Lecture 5

Restriction Endonucleases

Today we're going to begin a discussion of a very useful class of enzymes.

Restriction enzymes

Restriction enzymes are classified as endonucleases. Their biochemical activity is the hydrolysis ("digestion") of the phosphodiester backbone at specific sites in a DNA sequence. By "specific" we mean that an enzyme will only digest a DNA molecule after locating a particular sequence.

Here's an example: There is an enzyme called *BamHI* that searches for the sequence GGATCC in double-stranded DNA (by which I mean that the bottom strand is also present): When the sequence is located, the enzyme BamHI digests the phosphodiester backbone in two specific places - between the pair of G nucleotides on each strand.

That leaves us with a four nucleotide single stranded 5' end on each side after separation.

```
5'-GGATCC-3'   Bam HI     5'-G     -3'  +   5'- GATCC-3'
3'-CCTAGG-5'   ---->      3'-CCTAG -5'      3'-     G-5'
```

In reality there are more than six nucleotides on each strand, of course. The Bam HI just looks for the specific sequence it's interested in, and will accept no substitutes:

```
GAGGATACCACCAGGGTTACAGGATAGGAGTCAGGATCCAGAGGACCTAGGATACCTC
CTCCTATGGTGGTCCCAATGTCCTATCCTCAGTCCTAGGTCTCCTGGATCCTATGGAG
```

is digested by Bam HI (at the site shown in red) to give two fragments of DNA...

```
GAGGATACCACCAGGGTTACAGGATAGGAGTCAG     GATCCAGAGGACCTAGGATACCTC
CTCCTATGGTGGTCCCAATGTCCTATCCTCAGTCCTAG     GTCTCCTGGATCCTATGGAG
```

Restriction enzymes hydrolyze the phosphodiester backbone once on each strand (we say the strand is "nicked," perhaps to indicate that the cut isn't very deep). The bonds being broken by the enzyme are covalent. The hydrogen bonds responsible for base pairing are not broken by the restriction enzyme (however thermal energy is high enough at room temperature to separate BamHI fragments, for example).

	Which of these sequences have a BamHI site?
☐	GATTACCTAGGACTACAGGTACT CTAATGGATCCTGATGTCCATGA

☐	AACTATACCAGGATCCATCACCT TTGATATGGTCCTAGGTAGTGGA
☐	ATACTTACGAGGATCTCAGATCC TATGAATGCTCCTAGAGTCTAGG
☐	TTACATGCCATTCACCGGATCCT AATGTACGGTAAGTGGCCTAGGA

Requirements What does a restriction enzyme need in order to do its duty?

1. A double-stranded DNA sequence containing the recognition sequence.
2. Suitable conditions for digestion.

For example, BamHI has the recognition sequence: GGATCC and requires conditions similar to this:

10 mM Tris-Cl (pH 8.0)
5 mM Magnesium chloride
100 mM NaCl
1 mM 2-mercaptoethanol
Reaction conditions: 37 C

On the other hand, the enzyme Sma I has the recognition sequence: CCCGGG and requires conditions such as:

33 mM Tris-acetate (pH 7.9)
10 mM Magnesium acetate
66 mM Potassium acetate
0.5 mM Dithiothreitol
Reaction conditions: 25 C

Most restriction enzymes are used at 37 C, however Sma I is an exception. Other examples of temperature exceptions are Apa I (30 C), Bcl I (50 C), BstEII (60 C), and Taq I (65 C). Taq I, by the way, is a restriction enzyme from the same type of organism that produces Taq polymerase (*Thermophilus aquaticus*, or *Thermus aquaticus*). Restriction enzyme names are based on a species-of-origin.

For example:

BamHI (from Bacillus amyloliquifaciens (H))
Sma I (from Serratia marcescens S)
Mlu I (from Micrococcus luteus)
Hpa I (from Haemophilus parainfluenzae)

From what species do you think EcoRI is derived?

