A Novel Method for Nucleic Acid Sequence Determination

WILLIAM BAINS†§ and GEOFF C. SMITH‡

Department of Biochemistry† and School of Mathematical Sciences‡
University of Bath, Claverton Down, Bath BA2 7AY, U.K.

(*Received* 27 *May* 1988)

We describe a novel sequencing methodology which should be readily and completely automated. The method relies on fragmentation of a nucleotide or deoxynucleotide sequence into short fragments, and subsequent quantitation of the fragments by hybridization to oligo-deoxynucleotides on a solid support. The original sequence may be reconstructed from the resulting table of fragment frequencies. We present a specific protocol which would allow practical implementation of this approach.

The determination of nucleic acid sequences is one of the primary methods of molecular biology. The rapid methods of Sanger *et al.*, (1980) and Maxam & Gilbert (1977) have made "DNA sequencing" accessible to many laboratories. However, these methods are labour-intensive, and so resistant to full automation: consequently they are too slow (about 1 kilobase per postdoctoral worker per day) and expensive (about $1 per base) (Wada, 1987) for very large sequencing projects. Those semi-automated procedures that are available all rely on operator preparation and loading of electrophoretic gels (Smith *et al.*, 1986), and many are only methods for analysing autoradiographs (Anbalagan *et al.*, 1986; Elder *et al.*, 1986). If entire eukaryotic genomes are to be sequenced, a technology which would be more amenable to automation would be desirable. In this paper we describe the basis for a technology, called Fragmentation Sequencing, which might fill this role.

Basis

In other sequencing methods, each reaction product must be accurately characterized by size to provide complete information about a single base position. Fragmentation sequencing requires only quantitation of reaction products, each of which provides partial information about the entire sequence. The DNA is fragmented into all possible short sequences of a specified length ("fragments") and their number counted. The originating sequence is reconstructed from this data alone.

For exposition, we will consider fragmentation of a sequence into four-base fragments (tetranucleotides). For the purpose of analysis we shall assume that the terminal tetranucleotides are known. To attempt a reconstruction from the frequency table, we look to see if any overlaps are forced. For example, suppose that there is

§ To whom correspondence should be addressed.

0022-5193/88/230303+05 $03.00/0

a unique (non-terminal) tetranucleotide ending in CAG, and there is a unique (non-initial) tetranucleotide beginning in CAG. Overlapping juxtaposition of these tetranucleotides is forced.

We address ourselves to the question: What are the obstacles to the successful reconstruction of the sequence?

Given any tetranucleotide w we may attempt to grow a "forced subsequence" from seed w. We will obtain a subsequence of the original string maximal with respect to the property of being deduceable from the data. Let the first trinucleotide of this maximal subsequence be $L(w)$ and the last trinucleotide be $R(w)$.

We repeat the process with the remaining tetranucleotides successively until we have forced subsequences $w_1, \dots, w_n$. Suppose without loss of generality that $L(w_1)$ is the initial triplet and $R(w_1)$ is the final triplet. If $n = 1$ we have solved the problem.

For each $j(<n)$ there must exist two values $k(>1)$ such that $R(w_j) = L(w_k)$, and similarly for each $k(>1)$ there must exist at least two values of $j(<n)$ such that $R(w_j) = L(w_k)$. If this were not the case we could build longer forced subsequences. We now recast the problem. For each distinct value $L(w_i)$ and $W(w_i)$ we assign a symbol $A_1, \dots, A_n$. To each subwork w_i we assign the dyad of "A"s corresponding to $L(w_i)$ and $R(w_i)$.

Our original problem now reduces to finding the correct ordering of the "A"s, and then inserting between $L(w)$ and $R(w)$ corresponding to A_iA_j the relevant subsequence. (Other dyads A_iA_j may occur several, say M_{ij}, times.)

The total number of sequences compatible with our data is thus the number of possible "A"-sequences compatible with our "A"-dyad data, multiplied by

$$\prod_{i,j=1}^{m} (M_{i,j})!$$

An implementation of this is illustrated in Fig. 1. One tetranucleotide is chosen as the left-most tetranucleotide—AGCT in this example. Two tetranucleotides match the 3′ three bases: GCTC and GCTA. The former leads to the sequence AGCTC, which cannot be extended further, so this route fails with eight tetranucleotides unused. Thus the fifth base must be A. This can only be extended with an A and so on (GTCTAA is the start of a forced subsequence). Only one final sequence is compatible with the tetranucleotide table, regardless of the tetranucleotide used to start the algorithm.

Sequence Data Tests

The example in Fig. 1 shows that a unique reconstruction can be obtained even if there is internal repetition in the sequence. However, some types of repetition will cause values of $n > 1$, and hence ambiguous sequence deduction. This is illustrated in Fig. 2(a), where genomic DNA sequences have been fragmented into their component tetranucleotides and then the number of sequences which may be constructed from that tetranucleotide table counted. For sequences over 40 bases the numbers of possible sequences can be quite large. The tetranucleotide example is therefore not ideal for actual implementation.

