HUMAN MUTATION 19:361–373 (2002)

REVIEW ARTICLE

Determining Sequence Length or Content in Zero, One, and Two Dimensions

Mark W. Perlin* and Beata Szabady
Cybergenetics, Pittsburgh, Pennsylvania

For the Mutation Detection 2001 Special Issue

High-throughput assays are essential for the practical application of mutation detection in medicine and research. Moreover, such assays should produce informative data of high quality that have a low-error rate and a low cost. Unfortunately, this is not currently the case. Instead, we typically witness legions of people reviewing imperfect data at astronomical expense yielding uncertain results. To address this problem, for the past decade we have been developing methods that exploit the inherent quantitative nature of DNA experiments. By generating high-quality data, careful DNA-signal quantification permits robust analysis for determining true alleles and certainty measures. We will explore several assays and methods. In a one-dimensional readout, short tandem repeat (STR) data display interesting artifacts. Even with high-quality data, PCR artifacts such as stutter and relative amplification can confound correct or automated scoring. However, by appropriate mathematical analysis, these artifacts can be essentially removed from the data. The result is fully automated data scoring, quality assessment, and new types of DNA analysis. These approaches enable the accurate analysis of pooled DNA samples, for both genetic and forensic applications. On a two-dimensional surface (comprised of zero-dimensional spots) one can perform assays of extremely high-throughput at low cost. The question is how to determine DNA sequence length or content from nonelectrophoretic intensity data. Here again, mathematical analysis of highly quantitative data provides a solution. We will discuss new lab assays that can produce data containing such information; mathematical transformation then determines DNA length or content. Hum Mutat 19:361–373, 2002.

KEY WORDS: nucleic acid; transformation; Fourier transform; Laplace transform; DNA sequencing; fragment sizing; electrophoresis; DNA chip; short tandem repeat; STR, dideoxy terminator; genetics; forensics; mutation detection; SNP

INTRODUCTION

High-throughput assays are essential for the practical application of mutation detection in medicine and scientific research. Ideally, such assays would produce informative data of high quality with a low error rate at an ever-decreasing cost. With computer hardware, Moore's law of exponential reductions in the cost-per-bit of information processing [Moore, 1965] has been confirmed over many decades. However, the DNA revolution has not experienced such benefits from advancing technology. Rather, the cost-per-bit of acquiring DNA sequence information has been relatively constant.

This elevated cost may be artificial (e.g., due to business practices that maintain inflated prices for reagents and equipment) or natural (i.e., due to inherent limitations of current DNA assay technologies). Regardless, our primary objective in this paper is to demonstrate how new experimental methods can greatly reduce the inherent cost-per-bit of DNA sequence information.

Our key contribution here is to replace the classical one-dimensional size separation experiments (e.g., gel electrophoresis or mass spectrometry) with scalable zero-dimensional assays that can be performed in a single test tube. Importantly, such zero-dimensional assays can be

*Correspondence to: Dr. Mark W. Perlin, Cybergenetics, 160 North Craig Street, Suite 210, Pittsburgh, PA 15213.
E-mail: perlin@cybgen.com

DOI: 10.1002/humu.10087
Published online in Wiley InterScience (www.interscience.wiley.com).

readily carried out on two-dimensional arrays. This massively parallel approach vastly reduces the reagent volume and physical readout space for acquiring each bit of information.

Note that throughout this paper, we make consistent use of the mathematical term "dimension." Specifically, dimension refers to the number of coordinates in the underlying domain of a data function. A point has no coordinates, hence zero dimensions, as in a test tube or spot measurement. Data that occurs on a line has one domain coordinate, such as an electrophoretic lane or capillary's single sizing dimension. Data measured on a two-coordinate plane, such as the row and column of a microtiter plate, or the (x, y) data locations on an image, have a two-dimensional data acquisition domain.

A measured signal intensity is a data function that takes values on the underlying n-dimensional coordinate domain (n = 0, 1, 2). This function maps the coordinate domain points of a data acquisition experiment into their observed data values.

This paper introduces the novel concept of mathematical transform sequencing of linear DNA molecules. The first section describes DNA sequencing from several perspectives. The second section sketches mathematical transforms, and enumerates some of the desirable properties that they have. The third section introduces new sequencing and sizing assays based on DNA transforms.

SEQUENCING

The ability to separate DNA molecules by their length has enabled the DNA sequencing and fragment analysis that underlies the genetic revolution. This section briefly reviews classical DNA sequencing and fragment analysis from a perspective that sets the context for introducing new assays.

There is a general procedure (shown in Fig. 1A) for elucidating information about a DNA sequence:

1. In nature, there is sequence information residing in a nucleic acid. This information is often used in vivo for biological information processing. In vitro, the scientist's task is to derive information about the sequence.
2. A scientist transforms the sequence by chemical synthesis into a set of labeled mol-