Let's suppose that we use the enzyme BamHI to digest DNA, as in the previous example. The enzyme finds the sequence GGATCC on each strand (note that it reads the same on the complementary strand and so we say the sequence has a two-fold axis of symmetry, or is "palindromic") and nicks the phosphodiester backbone between the G nucleotides. The hydrogen bonds break naturally, from the energy of thermal motion in the solution (the word we use to describe this loss of base pairing is to say the strands "melt"), and the two fragments move away from each other.

```
5'-GAGGATACCACCAGGGTTACAGGATAGGAGTCAG-3'
3'-CTCCTATGGTGGTCCCAATGTCCTATCCTCAGTCCTAG-5'
```

and

```
5'-GATCCAGAGGACCTAGGATACCTC-3'
    3'-GTCTCCTGGATCCTATGGAG-5'
```

The two fragments have newly-exposed 5' and 3' ends. (And nearly all restriction enzymes leave a phosphate on the exposed 5' end. The enzyme Nci I, an exception to the rule, leaves a 3' phosphate).

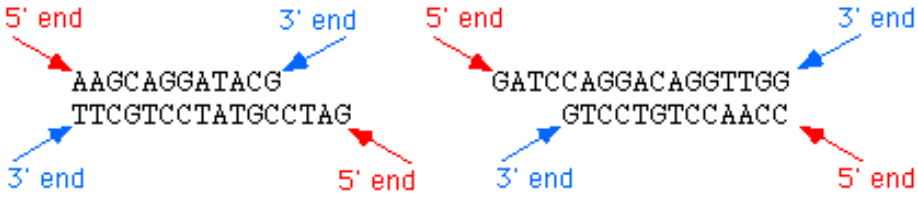
More on recognition sequences

Consider three enzymes with recognition sequences as indicated (a caret symbol (^) or asterisk (*) is often inserted to mark the place where the enzyme breaks the phosphodiester backbone)

Enzyme	Recognition sequence	Type of ends in product
BamHI	G^GATCC	5' overhang
SacI	GAGCT^C	3' overhang
SmaI	CCC^GGG	blunt
Take a look at the rebase database from N.E.B.		

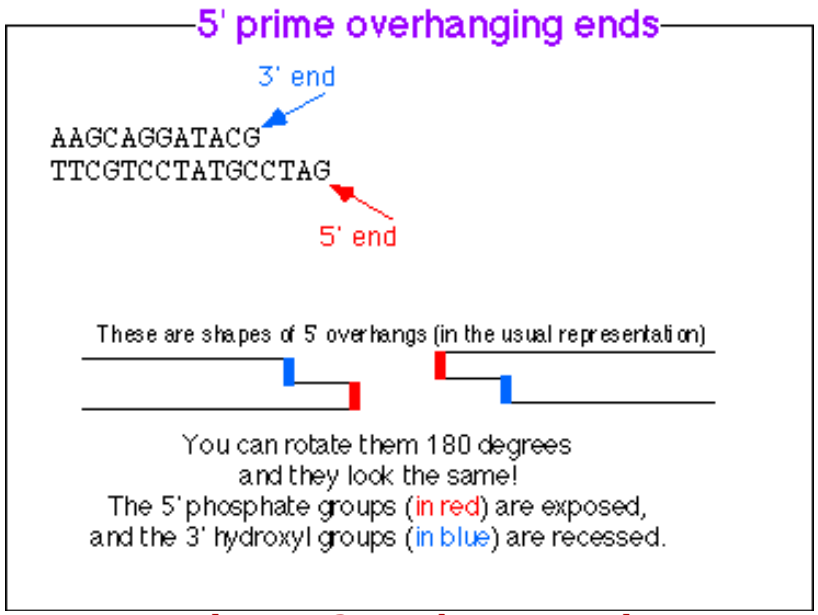
It is important that you recognize the differences between the three types of ends generated by restriction enzymes, and the three examples above are illustrative.

In the example of digestion with the enzyme BamHI, it's obvious that the newly created ends of the DNA do not line up evenly with each other. On each fragment, there is a four-nucleotide sequence 5'-GATC that hangs off the end and doesn't base-pair (because the other fragment has broken away and moved off).



