Proceedings of the Nutrition Society (1996), **55**, 605–612

DNA sequencing

BY W. D. REES
The Rowett Research Institute, Greenburn Road, Bucksburn, Aberdeen AB2 9SB

An organism has all its genetic information stored in a linear form in the DNA molecule(s) of its genome and the molecular biologist can access this by the technique of DNA sequencing. The primary sequence of a DNA molecule describes not only the sequence of proteins and RNA associated with any genes which are present but also how the genes are regulated. Knowledge of the sequence permits the production of recombinant proteins and can tell us a great deal about a gene's biological function. Indeed, with the recent rapid advances in sequencing and cloning technology, the sequence of a gene is often known long before the physiological function has been determined. Recent discoveries from the genome sequencing projects are set to revolutionize our understanding of human genetics and gene–nutrient interactions in health and disease.

INFORMATION CONTAINED IN A GENE SEQUENCE

Frequently the term gene sequence is used rather loosely and refers to the complementary DNA (cDNA) sequence of the gene. cDNA is produced from mRNA and is the sequence of the processed mRNA, whilst the genomic sequence describes the genomic copy of a gene, which in mammals is frequently interrupted with long intron sequences. Once cloned into a vector, cDNA does not differ from genomic DNA and can be sequenced in exactly the same way. The cDNA can be decoded using the genetic code to give the primary sequence of the corresponding protein. Protein sequencing is a complex process that often only reveals a few amino acids, whilst cDNA sequencing is a fast and simple process (provided that it has not proved difficult to clone the gene first!). Where a gene is of unknown function it may be possible to get an idea of the role of its product when there are sequence homologies with other genes of known functions. For example, protein kinases or DNA-binding proteins have highly conserved regions that are distinctive and easily identified in the DNA sequence. This process of reverse genetics has identified many new genes and even non-expressed pseudogenes.

Total genome sizes range from about 4 000 000 base pairs (bp) for a typical bacterium such as *Escherichia coli* to 2 800 000 000 bp for the haploid human genome, and a complete genome sequence represents a major task when a typical sequence reaction will sequence about 400 bp. Even sequencing the genomic copy of a mammalian gene can be limited by the very large amounts of non-coding DNA present in the introns. A typical 2 kb (2000 bases) mRNA may be coded for by a 100 kb gene and frequently only the cDNA and promoter regions are sequenced.

Where the complete genomic sequence of a gene is available it can be used to describe the patterns of exon usage and locate the intron–exon boundaries. The promoter region is the area of the gene where the transcriptional machinery binds to the DNA and starts the production of RNA, making it a key area in the control of gene expression. Many binding sites can be identified from sequence motifs and this can suggest the transcriptional activators that may be involved. This can often be supported by allied techniques

such as DNase footprinting that can identify regions that are protected from attack by nuclease by the binding of regulatory proteins. Such information is valuable in understanding gene–nutrient interactions.

The sequence also contains information on the evolutionary history of the gene. As time passes mutations will accumulate independently in different species. Information on the relatedness of the same gene in two species will reveal the distance between them in evolutionary time. These polymorphisms between two species or even two strains of the same species can be used to study the ecology of complex systems, for example gut microflora where conventional microbiological typing techniques can be complex and ambiguous (Avgustin *et al.* 1994).

PRINCIPLE

DNA is composed of two complementary chains of deoxyribonucleotides joined through phosphodiester linkages. The close similarity between the deoxyribonucleotides, which differ only in the heterocyclic base groups, complicates any sequencing method based on chemical cleavage of the phosphodiester bonds. However, Maxam & Gilbert (1977, 1980) devised a chemical protocol that was capable of sequencing significant lengths of DNA by chemical means. This is a complex procedure which is useful for some specialized applications, but has been replaced by the enzyme-based di-deoxy terminator method for most routine sequencing.

