Restriction Endonucleases, Vectors, Recombinant DNA

Major concepts

- Sequence-specific cleavage of DNA
    - Palindrome.
    - Endonuclease.
    - Restriction endonuclease.
    - Restriction site.
    - Site-specific cleavage.
    - Average frequency of cutting
    - Restriction mapping
- Recombinant DNA
    - Staggered cuts.
    - Sticky ends.
    - Recombination of DNA fragments
- Plasmid Vectors.
    - Inserting DNA sequences into vectors
    - PBR322 plasmid.
    - Disruption of antibiotic-resistance genes with DNA inserts
    - Multiple cloning sites (polylinkers).
    - Blue/white screening.

Recombinant DNA technology: This is the first of a series of lectures on gene cloning and other aspects of recombinant DNA technology. Recombinant DNA in this context refers to the creation of a new combination of DNA segments that are not found together naturally. Such technology is now widely used in many practical applications ranging from basic research on control of gene expression to forensic medicine to biotechnology. This lecture focuses on a group of highly specialized enzymes, the restriction endonucleases, which allow DNA to be cut at specific sites in a manner that permits the rejoining of cut ends to create new combinations. This opens the way for gene cloning by inserting the gene of interest into a self-replicating bacterial plasmid (a small circular DNA molecule with its own origin of replication). Later parts of the lecture discuss some of the techniques used to identify vectors containing cloned DNA sequences.

Restriction endonucleases: An *endonuclease* is an enzyme that can cleave the phosphodiester bonds of a nucleic acid at an internal site (as opposed to cleavage by an *exonuclease*, which can only remove nucleotides from one of the ends of a nucleic acid). Some endonucleases cut internal bonds of DNAs or RNAs randomly. However, *restriction* endonucleases cut both strands of a double stranded DNA only at specific restriction sites. There are many different restriction endonucleases, and each is highly specific for a restriction site, which usually consists of 4, 6, or 8 base pairs, with a few exceptions that will be discussed later.

Restriction sites: All of the restriction sites that we will be dealing with in this course are *palindromes* (they have the same double-stranded DNA base sequence in both

directions). As an example, the widely used *Eco* **RI** enzyme recognizes the sequence GAATT (read 5' to 3'). When read 5'- to 3'-, the sequence of the complementary strand is also GAATTC. Bacteria use restriction endonucleases as defense mechanisms, for example, against viral invasion. Each type of bacterium tends to have its own restriction enzymes and specific recognition sites. Foreign DNA is effectively destroyed by being cut at the recognition sites. Bacteria protect their genomes by modifying bases in the restriction sites in their own DNA, usually by methylation. Thus, each strain typically has both a restriction endonuclease and a DNA methylase with the same target specificities, as described in our textbook. The name "restriction" endonuclease was originally given to these highly specific nucleases because they "restrict" invasion by foreign DNAs, such as those of bacterial viruses. Strictly speaking, the palindrome-specific restriction endonucleases that we will be dealing with should be called type II restriction endonucleases, but the other types are not very useful for recombinant DNA technology, and are generally ignored.

Frequency of cutting: Because of their restriction site specificity, the restriction endonucleases cut DNA into fragments whose average length is determined by the number of base pairs in the restriction site (and to a lesser extent by the ratio of bases in the DNA). For DNA that has equal amounts of all four bases, each base has a probabilty of 1/4 at any particular position in the DNA sense strand. For a restriction site of 4 base pairs, the probability of random occurrence of that sequence is $(1/4)(1/4)(1/4)(1/4) = 1/256$. For 6 base pairs, the probability is 1/4,096, and for 8 base pairs it is 1/65,536. Thus, a restriction endonuclease with a 6 base pair restriction site would generate fragments whose average length is 4,096 base pairs. Such fragments are large enough to contain a complete gene (provided that the are no cut sites within the gene for the restriction endonuclease that is used).

Effect of base composition: For DNA whose base composition differs from 50% GT (which is equivalent to equal numbers of all four bases), it is necessary to calculate the probability of a site as the product of the probabilities of each of its components. For example, if a DNA is 66.7% GC (2/3 of its base pairs are GC) and one assumes random orientation of the base pairs, A and T will each have probabilities of 1/6 and G and C will have probabilities of 1/3. Thus the probability of GAATTC would be $(1/3)(1/6)(1/6)(1/6)(1/6)1/3) = 1/11,664$, as opposed to 1/4096 when all four bases are present in equal amounts. Thus, the average fragment length generated by *Eco* **RI** would be longer in a DNA with a higher GC content.

