Chapter 1

The tools for the job

1.1 Introduction

Cloning and manipulating genes requires the ability to cut, modify and join genetic material (usually DNA, but sometimes RNA) and check the parameters of the molecules, such as size, that are being manipulated. We will assume knowledge of the structure of the materials involved (DNA, RNA and so on) and start by describing the tools available for manipulating them. Many of the tools involved are enzymes that have important physiological roles in cells. To understand why they are useful for our purposes, we should be aware of their normal roles, too.

The choice of which enzyme is used for a particular purpose depends mainly on two considerations:

1. How easy (i.e. inexpensive) is it to purify? This will be determined by its abundance in the cell and by how easy it is to separate it from other undesirable activities.
2. How well does it do the job? This will depend upon its specificity ('accuracy') and specific activity ('speed') and upon the details of the reaction which it catalyses.

Other factors, such as stability, are also important.

Techniques of genetic manipulation can be applied to the production of the enzymes for genetic manipulation itself. It is possible to use cloned genes to prepare large quantities of these enzymes more easily, as well as to modify the genes to 'improve' their function, perhaps by slightly altering the properties of the enzymes they encode.

1.2 Cutting

Enzymes that break down nucleic acids are called nucleases. Those that break down RNA are called ribonucleases, or RNases, and those that break down DNA are called deoxyribonucleases or DNases.

There are two ways of breaking down a linear nucleic acid molecule: dismantling it bit by bit from the ends, or breaking it into pieces by cutting within the molecule. The former is called exonucleolytic activity (Greek *exo* = outside) and the latter endonucleolytic activity (Greek *endon* = within). Do not fall into the trap of thinking that *endo*nucleases work from the *end*s in! For cutting nucleic acid molecules into pieces, therefore, we will need endonucleases, and the most widely used ones are the restriction endonucleases.

1.2.1 Restriction endonucleases

Restriction endonucleases are part of the natural defence mechanisms of bacteria against incoming DNA, which may be from viruses or plasmids from a foreign population of cells. These enzymes were first recognized by their ability to restrict the growth of certain viruses in particular strains of *Escherichia coli*, and were named accordingly. (The verb restrict is now widely used by molecular biologists to mean 'cut with a restriction endonuclease'.) The restriction enzymes are associated with modifying enzymes, which methylate the DNA. Methylation protects the DNA from cleavage by endonucleases, and this stops the cell from degrading its own DNA. Invading DNA that has not been correctly methylated will be degraded unless it can be modified by the cell's methylating enzymes quickly enough, which happens only rarely.

DNA, once modified, remains protected even after replication. This is because semiconservative replication of a molecule methylated on both strands results in two daughter molecules that are hemimethylated (i.e. methylated on one strand), and hemimethylation is sufficient to confer protection against cleavage by an endonuclease. The non-methylated strand can then be modified before replication takes place again.

Three types of restriction/modification system are recognized. These are called Types (or Classes) I, II and III, and their key properties are summarized in Table 1.1. All the enzymes recognize particular DNA sequences, but only the Type II endonucleases cut within those recognition sequences. The recognition sites for a number of Type II enzymes are given in Table 1.2. These enzymes often make a 'staggered' cut to leave molecules that, although primarily double stranded, have short single-stranded ends. These are called sticky ends. Depending on the enzyme, either the 5′ end or the 3′ end may be left single stranded. The molecules generated have a phosphate group on the 5′ end and a hydroxyl group on the 3′ end. A small number of Type II enzymes cut just outside their recognition sites; for example, *Mbo*II cuts seven nucleotides 3′ to its recognition site of -GAAGA-. Others cleave within their recognition site, but at degenerate sequences; for example, *Mam*I cuts at -GATNN′NNATC-, where N can be any nucleotide. Nevertheless, both of these types of enzyme are still recognizably Type II on the basis of their biochemical properties. Except in the special cases just noted, all DNA molecules