The di-deoxy terminator method of Sanger & Coulson (1975) relies on the ability of DNA polymerase to synthesize a complementary strand of DNA from deoxy nucleoside triphosphates (dNTP) using a single-stranded DNA molecule as a template. This reaction is then spiked by the addition of a small amount of a di-deoxy nucleoside triphosphate (ddNTP). This is an analogue of the dNTP except that it lacks the hydroxyl group at C-3 (Fig. 1). The lack of the hydroxyl group does not prevent DNA polymerase from incorporating a ddNTP into the growing chain. However, the next dNTP in the chain cannot be added since there is no 3′hydroxyl to form the phosphodiester bond. By careful choice of the dNTP:ddNTP it is possible to construct a reaction that produces a series of single-stranded DNA molecules some of which are terminated on each occurrence of a particular base (Fig. 2).

MANUAL DNA SEQUENCING

The first step in the chain termination is to anneal a short synthetic oligonucleotide primer to the single-stranded template. All forms of DNA can be sequenced provided that the template DNA is pure and undegraded (free from nicks). The synthetic primer is

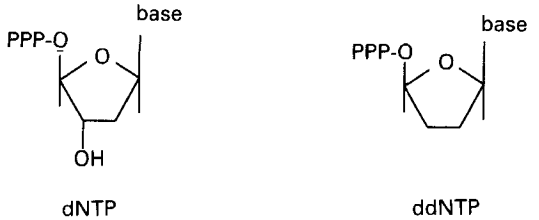

Fig. 1. Deoxy and di-deoxy nucleotides.

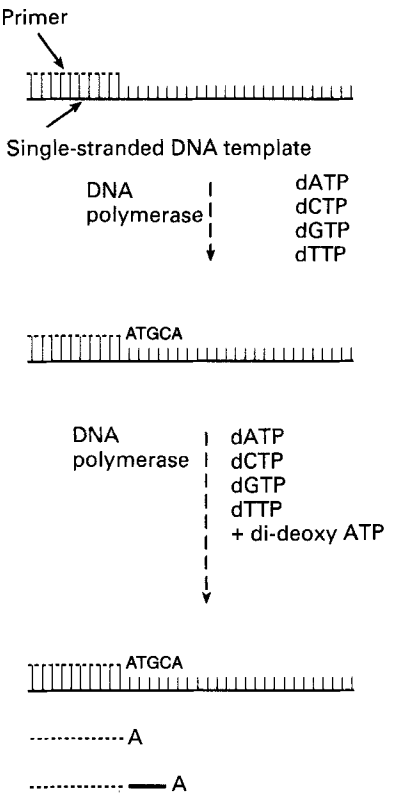

Fig. 2. The di-deoxy terminator reaction. The extension of a single-stranded template is initiated by the binding of a short synthetic oligonucleotide. DNA polymerase then extends the chain using deoxy nucleoside triphosphates (dNTP) in the reaction mixture. When the reaction is spiked by the addition of a small amount of di-deoxy nucleoside triphosphates (ddNTP) there is a possibility that either a conventional dNTP will be added extending the chain, or the ddNTP will be added terminating the chain.

usually an 18–20-mer oligonucleotide that forms a stable duplex with the template. The primer must be free of self complementarity, have no significant secondary structure and only bind to a single site in the template. A number of computer programs are available for primer design. DNA synthesizers for the production of oligonucleotides are commercially available, although expensive to operate. Where use is infrequent it can be an advantage to use one of the companies offering a complete oligonucleotide synthesis and purification service. Because of the high cost of primers it is common to clone the DNA into a plasmid or phage vector and use a sequence close to the polylinker as a primer site; thus, a number of different clones can be sequenced with the same two primers. PCR products can be used used as a template, but it is essential that the product is completely pure.