Since the one end that hangs over past the other has a free 5' end, we say that BamHI digestion creates a "5' overhanging end" which we sometimes call a "5' overhang."

Another term that means the same thing is to say that overhanging ends are "cohesive ends" or "sticky ends" meaning that they could hydrogen bond to other compatible complementary strands (compatible in the sense of Watson-Crick base pairing). By our usual convention of writing DNA, a 5' overhanging end has a characteristic shape.

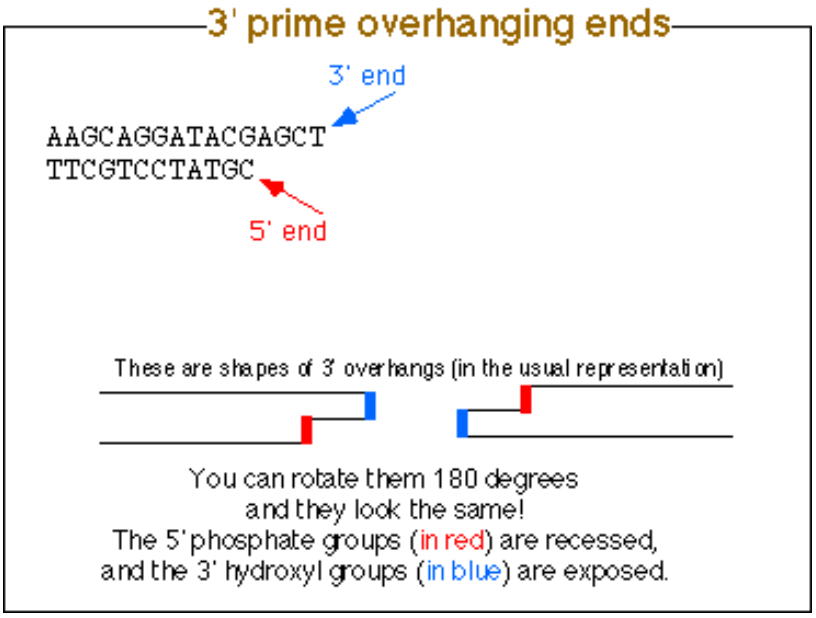


Some restriction enzymes leave a 3' overhanging end.

An example would be the enzyme Sac I:

Sac I searches for the sequence GAGCTC on each strand (once again, GAGCTC reads the same off of both strands because the sequence is palindromic). The enzyme breaks the phosphodiester bonds between the fifth and sixth nucleotides in the recognition sequence.

```
5'-GAGCTC-3'   Sac I    5'-GAGCT -3' +  5'-     C-3'
3'-CTCGAG-5'   ---->    3'-C     -5'    3'- TCGAG-5'
```

