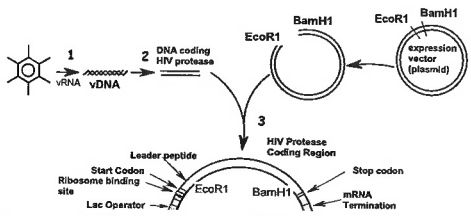
Lecture 38: Cloning, DNA Sequencing

III. Insertion of HIV protease coding region (codons) into the Expression Vector



A detailed view of the expression vector is shown below:

```
-Promoter for---Regulatory element---Ribosome--Start--Leader--GAATTC-------GGATCC--Stop---mRNA--------
-mRNA Syn.------for mRNA Syn---------binding---Codon--Peptide-CTTAAG-------CCTAGG--Codon--Termination-
                                     site                    EcoR1          BamH1
```

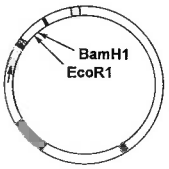
After insertion of the gene that encodes the HIV protease, the DNA will look like the following:

```
-Promoter for---Regulatory element---Ribosome--Start--Leader--GAATTC--HIV------GGATCC--Stop---mRNA---
-mRNA Syn.------for mRNA Syn---------binding---Codon--Peptide-CTTAAG--Protease-CCTAGG--Codon--Term.--
                                     site                    EcoR1            BamH1
```

Step A. Cut HIV protease DNA (PCR product) with EcoR1 and BamHI

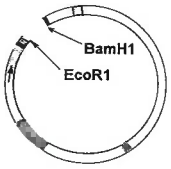
```
C-G-G-A-A-T-T-C-HIV--G-G-A-T-C-C-G     →     C-G-G     A-A-T-T-C-HIV--G     G-A-T-C-C-G
G-C-C-T-T-A-A-G-Prot-C-C-T-A-G-G-C           G-C-C-T-T-A-A     G-Prot-C-C-T-A-G     G-G-C
    EcoR1              BamH1
```

Step B. Cut Expression vector with EcoR1 and BamH1, discarding small segment of DNA between the EcoR1 and BamH1 sites:




```
-Start--Leader--G-A-A-T-T-C-------G-G-A-T-C-C--Stop---mRNA--
-Codon--Peptide-C-T-T-A-A-G-------C-C-T-A-G-G--Codon--Term.-
                      EcoR1             BamH1

-Start--Leader--G                G-A-T-C-C--Stop---mRNA--
-Codon--Peptide-C-T-T-A-A              G--Codon--Term.-
                     EcoR1             BamH1
```

Step C. Mix HIV protease fragment and expression vector:



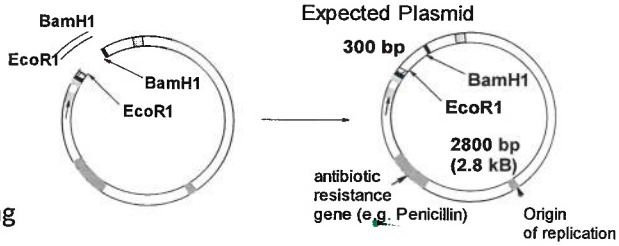

```
A-A-T-T-C--HIV------G
        G--Protease-C-C-T-A-G

-Start--Leader--G                              G-A-T-C-C--Stop---mRNA--
-Codon--Peptide-C-T-T-A-A                            G--Codon--Term.-
```

Cool to allow H-bonds to form due to cohesive (sticky) ends.

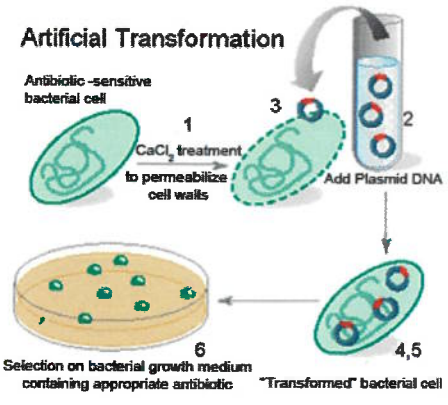
```
-Start--Leader--G A-A-T-T-C--HIV------G G-A-T-C-C--Stop--mRNA--
-Codon--Peptide-C-T-T-A-A G--Protease-C-C-T-A-G G--Codon—Term.-
```

