Sequencing two DNA templates in five channels by
digital compression

(MICHAEL NELSON*†, YANPING ZHANG†, DAVID L. STEFFENS‡, REINGARD GRABHERR§,
AND JAMES L. VAN ETTE†)

ABSTRACT By applying algebraic coding methods to the Sanger dideoxynucleotide procedure, DNA sequences of two templates can be determined simultaneously in only five reactions and data channels. A 5:2 data compression is accomplished by instantaneous source coding of nucleotide sequence pairs into one set of 5-bit block codes. A general algebraic expression, \( n^2 - 1 = 4^f \), describes conditions under which \( f \) DNA templates can be sequenced using \( n \) channels. Such compression sequencing is accurate and efficient, as demonstrated by manual 35S autoradiographic detection and automated on-line analysis using fluorescent-labeled primers. Symmetric 5:2 compression is especially useful when comparing two closely related sequences.

We have previously shown (1) that algebraic coding methods from information theory can be applied to DNA sequence analysis. In particular, if the presence or absence of DNA sequencing gel autoradiogram bands is treated as a 1 or 0, then three channels suffice to sequence a template by using a single label; alternatively, only three labels are needed for DNA sequencing in a single lane.

We demonstrate here a more efficient strategy for DNA sequencing based on source code compression of two sequences into 5-bit instantaneous block codes. This coding strategy resembles data compression techniques used in telecommunications and digital electronics (2, 3). Sequences of two DNA templates are determined simultaneously using five enzymatic reaction mixtures and data lanes, resulting in improved coding efficiency and 60% increased sequence throughput. By generalized extension of the method, the average code word length approaches 2.0 bits per nucleotide, equal to the source entropy of the four-letter genetic alphabet.

Background to Algebraic Coding Theory and DNA Sequence Analysis

In the original (4) and modified (5–7) enzymatic dideoxynucleotide DNA sequencing procedures, four spatially discrete or optically separable (8–10) data channels have been used for each DNA template. If the presence or absence of a signal at a specified gel migration distance is treated algebraically as a 1 or 0, and dideoxyguanine (ddG), ddA, ddT, and ddC reaction mixtures are loaded left to right on a DNA sequencing gel, then this coding arrangement defines a 4 \( \times \) 4 identity matrix. This coding matrix, shown in binary and expanded polynomial forms in Table 1, is called "8421 code," since the decimal numbers 8, 4, 2, and 1 are, in binary notation, 1000, 0100, 0010, and 0001, respectively (11, 12). The 8421 code used by Sanger et al. (4) for dideoxynucleotide sequencing has a coding efficiency (2, 3) of 0.5 (2 bits per nucleotide)/(4 bits per code word); and the algebraic distance between any pair of code words is 2, where distance is defined as the number of positions at which binary code words differ (13).

In standard four-lane dideoxynucleotide sequencing, the null set (0000) code word is undetectable and cannot be assigned to G, A, T, or C. However, there are \( 2^{4} - 1 = 15 \) detectable 4-bit code words, which may be arranged in \( 15 \times 14 \times 13 \times 12 = 32,760 \) different ways. Scientists do not normally consider most of these possibilities. Only \( 4 \times 3 \times 2 \times 1 = 24 \) of these sequencing arrangements are commonly employed, corresponding to row or column transpositions of the identity matrix. Code words are unnecessarily 4 bits long, and 11 of the 15 4-bit code words are unused. Clearly this is inefficient coding.

Efficient Coding Based on Communications Theory

Although four-channel DNA sequencing has proven satisfactory, other coding arrangements have been demonstrated (14–16) that allow improved ease of sample preparation, speed, or efficiency. However, these alternate codes have neither been interpreted in terms of information theory nor fully optimized as such.

In fact, DNA sequence analysis contains all of the essential features of a communications system, as described by Shannon and Weaver (17) in their classic treatise and more recently by others (13, 18, 19). As shown in Fig. 1, one may consider a DNA template to be an information source, which is then encoded when it is mixed into selected primer-dependent chain-terminating reaction mixtures. DNA polymerase produces encoded messages, in the form of labeled 3'-dideoxy-terminated polynucleotide chains, which are sent through defined channels of sequencing gels. Banding patterns are detected by autoradiography and decoded by the gel reader.