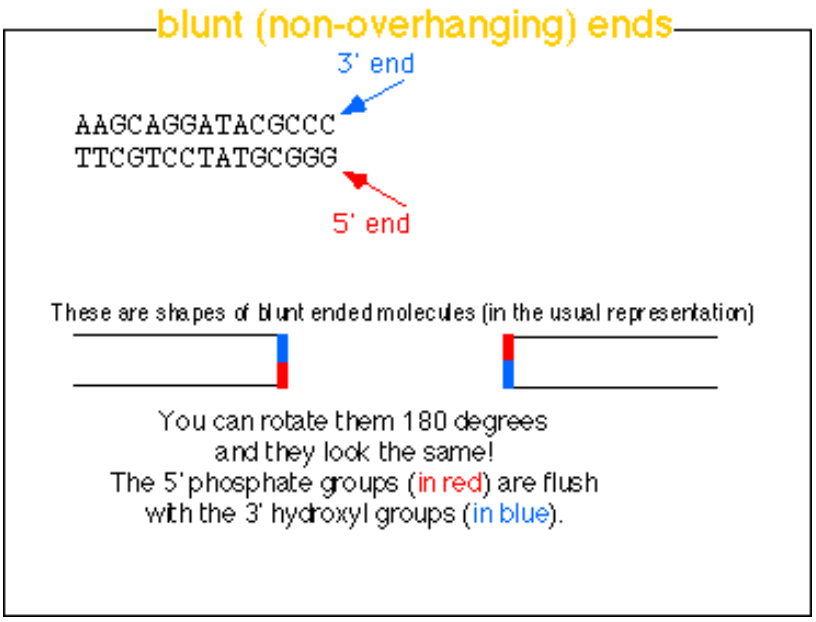


Some restriction enzymes leave a blunt end.

What do we call a DNA molecule that has ends that line up evenly with each other (i.e. neither end is overhanging)? We say the ends are "blunt" (meaning "not sharp") or "flush" (meaning "level or even").

For example, the enzyme Sma 1 cuts in the middle of the six nucleotide recognition sequence:

```
5'-CCCGGG-3'   Sma I      5'-CCC -3'  +  5'- GGG-3'
3'-GGGCCC-5'   ---->      3'-GGG -5'     3'- CCC-5'
```



Not all restriction enzymes recognize sequences that are palindromic.

For example, the enzyme Bsr I cuts as follows (where "N" can represent any nucleotide):

```
5'-ACTGGNN-3'  Bsr I     5'-ACTGGN      N-3'
3'-TGACCNN-5'  ---->     3'-TGAC      CNN-5'
```

The reason this is said to **not** be a palindromic sequence is that the two strands read **differently** in their antiparallel directions. The top strand is 5'-ACTGGNN while the bottom strand is 5'-NNCCAGTT. Compare that with the recognition sequence for an enzyme like Sma I (which is palindromic) and reads 5'-CCCGGG on the top and bottom strand, and you can see there is a difference. Restriction enzymes that do not recognize palindromic sequences might therefore be described in some references using multiple sequences, since the top and bottom strand read differently. That is useful if you are only scanning one strand of DNA for sites. For example the enzyme Bsr I, which was just described, is described by the sequences ACTGGN^ and C^CAGT

Some restriction enzyme sequences cut outside of their recognition sequence

An example of this is the enzyme Bsr I, just described.

```
5'-A C T G G N^N-3'
3'-T G A C^C N N-5'
```

On one strand, the enzyme breaks the phosphodiester backbone between the two unspecified "N" bases must 3' to the ACTGG. Here are some other examples of enzymes that have an extended "reach":

Mnl I

```
CCTCNNNNNNN^
GGAGNNNNNNN^
```

What type of DNA end does Mnl I leave?

Eco57I

CTGAAGNNNNNNNNNNNNNN^
GACTTCNNNNNNNNNNNNNN^

...sometimes written CTGAAG (16/14)

Ksp632I

CTCTTCN^
GAGAAGNNNN^

Here's an interesting trick. Ksp632I can be used, in combination with a single-stranded DNA nuclease such as mung bean nuclease or S1 nuclease to generate a 3 nucleotide deletion. Thus, you could use this trick to repetitively remove three nucleotides in a protein coding sequence (not changing the reading frame, but possibly introducing a mutation as well as a deletion). Don't you agree that site-directed mutagenesis is easier with PCR?

Some enzymes have split recognition sequences

Consider the enzyme Asp 700, with the restriction enzyme recognition sequence:

GAANN^NNTTC

The 6 nucleotides of recognition sequence is split - palindromic, with the four internal nucleotides not specified. What type of end does Asp 700 leave?

Can you figure out how the enzyme Dra III cuts, from the recognition sequence:

CAC(N3)^GTG ?

Would the overhanging end left by Dra III be the same at every site?

Some enzymes accept degenerate sequences

We've been using "N" nucleotides in our recognition sequences, but the N is obviously non-specific. There are enzymes that have partial degeneracies in their recognition sequence.