Naming of restriction endonucleases: Restriction endonucleases are named for the species and strain of bacteria they are derived from. The first letter is for the genus, the next two for the species designation, the fourth for the strain, and the Roman numerals that follow designate which enzyme from that strain. In addition, the first three letters, standing for genus and species are normally italicized. Thus, *Eco* **RI** is the first restriction endonuclease derived from *E. coli* strain RY13.

Problems with HTML portrayal of restriction endonuclease names: Because of problems with the on-screen display on some older browsers, I have inserted a space after

the italics so that the italic and non-italic portions of restriction endonuclease names will not overlap. Although this is needed for a clear on-screen presentation on some browsers, it results in the presence of a gap in the printed version of the notes (and on newer browsers) that is not a correct protrayal of the preferred style, and should not be used when not needed for clarity. The printed format used in the textbook is the correct way to write the names of restriction endonucleases. Thus, *Bgl* II looks better on screen, but *Bgl*II is the correct printed form. I also also sometimes use bold face type (***Eco*RI**) because regular italic type (*Eco* RI) tends not to be very clear on web pages, particularly when the display font size is set for 12 point or smaller.

List of selected restriction endonucleases: The review questions and problem sets that accompany this and subsequent lectures will require you to work with a substantial number of different restriction endonucleases. Although our current textbook has a fairly comprehensive list (table 9.1), I have retained the following list from previous notes (and added a few new ones this year) to be certain that you have access to all of the cut sites needed to solve the problems. In this list, the sequence is given for only one strand of the palindrome (the other is its reverse complement and is identical when read 5' to 3'). The cut site is shown by a vertical line (|) placed between the bases that are separated by the cut. Sticky ends (explained below) will result whenever the cut is not at the exact center of the sequence. Pu means any purine (A or G), Py means any pyrimidine (C or T). (A/T) means A or T (an AT base pair in either orientation). Note that *Eco* **RII** is unusual in that its recognition sequence contains an odd number of bases. There also exist numerous non-palindromic sites, but we will not be working with them in this course. For a more extended list of restriction endonucleases sold by one of the commercial suppliers, **click here** (also available from our "other genetics links" page).

- *Aac65* **I   G|GTACC**
- *Alu* **I   AG|CT**
- *Bam* **HI G|GATCC**
- *Bgl* **II   A|GATCT**
- *Cla* **I   AT|CGAT**
- *Eco* **RI   G|AATTC**
- *Eco* **RII |CC(A/T)GG**
- *Hae* **III   GG|CC**
- *Hin* **dII   GTPy|PuAC**
- *Hin* **dIII A|AGCTT**
- *Hpa* **II   C|CGG**
- *Kpn* **I   GGTAC|C**
- *Mbo* **I |GATC**
- *Not* **I   GC|GGCCGC**
- *Nst* **I   ATGCA|T**
- *Pst* **I   CTGCA|G**
- *Pvu* **I   CGAT|CG**
- *Sac* **I   GAGCT|C**
- *Sal* **I   G|TCGAC**
- *Sma* **I   CCC|GGG**

- *Xma* **I**   C|CCGGG

Restriction mapping: Cutting with restriction endonucleases can be used to break large pieces of DNA into smaller fragments. If two restriction endonucleases are used, each fragment produced by the first enzyme is likely to be cut into smaller pieces by the second enzyme. Electrophoresis and comparison of mobilities of the resulting fragments with those of "markers" of known size (see figures 9.24 and 9.25) can be used to determine the relative sizes of the fragments. By reversing the order in which the cuts are made, overlapping fragments can be aligned to generate a complete restriction map of the original DNA. This is illustrated for circular DNA from human mitochondria in the textbook (boxed example 9.1). This process can be continued with additional enzymes, creating a restriction map containing the locations of cut sites for multiple enzymes. Figure 9.2 is an example of this, except that it is limited to cut sites that only occur once in the entire circular genome of the plasmid.

Sticky ends: In many (but not all) cases, the DNA strand is not cut at the center of the restriction site. Thus, for example, *Eco* **RI** cuts its GAATTC recognition site between the G and the first A on each of the DNA strands (G|AATTC). This staggered cutting pattern leaves an overhanging segment of 4 base pairs AATT attached to the new 5' ends created by the cuts on each of the strands of all DNA fragments generated by *Eco* **RI**. These short unpaired segments are capable of forming transient double helical structures that can hold cut ends together long enough for DNA ligase to reseal them. If fragments cut from two different DNA molecules by the same restriction endonuclease are mixed, the ligation process will sometimes join the fragments in new combinations. For example, an isolated gene with sticky ends may be joined to the sticky ends of a circular plasmid that has been opened by a single cut with the same enzyme (figure 9.4). This will form a larger circular plasmid whose DNA sequence now includes the gene. As described in greater detail below, this is the basic process used to clone genes.