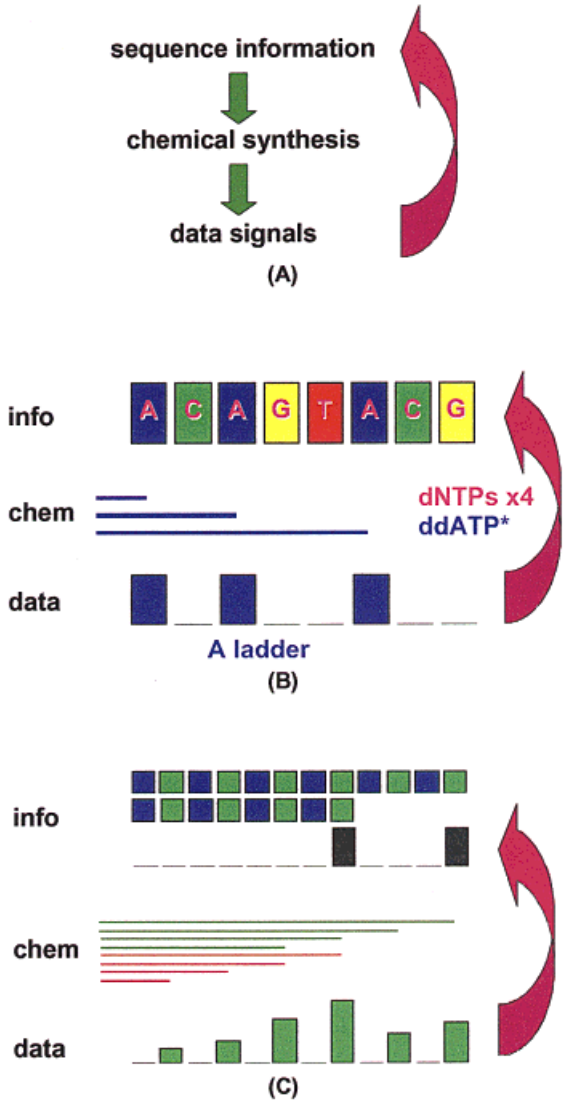

FIGURE 1. **A:** To characterize sequence information from DNA molecules, a chemical synthesis is performed, and data signals are produced from the labeled fragments. Analysis of these data signals recovers the unknown sequence information. **B:** In Sanger dideoxy terminator sequencing, the sequence information regarding the location of the A base is chemically synthesized into DNA fragments terminating at A. These fragments are then size-separated, and the positions of A-terminating fragments are observed in the data signals as a sequencing ladder. Since the ladder positions directly correspond to the sequence positions, the analysis is straightforward. **C:** In STR fragment sizing, two DNA fragments (from the maternal and paternal chromosomes) of unknown length are present in a sample. These are chemically transformed by PCR amplification into corresponding DNA fragments. Due to PCR artifacts, more than two fragments are typically generated. The fragments are observed in a size separation assay that produces a quantitative band for every fragment, with the data faithful to the fragment length and concentration. Analysis of these data then infers the sequence information, i.e., the length of the two alleles. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

ecules. Assaying these synthesized molecules is easier than studying the original nucleic acid template.
3. Via some assay, the scientist obtains data signals from the labeled synthesized molecules that are useful in reconstructing the sequence. Scientists perform a computer (or other) analysis of these signals to derive useful information about the DNA sequence.

Sanger Sequencing

The first automatable DNA sequencing technique was Sanger's method [Sanger et al., 1977]. Sanger sequencing uses dideoxy terminators to chemically synthesize a set of corresponding DNA fragments. Suppose that we want to know where one of the four base pairs, say adenosine (A), resides:

- Perform chemical synthesis on a DNA template molecule in the presence of all four incorporating precursor nucleotides (dNTPs), and the dideoxy precursor nucleotide ddATP. This operation will form DNA fragments that show early termination at just those locations where there is an "A" in the DNA sequence.
- Size separation (and detection) of these terminated fragments by gel electrophoresis will create a ladder of DNA fragment bands. The presence of a ddATP terminating ladder band at a gel-ladder position indicates that the nucleotide A occurs at the corresponding position in the DNA sequence. Conversely, the absence of an A band implies that a non-A base resides at the corresponding position in the DNA sequence.

For example, consider the small case (shown in Fig. 1B):

1. sequence information: ACAGTACG;
2. chemical synthesis: {A, ACA, ACAGTA};
3. Data ladder: 10100100.

The sequence information ACAGTACG for A is in one-to-one correspondence (via the identity transformation) with the chemical synthesis products {A, ACA, ACAGTA}. And the chemical termination products {A, ACA, ACAGTA} are in one-to-one correspondence (via the identity transformation) with the observed data ladder 10100100. Therefore, the inverse operation (proceeding from the observed data ladder to infer the unknown sequence information) is quite simple, since the inverse of the identity is again the identity.

A characteristic function indicates which elements do ("1") or do not ("0") belong to a set or a sequence location. So, proceeding inversely back from the data (using the identity), the data ladder 10100100 is the characteristic function of the lengths of the terminated fragments {A, ACA, ACAGTA}. And, continuing back from the fragments (using the identity), the characteristic function 10100100 provides the sequence information describing where base A appears in the sequence ACAGTACG.

The situation is so straightforward here that describing the "identity transformation" and "inverse operation" and so on add little to understanding Sanger's beautiful method (which remains the cornerstone of modern DNA sequencing). However, in later sections we shall encounter more complex mathematical transformations of sequence into data, and these basic concepts will assume a greater utility.

STR Sizing

Microsatellite markers, or short tandem repeats (STRs), are an abundant class of polymorphisms in mammalian genomes that have importance in genetics, forensics, and medicine [Weber and May, 1989]. At each marker location there are two alleles, each corresponding to the maternal and paternal chromosomes. The standard PCR assay for STRs comprises a labeled forward primer, and an unlabeled reverse primer, in principle generating two fragments (one for each allele). However, in practice, PCR artifacts (such as stutter/slippage and peak imbalance) produce a more complex pattern of multiple DNA fragments (shown in Fig. 1C). After electrophoretic separation of the DNA fragments, the quantitative pattern is analyzed to infer the two underlying alleles.

Automated Analysis

With the advent of highly quantitative DNA analysis instruments (e.g., the automated fluorescent DNA sequencer), automated analysis of quantitative fragments is now possible. A key example is the large-scale automation of DNA sequence data and its assembly into larger con-

tiguous regions. Such computer-led automated analysis enabled the sequencing of the human genome.

Another application is the automated allele calling of STR data [Perlin et al., 1995]. By precisely modeling the expected PCR-allele patterns, a computer can examine the quantitative STR data and mathematically deconvolve the complex pattern back into the two peaks of the underlying allele fragments (shown in Fig. 1C; right upward arrow). This inverse transformation "stutter deconvolution" has been used effectively in genetic [Pálsson et al., 1999] and forensic [Perlin, 1999] projects.

In forensics, it is common to encounter crime stains that contain mixed DNA from more than one individual (e.g., rape-kit evidence). In such cases, it is useful to determine the DNA profile of an unknown suspect contained in the mixture. Highly quantitative STR data can be generated from the crime stain, via PCR and read out on an automated DNA sequencer. Then, computer-based "mixture deconvolution" can solve the inverse transformation problem to mathematically extract the DNA profile of an unknown perpetrator [Perlin and Szabady, 2001].

Other Gel-Free Methods

Sequencing by hybridization. There are DNA sequencing methods that do not use size separation. One such approach is "sequencing by hybridization" (SBH), which probes arrayed DNA sequences with oligonucleotides in order to ascertain information about the sequence [Drmanac et al., 1993; Southern et al., 1991]. Hyseq's system probes oligos against arrayed samples, whereas Affymetrix chips [Fodor et al., 1991] probe the sample against arrayed oligos. SBH works best with known sequence variations (e.g., gene mutations) for which a set of informative oligos can be manufactured. The gene chips may have less utility when more flexible DNA sequencing is required, or in resolving complex mixtures that contain closely related gene sequences.