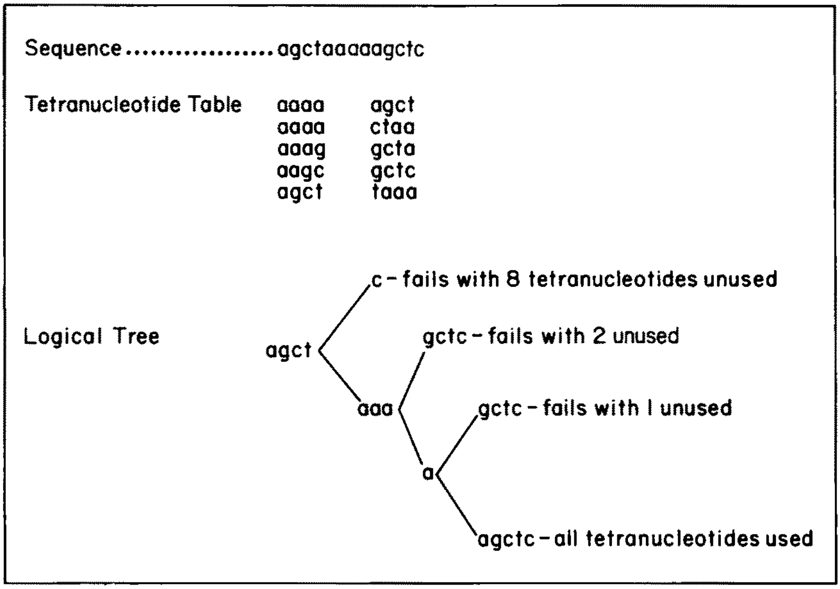
Sequence.................agctaaaaagctc
Tetranucleotide Table
aaaa agct
aaaa ctaa
aaag gcta
aagc gctc
agct taaa
Logical Tree
agct
c - fails with 8 tetranucleotides unused
aaa
gctc - fails with 2 unused
a
gctc - fails with 1 unused
agctc - all tetranucleotides used

FIG. 1. Example of logic used to identify sequences which are compatible with a tetranucleotide table.

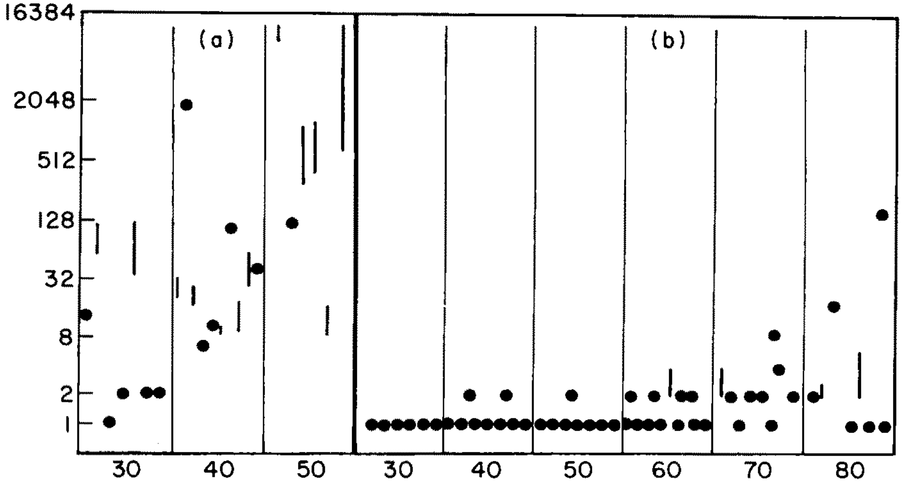

FIG. 2. Observed number of sequences compatible with the fragments produced from randomly selected segments of human genomic DNA. x-axis: length of original DNA segment. y-axis: number of sequences found. In each case a genomic sequence from the human beta globin locus (GenBank sequence HUMHBB) was fragmented into its n-base fragments and the table of those fragments reconstituted into all compatible sequences. The top end of the bars shows the total number of sequences found, the bottom end the number which have the correct left- and right-most tetranucleotides. A circle indicates these two numbers are identical. Fig. 2(a): number of sequences which can be reconstituted from tetranucleotide tables. Fig. 2(b): number of sequences which can be reconstituted from ambiguous hexanucleotide tables.

Ambiguous Hexanucleotide Fragmentation

Of several methods to remove the problem of disambiguation implied by Fig. 2(a), the most likely to yield a workable experimental protocol is the use of ambiguous hexanucleotides. We construct a frequency table of hexanucleotides such that the middle two bases are not specified: each entry is a pool of the frequencies of 16 hexanucleotides with all possible central dinucleotides. If such tables are used in the reconstruction experiment described above, few sequences shorter than 70 bases yield more than one sequence consistant with the hexanucleotide table. The reason for this is that ambiguous hexanucleotides sample the sequence over a larger range than unambiguous tetranucleotides, and consequently require more complex repetition to generate ambiguities in reconstruction. Thus if ambiguous hexanucleotide frequencies could be measured in sequences 70–80 bases long, these fragments could be used to reconstruct the original sequence.