An example of this is Aha II, recognizing the sequence GR^CGYC, where R = G or A, and Y = T or C. For Aha II then, the following are all acceptable recognition sequences:

GG^CGCC

GA^CGCC

GG^CGTC

GA^CGTC

Additional examples follow:

Hind II
GTY^RAC

Hae II
RGCGC^Y

DraII
RG^GNCCY

Not all restriction enzymes recognize six-nucleotide pair sequences.

You may have already noticed this is true, but here are two examples:

```
5'-TTAATTAA-3'   Pac I    5'-TTAAT     TAA-3'
3'-AATTAATT-5'   ---->    3'-AAT     TAATT-5'

5'-NAATTN-3'   Tsp509 I     5'-N       AATTN-3'
3'-NTTAAN-5'   ---->         3'-NTTAA       N-5'
```

Note that there are 8 nucleotides that specify the location of a PacI restriction site, and 4 that specify the location of a Tsp509I site.

I can hear you thinking:

"Wait a minute! The Tsp509I has 6 nucleotides in its sequence, not 4!"

That's true, but since the two "N" nucleotides could be anything, they don't really add to the specificity of the recognition sequence. They are included only as "placeholders" to help indicate where the phosphodiester bonds are broken. The following are all valid Tsp509I sites (I've highlighted the central four nucleotides in red):

```
AAATTG    TAATTA    CAATTC    GAATTA    AAATTC    etc.
TTTAAC    ATTAAT    GTTAAG    CTTAAT    TTTAAG
```

You may be amazed to notice that the recognition sequence for PacI (TTAATTAA) also has a Tsp509I site in it!

A calculation to ponder:

The AATT sequence is something that occurs by chance pretty frequently. If a DNA sequence is evenly made up of G, A, T, and C nucleotides (i.e. 25% of each), we would expect to find the sequence "AATT" by chance about every 256 nucleotides on the average. Why is that? Because if we point to a nucleotide in a sequence at random, the chances would be one in four that it would be "A" (the first nucleotide in the recognition sequence). The chance that the next nucleotide is also "A" is also 1 in 4; the chance that the nucleotide after that is "T" is 1 in 4; and the chance that the next one is also "T" is also 1 in 4. Therefore, the chance that we have randomly pointed to a sequence that reads "AATT' is:

$$(1/4) \times (1/4) \times (1/4) \times (1/4) = 1/256$$

Any recognition sequence that was four nucleotides in length could be found every 256 nucleotides (on the average) in this simple scenario. In actuality, sequences are usually not evenly made up of G, A, T, and C nucleotides, which skews the statistics a bit. In addition, certain short sequences may be more or less common in the DNA, which will also affect the frequency with which a recognition sequence is found. The dinucleotide CG is very uncommon in mammalian DNA, which makes it less likely that you will find a recognition sequence for the enzyme Hpa II (C^CGG).

Longer recognition sequences lead to lower probability of having a site at any point in a DNA strand. In our simplistic scenario where every nucleotide is evenly distributed in DNA, you would expect to find a PacI site every 65,000 nucleotides (on the average). That's because there's a one in four chance that each of the eight nucleotides (taken individually) in a random sequence is just right for PacI. The chances of being lucky eight times in a row is one over four to the eighth power, or about one over sixty five thousand.

Enzymes with recognition sequences from 4 to 8 nucleotides in length each have uses in genetic engineering.

6-cutters (i.e. enzymes that have recognition sequences specified by six nucleotides) are good for day-to-day cloning work: You are using 6-cutters in the experiments you are performing in the lab, because they cut frequently enough that there are one or two sites in the plasmid, but infrequently enough

that they do not cut into the essential elements such as the origin of replication or ampicillin resistance gene. An example of a 6-cutter is HindIII (A^AGCTT) which cuts the genome of bacteriophage lambda (48 kbp) at 7 sites.