Sequencing by synthesis. Another gel-free approach is adding one base to a nascent DNA strand, detecting which base was added, and then repeating the process (synthesis + detection) until the sequence is determined [Cheeseman, 1994; Metzker et al., 1994]. There is a new commercial variation "pyrosequencing" in which each step fills in the appropriate nucleotide for its full extent in the template [Ronaghi et al., 1996]. These potentially powerful methods suffer from an instrumentation constraint: the biochemical synthesis and the physical detection must be combined into a single complex DNA sequencing device. Decoupling the two processes might permit the use of simpler off-the-shelf instrumentation, and allow more parallelization at a lower cost.

TRANSFORMS

While undeniably powerful, current sequencing methods do not scale up to miniaturized, inexpensive DNA assays. Indeed, the cost of genetic information has stayed relatively constant over the past decade at about $1.00 per bit. (A "bit" can be defined here as the information content of a DNA sequencing or fragment-sizing experiment.)

A worthy societal goal is widespread dissemination of genetic information technology to many application areas, thereby reducing crime, hunger, and disease. Therefore, our technological goal is to achieve massive parallelism in genetic experimentation, reducing the cost to under $0.01 per bit. For this to happen, entirely different (and scalable) DNA assays are required. Our approach is to base such new assays on mathematical transformation of DNA sequence information.

Visual Definitions

For a detailed introduction to mathematical transforms, the reader is referred to more specialized texts [Papoulis, 1962; Rudin, 1974]. For our purposes, a visual introduction to the transform concept will suffice.

Consider a mathematical function, such as a DNA sequencing ladder for one base, "10100100." Here, the function f(t) defined by the ladder is: $f(1) = 1$, $f(2) = 0$, $f(3) = 1$,…, $f(8) = 0$. Now, consider a second function g(s, t), which is a decay curve. At every value of t, we define the decay function $g(s, t) = 2^{-st}$, where s is a decay constant; with larger s, g decays faster over t. For $s = 0.5$, the decay curve g(0.5, t) is shown overlain on the ladder function f(t) (Fig. 2A).

As the curve g becomes smaller with increasing t, so too will the product of the numbers f(t) and g(s, t). Multiplying f(t) with g(s,t) at all ladder locations t, and adding up the product terms, gives one number, the sum:

$$h(s) = \sum_{t=1}^{n} f(t) \times g(s,t)$$

This sum, h(s), is called the "transform" of the function f(t) with respect to the kernel g(s,t). For example, with s = 0.5, the transform value h(0.5) is 1.1857 (Fig. 2B). And, with s = 1.0, the steeper decay curve g(1,t) gives the smaller transform value h(1.0) of 0.6406. This is because the smaller coefficients of the more rapid decay curve include less of each ladder signal where f(t) = 1 (as shown in Fig. 2C), and so the total sum of the products is reduced as well.

Transforms have very useful information properties, which enable their application to DNA sequencing. After describing (and naming) some common transforms, we will enumerate some of these properties.

Fourier and Other Transforms

Transforms are ubiquitous in biology, nature, science, technology, and engineering [Bracewell, 1978]. For example, the Fourier transform provides a frequency representation of data signals. The human senses are based on it. The cochlea in our ear transforms sound waves into frequency components: we hear in frequencies, not in temporal wave signals. Our retina and brain build a frequency representation of image: we visually process in frequency components, not in spatial images. Even our sense of touch is based on a neuronal Fourier analysis. And much of our everyday technology (telephone signal encoding, magnetic resonance imaging, etc.) is engineered around the Fourier transform.

Fourier transforms represent signals by breaking them down into their component frequencies (Fig. 3A). Most any data signal can be represented this way, even DNA ladders. The Fourier transforms of each of the four DNA ladders from the sequence ACAGTACG above are shown (color coded by base in Fig. 3B).

The Hadamard transform is useful in coding (and decoding) signal processing applications. Interestingly, its mathematical matrix is very simple, containing only positive and negative ones (as shown in Fig. 4A). The Laplace transform, based on varying decay curves, is used throughout engineering (Fig. 4B).

We have developed DNA transform sequencing methods using all of these mathematical transforms. However, in the remainder of this