Four separate chain-extension reactions are then set up, each containing one of the four ddNTP, to produce four sequence ladders each terminated at the added ddNTP. There will only be a small amount of each band present at the end of the reaction, thus a small amount of radioactive dNTP is also added to the reaction so that the newly-produced chains can be visualized. The reaction products are then separated from one

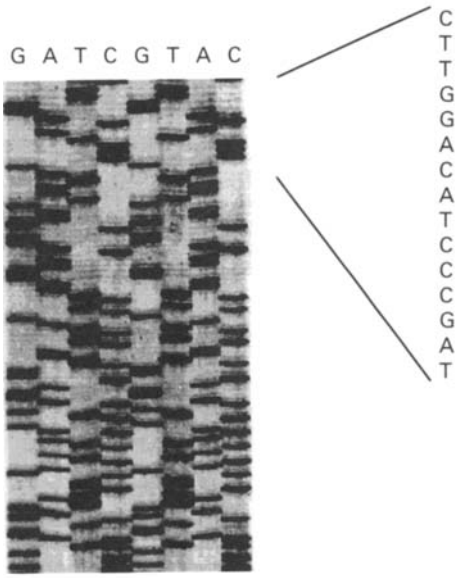

Fig. 3. Part of a typical sequencing gel. The products of four sequencing reactions spiked with di-deoxy (dd) nucleoside triphosphates, ddATP, ddCTP, ddGTP and ddTTP were separated on a 80 g polyacrylamide/l gel, dried, and autoradiographed. The longest chain extension (e.g. the largest) is at the top of the gel and the smallest at the bottom. The sequence can be read from the bottom to the top of the gel (going from 5′ to 3′).

another by running them side by side on a thin polyacrylamide gel. The gel contains urea and is run at a high temperature to separate the newly-synthesized strands from the templates and prevent the formation of secondary structure. The electrophoresis must be carefully controlled to separate bands that are just one nucleotide longer than the next. After drying, the gel is autoradiographed by placing it next to a sheet of X-ray film and 48 h later the film is developed to reveal a pattern of bands similar to the example shown in Fig. 3.

The sequence of the template can then be read, as each band is one nucleotide longer than its predecessor. The band that has moved the furthest represents the smallest product, being the first nucleotide and closest to the primer. The next larger band is the next nucleotide in the sequence and the column in which it appears is the nucleotide at that position. This is continued up on the gel until the bands are so close to one another that the sequence cannot be read further. This method is very effective and allows 250–400 bp of sequence to be read. All the reagents are readily available in kits supplied by major biotechnology companies. Two major items of equipment are required, the special large acrylamide gel tank and a high-voltage power supply.

AUTOMATED DNA SEQUENCING

Manual sequencing is quite laborious and the throughput of samples can be increased by automation. Robotic sample preparation can be employed to simplify the preparation of templates although the high cost and complexity of these workstations confine them to major sequencing centres. Automated sequencers are more common and give a major

increase in efficiency by replacing the gel drying and autoradiography steps with a non-radioactive continuous detection system that allows computerized sequence reading.

One of the most widely used methods conjugates a different coloured fluorochrome to each of the four ddNTP (Prober *et al.* 1987). The chain-extension reaction incorporates them in exactly the same way as the ddNTP to produce a ladder of DNA with the four colours indicating the final nucleotide in the chain. The four colours can be analysed simultaneously so that only one reaction is required and only one gel track is used for each sequence reaction. The samples are run on a conventional gel that is scanned by a laser. As the product bands pass the laser beam the fluorophore is excited, producing a characteristic emission depending on which of the four terminators has been incorporated. The fluorescent emissions are detected by a photomultiplier tube that passes its signal to a computer for processing. The complete scan for each of thirty-two lanes on the gel is then analysed by a computer program that automatically generates the sequence. The read lengths are similar to those of the manual method, but with one gel reading more than 10 kb of sequence the throughput of samples is dramatically increased.

