DNA sequencing of four bases using three lanes

Michael Nelson*, James L.Van Etten[1] and Reingard Grabherr[1]
Department of Biochemistry and Molecular Biology, University of Chicago, 920 East 58th Street, Chicago, IL 60637 and [1]Department of Plant Pathology, University of Nebraska, Lincoln, NE 68583-0722, USA

Received November 14, 1991; Revised and Accepted February 5, 1992

ABSTRACT

A three-lane DNA sequencing strategy is described which is based on redundant binary coding principles from communications theory. Three-lane sequencing is an efficient, accurate, and flexible strategy, suitable for large-scale automated DNA sequencing. Communications theory and algebraic coding principles can also be applied to sequencing other informational macromolecules.

INTRODUCTION

Rapid DNA sequencing based upon either chemical cleavages (1) or chain-terminating dideoxynucleotides (2) uses four or more data lanes for electrophoretic separations. We describe here a streamlined method for analysing DNA sequences, based upon redundant binary encryptment of G, A, T, and C, using principles from communications theory (3). A practical consequence of this strategy is that the four DNA bases can be read unambiguously using three data lanes on DNA sequencing gels. Furthermore, since DNA bases are read in two or more data lanes, three-lane sequencing is inherently more reliable than non-redundant four-lane methods.

Binary encoding of G, A, T, and C

In traditional DNA sequencing methods, each of four gel lanes contains labeled DNA chains which terminate in G, A, T, or C. For example, enzymatic dideoxy sequencing employs unique data lanes for chain termination reactions of ddG, ddA, ddT, and ddC. Therefore G, A, T, or C bases appear as autoradiogram bands which occur in a single data lane (2).

On DNA sequencing gel autoradiograms the presence or absence of bands represents a two-dimensional arrangement of information 'bits'. In any particular data lane the absence of a band can be treated mathematically as a **0** and the presence of a band as a **1**. Thus in a typical four-lane dideoxy sequence array, G, A, T, and C lanes can be arranged so that

$$G = (1,0,0,0),\ A = (0,1,0,0),\ T = (0,0,1,0),\ C = (0,0,0,1).$$

That is, 'G' is read as a band which occurs only in lane one, 'A' is a band only in lane two, 'T' is a band only in lane three, and 'C' is a band only in lane four.

In contrast to enzymatic dideoxy sequencing, the original chemical DNA sequencing method of Maxam and Gilbert (1) used base-specific cleavage reactions for G, A + G, C, and C + T, arranged in four data lanes. Alternate chemical cleavages for A>C, A>G, A, C, or 5mC + T (4, 5, 6, 7, 8) also have required four or more data lanes in DNA sequencing gels. Chemical DNA sequencing typically employs G, A + G, C, C + T, and A>C reactions in five data lanes (4). 'G' is read as a band in lanes one and two, 'A' is read as a band in lanes two and five, 'T' is a band only in lane four, and 'C' is a band in lanes three and four with a fainter band in lane five. In binary code this five-lane encryptment is:

$$G = (1,1,0,0,0),\ A = (0,1,0,0,1),\ T = (0,0,0,1,0),\ C = (0,0,1,1,1).$$

At a particular mobility position, the number of information bits (**0** = absence of band; **1** = presence of band) on a sequencing gel array of **n** data lanes is $\mathbf{2^n}$. However, in order to read sequences from an autoradiogram, at least one data lane must contain a band at a given mobility position. Unreadable binary codes such as **(0,0,0,0)** in four lanes or **(0,0,0,0,0)** in five lanes are therefore not allowed. Disallowing null set encoding, there are $\mathbf{2^n-1}$ possible binary coding arrangements when **n** data lanes are used. Therefore the condition $(2^n-1) \geq 4$ must be satisfied to uniquely encode all four DNA bases. Solving for **n**, three data lanes are sufficient to distinguish G, A, T, and C bases. In fact, up to seven (2^3-1) different bases may be uniquely encoded using the following three-bit codes:

(0,0,1)	**(0,1,1)**	**(1,1,1)**
(0,1,0)	**(1,1,0)**	
(1,0,0)	**(1,0,1)**	

These same three-bit codes are described by Shannon and Weaver (3, page 101) in their classic treatise, 'The Mathematical Theory of Communication'. In digital electronics and telecommunications, these linear block codes have been examined in detail with respect to information content, signal redundancy, and error rate (9).

Choice of binary codes based on signal redundancy, error rate, and G + C content

Since DNA has four bases, any four of the above seven three-bit codes permit reading of sequences in a three-lane data array.