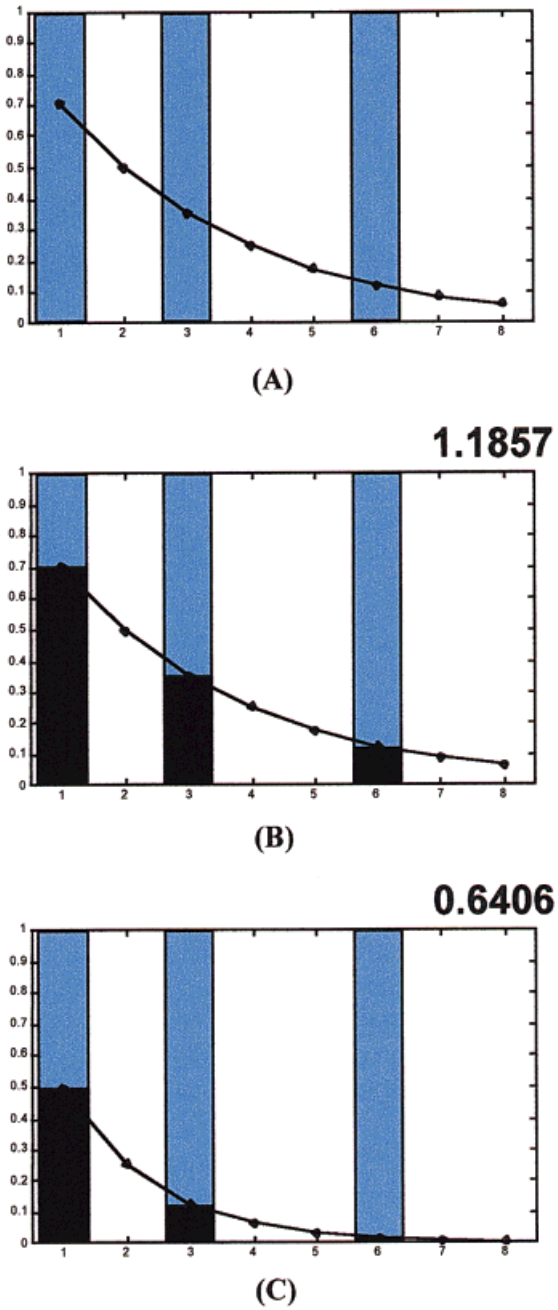

FIGURE 2. **A:** The positions of the A base in the sequence (i.e., the A ladder) are shown as light bars. This ladder can be viewed as a characteristic function, taking values 0 or 1. Overlain upon the sequence is a decay curve. **B:** Decay constant = 0.5. In the presence of a decay curve, there is a reduced amount of potential signal at every position. The value of the ladder (0 or 1) is multiplied at every position by the value of the decay curve. Adding up all these product terms (the 1 value times the decay value at the position) produces a single number. Here, the sum of the products equals 1.1857. **C:** Decay constant = 1.0. With a different decay curve, there is a different amount of signal observed at each ladder position. Here, the more rapid decay curve reduces the sum to 0.6406. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

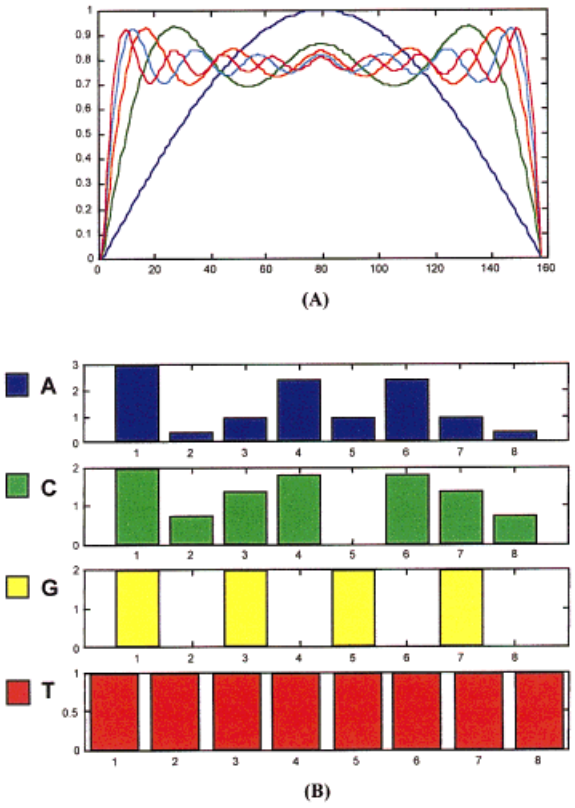

FIGURE 3. **A:** The Fourier transform (FT) is found throughout nature, science, and engineering. The FT describes the frequency content of a signal. Interestingly, the information contained in a signal representation, and its corresponding frequency representation, is identical. The FT permits effortless mathematical transformation back and forth between the two equivalent representations. **B:** The discrete FT of the sequencing ladder "ACAGTACG" is shown for each of the four component bases. The information content of each base's FT is identical to that of the original base ladder, and can be used to mathematically determine the original DNA sequence. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

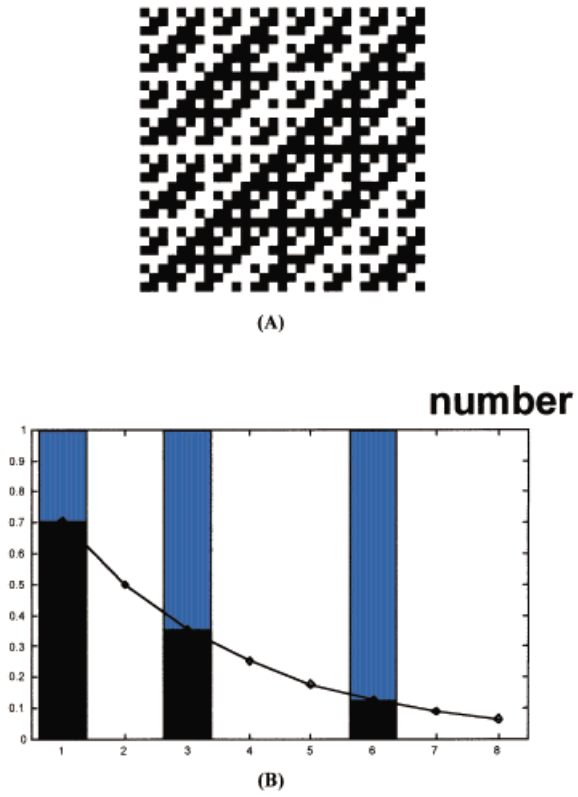

FIGURE 4. **A:** The Hadamard transform is created from only +1 or –1 values, and provides a stable way to code and decode sequence information. The pattern of pluses and minuses for a Hadamard matrix is shown. **B:** The Laplace transform is formed by applying all possible decay curves to a signal. Each decay curve produces a single number. From these numbers, one can recover information about the original sequence. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

paper, we shall focus solely on Laplace transform implementations.

Information Properties

Mathematical transforms can have many very useful information-processing properties. Here are some key properties, elaborated in the context of DNA:

Equivalent information. Suppose a function takes values at n points. (Think of a 0/1 DNA ladder representing n fragment lengths.) Then the data from certain n transform experiments contain equivalent information to the original function. That is, we need never interrogate the original function. We can infer the function (e.g., the DNA sequence) entirely by analyzing the transform experiments. Proceeding backward from transform data values, to infer the original function, is known as an "inverse transform."

Information reduction. If a function takes values at n points, then performing only k transform experiments (where k is much less than n) can provide the information we seek. For example, in some situations we can estimate a mean value in one transform experiment, and a variance value in a second experiment. If all we desire is the mean and variance of a distribution (e.g., for a Gaussian model), then just $k = 2$ experiments suffice for our desired level of knowledge of the function. This property can, in some circumstances, greatly reduce the required number of experiments.

Linear additivity. For transform T, this property means that $T(a + b) = T(a) + T(b)$. For

example, consider PCR amplification and sequencer detection as a transformation "STR," with {a, b} as DNA templates. Then, within a large linear range, STR(a + b) = STR(a) + STR(b). This linearity property lets us arithmetically combine data, thereby exploiting highly quantitative experiments on DNA molecules and their fragment products.