Table 1.1. Characteristics of restriction and modification systems

	Class I	Class II	Class III
Composition	Multienzyme complex with R (endonuclease), M (methylase), and S (specificity) subunits, e.g. as R_2M_2S	Separate enzymes; endonuclease is a homodimer, methylase a monomer	M subunit provides specificity; on its own, functions as methylase; as heterodimer with R subunit, functions as methylase-endonuclease
Cofactors[a]	Mg^{2+}, ATP, SAM (needed for cleavage and methylation)	Mg^{2+}, SAM (for methylation only)	Mg^{2+}, ATP (for cleavage), SAM (needed for methylation; stimulates cleavage)
Recognition sites	Asymmetric, bipartite, may be degenerate, e.g. *Eco*K (AACN$_6$GTGC)	Symmetric, may be bipartite, may be degenerate (Table 1.2)	Asymmetric, uninterrupted, 5–6 nt long. E.g. *Eco*P15-CAGCAG. Two copies in opposite orientation, but not necessarily adjacent, needed for cleavage; one for methylation
Cleavage	Variable distance (100–1000 nt) from recognition site	Within recognition site, except for Class IIs (**s**hifted cleavage), which cleaves outside, at a defined distance	25–27 nt from recognition site
Number of systems characterized	Several, grouped into a few families. e.g. K, includes *Eco*B, *Eco*D, *Eco*K, and others	Hundreds	Few

[a] ATP: adenosine triphosphate; SAM: *S*-adenosyl methionine.

resulting from cutting with a given Type II enzyme will have the same sequences at their ends. That will not be true with the Types I and III enzymes, as they cut outside their recognition sites. Because molecules that are cut with Type II enzymes generally have the same ends, such molecules can base-pair with each other, and, as we shall see, be covalently joined by a DNA ligase. Some Type II enzymes give clean cuts rather than staggered ones, cutting both strands at the same place (see Table 1.2). This gives double-stranded or blunt ends on the molecules. That is not a problem, since blunt-ended

Table 1.2.	Examples of recognition sequences of Type II restriction endonucleases[a]		
*Apa*I	G GGCC′C C′CCGG G	*Aha*III	T T T′A A A A A A′T T T
*Bam*HI	G′GATC C C CTAG′C	*Bgl*II	A′GATC T T CTAG′G
*Bsp*120I	G′GGCC C C CGG′C	*Dpn*I	GA′ TC CT′AG
*Dra*I	T T T′A A A A A A′T T T	*Eco*RI	G′A A T T C C T T A A′G
*Hinc*II	GTPy′PuAC CAPu′PyTG	*Hind*III	A′AGCT T T TCGA′A
*Hpa*II	C′CG G G GC′C	*Mae*III	′GTNAC CANTG′
*Not*I	GC′GGCC GC CG CCGG′CG	*Pvu*II	CAG′CTG GTC′GATC
*Sal*I	G′TCGA C C AGCT′G	*Sau*3A	′GATC CTAG′
*Sph*I	G CATG′C C′GTAC G	*Taq*I	T′CG A AGC′T
*Xba*I	T′CTAG A A GATC′T		

[a] ′=cleavage site; N=any nucleotide; Py and Pu=pyrimidine and purine nucleotides respectively.

molecules can also be joined by ligase. The following features of cleavage by Type II enzymes are also important:

1. **Recognition sites generally read the same on both strands** (as long as the same polarity, e.g. 5′ to 3′, is read). Such sequences are often described as palindromes. It is not necessary for recognition sequences to be palindromic for all molecules cut with the same enzyme to be able to reanneal, although it does increase the number of configurations in which reassociation can take place. For example, Figure 1.1 shows how two molecules cut with the enzyme *Hin*dIII (recognition sequence -AAGCTT-) can reanneal, with either end of the right-hand molecule annealing with the left-hand one. Two molecules cut with an enzyme with a non-palindromic recognition sequence could also reanneal, but fewer orientations are possible.
2. **Most enzymes have recognition sites of four or six nucleotides.** If all nucleotides occurred with equal frequencies (both in the DNA to be cut and in the enzyme recognition sites) and at random, a particular four-nucleotide motif would be expected to occur on average once every 4^4 (i.e. 256) nucleotides. So the average length of fragments generated by enzymes with such sites