* To whom correspondence should be addressed at present address: Department of Plant Pathology, University of Nebraska, Lincoln, NE 68583-0722, USA

Maxam and Gilbert (1) were aware of some of these alternatives when they wrote, 'In principle, one could sequence DNA with three chemical reactions, each of a single base specificity, using the absence of a band to identify the fourth position... This would be a non-redundant method... subject to considerable error'. In practice such null set encoding **(0,0,0)** is disallowed as described above. However, it is possible to arrange DNA sequencing reactions in many ways not described by Maxam and Gilbert. Using enzymatic or chemical reactions, one can arrange sequencing reactions in **$(2^n-1)!/3!$ = 7!/3! = 840** different ways in a three-lane data array, so that G, A, T, and C have unique three-bit codes. In other words, if G is chosen from seven possible three-bit codes, then there are six remaining choices for A, five for T, and four for C. Altogether, there are **(7×6×5×4) = 840** possible ways to sequence four DNA bases in three lanes. However as we show below, some of these 840 arrangements are preferred.

As a consequence of communications theory, highly redundant codes are used in telecommunications for complex data transmissions (3). Likewise, in military intelligence, radio signals are most accurately transmitted as a small number of repeated bursts (10). With respect to DNA sequence analysis, three-bit arrays are sufficient to discriminate the four nucleotides, regardless of which three-bit message codes are chosen. However, the most unambiguous three-bit coding choices for reading an autoradiogram are those in which bands are present in *two or more* lanes, i.e.,

G = (0,1,1), A = (1,0,1), T = (1,1,0), C = (1,1,1).

In this arrangement, characteristic bands appear in two of three lanes for G, A, or T, whereas C gives a band in all three lanes (Figure 1B). Band redundancy reduces potential sequencing errors, and furthermore requires neither more reaction samples nor more data lanes. Instead, one simply uses three reactions containing appropriate mixtures of dNTPs and ddNTPs.

The non-random sequence character of a DNA sample should be considered when arranging reactions for three-lane sequencing. In particular, it should be recognized that, for G + C rich DNAs or those containing G + C compression zones, coding choices such as

G = (1,1,0), A = (1,0,1), T = (0,1,1), C = (1,1,1) or
G = (1,1,1), A = (1,0,1), T = (0,1,1), C = (1,1,0)

give maximal redundancy of G or C three-bit codes, and consequently the lowest error rates. On the other hand, for A + T rich sequences, coding choices where A and T are triply redundant such as

G = (0,1,1), A = (1,1,1), T = (1,0,1), C = (1,1,0) or
G = (1,0,1), A = (0,1,1), T = (1,1,1), C = (1,1,0)

can be read more accurately. Similarly, it is possible to choose optimal binary codes for redundant reading of unusual DNA sequences such as repeated dinucleotides, polypurine runs, or polypyrimidine tracts. However the **(1,1,1)** code should be used with caution as described below.

Detecting sequencing errors: Avoiding the (1,1,1) code

Whereas extra bands in four-lane sequencing gels produce ambiguous reading errors, which may be detected as such, in three-lane sequencing extra bands are sometimes misread and thus undetected. While these undetected errors may at first seem to be a disadvantage to three-lane sequencing, it is not necessarily