ASSAYS

We would like to obtain DNA sequence and length information without performing a size-separation experiment. The usual Sanger size-separation approach implements a simple identity transform. While easy to interpret, this identity operation imposes certain costs and constraints. Therefore, a different mathematical transform is needed, one that can enable a zero-dimensional (e.g., test tube) DNA assay. Such 0-D assays are scalable to two-dimensional massively parallel DNA array experiments.

Terminators Induce Laplace Transform

A typical DNA assay experiment proceeds in three stages: 1) amplify a sequence of DNA template, usually by using PCR; 2) extend new sequences, for example, a template-directed polymerase extension done in presence of certain nucleotides and analogs; and 3) detect the labeled fragments, e.g., by quantitative fluorescence measurements.

Consider dideoxy Sanger-style sequencing of the four component ladders. Here, the extension is done in the presence of all four dNTPs, along with a small amount of labeled dideoxy analog precursor "ddNTP*" for every base. And, the fluorescent (or other) detection follows an initial electrophoretic size separation.

What is the probability of extending the nascent DNA chain at every base addition step, rather than terminating? It depends on the quantity of ddNTP* used:

$$p = \frac{[dNTP]}{[dNTP] + \alpha_N [ddNTP^*]}$$

where α_N accounts for the incorporation efficiency of each ddNTP* relative to its dNTP, so that α_N[ddNTP*] is the effective concentration of ddNTP*. At higher concentrations of ddNTP*, a noticeable decay curve is introduced into the set of terminated fragments (Fig. 5). The greater the concentration [ddNTP*], the more rapid the decay, and the less the amount of observed label from each fragment.

Now, if the fragments were not size separated, then a total amount of label would be detected from the set of all fragments in the tube. The

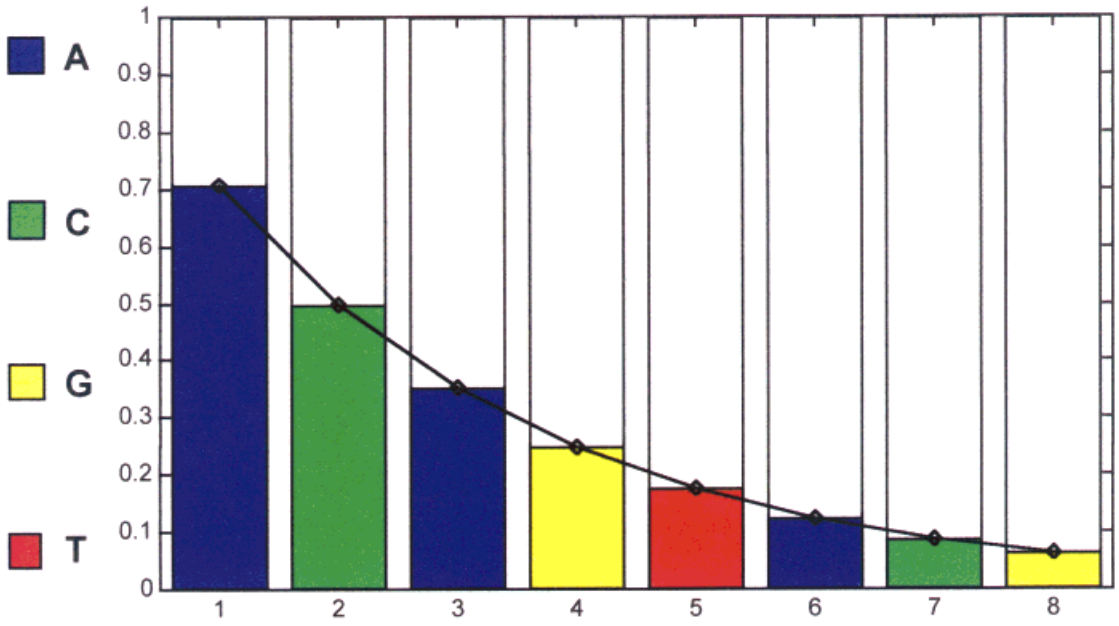

FIGURE 5. **The relative amounts of terminated fragments produced for the DNA sequence "ACAGTACG" in the presence of ddNTP*, with decay constant s = 0.5 (extension probability p = 0.707). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]**

greater the concentration of dideoxy terminators [ddNTP*], the more rapid the decay, and the less the amount of total observed label from all the fragments. But this total detected amount (in the presence of a decay curve) is precisely the Laplace transform. The detected value g(s) depends only on the actual DNA sequence f(t), and on the decay factor s.

(Note that the decay factor s is mathematically determined by the extension probability p via the relation: $s = -\log(p)$, e.g., a low extension probability creates a very rapid decay.)

Therefore, useful information can be obtained from the DNA sequence f(t) without performing any size separation: the sequence f(t) does not need to be directly observed. Any desired decay factor s can be achieved by setting the incorporation probability p via the appropriate set of effective dideoxy terminator concentrations $\{\alpha_N[\text{ddNTP*}]\}$. Different labels are used for each base. By performing multiple experiments over a range of different decay factors, the experiments create Laplace transform values g(s) for the function. Inverse transformation (or other methods) can then obtain use this set of data values {g(s)} to derive useful information about the content of the DNA sequence f(t).

The terminators are used to introduce a certain amount of decay, or "friction" into the template-directed extension reaction. Therefore, we dub the method "friction sequencing."

Friction Genotyping

For STR-fragment-length analysis, we can adapt the experiment to detect the number of repetitive units in each allele fragment. For example, with a "CA"-repeat unit, the extension can be done in the presence of all four dNTPs, along with varying quantities of labeled dideoxy adenosine analog precursor "ddATP*." Importantly, no electrophoretic size separation is required.

We describe here some initial "friction genotyping" experiments that we conducted using synthetic CA-repeat STR-like templates having 1, 2, and 3 repeat units (essentially, $(CA)_nG$; n = 1, 2, 3). We employed capillary electrophoresis to carefully scrutinize the generated peaks, and quantitate the peak components. The laboratory details are provided in the Appendix section.

Incorporation efficiency. Using a $(CA)_1G$ template, we determined that a ddATP:dATP ratio of 2:1 (i.e., 100 pM ddATP and 50 pM dATP) roughly corresponds to an extension probability of 0.5. This was done by checking for roughly equal heights (in the 5′ strand end NED dye (PE Biosystems, Foster City, CA)) of the $(CA)_1$-ddATP-terminated product and the $(CA)_1G$-dATP-product, as shown in Figure 6A.