Step D. Use **DNA ligase**, to rejoin DNA fragments.

```
--G-A-A-T-T-C--HIV------G-G-A-T-C-C--Stop
--C-T-T-A-A-G--Protease-C-C-T-A-G-G--Codon
      EcoR1                    BamH1
```



Step E - Transformation. After the plasmid is ligated, it would be transformed into bacteria by breaching the bacterial membrane using divalent ions. **Typically, one bacterial cell acquires one plasmid.** Once the plasmid is inside the cell, the following two DNA sequence elements are important for maintaining the plasmid in the cell:

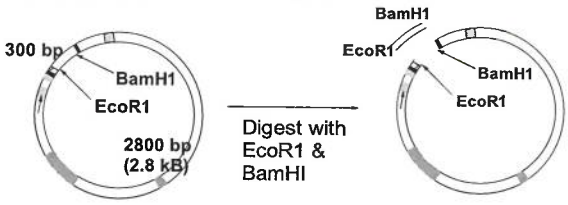


- **Antibiotic resistance gene**: This is a DNA sequence (gene) that will produce a protein that confers resistance to an antibiotic (e.g. penicillin). By growing the bacteria in the presence of the antibiotic only those bacteria that contain the plasmid grow – this is referred to as **selection**. Common antibiotics are:
  - Ampicillin
  - Kanamycin
  - Tetracycline
  - Chloramphenicol
- **Origin of DNA replication**: This is required so that the plasmid will be copied along with the chromosome of the bacteria when the bacterial cell divides. Often there are many copies of the plasmid in the cell, with all copies being identical.

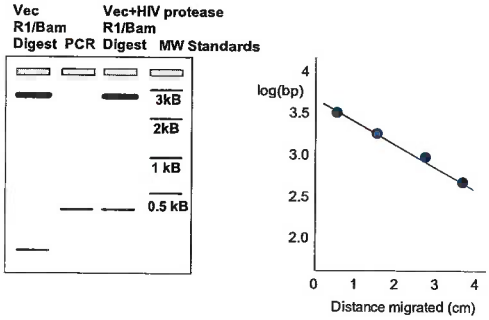
Step F - Verification of the DNA construct.

Since there are many ways the DNA fragments can ligate together it is important to isolate the plasmid DNA from a colony and verify that it is correct.

This is typically done by measuring fragment sizes after digestion with restriction enzymes using gel electrophoresis using gels made from agarose and then DNA sequencing to verify that the construct is correct.



In DNA gel electrophoresis, the dependence on migration distance is the same as with SDS-PAGE, a plot of log(MW) versus distance in linear



Lane 1: original vector cut with R1 and BamH1.
Lane 2: PCR product (300 basepairs)
Lane 3: Correct plasmid cut with R1 and BamH1.
Lane 4: Molecular weight standards

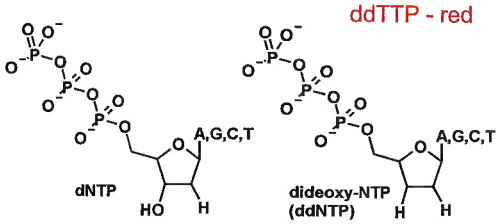
Sanger DNA Sequencing: The start of the sequence is defined by the location of primer annealing. You would use a DNA primer that anneals to the plasmid upstream (to the 5' side) of the DNA that we wish to sequence, the sequence begins with the base that is added to the 3' end of the primer.

- The order that a DNA polymerase adds bases is determined by:
  - random termination of chains with a known base using dideoxynucleotides, therefore the last base incorporated at the 3' end of the fragment is known.
  - the position of that base is measured by separating the synthesized molecules by size.