Isoschizomers: In certain cases, two or more different enzymes may recognize identical sites. Enzymes from different sources that recognize the same site and cut it either the same way or differently are called isoschizomers. *Sma* **I** and *Xma* **I** in the list above are an example of isoschizomers that cut the same site in different ways. For a massive list of isochizomers, **click here** and then click on "isoschizomer list" (long table, may load slowly).

Matching sticky ends: It is possible to have an overhang at the 5'-end or the 3'-end, or to cut straight across the middle of the recognition site, leaving blunt ends. Thus, for example, DNA cut with *Eco* **RI** will all have an AATT overhang at the 5'-end of each strand, whereas DNA cut with *Kpn* **I**, will have a GTAC overhang at the 3'- end of each strand, and DNA cut with *Sma* **I** will have blunt ends. In order to use sticky ends for joining recombinant DNA molecules, it is necessary to have the same type of overhang on both fragments. When two fragments with compatible sticky ends encounter each other, the single-stranded overhangs will base pair in an antiparallel fashion to hold the fragments together long enough for DNA ligase to form new covalent phosphodiester linkages. Thus, it is easy to join pairs of fragments cut with the same enzyme. However,

it is not possible to join a 5'- overhang and a 3'- overhang, even if the cut site sequences are identical, as in the case of the isoschizomers **Acc65 I** and **Kpn I**, whose cut sites are G|GTACC and GGATC|C, respectively.

Regeneration of cut sites: Ligating together two fragments that have been cut with the same restriction endonuclease generates a new cut site for that enzyme. This makes it easy to recover a cloned sequence by cutting it out of the vector with the same restriction endonuclease that was originally used to prepare it for cloning. In certain cases, two different restriction endonucleases may generate identical sticky ends even though their cut sites are not identical. A good example of this is **Bam HI**, whose cut site is G|GATCC, and **Bgl I**, whose cut site is A|GATCT. Both generate 5'- overhangs of GATC, which allows DNA fragments generated by these two enzymes to be ligated readily. However, the resulting double stranded link

GGATCT
CCTAGA

is no longer a palindrome and cannot be cut with either of the enzymes.

Vectors: In order to achieve replication of cloned genes, it is necessary to insert them into self-replicating genomes that are referred to as vectors. Although bacterial plasmids are widely used as vectors, as described below, there are also many other types of vectors. To be optimally useful, a vector must possess the following properties.

1. The vector must contain an origin of replication (ori^+), which allows the DNA to replicate itself and the DNA it carries independently of the host DNA.
2. To be useful for cloning, circular vector genomes must be capable of being opened by restriction endonucleases without any loss of vector DNA. This requires the presence of "unique" restriction sites that are present only once in the entire circular vector DNA. Cutting with a restriction endonuclease that is specific for such a site opens the circle and "linearizes" the vector without any loss of vector DNA (figure 9.3). This allows the complete vector to be reconstituted simply by rejoining the cut ends. It also allows a foreign DNA sequence to be inserted between the cut ends to create a new larger circular vector that contains an inserted sequence, such as a cloned coding sequence (figure 9.4). It should be noted, however, that there are some types of vectors that do not have circular genomes.
3. Most vectors code for some kind of selectable marker (such as antibiotic resistance) so that the presence of the vector is required for the host organism to survive under selective conditions.
4. Ideally, there should also be a second marker that makes it possible to distinguish between vectors that contain cloned DNA and those that do not.

Plasmids: Bacterial plasmids are small circular DNAs that have their own origins of replication and are capable of autonomous replication within bacterial cells. Plasmids that carry appropriate genes are capable of making bacteria resistant to antibiotics, which

makes it possible to select for bacteria that have taken up the plasmids. Because of their small size, plasmids often have only a single cut site for a particular restriction endonuclease, which allows the circles to be opened for adding a foreign DNA without risk of losing parts of the plasmid. They can also be modified in a variety of ways, including the addition of multiple cloning sites (described below). Another trick that is sometimes used is to eliminate naturally-occurring cut sites (or generate new cut sites) by changing the third bases of codons in ways that eliminate (or generate) restriction endonuclease cut sites without altering the amino acid coding of genes carried in the vector. Such "mutations" are silent with regard to amino acid coding, but not with regard to cleavage of the DNA by restriction endonucleases.