SEQUENCING STRATEGY

Sequencing reads of 250–400 bp are short in comparison with the length of genes or chromosomes and it is necessary to integrate the sequence information from a number of different sequence reactions (Fig. 4). For short lengths, such as cDNA clones, it is possible to use the sequence of the furthest end of the first reaction to design a synthetic primer for the next and so on until the whole sequence has been determined. Otherwise a long section of DNA is fragmented into a number of shorter fragments using two or more restriction enzymes that cut the DNA relatively frequently. These fragments are subcloned into vectors and sequenced using primers that prime in the vector sequence. The overlapping parts of the sequence data can then be aligned to construct a full sequence of the target DNA. Where a gene of interest is closely linked (i.e. inherited with an identical frequency and, therefore, physically located close to the target gene on the chromosome) it is possible to find the unknown gene by 'chromosome walking'. By sequencing and re-assembling overlapping clones derived from the region around the marker it is possible to identify the sequence motifs that describe the presence of new genes, one of which may be the gene of interest. It is by the use of this technique that a number of important genes have been identified, the best known example being the gene for cystic fibrosis.

DATA ANALYSIS

Large quantities of sequence data are very unwieldy and difficult to analyse by manual methods. Fortunately, because of their linear nature, sequence data are ideal for computer analysis and a variety of software packages are available.

All sequence data are collected in three principal databases; EMBL, The European Molecular Biology Laboratory at Heidelberg; GenBank in the USA; DDBJ in Japan. Data is exchanged among the three databases daily. In the UK copies of the EMBL and GenBank databases are held on the BBSRC SEQNET computer at Daresbury and can be accessed through JANET or the internet. Besides the sequence database, SEQNET also provides all major software packages (GCG, Serratus, Staden, Pearson, Blast and

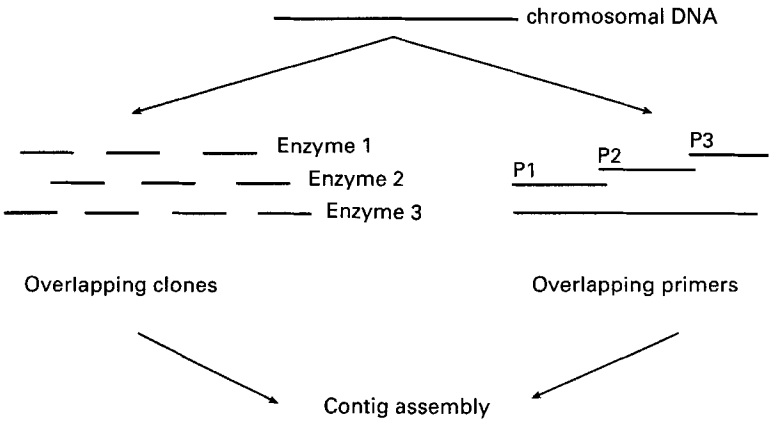

Fig. 4. Sequencing strategy. A long stretch of DNA can either be fragmented using restriction enzymes and then each of the fragments subcloned and sequenced, or using data from the previous sequence a new primer can be designed to sequence onward from the end of the previous read.

Phylip) as well as a host of other utilities for protein analysis. More information is available on the SEQNET home page of the World Wide Web (http://www.dl.ac.uk/SEQNET/home.html). An expressed sequence tags (EST) database (see p. 611) is also available from the National Center for Biotechnology Information (NCBI), at the National Institutes of Health in the USA (Boguski *et al.* 1993). The databases are available on CD-ROM or can be accessed on line through the internet (http://www.ncbi.nlm.nih.gov).

Information is organized in the databases in order of species and each sequence is assigned its own unique accession number. It is usually a condition of publication that any new sequence data should be deposited with one of the databases. In the case of an unknown gene the first step is usually to compare the sequence with all the sequences in one of the databases to see if the gene has any homologies with previously identified genes. Programs can do this with either the DNA sequence (FASTA) or the deduced protein sequence (SWISSPROT). The location of coding regions can be performed using programs that identify patterns and signal sequences.