For dideoxynucleotide sequencing, encoding is accomplished by judiciously mixing templates, primers, all four dNTPs, one or more ddNTP(s), and DNA polymerase. By encoding DNA sequences in binary channels, powerful algebraic coding methods developed over the past 40 years for telecommunications, computer science, and digital electronics (2, 3, 18–20) can be applied to DNA sequence analysis. Such algebraic coding techniques can be used to increase the speed, efficiency, or accuracy of dideoxynucleotide sequencing.

Channel and Source Coding

Having recognized that DNA sequence analysis is a form of communication in which G, A, T, and C are encoded as

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.

Abbreviations: ddG, ddNTP, etc., dideoxyguanine, dideoxynucleotide triphosphate, etc.
binary code words (1), we can improve coding efficiency using two classic methods of algebraic coding theory: channel coding or source coding (2, 18, 19, 21).

In channel coding, as described (1), for three-lane DNA sequencing, coding efficiency is optimized by selecting the smallest number \( n \) of binary channels to satisfy the inequality, \( 2^n - 1 \geq b \), where \( b \) is the number of code words for DNA bases in the four-member source alphabet, \( S_4 = \{ G, A, T, C \} \).

\[
2^n - 1 \geq 4.
\]

Solving for \( n \), we may calculate that \( n \geq 3 \). In other words, three binary communication channels (data lanes) are sufficient for encoding four DNA bases.

In three-lane sequencing, coding efficiency is limited by the requirement that all code words have a fixed length of three \( (n = 3) \). This limitation can be removed, however, if variable length code words are employed. For example, if \( G, A, T, \) and \( C \) code words have length two or three in equal proportions, then it is still possible to satisfy Eq. 1: \( 2^{2.5} - 1 \geq 4 \). In other words, if the average number of binary data channels employed per template is 2.5, then \( G, A, T, \) and \( C \) can still be uniquely encoded. If \( f \) is the number of different DNA templates to be sequenced, then this same average number of channels per template, \( n_{\text{avg}} = n/f = 2.5 \), can be achieved when sequences of two DNA templates \( (f = 2) \) are encoded using five data channels \( (n = 5) \):

\[
2^5 - 1 \geq 4^2.
\]

In contrast to channel coding, in source coding, code words are assigned different lengths according to their frequency of occurrence (22–24). For example, in the original 1837 Morse code, the frequent letter “e” is encoded by the shortest “-...” symbol, whereas “q” is encoded by the “-...-...” symbol (22, 23). In this fashion it is possible to compress more code words into a limited number of communication channels (2, 19, 24).

Source Code Compression of Two DNA Sequences

Source coding of DNA sequences need not strictly be a function of the frequency of source alphabet symbols (G+C content). Code word assignments are constrained by peculiarities of DNA base chemistry, secondary structure, algebric distance from assigned code words to probable sequencing errors, and the ease of encoding and decoding. Source coding for dideoxynucleotide sequencing is, therefore, necessarily nonideal, since not all code words can be assigned. In this sense, the assignment of 5-bit DNA code words for source code compression resembles Braille code, where 6-bit code words are assigned primarily by their ease of use (25).

Guidelines for Code Construction. The following guidelines have been used to assign code words for DNA source alphabet symbols. (i) Efficient coding can be achieved when two DNA templates \( (f = 2) \) are read simultaneously in \( n \) data channels under the condition \( 2^n - 1 \geq 4^n \), so that \( n \geq 5 \). (ii) Null set encoding \((00000)\) of DNA bases is disallowed because this code word is undetectable. (iii) Full set encoding of DNA bases on one or both strands is disallowed: 01111, 11110, 10111, 11101, and 11111 are error codes for premature termination by DNA polymerase. (iv) Since DNA secondary structure complications are more probable in G+C-rich or polypropylene regions (26, 27), G is redundantly encoded and its algebric distance to C is maximized. (v) The G+C content determines the choice of A, T, and C code words. T should be assigned the 3-bit 001- prefix and 100 suffix when the G+C content is \( \geq 50\% \), whereas C should be assigned the 3-bit 001- prefix and 100 suffix when the G+C content is \( \leq 50\% \). (vi) Code words for the four DNA bases share no common prefixes \((00-, 01-, 10-, 11-)\), so that instantaneous code recognition (2, 13) is achieved. (vii) Symmetrical codes allow DNA sequences to be most easily encoded and decoded.