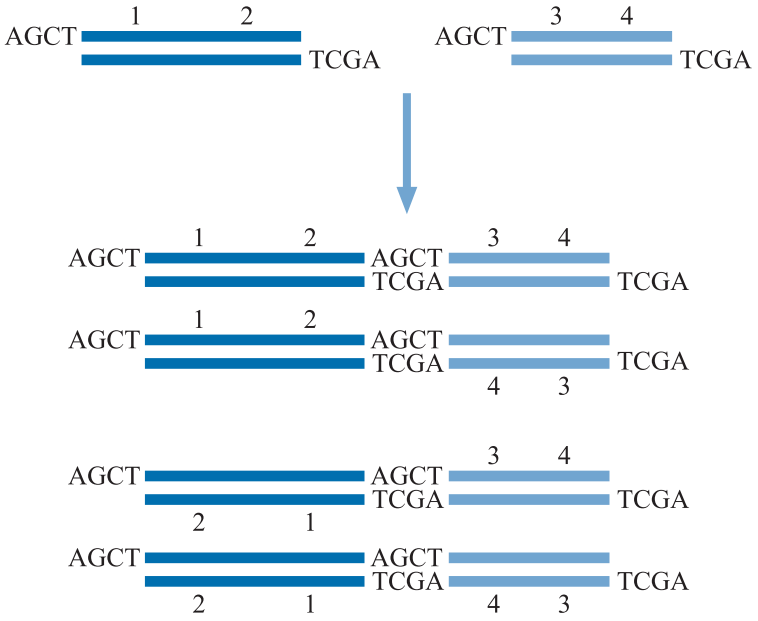

Fig I.1 Annealing of two molecules cut with *Hin*dIII. 1, 2, 3 and 4 represent arbitrary points on the molecules. Note that the right-hand molecule can anneal in two possible orientations to either end of the left-hand molecule, because of the palindromic nature of the *Hin*dIII cleavage site.

would be 256 base-pairs. Similarly, enzymes with a six-nucleotide recognition sequence would generate fragments with an average size of 4^6 (i.e. 4096) base-pairs. In practice, that does not happen for the following reasons:

(a) The bases do not occur with equal frequencies in the recognition sites.

(b) The bases do not occur with equal frequencies in the DNA to be cut, and the frequencies can vary over the genome.

(c) The bases do not occur at random, e.g. certain dinucleotides are favoured and others avoided. The degree of non-randomness often varies over a genome.

3. **Different enzymes can recognize the same sequence.** For example, *Dra*I and *Aha*III both recognize and cut at -TTT′AAA-. They are said to be isoschizomers. Enzymes with the same recognition sequence do not necessarily cut at the same position within it, though. For example *Apa*I recognizes and cuts at -GGGCC′C-, whereas *Bsp*120I recognizes and cuts at -G′GGCCC-.
4. **Different enzymes can generate the same ends.** For example, the enzyme *Sau*3AI produces the ends GATC-, and *Bam*HI does the same. This means that molecules produced by digestion with *Sau*3AI will be able to anneal and be ligated to molecules produced with *Bam*HI. Given that blunt-ended molecules can also be ligated, molecules cut with any enzymes that give blunt ends can be compatible. Notice, though, that ligation of molecules cut with different enzymes may not regenerate the original recognition sites used. It may also be possible to ligate (though at a lower efficiency) molecules whose sticky ends are nearly, but not fully complementary.

5. **Cutting can be influenced by other factors**. The most important are:
   (a) methylation;
   (b) the buffer used;
   (c) secondary structure in the substrate.