THE HUMAN GENOME PROJECT

The ultimate analysis of gene–nutrient interactions requires an understanding of not just all the nutrients required by an organism but also all the genes expressed by it. The complete sequence of the human genome will identify and describe all the genes that are responsible for human life. The sequencing of the human genome represents a major undertaking and consortiums of different laboratories are working to compile the sequences of individual chromosomes. As described previously each chromosome is subdivided into large fragments which are cloned in yeast artificial chromosomes (YAC). Each YAC clone is then further subdivided, sequenced, and the complete sequence re-assembled from the overlapping contigs. All these data are then to be re-assembled in a major computer database.

An altogether different approach has been taken by commercial companies with an eye to patenting human genes of commercial importance. The advent of low-cost, high-accuracy and high-throughput automated sequencing has made it possible to randomly sequence all the genes expressed by an individual tissue. These sequences are known as EST. For example, mRNA isolated from the brain is converted to cDNA and cloned into a suitable vector to produce several million clones. Each of these contains an insert corresponding to an mRNA expressed in the brain. The clones are then sequenced with an automated sequencer to produce a database describing partial sequences of all the genes expressed in the brain. Computer analysis is then carried out to eliminate duplicates and show homology with existing sequences. Previously-unidentified sequences are those of new genes which may be of importance in brain function. It is argued that this is a more efficient method of mapping and sequencing the genome, since it describes the 1% of the total DNA that is expressed as RNA.

Considerable progress has already been made and at the end of March this year (1995) the NCBI database already contained about 85 000 human EST and was growing at about 4000–6000 sequences weekly. Commercial companies are keen to patent new genes that have been 'discovered' and there is a vigorous debate on both the legal and ethical aspects of this approach. Large-scale EST sequencing is mechanical science using well-established technology to generate sequence data. The innovative component is the identification of gene function, rather than simply sequence; if a sequence of unknown function can be patented this will reward the routine and discourage novel approaches. The ethical, moral and legal implications of the EST approach has found current patent law wanting and is a good example of science progressing ahead of the legislation.

EST are much more significant, however, than mere sequences, and can be used to generate a high-resolution human genetic map. This will be of great importance in identifying genes associated with particular organs and disease states and will greatly enhance our understanding of gene–nutrient interactions. Each EST is produced from its corresponding gene which has a particular position on a chromosome. Thus, in our example, brain EST can be used to describe the chromosomal loci of genes expressed in the brain. The chromosomal position of an EST is known as the sequence tagged site (STS) and it is planned to identify the chromosomal locations of more than 50 000 cDNA in the next few years (Stewart, 1995). Human geneticists will then be able to compare the inheritance of STS markers with known human traits or diseases; for example, the inheritance of specific brain STS markers may be associated with specific neurological disorders indicating that genes of interest are located at or near that chromosomal locus. More information on this programme is available on the Lawrence Livermore National Laboratory home page (http://www.bio.llnl.gov/bbrp/genome/genome.html).

CONCLUSIONS

Modern methods for the automated sequencing of DNA are about to revolutionize human and animal genetics. A combination of advanced cloning techniques and genome mapping is about to identify most if not all the genes involved in all aspects of animal life. The ability to separate genetic components from nutritional effects in complex disease states is set to enhance our understanding of the inter-relationship between genes and nutrition in all aspects of development, growth and pathology.

REFERENCES

Avgustin, G., Wright, F. & Flint, H. J. (1994). Genetic diversity and phylogenetic relationships among strains of Prevotella (Bacteroides) ruminicola from the rumen. *International Journal of Systematic Bacteriology* **44,** 246–255.

Boguski, M. S., Lowe, T. M. J. & Tolstoshev, C. M. (1993). dbEST a database for 'expressed sequence tags'. *Nature Genetics* **4,** 332–333.

Maxam, A. M. & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences, USA* **74,** 560–564.

Maxam, A. M. & Gilbert, W. (1980). Sequencing end labeled DNA with base specific chemical cleavages. *Methods in Enzymology* **65,** 499–560.

Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., Jensen, M. A. & Baumister, K. (1987). A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238,** 336–341.

Sanger, F. & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94,** 441–448.

Stewart, A. (1995). The human gene map initiative. *Genome Digest* **2,** 1–4.

Printed in Great Britain