8-cutters are good for carving up chromosomes into specific pieces that are still quite large. PacI might cut the *E. coli* chromosome into only about 20 pieces, for example, whereas BamHI might cut it into about 300 pieces. Example of where an 8 cutter might be of use: Suppose you were trying to obtain a specific fragment of a yeast chromosome, for example. Then, it would be impractical to use a 6-cutter enzyme because you would generate too many small fragments during your digestion. An 8-cutter might cut infrequently, generating larger and more useful products. An example of an 8-cutter is NotI (GC^GGCCGC) - the NotI recognition sequence is not present in the genome of bacteriophage lambda.

4-cutters are good for experiments where you want the possibility of cleavage at many potential sites. For example, if you want to gather a collection of random DNA fragments, some of which may contain a gene you want, you can perform a partial digestion (not all sites are cleaved due to limitation in enzyme activity) using a 4-cutter. One that is commonly used for this purpose is Sau3AI, which cleaves:

```
5'-NGATCN-3'    Sau3AI      5'-N      GATCN-3'
3'-NCTAGN-5'    ---->       3'-NCTAG      N-5'
```

There are 116 Sau3AI sites in the genome of bacteriophage lambda.

Star activity A cautionary note about restriction enzyme buffer conditions is appropriate at this point.

Some restriction enzymes exhibit what are called "relaxed" specificity or "star" activity when they are used under the wrong buffer conditions.

For example, EcoRI, which normally only recognizes the sequence G^AATTC, will exhibit EcoRI-star activity if the ionic conditions are too low (e.g. below 50 mM), if the concentration of glycerol is too high, or if too many units of enzyme are added. These "star" activities of EcoRI are N^AATTN and R^RAYY.

BamHI (G^GATCC), which we have been discussing, has star activity under similar non-optimal conditions, and will cause recognition at G^GATCN or G^RATCC.

Isoschizomers Restriction enzymes that recognize the same sequence and are derived from different organisms are called isoschizomers. They may have different sites of specific cleavage and still be isoschizomers. For example, Sma I and Xma I are isoschizomers, even though they leave different ends. PspAI is also an isoschizomer, and it yields exactly the same result as Xma I.

Sma I (*Serratia marcescens* SB)

CCC^GGG

Xma I (*Xanthomonas malvacaerum*)

C^CCGGG

PspAI (*Pseudomonas* species)

C^CCGGG

You might want to have both Sma I and Xma I in your freezer, because they leave different DNA ends, but there would be no point in having both Xma I and PspAI - they are identical in function. They are not identical in price however (Xma I being more expensive than either Sma I or PspAI).

Methylation issues We have already discussed in a previous class the matter of Dpn I enzyme, which digests the sequence GATC only if the A is methylated. There are several isoschizomers of Dpn I (we ignore the methylation issue when deciding if two enzymes are isoschizomers - we only consider the primary DNA sequence).

Dpn I

```
Methyl
     |
    GA^TC
```

or both A and C methylated:

```
Methyl  Methyl
     |  |
    GA^TC
```

but not with A unmethylated

Mbo I

```
^GATC
```

or

```
      Methyl
       |
  ^GATC
```

but not if the A is methylated, or the C is hydroxymethylated

Sau3AI

```
^GATC
```

or:

```
  Methyl
   |
^GATC
```

will both digest, but not if the C is methylated:

```
    Methyl
     |
^GATC
```

Remember that there is a "dam" methylase in some strains of E. coli that methylates A residues in GATC. Could DNA from that strain be digested with Mbo I? Could it be digested with Sau3AI?

Suppose you have DNA from a eukaryotic species that performs some methylation of C bases in GATC. Could you digest DNA from that organism with Dpn I? With Mbo I? With Sau 3AI?

What would happen if you prepared a plasmid that was grown in a dam+ strain of E. coli, then moved (transfected) the DNA into human cells (which do not have the capability of methylating A bases). If the plasmid could replicate in the human cell, would you expect that plasmid extracted from the cell after a few weeks would be resistant or sensitive to Dpn I? Mbo I? Sau3AI?

Stan Metzenberg
Department of Biology
California State University Northridge
Northridge CA 91330-8303
stan.metzenberg@csun.edu