   Methylation of bases in DNA may result from the modification activity of a restriction–modification system, or from the activity of one of many independent methylases. Methylated bases commonly encountered include N^6-methyladenine, 5-methylcytosine, 5-hydroxymethylcytosine and N^4-methylcytosine. Restriction enzymes will generally not cut molecules where particular bases within their recognition site are methylated. Methylation at certain positions within the recognition site may not affect cleavage, and for some enzymes methylation at other positions may actually be required for cleavage. For example, cleavage by *Bam*HI is inhibited by methylation at the internal C of the -GGATCC- recognition site, but not by methylation of the other C or the A, whereas cleavage by *Apy*I is inhibited by methylation of the first C of its recognition site (-CCAGG- or -CCTGG-) and requires methylation of the second C.

   The specificity of some enzymes is affected by the buffer used. For example, the enzyme *Eco*RI normally cuts at the sequence -GAATTC-, but the specificity is relaxed in the presence of glycerol at concentrations greater than 5% v/v, and cutting can take place at -AATT- or -PuPuATPyPy-. This is often refered to as star activity, denoted *Eco*RI*. The extent of cutting can be modified by certain compounds. One example is ethidium bromide, shown in Figure 1.2. This molecule, which is also used for visualizing nucleic acids in gels, can be intercalated (inserted) between the bases in a double-stranded DNA molecule. This interferes with the action of restriction endonucleases, allowing cutting in one strand only.

   Some sites are cut much less efficiently than others within the same molecule. This may be due to secondary structures in the DNA that interfere with recognition or cleavage by the endonuclease.
6. **Enzyme activities are measured in units.** A unit is the amount of an enzyme required to digest a standard amount (usually 1 μg) of a standard type of DNA (often bacteriophage lambda, or a specified plasmid) in a given time (usually 1 h) under given

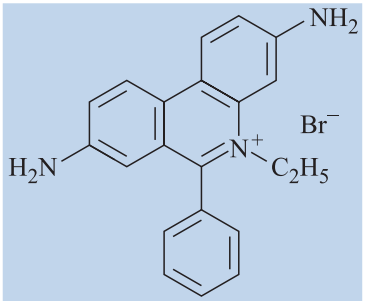

Fig I.2 Ethidium bromide.

conditions (temperature, pH, etc.). Digesting DNA molecules containing many sites may, therefore, require more units of enzyme than the amount required to digest the same mass of a DNA containing fewer sites.

7. **Restriction endonuclease preparations used for cloning must be free of other nucleases.** If not, the ends of the molecules generated might be degraded by exonucleases, and reannealing would be prevented. Contaminating endonuclease activity would cut the molecules into fragments with no (or the wrong) single-stranded ends, which would cause a similar problem. Manufacturers, therefore, usually test enzyme preparations by incubating DNA with a large excess of the enzyme, and determining what proportion of the products can be religated and whether the religated molecules can still be cut with the enzyme. The higher the proportion of correct religation, the 'cleaner' the enzyme preparation. Enzyme preparations can also be tested for the presence of exonuclease by incubation with DNA molecules that are radioactively labelled at their ends. The presence of exonuclease is indicated by the release of the radioactive label from the substrate DNA. Low levels of contaminating nucleases are not a problem in simple restriction enzyme mapping, though.
8. **Partial digestion may be useful.** Sometimes we deliberately do not carry out digestion with a restriction enzyme to completion. For example, we might need to fragment total DNA (often called genomic DNA) prepared from an organism into pieces of roughly the same size, say 10 kbp, so that every sequence in the organism is represented in the collection of 10 kbp fragments. Simply cutting to completion with an enzyme with a six-nucleotide recognition sequence and taking the fragments produced that were approximately 10 kbp would not be suitable. A lot of the DNA would only be cut into smaller (or larger) pieces and would never be represented in the 10 kbp size class. Therefore, a better method is to use an enzyme that cuts very frequently (e.g. at a four-nucleotide recognition site), but to adjust either the reaction time allowed or the ratio of enzyme to DNA in the reaction so that only a few of the possible sites are cut. In this way, the average size of the fragments can be raised to 10 kbp or whatever else is required; and if all sites have a more or less equal chance of being cut, then all regions of the DNA can be represented among the 10 kbp fragments (unless the distribution of sites in a particular region is grossly abnormal). This approach is very useful in constructing genomic libraries (see Chapter 5).
9. **Cleavage sites can be determined using standard molecules.** A selection of DNA molecules whose sequence is known completely are digested with the enzyme. Samples are also digested by combinations of the enzyme under test with others that have known cleavage sites. The sizes of the fragments generated are measured by gel electrophoresis, and a computer