Five-Bit DNA Source Code Construction. Let \( G_1, A_1, T_1, \) and \( C_1 \) denote guanine, adenine, thymine, or cytosine in DNA sequence 1; and let \( G_2, A_2, T_2, \) and \( C_2 \) denote guanine, adenine, thymine, or cytosine in DNA sequence 2, respectively. Then \( S_1 \) is the source alphabet \( \{G_1, A_1, T_1, C_1\} \) for sequence 1; and \( S_2 \) is the source alphabet \( \{G_2, A_2, T_2, C_2\} \) for sequence 2. Pairwise combinations of \( \{G_1, A_1, T_1, C_1\} \times \{G_2, A_2, T_2, C_2\} \) define a 16-member source alphabet, \( S_{1x2} = \{G_1G_2, A_1G_2, T_1G_2, C_1G_2, G_1A_2, A_1A_2, T_1A_2, C_1A_2, G_1T_2, A_1T_2, T_1T_2, C_1T_2, G_1C_2, A_1C_2, T_1C_2, C_1C_2\} \). Five binary channels can encode 2\(^5\) different 5-bit block codes to which these 16 source code symbols can be assigned \((2^5 - 1 \geq 4^2)\), but no provisions are made for aberrant electrophoretic mobilities using \( S_{1x2} \).

To account for aberrant gel mobilities, let 0\(_1\) and 0\(_2\) denote the absence of a band in DNA sequence 1 and sequence 2, respectively. A superior 5-bit code can be constructed if pairwise combinations of \( \{G_1, A_1, T_1, C_1, 0_1\} \times \{G_2, A_2, T_2, C_2, 0_2\} \) are used to define a larger 25-member source alphabet, \( S_{1x2} \), which can be encoded using 25 of the binary numbers 0 through 31. This set of 5-bit block codes is also known as “Baudot telecode type” since 2\(^5\) = 32 binary symbols are sufficient to represent the 26 letters of the alphabet and six punctuation marks (12, 28).

Table 1. Banding patterns in Sanger dideoxynucleotide DNA sequence analysis can be expressed as binary numbers

<table>
<thead>
<tr>
<th>Base</th>
<th>ddG</th>
<th>ddA</th>
<th>ddT</th>
<th>ddC</th>
<th>Binary notation</th>
<th>Expanded polynomial expression</th>
<th>Decimal notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1000</td>
<td>((1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 0 \times 2^0))</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0100</td>
<td>((0 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 0 \times 2^0))</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0010</td>
<td>((0 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 0 \times 2^0))</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0001</td>
<td>((0 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0))</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. DNA sequence analysis contains all of the features of a communication system: (i) an information source (DNA template), (ii) a process of encoding (ddNTPs, primers, and channels selected), (iii) a sender (DNA polymerase), (iv) a set of communications channels (gel lanes), (v) a receiver (X-ray film), and (vi) a decoder (the gel reader).
In accordance with the above guidelines, 5-bit (Baudot) code words are assigned to $S_{1 \times 2}$ source alphabet symbols shown in Table 2. For example, bands in channels 1, 4, and 5 (10011) at a given mobility position indicates that sequence 1 = A and sequence 2 = G. Sequence 1 can also be read left-to-right as a three-lane sequence in channels 1, 2, and 3; sequence 2 can be read right-to-left in channels 3, 4, and 5. A, T, and G in sequences 1 and 2 can be read as 2-bit code words (A = 10 or 01; T = 01 or 10; G = 11), whereas C is encoded using a 3-bit code word (C = 001 or 100).

**Experimental: Sequencing Two DNA Templates in Five Channels**

By using five mixed dideoxynucleotide reaction mixtures, DNA sequencing was carried out by a modification of the method of Tabor and Richardson (6) or the automated fluorescent sequencing protocol of Steffens and coworkers (29, 30). Results from (i) dideoxynucleoside 5'-[α-[35S]thio]triphosphate sequencing (Fig. 2), (ii) automated on-line DNA sequencing using near infrared fluorescence detection (Fig. 3), and (iii) DNA sequence comparisons using 5:2 compression (Fig. 4) are shown below.