Friction experiments. For the key experiments, we performed 18 reactions. We used three (approximate) extension probabilities: p = 0.25 (300 pM ddATP), 0.50 (33 pM ddATP), and 0.75 (33 pM ddATP). These were done for all six possible genotypes (two alleles selected from three choices), using the template combinations: 1, 2, 3, 1+2, 1+3, 2+3, where "n" denotes the template for $(CA)_nG$, and "m+n" denotes equimolar quantities of the $(CA)_mG$ and $(CA)_nG$ templates.

Peak data. The multicolor electrophoretograms are shown for a homozygotic genotype (template 2, $(CA)_2G$) experiment in Figure 6B, and for a heterozygotic genotype experiment (templates 1+2, $(CA)_1G + (CA)_2G$) in Figure 6C. The peak heights were tabulated for each dye from the GeneScan data, and used as estimates of DNA concentration.

Unique signatures. For each experiment, the ratio of the JOE (3′ terminator) signal to the NED (5′ strand) signal was computed from the fluorescent data. For a single DNA fragment, this ratio decreases exponentially with the fragment length. For two fragments, the ratio can be predicted by theory or calibrated from the data. For each STR genotype, these ratios recorded for different ddATP friction experiments can be used as a signature for calling the genotypes. The signatures of the six genotypes in this test system are shown in Table 1A.

Distinguishability. How distinguishable are the cluster signatures from each other? To determine this, we computed the Euclidean distances between all signature pairs. The results shown in Table 1B indicate that one can distinguish the signatures from one another, and thereby robustly call the genotypes.

Linear additivity. A useful check on our data is how well it conforms to our linear transform model. For example, we predict (and observe) that the heterozygotic genotype curve in Figure 6C can be formed by adding together the curves

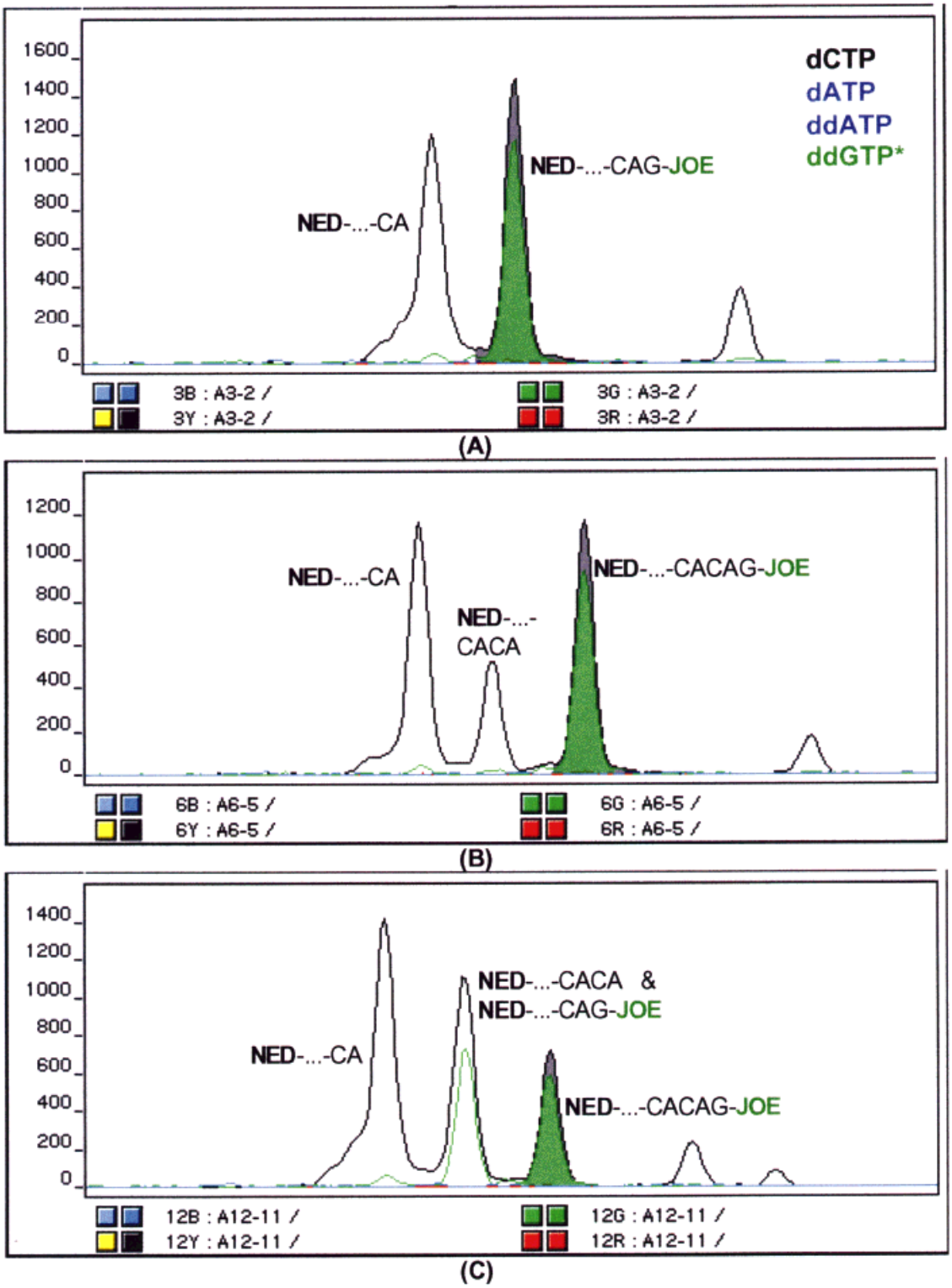

FIGURE 6. **A:** ABI/310 readout of the sequence extension of the $(CA)_1G$ template using 100 pM of ddATP relative to 50 pM of dATP. The 5′ strand end label (NED, shown in black) shows that the two peaks have roughly equal height. **B:** ABI/310 readout of the sequence extension of the $(CA)_2G$ template using 100 pM ddATP and 50 pM dATP. **C:** ABI/310 readout of the sequence extension of the combined $(CA)_1G$ and $(CA)_2G$ templates using 100 pM ddATP and 50 pM dATP. This signal combines the signals from the individual alleles. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