analysis allows you to infer possible recognition sites from the sequences on the grounds that they are the only sites that would generate fragments of the observed sizes. More accurate measurement (to the exact number of nucleotides) of the sizes of molecules generated by digestion allows the actual cleavage site within the recognition sequence to be inferred. These accurate size measurements are also done electrophoretically, using the products of DNA sequencing reactions as size markers.

10. **Nomenclature follows a simple convention.** Once an enzyme has been characterized, it must be given a name. The convention is that names start with three letters (italicized): the first letter of the genus and the first two letters of the species of the source cells. Where relevant, they are followed by an indication of the strain and then a number (in Roman numerals) indicating which one of the enzymes from that strain the name refers to. For example, the enzymes *Eco*RI and *Eco*RII refer to the first and second activities isolated from strain R of *E. coli*. Often, names are abbreviated, so *Eco*RI is often referred to colloquially just as 'RI' (pronounced 'R-one'), *Bam*HI as 'Bam' and so on. According to the general conventions for enzyme nomenclature, Types I, II and III restriction endonucleases are classified as 'endodeoxyribonucleases producing 5′-phosphomonoesters' and classified as EC 3.1.21.3, EC 3.1.21.4 and EC 3.1.21.5, respectively.
11. **Other enzymes can cut DNA molecules at specific sequences.** During the course of infection of *E. coli* cells by bacteriophage lambda, copies of the phage genome are cut at a specific 16-nucleotide site, called *cos*, leaving a 12-nucleotide single-stranded overhang. The cleavage is carried out by a phage enzyme called the terminase. *Cos* sites are sometimes introduced into large DNA molecules to allow cleavage at a single site when conventional restriction enzymes would cut the molecule into several pieces.

1.2.2 DNase

In some instances, restriction endonucleases are unsuitable for cutting DNA. That might be so if the DNA has a very abnormal base composition, although such a wide variety of enzymes is now available, with so many recognition sites, that this is rarely a problem. A more common problem is when it is necessary to break DNA into a random collection of fragments with a mean size of only a few hundred base-pairs. Partial digestion with a four-nucleotide-recognizing enzyme is not suitable; nearly every site would have to be cut to get the required average size, and this would mean that some sequences would be represented only on fragments either much smaller or much larger than the required size range. The problem can be avoided using a DNase such as DNaseI, which has very little (and for this purpose essentially no) sequence specificity. Again, careful adjustment of either the enzyme:DNA ratio or the incubation

time is necessary to ensure the optimal distribution of fragment sizes. One problem with the use of DNase is that the ends of the molecules produced do not have a unique single-stranded sequence. Also, not all the ends are blunt. This makes cloning of the fragments difficult, but the problem can be solved by rendering all the ends blunt with a suitable DNA polymerase.

1.2.3 Physical stress

In addition to enzymatic means, we can use physical shearing to cleave DNA at random. We can accomplish this in several ways. For example, we can simply stir a solution or force it through a narrow opening such as a syringe needle or a pipette tip, or we can use sonication (which provides high-frequency vibrations). In practice, sonication is the preferred method, since it is the easiest to control and is often more reproducible than DNaseI treatment. Different kinds of sonicator are available. In the simplest form, a metal probe is dipped into the solution and vibrates at high frequency. This has the disadvantage that the probe can be a cause of cross-contamination between DNA preparations unless it is carefully cleaned. An alternative instrument is the cup-horn sonicator, where the solution to be sonicated is retained in a tube that floats in a small volume of water. The probe is dipped into the surrounding water and vibrations are transmitted through the water to the tube containing the sample. With shearing, as with DNase treatment, there is no control over the sequences at the ends of the fragments produced.