**Discussion**

If DNA sequence analysis is treated as a form of communication (1, 17), then algebraic coding techniques can improve the efficiency and throughput of the Sanger dideoxynucleotide procedure. We have shown that source code compression can be applied to dideoxynucleotide DNA sequence

<table>
<thead>
<tr>
<th>DNA source code symbol</th>
<th>Five-bit binary code</th>
<th>DNA-sequencing gel-banding pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>G₂G₂</td>
<td>11011</td>
<td></td>
</tr>
<tr>
<td>G₂A₂</td>
<td>11001</td>
<td></td>
</tr>
<tr>
<td>G₂T₂</td>
<td>11010</td>
<td></td>
</tr>
<tr>
<td>G₂C₂</td>
<td>11100</td>
<td></td>
</tr>
<tr>
<td>G₂G₂</td>
<td>11000</td>
<td></td>
</tr>
<tr>
<td>A₂G₂</td>
<td>10011</td>
<td></td>
</tr>
<tr>
<td>A₂A₂</td>
<td>10001</td>
<td></td>
</tr>
<tr>
<td>A₂T₂</td>
<td>10010</td>
<td></td>
</tr>
<tr>
<td>A₂C₂</td>
<td>10100</td>
<td></td>
</tr>
<tr>
<td>A₂G₂</td>
<td>10000</td>
<td></td>
</tr>
<tr>
<td>T₂G₂</td>
<td>01011</td>
<td></td>
</tr>
<tr>
<td>T₂A₂</td>
<td>01001</td>
<td></td>
</tr>
<tr>
<td>T₂T₂</td>
<td>01010</td>
<td></td>
</tr>
<tr>
<td>T₂C₂</td>
<td>01100</td>
<td></td>
</tr>
<tr>
<td>T₂G₂</td>
<td>01000</td>
<td></td>
</tr>
<tr>
<td>C₂G₂</td>
<td>00111</td>
<td></td>
</tr>
<tr>
<td>C₂A₂</td>
<td>00101</td>
<td></td>
</tr>
<tr>
<td>C₂T₂</td>
<td>00110</td>
<td></td>
</tr>
<tr>
<td>C₂C₂</td>
<td>00100</td>
<td></td>
</tr>
<tr>
<td>C₂G₂</td>
<td>00100</td>
<td></td>
</tr>
<tr>
<td>0₂G₂</td>
<td>00011</td>
<td></td>
</tr>
<tr>
<td>0₂A₂</td>
<td>00001</td>
<td></td>
</tr>
<tr>
<td>0₂T₂</td>
<td>00010</td>
<td></td>
</tr>
<tr>
<td>0₂C₂</td>
<td>00010</td>
<td></td>
</tr>
<tr>
<td>0₂G₂</td>
<td>00000</td>
<td></td>
</tr>
</tbody>
</table>

DNA bases in sequence 1 are {G₂, A₂, T₂, C₂}; bases in sequence 2 are {G₂, A₂, T₂, C₂}. 0₂ or 0₂ refers to the absence of a base in sequence 1 or sequence 2, respectively. If the presence or absence of autoradiogram bands is treated algebraically as 1 or 0, then all pairwise combinations of {G₂, A₂, T₂, C₂, 0₂} x {G₂, A₂, T₂, C₂, 0₂} can be encoded using 5-bit code words, corresponding to 25 of the binary numbers 0 through 31.