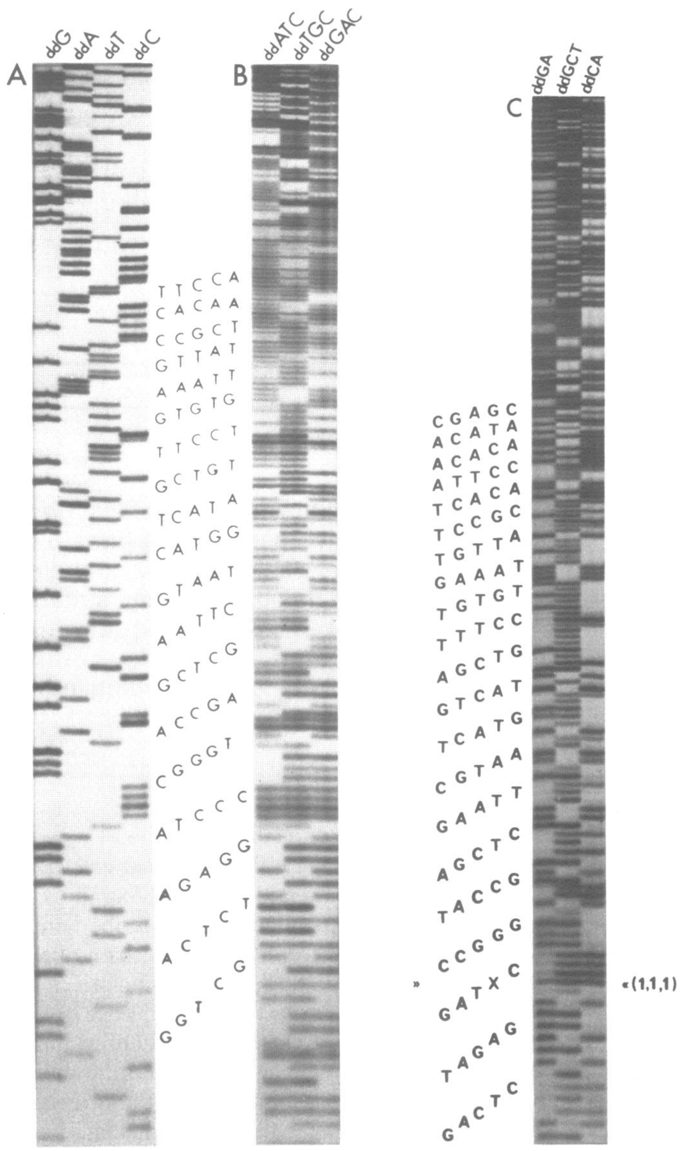

Figure 1. Sequencing of M13mp18 ssDNA using a (−40) primer. **(A)** Standard four-lane dideoxy sequencing: **G = (1,0,0,0), A = (0,1,0,0), T = (0,0,1,0), C = (0,0,0,1)**. **(B)** Three-lane dideoxy sequencing using mixtures of three ddNTPs in each reaction: **G = (0,1,1), A = (1,0,1), T = (1,1,0)**, and **C = (1,1,1)**. All four DNA bases are redundantly encoded. **(C)** Three-lane dideoxy sequencing using two ddNTPs in lane 1, three ddNTPs in lane 2, and two ddNTPs in lane 3: **G = (1,1,0), A = (1,0,1), T = (0,1,0), C = (0,1,1)**. G, A, and C are redundantly encoded. Note: Premature termination by DNA polymerase is detected as **(1,1,1)**. The dideoxy sequencing kit (Sequenase Version 2.0; US Biochemicals, Cleveland, OH) was used in the following manner. For annealing, 1 μg M13mp18 ssDNA, 1 μl 5×reaction buffer, and 1 μl (−40) primer were mixed in a volume of 5 μl, held at 65°C for 2 minutes, and then the temperature was reduced to 23°C over a period of 30 minutes. For labeling, 5 μl template-primer, 1 μl 0.1 M dithiothreitol, 2 μl (1/5 diluted) labeling mix, 1 μl α^{35}S-dATP (1000−1500 Ci/mMol), and 2 μl (1/8 diluted) Sequenase, were mixed (final volume 11 μl) and incubated for 3 minutes at 23°C. Then 3.3 μl samples were transferred to prewarmed (37°C) tubes containing 5 μl of ddNTP/dNTP extending mixtures and incubated for five minutes at 37°C. In Figure 1B, lane 1 contains (1 μl ddA + 1 μl ddT + 1 μl ddC + 2 μl extending mix); lane 2 contains (1 μl ddG + 1 μl ddT + 1 μl ddC + 2 μl extending mix); lane 3 contains (1 μl ddG + 1 μl ddA + 1 μl ddC + 2 μl extending mix). When only two ddNTPs were used (Figure 1C, lanes 1 and 3), 1 μl of distilled water replaced the third ddNTP. Five μl stop buffer was added to each reaction, samples were heated to 90°C for three minutes, cooled on ice, and 3 μl of each reaction was loaded onto a 6% polyacrylamide-40% (w/v) urea DNA sequencing gel. Separation was carried out at constant power (60 W), approximately 1700 V, at about 70°C.

so, depending on the binary codes chosen for the four DNA bases and the non-random source of such errors. It is important to distinguish ambiguities and misreading errors (white noise) from probable *non-random* errors. We will examine these different types of errors, and default codes for detecting them, in a separate manuscript (Grabherr and Nelson, manuscript in preparation), since this is a complex subject. For purposes here, two points should be recognized: (i) More, rather than fewer, data channels are needed for increased error detection capability and (ii) *undetected misreading errors* are the price that is paid for increased signal redundancy in a three-bit code. In particular, when any one of the four DNA bases is encoded as **(1,1,1)**, as in Figure 1B, premature DNA polymerase terminations are undetected (see Figure 1C) . In practice, since premature terminations (bands in all lanes) are a common source of DNA sequence ambiguity, the **(1,1,1)** code should be avoided as a message code. Therefore the limited condition $(2^n-2) \geq 4$ must be fulfilled in order to avoid null set **(0,0,0)** and full set **(1,1,1)** encoding. Three of the six remaining three-bit codes allow redundant base encryptment for DNA sequencing: **(1,1,0)**, **(1,0,1)**, and **(0,1,1)**. As described above, G and C should be encoded redundantly in most situations, while the remaining redundant three-bit code can be used for either A or T; for example,

(i) G = **(1,1,0)**, A = **(0,1,0)**, T = **(1,0,1)**, C = **(0,1,1)**or
(ii) G = **(1,1,0)**, A = **(1,0,1)**, T = **(0,1,0)**, C = **(0,1,1)**.