Priming: We choose a primer that anneals on the plasmid adjacent to the place where the DNA was inserted, therefore we can sequence the DNA without knowing anything about the sequence of the insert. In the example shown on the right the primer anneals upstream from the EcoR1 site that was used to insert the HIV protease gene (the sequence of the plasmid is known)

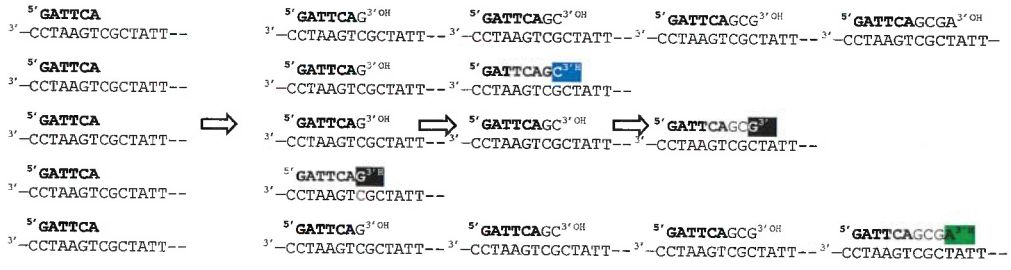
ddATP - green
ddCTP - blue
ddGTP - black
ddTTP - red

Which base was added: The DNA fragments that are generated will end with a known, **colored**, base. This is accomplished by including a small amount (usually ~1%) of a dideoxy nucleoside triphosphate in the reaction with normal dNTPs. Each type of dideoxy (A,G,C,T) has a different color that comes from special fluorescent properties of the dideoxybase, allowing identification of the base.



What is the consequence of missing a 3'-OH on the dideoxynucleotide?

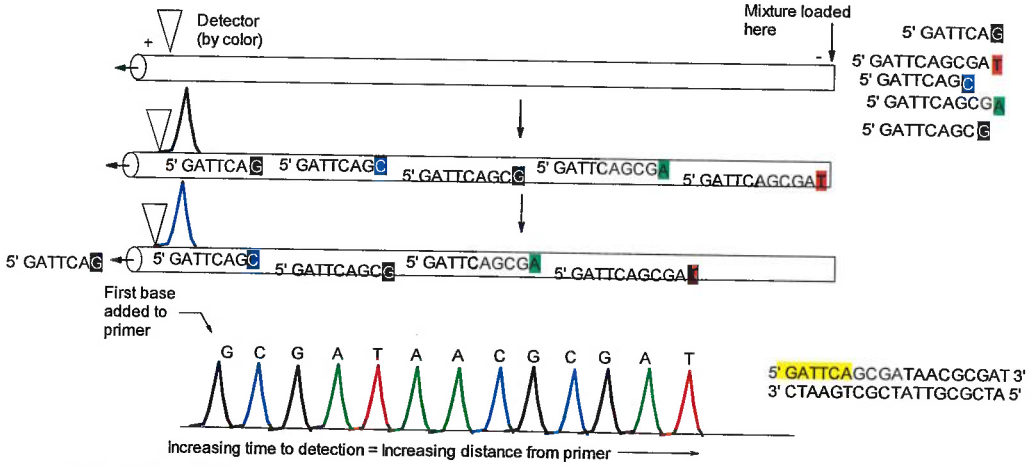
Example: Consider elongation of a collection of five primer-templates, assuming a ratio of dNTP to ddNTP of 4:1 – the chance of termination by a ddNTP is ~20% - one in four additions of a base will terminate (The usual ratio is 1:100, i.e. most additions will not terminate, such that it is possible to generate ***all*** possible fragments for ~1000 bases). For the following five primer/templates, one never terminates, one after the 4th base, one after the 3rd, one after the 2nd and one after the first.



After completion of synthesis the following *fluorescent* fragments would be generated.

5′ **GATTCAG**3′ 5′ **GATTCAGC**3′ 5′ **GATTCAGCG**3′ 5′ **GATTCAGCGA**3′ 5′ **GATTCAGCGAT**3′
5′ **GATTCAGCGATA**3′ **(Primer is bold**, non-bold represent dNTPs added before termination)

Measuring the location of the added dideoxyNTP: The position of each colored base, relative to the primer, is measured by separation of the DNA molecules by size using a technique called capillary DNA electrophoresis. The DNA fragments are loaded on the right, and migrate to the left. At the left end a laser excites the fluorescence and a detector read the emitted light and determines the base based on the wavelength of the emitted light. The smaller molecules pass the detector first.



Output (Time point the molecule passes detector = same order of base addition to the primer by polymerase.)

DNA sequence to Amino Acid Sequence:

A codon is a series of three nucleotide bases that encode a single amino acid.