TABLE 1. Friction Genotyping Signature Analysis Results

A. Each column is the signature observed for a unique pair of DNA fragment lengths

ddATP	11	22	33	12	13	23
33	0.6180	0.5899	0.5680	0.6178	0.6158	0.6149
100	0.4337	0.3294	0.2753	0.4095	0.3773	0.3135
300	0.2147	0.1100	0.0655	0.1658	0.1434	0.0847

B. The pairwise Euclidean distances between the genotype signatures

L2 norm	1	2	3	1+2	1+3	2+3
1	0.000	0.151	0.223	0.055	0.091	0.177
2	0.151	0.000	0.073	0.102	0.064	0.039
3	0.223	0.073	0.000	0.175	0.137	0.064
1+2	0.055	0.102	0.175	0.000	0.039	0.126
1+3	0.091	0.064	0.137	0.039	0.000	0.087
2+3	0.177	0.039	0.064	0.126	0.087	0.000

C. For each heterozygotic allele pair, its observed signature is shown (left) together with the average (right) of the two observed signatures of its component alleles

ddATP	12	(11+22)/2	13	(11+33)/2	23	(22+33)/2
33	0.6178	0.6039	0.6158	0.5930	0.6149	0.5789
100	0.4095	0.3816	0.3773	0.3545	0.3135	0.3024
300	0.1658	0.1623	0.1434	0.1401	0.0847	0.0877

of the homozygotic genotypes of Figures 6A and 6B. This hypothesis can be tested by comparing each observed heterozygote signature with the average of the observed signatures of its homozygote components. These comparisons are shown in Table 1C. The observed data are consistent with linear additivity.

PCR artifacts. PCR stutter and relative amplification artifacts can complicate STR allele determination. However, our previous work on stutter deconvolution showed that such data artifacts are reproducible. That is, the expected value of the distorted peak distribution can be calibrated as a linear matrix effect, with the linear correction incorporated into the analysis. Since the PCR effect is linearly additive, these changes to the peak signal f(t) do not alter the linear Laplace transform mathematics.

Summary. Much information can be computed from this test data set. We determined the relative efficiency α of ddATP incorporation to be 0.41, relative to dATP; we were able to estimate the extension probability p for each ddATP amount used; and we were able to check other model assumptions against the data. This compatibility of data and model here suggest that our DNA transform sequencing approach might eventually develop into a useful and robust gel-free DNA sequencing assay.

High-Density Array Formats

Why devise a zero-dimensional DNA assay that can be carried out in a tube or well? So that the assay can be miniaturized, and scaled up with vast numbers of inexpensive experiments conducted in parallel. This scale-up can be approached in the near-term using two-dimensional microtiter plates. For long-term gains, a two-dimensional dense array format (i.e., DNA chip) would work well.

A format I array has the PCR products of many samples spotted onto an array [Lehrach et al., 1990; Pevzner and Belyi, 1997]. This format is useful when thousands of samples are available in advance—as in large genetic or forensic database projects. For each marker, several friction transform experiments are done across the entire array. For each array-based transform experiment, a set ratio of terminators is used to determine the decay factor. Quantitative fluorescent detection on a flat-bed laser scanner acquires the data in parallel from all the samples at once. Subsequent computer analysis then applies the inverse transformation to recover information (content, length, etc.) about the unknown DNA sequences.

A format II array [Fodor et al., 1991], conversely, has the marker DNA (and possibly the different decay reagents) gridded at different locations on the array surface. Probing a single PCR sample against the array can then answer many questions simultaneously about DNA sequences or alleles at multiple chromosomal locations in the sample. This would be useful, for example, with a fixed set of markers and one unknown sample, as in forensic analysis.

CONCLUSION

The high cost of genetic information limits current research and expectations for clinical application. The total data acquisition cost for a DNA fragment sizing experiment is about $1.00 for each genotype, a dollar per bit. Similar costs are incurred with gene sequencing for mutation analysis. For large-scale efforts (e.g., gene discovery or population screening) these costs all but prohibit rapid progress. In cancer genetics, this high cost-per-bit limits the widespread use of assays for genetic polymorphism and mutation detection.

A major cost factor in DNA sizing assays is their current reliance on one-dimensional (1-D) size separation technologies. These assays use the "lane" as the readout pathway. However, there are practical limitations on the degree of multiplexing within each lane, as well as on the number of lanes per run. Recently, DNA arrays comprised of a 2-D arrangement of 0-D dots have been used to replace certain DNA size-separation assays. By packing in many dots, these arrays can provide a vast increase in data density, relative to lane-based methods. When the biochemistry can be performed directly on the array surface, this density can translate into an equivalent reduction in the genetic cost-per-bit.

We have developed novel methods for characterizing DNA fragments based on "DNA transform sequencing." Our approach exploits the chemistry of DNA sequencing to obtain numerical values that provide information about the sequence. Friction genotyping can be used to size DNA fragments in a 0-D "lane-free" format, without performing a size separation. Friction sequencing can be used for interrogating DNA content. Our gel-free methods 1) enable massively parallel array-based DNA analysis, 2) decouple the biochemistry from the signal detection, and 3) may provide orders of magnitude cost reduction relative to current assays in certain applications.

APPENDIX

Friction Genotyping Methods and Materials

Templates. To test out this DNA-sizing approach in the laboratory, we designed an experiment that used synthesized CA-repeat oligonucleotide templates. The three templates contained $(GT)_n$, n = 1, 2, 3, and were 5′ biotinylated for purification steps. The sequencing primer was fluorescently labeled (NED dye, PE Biosystems) on the 5′ end in order to estimate quantities related to the number of DNA strands. A poly-A tail was added for better sequencer detection. The sequences used were:

5′-NED-A_{10}-GTTTTCCCAGTCACGA-3′
3′-CAAAAGGGTCAGTGCT-
$(GT)_n$-CCAA-Biotin-5′

Extension from the sequencing primer forms a $(CA)_n$ subsequence, followed by a G. We shall loosely refer to the biotinylated "…GCT-$(GT)_n$-CCA…" template by its complementary "$(CA)_n$G" name.

Nucleotides. In the Sequenase (USB, Cleveland, OH) extension reaction, we used the nucleotide precursors: dCTP; dATP and ddATP (Amersham, Piscataway, NJ), in predetermined ratios; and ddGTP-JOE, labeled with the fluorescent JOE dye (NEN Life Science Products, Boston, MA). The ddATP:dATP ratio was set to achieve a desired extension probability p. No TTP precursors were used. Thus, sequence termination could occur by either ddATP, which prematurely terminated the $(CA)_n$G sequence; or ddGTP, which labeled and terminated the full-length $(CA)_n$G sequence.