1.3 Modification

1.3.1 Phosphatases

Phosphatases are enzymes that hydrolytically remove phosphate groups from DNA molecules, replacing them with hydroxyl groups. The terminal phosphate groups left by restriction enzymes are needed for most ligation reactions, and the application of phosphatase in blocking unwanted ligation reaction will be described in Chapter 3. Widely used preparations come from calf intestines, shrimps and the Antarctic psychrophilic (cold-loving) bacterium TAB5. Many phosphatase preparations, especially the last one, can be readily inactivated by heating. This is useful when we want to terminate phosphatase activity prior to a ligation.

1.3.2 Polymerases

We will meet four classes of DNA or RNA polymerases: DNA-dependent DNA polymerases, RNA-dependent DNA polymerases, DNA-dependent RNA polymerases and template-independent polymerases. There are also RNA-dependent RNA polymerases, but they are less important for our purposes.

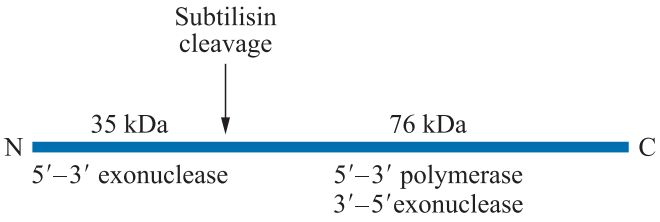

Fig I.3 DNA polymerase I. The locations of the activities and the cleavage site of subtilisin are indicated.

1. **DNA-dependent DNA polymerases.** These enzymes synthesize a DNA strand in a 5′-3′ direction using a DNA template. They can also have 5′-3′ and 3′-5′ exonuclease activities, and all these activities can be exploited in various ways. The preparations used come from bacteria, such as *E. coli* and *Thermus aquaticus*, and from bacteria infected with viruses such as T4 and T7. The *E. coli* enzyme that is widely used is DNA polymerase I, which normally has important roles in DNA repair and the replacement with DNA of the RNA primers used for DNA synthesis. The enzyme has 5′-3′ polymerase, 3′-5′ exonuclease (serving a proof-reading function), and 5′-3′ exonuclease activities. These are essentially located on different domains of the molecule. Cleavage with the protease subtilisin generates an N-terminal fragment of 35 kDa containing the 5′-3′ exonuclease activity and a C-terminal one of 76 kDa with the polymerase and 3′-5′ exonuclease activities (Figure 1.3). The 76 kDa piece is sometimes called the Klenow fragment and the intact molecule the Kornberg enzyme.

The 5′-3′ DNA polymerase activity allows a complementary DNA strand to be synthesized using a suitable template. This template might be a large piece of single-stranded DNA with a small primer annealed, or it might be a restriction fragment with a recessed 3′ (i.e. overhanging 5′) end. Incubation of either of these templates with DNA polymerase and the correct deoxynucleoside triphosphates would result in the filling in of the single-stranded region (the 'recessed end') to produce a blunt-ended molecule. This is often called end filling. Note that the 5′-3′ exonucleolytic activity of the Kornberg enzyme could also produce such a molecule by degradation of the overhanging 5′ end rather than by synthesis of its complement. A recessed 5′ end cannot be rendered blunt by the polymerase activity, because synthesis would have to be in the 3′-5′ direction, which is not possible. In that case, the 3′-5′ exonucleolytic activity could render the ends blunt by degradation of the overhanging 3′ end. Rendering overhanging ends blunt by any of these means is termed polishing, and is summarized in Figure 1.4. *E. coli* DNA polymerase is also used for DNA sequencing.

Thermostable DNA polymerases are particularly important for amplification of DNA by the polymerase chain reaction (PCR), as described in Chapter 2. They are isolated from extremely thermophilic bacteria, often growing in hyperthermal oceanic vents, such as *Thermus aquaticus*, *Thermococcus litoralis* and *Pyrococcus furiosus*.