1. Three DNA bases specify a single amino acid. These are called a 'codon'. For example, the following codon is translated as: TGG =
2. The first codon in all genes that encode proteins is ATG (AUG in the RNA), coding for the amino acid methionine. *HIV protease does not start with a Met because it is produced from a longer peptide by proteolysis.*
3. Special codons (termination codons) indicate the end of the protein. These are UAA, UAG, UGA. *The HIV protease sequence lacks a stop codon because its carboxy terminus is produced by proteolysis.*
4. The "reading frame" must be defined during the translation of the mRNA to protein. The reading frame is the base that is taken to be the first base of the codon. The rest of the codons are obtained by taking 3 bases at a time. Without knowledge of the reading frame the above sequence could be punctuated in any one of the following three ways, giving three completely different sequences. There is only one correct reading frame. The reading frame from a DNA sequencing experiment can only be unambiguously established by comparing the protein sequence predicted from the DNA to the protein sequence determined by chemical methods.

Codon Table:

5' Base	Middle Base				3'
	T	C	A	G	
T	Phe	Ser	Tyr	Cys	T
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	**Term**	**Term**	A
	Leu	Ser	**Term**	Trp	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	T
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	T
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Frame 1		*Frame 2*		*Frame 3*
CCT CAG ATC	***or***	C CTC AGA TC	***or***	CC TCA GAT C
Pro Gln Ile		Leu Arg Ser		Ser Asp-

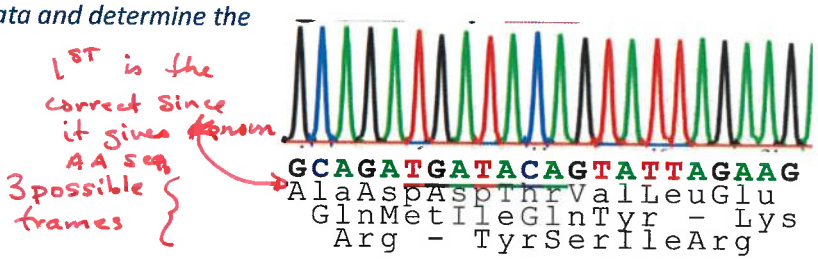
Reading and Interpreting DNA Sequencing:

```
Region of HIV DNA Coding for HIV protease.
   5'-ggagccgatagacaaggaactgtatcctttaacttcCCTCAGATCACTCTTTGGCAA 57
                                                ProGlnIleThrLeuTrpGln 7
 58CGACCCCTCGTCACAATAAAGATAGGGGGGCAACTAAAGGAAGCTCTATTAGATACAGGA 117
  8ArgProLeuValThrIleLysIleGlyGlyGlnLeuLysGluAlaLeuLeuAspThrGly 27
118GCAGATGATACAGTATTAGAAGAAATGAGTTTGCCAGGAAGATGGAAACCAAAAATGATA 177
 28AlaAspAspThrValLeuGluGluMetSerLeuProGlyArgTrpLysProLysMetIle 47
178GGGGGAATTGGAGGTTTTATCAAAGTAAGACAGTATGATCAGATACTCATAGAAATCTGT 237
 48GlyGlyIleGlyGlyPheIleLysValArgGlnTyrAspGlnIleLeuIleGluIleCys 67
238GGACATAAAGCTATAGGTACAGTATTAGTAGGACCTACACCTGTCAACATAATTGGAAGA 297
 68GlyHisLysAlaIleGlyThrValLeuValGlyProThrProValAsnIleIleGlyArg 87
298AATCTGTTGACTCAGATTGGTTGCACTTTAAATTTTcccattagccctattgagact 354 -3'
 88AsnLeuLeuThrGlnIleGlyCysThrLeuAsnPhe
```

Example 1 – Sequencing primer was ACAGGA (1st yellow highlight):

PRIMER 5'ACAGGA
TEMPLATE CTATGTCCTCGTCTACTATGTCATAATCTTCTTTACT...

Predict the Sanger sequencing data and determine the correct reading frame.



Example 2 – A part of the DNA sequence of a mutant and wild-type HIV protease gene is shown below. *What is the mutation?*

i. Locate sequence differences.
ii. Read wild-type and mutant sequences.

silent mutation no change in A.A. Sequence.

	Val	Leu	Glu	Glu	Met
Mutant:	TA	CTA	GAG	GAA	A
WT:	TA	TTA	GAA	GAA	A
	Val	Leu	Glu	Glu	Met