pBR322 plasmids: An artificially engineered plasmid, designated pBR322 (textbook figure 9.2), was introduced in 1977 and widely used for many years. pBR322 carries genes conferring resistance to ampicillin and tetracycline and has a number of unique restriction endonuclease cut sites. If a cut is made with *Pst* **I** in the middle of the ampicillin resistance gene, a foreign gene that has been cut out of its genome with *Pst* **I** can be inserted into the plasmid because it has the same sticky ends as the opened plasmid. This makes a larger circle, but if the insert is not too large, the plasmid is still capable of replication, thus providing a means for replication of the cloned gene.

Screening: To screen for recombinant PBR322 plasmids, rejoined plasmids with a cloned DNA in the middle of the ampicillin resistance gene are initially infected into bacteria and plated on nutrient agar containing tetracycline. Only those bacteria that have taken up plasmids with intact tetracycline resistance genes can form colonies. Replica colonies (figure 7.4) are transferred to ampicillin plates. Bacteria containing religated plasmids with no cloned DNA inserts will multiply, but those with plasmids that have inserts disrupting the ampicilin resistance gene will not. Comparing the growth patterns (see figure 9.6) allows one to pick colonies from the tetracycline plates that lack ampicillin resistance and thus contain plasmids carrying cloned foreign DNA (figure 9.5). Alternatively, this scheme can be reversed by cloning into the middle of the tetracycline resistance gene with *Bam* **HI** and then selecting for bacteria that are resistant to ampicillin and sensitive to tetracycline.

Blunt-end ligation: As an alternative to restriction endonuclease cloning, which may cut within sequences that one wishes to preserve, it is possible to do blunt-ended cloning of randomly sheared DNA (whose random breaks should sometimes be located outside of the desired sequences). Deoxynucleotide terminal transferase can be used to add short poly-T sequences at the 3'-ends of an opened vector and to add short poly-A sequences to the 3'-ends of the DNA to be cloned. Repair enzymes and ligase can be used to seal the ends. Alternatively, random blunt end ligation can be done with the phage enzyme T4 DNA ligase. Blunt ends can be generated by shearing the DNA, or by cutting with blunt end restriction endonucleases, such as *Hin* **dII** or *Sma* **I**.

Linker DNA: Short segments of synthetic DNA containing cut sites for any given enzyme can be ligated into an existing cut site or blunt-end ligated onto the ends of linear

DNA fragments. Such linkers allow the use of restriction sites that do not occur naturally in vectors or in sequences to be cloned.

Multiple cloning sites: Most sophisticated modern vectors contain multiple cloning sites, which consist of a short stretches of artificially synthesized DNA containing cut sites for a number of different restriction endonucleases located side by side (figure 9.7). This allows selection of a restriction endonuclease for cloning that does not cause internal cuts in the gene being cloned. It also allows different cuts to be made at the two ends of the gene in order to force it into the vector in a particular orientation. In many cases, expression of the gene is driven by a promoter in the vector, and it is necessary to insert the gene so that it is read in a forward direction. One additional advantage of using different cuts at the two ends is that the vector cannot religate without an insert because its two ends no longer have matching sticky ends.

Blue-white screening: A more sophisticated screening technique that does not require replica plating is now widely used. This system takes advantage of the fact that a functional beta-galactosidase enzyme can be generated from separate N-terminal and C-terminal fragments of the enzyme protein. A bacterial strain that synthesizes only the C-terminal part of the enzyme is used. The pUC19 plasmid contains an ampicillin-resistance gene plus an engineered gene that has an AUG start codon followed by a complex multiple cloning site, followed in frame by the N-terminal part of beta-galactosidase (figure 9.7). When the bacterial lac operon is stimulated with IPTG, and the intact vector is present, the two separately coded halves of the beta-galactosidase enzyme combine to form a functional enzyme, which is capable of hydrolyzing a complex synthetic galactoside called X-gal to generate a blue color. The presence of the multicloning site adds a few amino acids to the N-terminal fragment, but does not disrupt its ability to combine with the C-terminal part to form a functional enzyme. However, a much larger cloned insert disrupts production of the N-terminal part of the enzyme. This causes IPTG-treated colonies to stay white in the presence of X-gal. Bacteria that do not contain vectors are unable to form colonies in the ampicillin-containing selective medium. Colonies that are resistant to ampicillin and remain white with X-gal are thus likely to contain cloned gene