The second of these redundant arrangements (ii) is employed in Figure 1C: Mixed dideoxy reactions contain (ddG + ddA) for lane one, (ddG + ddT + ddC) for lane two, and (ddA + ddC) for lane three.

A technically simpler arrangement, in which mixtures of only two ddNTPs are used in each of three reaction samples, allows redundant encoding of G and C. For example, the encryptment,

G = **(1,1,0)**, A = **(1,0,0)**, T = **(0,0,1)**, C = **(0,1,1)**,

allows DNA sequences to be deduced using the original Maxam and Gilbert (1) reading scheme, although only three reactions, (ddG + ddA), (ddG + ddC), and (ddT + ddC) and three data lanes are required. However, in this arrangement neither A nor T is redundantly encoded.

Finally, according to Shannon and Weaver (3, page 39), 'by sending information in a redundant form, the probability of errors can be reduced'. We have carried out dideoxy sequencing with triply redundant base encoding in a *six lane* array which allows systematic correction of single errors and detection of double errors (Grabherr and Nelson, manuscript in preparation).

High redundancy dideoxy DNA sequencing using three lanes

Based upon the above considerations, M13mp18 single-stranded DNA was sequenced on three-lane DNA sequencing gels using two different redundant binary encryptments. Dideoxy sequencing reactions employed three different mixtures of dNTPs and chain terminating ddNTPs as described in the legends to Figure 1B and 1C.

DISCUSSION AND CONCLUSIONS

Special applications of three-lane DNA sequencing

If the presence or absence of bands on gel autoradiograms is treated as *digital information* (11), as ones or zeros mathematically, then communications theory (3) and its powerful algebraic coding methods (9) can be applied to DNA sequence analysis. In algebraic terms, we have demonstrated that the minimal DNA message code has three information bits per symbol when **n** data lanes are used under the condition $(2^n-1) \geq 4$. Therefore, three binary channels suffice to define four unique DNA message codes. More generally, the reading of DNA sequence data can be treated as a form of communication, a perspective which has mathematical and practical consequences.

In practice, based upon an understanding of binary codes and signal redundancy, it is possible to (i) reduce the number of DNA sequencing reaction samples and gel lanes, (ii) reduce sequence ambiguities by redundant encoding and consideration of non-random sequence characteristics. Three-lane sequencing is therefore an efficient, accurate, and flexible strategy.

In some special applications, streamlined three-lane sequencing is technically advantageous. The autoradiograms presented in Figure 1 show that both standard and three-lane sequencing permit reading equivalent lengths of DNA sequence. However, employing only three reactions and three data lanes, one can read at least as much DNA sequence for less effort.

Accordingly, the more efficient three-lane method may be especially useful in large-scale sequencing efforts where many reactions, data lanes, and gels are run; for example, in (i) 'genomic sequencing' (12), (ii) 'multiplex sequencing', (13, 14), where gels are repeatedly blotted to labeled probes, or (iii) automated on-line DNA sequencing using multiple lanes and radiometric (15, 16) or fluorescent (17) detection. In these macrosequencing applications, the three-lane method gives 33% greater throughput with increased signal redundancy. Furthermore, since signal redundancy results in superior signal-to-noise characteristics, it may be possible to read longer sequences per data lane *and* more lanes per gel, resulting in substantially increased performance.

The three-lane strategy may also be helpful in sequencing DNAs that have unusual sequence characteristics, such as very high or very low G + C contents, scattered G + C compression zones, or polypurine tracts. In these cases, DNA sequencing reactions can be redundantly arranged to minimize reading ambiguities.

As described above, as many as seven different DNA bases can be discriminated using three-bit binary codes. Therefore, by judicious selection of chemical reactions and binary codes, it is possible to identify G, A, T, C, and methylated bases such as 4mC, 5mC, and 6mA using only three gel lanes.