Practical Implementation

Labelled hexanucleotides could be generated from single-stranded DNAs or RNAs by Klenow polymerase or reverse transcriptase copying, respectively. Fragmentation of products into short oligonucleotides enzymatically or by chemical degradation would yield a mixture of lengths of oligonucleotide. This would be hybridized to a panel of the 256 ambiguous hexanucleotides on solid support: conditions would be chosen so that only exactly matching sequences hybridized and hence that penta-nucleotides would not interfere with hexanucleotide quantitation (Thein & Wallace, 1986). Site specificities of fragmentation could be removed by suitable combination of fragmentation regimes or compensated after quantitation if the nature of the bias is known. Sonication, for example, fragments DNA essentially at random. Hexanucleotides can hybridize specifically to single stranded DNA (Thein & Wallace, 1986; Feinberg & Vogelstein, 1983, 1984), and the different melting temperatures of A-rich and G-rich hexanucleotides can be abolished by carrying out the hybridization in tetramethylammonium chloride (Wood *et al.*, 1985). Thus there is no methodological constraint on this aspect of the procedure. The hybridization and subsequent quantitation of label by Cherenkov counting could be carried out entirely automatically.

Accuracy Constraints

It is important to show that the accuracy of measurement required is achievable by present hybridization methods. The method requires the relative amount of labelled DNA bound to each of 256 oligonucleotides to be measured: for a 70–80 base fragment, this means that the majority of oligonucleotides will have no labelled DNA bound. The amount of label bound to the others will depend on the sequence to be determined. One-hundred 70-base sequences from the human beta globin locus were fragmented into their component ambiguous hexanucleotides: only 0·74% of hexanucleotides were represented more than three times in any one sequence. (Of this 0·74%, 59% were represented four times). Thus for the method to be practical, it must be possible to quantitate the amount of hybridization so as to distinguish between 0, 1, 2, 3 and more relative units of product: with the exception of the sequencing of exactly repeating satellite DNAs the absolute amounts are not required. The ability of oligonucleotide hybridization to distinguish between the presence of 0, 1 and 2 copies of specific sequences in human genomic DNA by hybridization to Southern blots (see, for example, Conner *et al.*, 1983; Kidd *et al.*, 1983; Bos *et al.*, 1985) suggests that the desired accuracy will be easy to attain in the less technically exacting hybridization of these oligonucleotides to cloned DNAs.

Conclusion

To sequence large genomes in acceptable time and cost limits, a methodology that is more suitable for automation than existing methods is needed (Wada, 1987). Fragmentation sequencing removes the necessity for electrophoretic separation of products, the most labour- and skill-intensive part of existing procedures. It should therefore be readily automatable and, when coupled with existing rapid methods for generating labelled single stranded DNA for m13 clones (Sanger *et al.*, 1980; Messing, 1983), should consequently improve both the cost and the speed of DNA sequencing.

REFERENCES

ANBALAGAN, R., WARNER, M., WILDING, P., SUMMERS, M. R. & AVDALOVI, N. (1986). *Fed. Proc.* **45**, 1851

BOS, J. L., TOKSOZ, D., MARSHALL, C. J., VERLAAN-DE VRIES, M., VEENEMAN, G. H., VAN DER EB, A. J., VAN BOOM, J. H., JANSSEN, J. W. G. & STEENVOORDEN, A. C. M. (1985). *Nature, Lond.* **315**, 726-730

CONNER, B. J., REYES, A. A., MORIN, C., ITAKURA, K., TEPLITZ, R. L. & WALLACE, R. B. (1983). *Proc. natn. Acad. Sci.* **80**, 278-282

ELDER, J. K., GREEN, D. K. & SOUTHERN, E. M. (1986). *Nucl. Acids Res.* **14**, 417-424

FEINBERG, A. & VOGELSTEIN, B. (1983). *Anal Biochem* **132**, 6-13

FEINBERG, A. & VOGELSTEIN, B. (1984). *Anal Biochem* **137**, 266-267

KIDD, V. J., WALLACE, R. B., ITAKURA, K. & WOO, S. L. C. (1983). *Nature, Lond.* **304**, 230-233

MAXAM, A. & GILBERT, W. (1977). *Proc. natn. Acad. Sci.* **74**, 560-564

MESSING, J. (1983). *Method in Enzymology* **101**, 20-78

SANGER, F., COULSON, A. R., BARRELL, B. G., SMITH, A. J. H. & ROE, B. A. (1980). *J. molec. Biol.* **143**, 161-178

SMITH, L. M., SANDERS, J. Z., KAISER, R. J., HUGHES, P., DODD, C., CONNELL, C. R., HEINER, C., KENT, S. B. H. & HOOD, L. E. (1986). *Nature, Lond.* **321**, 674-679

THEIN, S. L. & WALLACE, R. B. (1986). In: *Human genetic disease: a practical approach* (ed. K. E. Davies), pp. 33-50, IRL Press.

WADA, A. (1987). *Nature, Lond.* **325**, 771-772

WOOD, W. I., GITSCHIER, J., LASKY, L. A. & LAWN, R. M. (1985). *Proc. natn. Acad. Sci.* **82**, 1585-1588