**Fig. 2.** Three-lane and 5:2 compression sequencing using mixed ddNTPs. Sequence 1 (pSP1H1) and sequence 2 (pSP1H3) are pBlueScript KS(−) phagemid clones containing different DNA inserts. Lanes 1–4 (sequence 1) and lanes 16–19 (sequence 2) are standard four-lane dideoxynucleotide sequence tracts. Lanes 5–7 (sequence 1) and lanes 13–15 (sequence 2) are three-lane sequence tracts (see ref. 1). Lanes 8–10 (sequence 1) and lanes 10–12 (sequence 2) are 5:2 source code compression sequencing tracts using five mixed ddNTP reaction mixtures and single-stranded templates as shown. The dideoxynucleotide kit (United States Biochemical Sequenase, Version 2.0) was used in the following manner. For annealing, 1 μg of single-stranded DNA, 2 μl of 5× reaction buffer, 25 ng of T3 primer, and water were mixed in a total volume of 10 μl, incubated at 65°C for 2 min, and cooled to 23°C over a period of 30 min. For labeling, 10 μl of template-primer, 1 μl of 0.1 M diithiothreitol, 2 μl of 1:5 diluted labeling mixture, 0.5 μl of deoxyadenosine 5'-[α-[35S]thio]triphosphate (1000–1500 Ci/mmol; 1 Ci = 37 GBq), and 2 μl of Sequenase diluted 1:8 were mixed in a final volume of 15.5 μl and incubated for 2 min at 23°C. Then 3.5-μl samples were transferred to prewarmed (37°C) tubes containing 2.5 μl of ddNTP/ddNTP extending mixtures (or 2 μl of each ddNTP/ddNTP combination when two ddNTPs were used) and incubated for 5 min at 37°C. Stop buffer (4 μl) was added to each reaction mixture. Samples in lane 10 were mixed (2 μl from each C reaction mixture) at this step. Samples were heated at 90°C for 3 min and 2.5 μl of each reaction mixture was loaded onto a 40-cm-long 0.1-mm-thick 6% polyacrylamide/urea DNA sequencing gel. Separation was carried out at constant power (60 W) and ~1500 V for 2.5 hr.

analysis in both manual and automated formats. In addition, DNA sequence comparisons by 5:2 compression sequencing are discussed below. In addition, certain theoretical limits of DNA code compression are considered.

**35S-Labeled Dideoxynucleotide DNA Sequencing Using 5:2 Source Code Compression.** As shown in Fig. 2, it is possible to simultaneously sequence two DNA templates by using only five mixed dideoxynucleotide reaction mixtures and five data channels, as detected by 35S autoradiography. Sequences from two pBlueScript KS(−) single-stranded plasmid DNA templates could be accurately read to at least 450 nt from a T3 sequencing primer. Light and dark bands seen in standard four-lane sequencing gels were faithfully reproduced in three-lane and 5:2 compressed sequence tracts.
Two primers, sequencer A Compression. patterns.


Note the reflective symmetry in the first 4,000 bases of these two sequences: polylinker regions of both plasmids are the same, but sequences diverge thereafter into asymmetric patterns.

Automated Fluorescent Sequencing Using 5:2 Source Code Compression. A prototype LI-COR model 4000 automated fluorescent sequencer (29, 30) was used to read two sequences to a length of >500 nt apiece by using five channels. Two primers, labeled with the near infrared dye IR-144, were used to prime an M13mp18 single-stranded DNA clone containing a 1.2-kb insert by using a Taq polymerase sequencing protocol designed to minimize DNA secondary structure (7, 29). Labeled dideoxynucleotide reaction products were detected on-line by near infrared (820 nm) fluorescence during electrophoresis. Strong and weak signal amplitudes from three-lane and 5:2 compression tracts were faithfully reproduced from parallel control four-lane sequencing (Fig. 3).

DNA Sequence Comparisons by 5:2 Compression. The use of symmetric 5:2 compression has an important practical consequence: one can quickly detect sequence variations as asymmetric patterns in an otherwise symmetrical background. As illustrated in Fig. 4, six base changes between two closely related sequences, the miCviI1 DNA methyltransferase gene and a nonfunctional pseudogene (31), are easily detected. This strategy should be especially useful in comparing sequences of wild-type and site-directed mutants or genetic variants. One focuses immediately on sequence differences: the burden of reading and then comparing two separate four-lane sequences is eliminated.

Theoretical Compression Limit. Calculation from DNA source entropy. Coding efficiency is defined as the source entropy divided by the average code word length (2, 17, 21) and maximal coding efficiency is reached when the average code word length and source entropy are equal: \( H(S)/(n/f) < 1.0 \). For random DNA, the source entropy is 2 bits per nucleotide (32, 33). Therefore, regardless of which block coding scheme is employed, the average number of binary channels per nucleotide code word is never less than 2 (2, 17, 21).