Extension products. The result of a sequencing reaction is a set of 5′ labeled molecules (n = 1,2,3): 5′-NED-A_{10}-GTTTTCCCAGTCACGA-$(CA)_n$-3′, along with a full-length molecule labeled at both the 5′ and 3′ ends: 5′-NED-A_{10}-GTTTTCCCAGTCACGA-$(CA)_3$-G-JOE-3′. The ratio of the observed total JOE to NED fluorescent dye intensities is, therefore, a measure of the fraction of full-length molecules (relative to all the molecules). This fraction is a function of the extension probability p used in the mathematical analysis, and the functional form relating the p we set to the ratio we observe, is precisely the Laplace transform, from which can determine the DNA sizes.

Immobilization. Reacti-Bind™ streptavidin-coated polystyrene strip plates (Pierce, Rockford, IL), were used, with Blocker™ BSA. The plates were washed 3 times with 200 μL of TBS buffer (25 mM TRIS and 150 mM NaCl; pH = 7.2) by shaking at room temperature. To immobilize the template, we added 3 μL Binding Buffer (5 mM EDTA, 5× Denhardt's and 0.1% Tween 20 in

TBS) and 1 μL [1 μM] biotinylated sequencing template (1 pM) (Gibco BRL, Life Technologies, Rockville, MD). The solution was incubated at room temperature for 15 min, and then washed 3 times (repipetting 3 times) with 200 μL washing buffer (0.3% Tween 20 in TBS).

Extension. We combined 2 μL [5×] of Sequenase reaction buffer (USB Corporation, Cleveland, OH) with 1 μL [1 μM] (1 pM) of the NED-labeled sequencing primer. These were incubated at 65°C for 6 min in a thermal cabinet (Biometra, OV/5), and then further incubated at 37°C for 25 min. Additional reagents were then added, including: 1 μL [50 μM] (50 pM) dATP (Promega, Madison, WI); 2.5 μL [20 μM] (50 pM) dCTP (Promega, Madison, WI); 5 μL [10 μM] (50 pM) ddGTP-JOE (NEN Life Sciences, Boston, MA); 1 μL [10 U/μL] Sequenase (USB Corporation, Cleveland, OH); x μL [100 μM] ddATP (variable) (Amersham, Piscataway, NJ); deionized water (variable), filling to 17.5 μL total volume. For sequencing extension, the reaction mixture was incubated at room temperature for 25 min. Washing was done 3 times with 200 μL of washing buffer.

Denaturation. To remove the nonbiotinylated strand, we added 20 μL of deionized formamide, denaturating on a heatblock at 95°C for 5 min. Then, 2 μL of this sample was added to 12 μL of deionized formamide prior to loading onto an ABI/310 automated DNA sequencer.

Detection. The ultimate goal of "friction sequencing" is to replace size-separating DNA sequencers with whole-sample fluorescence measurements. During assay development, however, to best understand the sequencing extension products, we size-separated the products on an ABI/310 single capillary genetic analyzer (PE Biosystems, Foster City, CA). A 14 μL loading volume was used, with the POP4 gel, an STR capillary, and filter set F. The run time was 20 min, at a run temperature of 60°C. The peak heights and areas were estimated using PE's GeneScan software. Initial calculations were done in Microsoft Excel on an Apple Macintosh computer.

ACKNOWLEDGMENTS

We very much appreciate Dr. Michael B. Gorin's very helpful discussions throughout the course of this research. I also thank Dr. Richard Cotton for inviting me to speak on this topic at the HUGO Mutation Detection 2001 meeting, and thank the organizers for an extremely well-run meeting.

REFERENCES

Bracewell RN. 1978. The Fourier transform and its applications. New York: McGraw-Hill Book Co.

Cheeseman PC. 1994. Method for sequencing polynucleotides. US Patent #5,203,509; filed February 27, 1991, published April 12, 1994.

Drmanac R, Drmanac S, Strezoska Z, Paunesku T, Labat I, Zeremski M, Snoddy J, Funkhouser WK, Koop B, Hood L. 1993. DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing. Science 260:1649–1652.

Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. 1991. Light-directed spatially addressable parallel chemical synthesis. Science 251:767–773.

Lehrach H, Drmanac A, Hoheisel J, Larin Z, Lennon G, Monaco AP, Nizetic D, Zehetner G, Poustka A. 1990. Hybridization fingerprinting in genome mapping and sequencing. In: Davies KE, Tilghman SM, editors. Genetic and physical mapping I: genome analysis. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory. p 39–81.

Metzker ML, Raghavachari R, Richards S, Jacutin SE, Civitello A, Burgess K, Gibbs RA. 1994. Termination of DNA synthesis by novel 3′-modified-deoxyribonucleside 5′-triphosphates. Nucleic Acids Res 22:4259–4267.

Moore GE. 1965. Cramming more components onto integrated circuits. Electronics 38:114–117.

Pálsson B, Pálsson F, Perlin M, Gubjartsson H, Stefánsson K, Gulcher J. 1999. Using quality measures to facilitate allele calling in high-throughput genotyping. Genome Res 9:1002–1012.

Papoulis A. 1962. The Fourier integral and its applications. New York: McGraw-Hill.

Perlin MW, Lancia G, Ng S-K. 1995. Toward fully automated genotyping: genotyping microsatellite markers by deconvolution. Am J Hum Genet 57:1199–1210.

Perlin MW. 1999. Computer automation of STR scoring for forensic databases. In: First International Conference on Forensic Human Identification in the Millennium, Oct. 25–27. London, UK: Forensic Science Service.

Perlin MW, Szabady B. 2001. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. J Forensic Sciences 46:1372–1377.

Pevzner P, Belyi I. 1997. Software for DNA sequencing by hybridization. Comput Appl Biosci 13:205–210.

Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 242:84–89.

Rudin W. 1974. Real and complex analysis. New York: McGraw-Hill.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467.

Southern EM, Maskos U, Elder JK. 1991. Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. Genomics 13:1008–1017.

Weber J, May P. 1989. Abundant class of human DNA polymorphisms that can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396.