Communications theory predicts that the ability to define message codes is an exponential function of the number of channels over which signals are transmitted. Since these communication channels can be *spatially and/or optically* discrete, a three label/single lane arrangement is algebraically identical to the single label/three lane arrangement described above. Two problems with the widely used four fluor/single lane automated DNA sequencing instrument (18) are the aberrant electrophoretic mobility and spectral overlap of 4-fluoro-7-nitrobenz-2-oxa-1,3-diazole labeled DNA fragments with the other fluorescent-labeled DNA fragments. Both of these problems can be eliminated in the following three fluor/single lane dideoxy sequencing configuration: ddG is enzymatically incorporated with both fluorescein- and tetramethylrhodamine-labeled sequencing primers **(1,1,0)**, ddA is used with fluorescein-labeled primer alone **(1,0,0)**, ddT is used with tetramethylrhodamine-labeled primer alone **(0,1,0)**, and ddC is used with Texas Red-labeled primer alone **(0,0,1)**. This three fluor/single lane configuration should be superior to the standard four fluor/single lane arrangement in terms of sensitivity, spectral resolution, and electrophoretic mobility problems.

Finally, the principles of algebraic coding and communications theory outlined here have implications for analysing macromolecules other than DNA. For example, protein sequencing requires discrimination of twenty different amino acids. In a gel array of **n** lanes, binary codes are unique for each amino acid when $\mathbf{(2^n-1) \geq 20}$. Solving for **n**, the number of data lanes required to unambigously read twenty amino acids is five. Therefore, in principle it is possible to sequence proteins using end-labeling (19, 20) and as few as five semi-selective protein cleavages (21, 22) on a five-lane denaturing gel.

ACKNOWLEDGEMENTS

We thank Les Lane, Roy French, Michael McClelland, and Myron Brakke for comments on the manuscript and Hagsdan Piperman for secretarial assistance. This work was supported in part by Public Health Service Grant GM-32441 from National Institute of General Medicine to J.V.E.

REFERENCES

1. Maxam,A.M. and Gilbert,W. (1977) *Proc. Nat. Acad. Sci. USA* **74**, 560–564.
2. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Nat. Acad. Sci. USA* **74**, 5463–5467.
3. Shannon,C.E. and Weaver,W. (1949) *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL.
4. Maxam,A.M. and Gilbert,W. (1980) *Methods Enzymol.* **65**, 499–559.
5. Rubin,C.M. and Schmid,C.W. (1980) *Nucleic Acids Res.* **8**, 4613–4619.
6. Friedmann,T. and Brown,D.M. (1979) *Nucleic Acids Res.* **8**, 615–622.
7. Iverson,B.L. and Dervan,P.B. (1987) *Nucleic Acids Res.* **15**, 7823–7830.
8. Fritsche,E., Hayatsu,H., Igloi,G.I., Shigeru,I. and Kossel,H. (1987) *Nucleic Acids Res.* **15**, 5517–5527.
9. Blake,I.F. (1972) *Algebraic Coding Theory: History and Development*, Dowden, Hutchinson & Ross, Inc., Stroudsburg, PA.
10. Golay,M.J.E. (1949) *Proc. Inst. Radio Engineers* **37**, 657.
11. Ambler,R.P. (1976) In Markham,R. and Horne,R.W. (eds), *Structure-function Relationships of Proteins*, North-Holland Publishing Company, Amsterdam, pp. 1–14.
12. Church,G.M. and Gilbert,W. (1984) *Proc. Nat. Acad. Sci. USA* **81**, 1991–1995.
13. Church,G.M. and Kiefer-Higgins,S. (1990) *Science* **240**, 185–188.
14. Chee,M. (1991) *Nucleic Acids Res.* **19**, 3301–3305.
15. Beck,S. and Pohl,F.M. (1984) *EMBO J.* **3**, 2905–2909.
16. Nelson,R.M. and Danby,P.C. (1987) British Patent GM120 941B.
17. Brumbaugh,J., Middendorf,L.R., Grone,D.L. and Ruth,J.L. (1988) *Proc. Nat. Acad. Sci. USA* **85**, 5610–5614.
18. Smith,L.M., Kaiser,R.J., Sanders,J.S. and Hood,L. (1987) *Methods Enzymol.* **155**, 260–301.
19. Jay,D.G. (1984) *J. Biol. Chem.* **259**, 15572–15578.
20. Jue,R.A. and Doolittle,R.F. (1985) *Biochemistry* **24**, 162–170.
21. Spande,T.F., Witkop,B., Degani,Y. and Patchornik,A. (1970) *Adv. Protein Chem.* **24**, 97–260.
22. Witkop,B. (1961) *Adv. Protein Chem.* **16**, 221–321.