Derivation of a DNA sequence compression theorem. We have shown that a coding efficiency of 2.5 bits per nucleotide can be achieved \((n/f = 2.5)\) when five channels are used to sequence two DNA templates. However, it is possible to sequence more than two templates at a time. For example, three templates \((f = 3)\) can be sequenced using seven

FIG. 3. Automated three-lane and 5:2 compression sequencing using on-line near infrared fluorescence detection (29, 30) and mixed dideoxynucleotide reaction mixtures. Standard four-lane dideoxynucleotide sequencing reaction mixtures (lanes 1–4 and 16–19), three-lane sequencing reaction mixtures (lanes 5–7 and 13–15), and 5:2 compression sequencing (lanes 8–12) were carried out using a TaqTrack sequencing kit (Promega) as described (29). Template was single-stranded DNA from an M13mp18 recombinant clone containing a 1.2-kb pBR322 insert. Sequence 1 and sequence 2 were defined by two M13 universal forward 19-mer primers that hybridized at different locations on the template DNA, resulting in sequence tracts offset by 12 nt. Primers were covalently coupled to the infrared fluorophore IR-144 (Research Organics) via a 10-atom spacer arm to a thymidine near the 5' end of the primer. IR-144-dye-labeled DNA fragments resulting from dideoxynucleotide termination products were imaged using a prototype LI-COR model 4000 DNA Sequencer.

FIG. 4. Mirror-image sequencing of wild-type and mutant genes. Sequence 1 and sequence 2 are a miCviI1 DNA methyltransferase gene and a nonfunctional pseudogene (31) that differs from the wild-type gene at six positions (arrows). Note: sites of mutation are readily identified as asymmetric five-lane banding patterns. Five mixed dNTP reactions were carried out as described in Fig. 2. Lanes: 1–3, template for wild-type sequence 1; 3–5, template for mutant sequence 2.
channels because all $4^3$ three-way combinations of \{G_1, A_1, T_1, C_1\} × \{G_2, A_2, T_2, C_2\} × \{G_3, A_3, T_3, C_3\} can be encoded in 7-bit block codes:

$$2^n - 1 \geq 4^f \text{ or } n \geq 7. \quad [3]$$

By comparing Eqs. 1, 2, and 3, we may generally describe the relationship between the number $n$ of binary channels and the number of DNA fragments $f$ to be sequenced:

$$2^n - 1 \geq 4^f. \quad [4]$$

From this DNA sequence compression theorem (Eq. 4), it is possible to calculate (i) the number $n$ of channels needed to sequence $f$ templates, (ii) the number of DNA fragments $f$ that can be sequenced in $n$ channels, and (iii) the average code word length $n/f$, as shown in Table 3. The lower limit $(n/f \to 2)$ calculated from our DNA sequence compression equation is identical to that derived independently from a DNA source entropy calculation. In other words, if many sequences are compressed using a large number of channels, then slightly more than 2 bits per nucleotide and 2 channels per DNA template.

## Conclusions

Five conclusions result from these experiments. (i) Source code compression has been applied to enzymatic dideoxynucleotide DNA sequence analysis. Two DNA sequences >450 nt were simultaneously read using only five mixed dideoxynucleotide reaction mixtures and five data channels, as detected by $^{35}$S autoradiography. (ii) Digital compression increases the efficiency and throughput of multichannel automated fluorescent DNA sequencing. Two DNA sequences were deduced simultaneously using only five mixed dideoxynucleotide reactions and channels, where eight reaction mixtures and channels were previously employed. (iii) A general algebraic result, $2^n - 1 \geq 4^f$, describes the conditions under which $f$ templates can be sequenced using $n$ binary channels. Therefore, to sequence $f$ templates by code compression, $n \geq 2f + 1$ channels are required. (iv) The average code word length reaches a limit of $n/f \to 2$, equal to the source entropy of the four-letter genetic alphabet. (v) Digital compression may be especially useful when two similar DNA sequences, such as site-directed mutants or genetic variants, are to be compared. In $5.2$ compression sequencing with instantaneous coding, base changes are rapidly identified as asymmetries in an otherwise symmetrical five-channel pattern.

Based on these conclusions two further predictions can be made. (vi) The algebraic coding principles used here do not strictly depend on whether channels are physically separate or optically discrete. If combinations of spatially and/or optically discrete channels are employed, then it is theoretically possible to sequence $f$ DNA fragments by using $n$ channels and $f$ distinguishable labels under the condition, $2^n - 1 \geq 4^f$. Such DNA sequencing algebra is generally useful in designing sequencing strategies and automated instruments.

We thank Lyle Middendorf, Les Lane, Roy French, Myron Brakke, and Michael McClelland for critical reading of this manuscript. Parts of this work were supported by U.S. Public Health Service Grant GM-32441 to J.L.V.E.