis. *J. Biomol. Struct. Dyn.*, **7**, 63–73.
9. Pevzner,P.A. and Lipshutz R.J. (1994) Towards DNA sequencing chips. In Privara,I., Rovan,B. and Ruzicka,P. (eds), *Proceedings of the 19th Symposium on Foundations of Computer Science*. Springer, Berlin, pp. 143–158.
10. Frieze,A.M., Preparata,F.P. and Upfal E. (1999) Optimal reconstruction of a sequence from its probes. *J. Comput. Biol.*, **6**, 361–368.
11. Preparata,F.P. and Upfal,E. (2000) Sequencing-by-hybridization at the information-theory bound: an optimal algorithm. *J. Comput. Biol.*, **7**, 621–30.
12. Ben-Dor,A., Pe'er,I., Shamir,R. and Sharan,R. (2001) On the complexity of positional sequencing by hybridization. *J. Comput. Biol.*, **8**, 361–371.
13. Pe'er,I., Arbili,N. and Shamir,R. (2002) A computational method for resequencing long DNA targets by universal oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **99**, 15492–15496.
14. Head,S.R., Rogers,Y.H., Parikh,K., Lan,G., Anderson,S., Goelet,P. and Boyce-Jacino,M.T. (1997) Nested genetic bit analysis (N-GBA) for mutation detection in the p53 tumor suppressor gene. *Nucleic Acids Res.*, **25**, 5065–5071.
15. Ahrendt,S.A., Halachmi,S., Chow,J.T., Wu,L., Halachmi,N., Yang,S.C., Wehage,S., Jen,J. and Sidransky,D. (1999) Rapid p53 sequence analysis in primary lung cancer using an oligonucleotide probe array. *Proc. Natl Acad. Sci. USA*, **96**, 7382–7387.
16. Kozal,M.J., Shah,N., Shen,N., Yang,R., Fucini,R., Merigan,T.C., Richman,D.D., Morris,D., Hubbell,E., Chee,M. *et al.* (1996) Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nature Med.*, **2**, 753–759.
17. Guo,Z., Gatterman,M.S., Hood,L., Hansen,J.A. and Petersdorf,E.W. (2002) Oligonucleotide arrays for high-throughput SNPs detection in the MHC class I genes: HLA-B as a model system. *Genome Res.*, **12**, 447–457.
18. Cutler,D.J., Zwick,M.E., Carrasquillo,M.M., Yohn,C.T., Tobin,K.P., Kashuk,C., Mathews,D.J., Shah,N.A., Eichler,E.E., Warrington,J.A. *et al.* (2001) High-throughput variation detection and genotyping using microarrays. *Genome Res.*, **11**, 1913–1925.
19. Tillib,S.V. and Mirzabekov,A.D. (2001) Advances in the analysis of DNA sequence variations using oligonucleotide microchip technology. *Curr. Opin. Biotechnol.*, **12**, 53–58.
20. Gunderson,K.L., Huang,X.C., Morris,M.S., Lipshutz,R.J., Lockhart,D.J. and Chee,M.S. (1998) Mutation detection by ligation to complete n-mer DNA arrays. *Genome Res.*, **8**, 1142–1153.
21. Drmanac,S., Kita,D., Labat,I., Hauser,B., Schmidt,C., Burczak,J.D. and Drmanac,R. (1998) Accurate sequencing by hybridization for DNA diagnostics and individual genomics. *Nat. Biotechnol.*, **16**, 54–58.
22. Drmanac,R., Drmanac,S., Baier,J., Chui,G., Coleman,D., Diaz,R., Gietzen,D., Hou,A., Jin,H., Ukrainczyk,T. *et al.* (2001) DNA sequencing by hybridization with arrays of samples or probes. *Methods Mol. Biol.*, **170**, 173–179.
23. Lebed,J.B., Chechetkin,V.R., Turygin,A.Y., Shick,V.V. and Mirzabekov,A.D. (2001) Comparison of complex DNA mixtures with generic oligonucleotide microchips. *J. Biomol. Struct. Dyn.*, **18**, 813–823.
24. Drmanac,R. and Drmanac,S. (2001) Sequencing by hybridization arrays. *Methods Mol. Biol.*, **170**, 39–51.
25. Drmanac,R., Drmanac,S., Chui,G., Diaz,R., Hou,A., Jin,H., Jin,P., Kwon,S., Lacy,S., Moeur,B. *et al.* (2002) Sequencing by hybridization (SBH): advantages, achievements and opportunities. *Adv. Biochem. Eng. Biotechnol.*, **77**, 75–101.
26. Liu,Y., Hansen,E., Penney,R., Gelfand,C.A. and Boyce-Jacino,M.T. (2001) A universal assay for DNA sequence analysis and SNP genotyping. Poster presented at the 13th International Conference on Genome Sequencing and Analysis, San Diego, CA, USA.
27. Head,S.R., Goelet,P., Karn,J. and Boyce-Jacino,M. (2001) Patent US 6,322,968.
28. Head,S.R., Goelet,P., Karn,J. and Boyce-Jacino,M. (2002) Patent US 6,337,188.
29. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.
30. Pe'er,I. and Shamir,R. (2000) Spectrum alignment: efficient resequencing by hybridization. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 260–268.