Lecture 37: Cloning and DNA Sequencing
Insertion of HIV protease Gene into the Expression Vector

The final product, after PCR:

5' CCGAATTCCCTAGATCCTTTCGCAAA...TTAATTGGATCCGG 3'
3' GGCTTAAGGAATCTGAGAAAGCTGTT...AATTACGCTTGCGG 5'
EcoRI ProIlnIleThrLeuTrpPhe BamH1

A detailed view of the expression vector is shown below:

Promoter for--Regulatory element--Ribosome--Start--Leader--GAATTC--HIV--GAATTC--Stop--mRNA--
--mRNA Syn.--for mRNA Syn.---binding--Codon--Peptide--CTTAAG--Codon--Termination--
--site--Codon--Termination--Site--Codon--Termination--Site--Codon--Termination--Site--Codon--Termination--Site

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

EcoRI BamH1

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

EcoRI BamH1

Step C. Mix HIV protease fragment and expression vector:


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

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


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.

- **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 the cell and verify that it is correct, either by DNA sequencing, or by measuring fragment sizes after digestion with restriction enzymes.
Sanger DNA Sequencing: The order of the addition of bases by a polymerase, beginning with a primer = sequence of the DNA downstream from the primer.

- 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
  - the position of termination 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 EcoRI site that was used to insert the HIV protease gene (the sequence of the plasmid is known).

**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.

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

**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. Usual ratio is 1:100, i.e. most additions will not terminate.

```
5' GATTCA 3' | 5' GATTCAGOH 3' | 5' GATTCAGOH 3' | 5' GATTCAGGCGOH 3' | 5' GATTCAGCGA OH 3'
```

After completion of synthesis the following fragments would be generated.

**Measuring the location of the added dideoxyNTP:** The position of colored base is measured by separation of the DNA molecules by size. The smaller molecules pass the detector first.

Detector (by color)

Output (Time the molecule passes detector = same order 5' - 3' as the non-template strand)

First base added to primer

Increasing detection time = Increasing distance from primer

3' CTAAGTCGGCTATTGCCGCTA 5'

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

   \[ TGG = \text{Trp} \]

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.

   **Frame 1**
   
   \[ \text{CCT CAG ATC} \text{ or } \text{CCT CAG ATC} \text{ or } \text{CCT CAG ATC} \]
   
   **Frame 2**
   
   \[ \text{Pro Gln Ile} \text{ or } \text{Leu Arg Ser} \text{ or } \text{Ser Asp} \]
   
   **Frame 3**
   
   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.

**Region of HIV DNA Coding for HIV protease.**

\[ 5' - \text{ggagccgataagcaaggaactgtatcccttaacttgccttcctcagatcactctctttggca} \]

**A**

\[ \text{Pro Gln Ile Thr Leu Trp Glu} \]

\[ \text{Arg Pro Leu Val Thr Ile Leu Ile Gly Gly Gly} \text{ Glu Ala Leu Leu Asp Thr Gly} \]

**C**

\[ \text{Ala Asp Asp Thr Val Leu Glu Met Ser Leu Pro Gly Arg Trp Pro Gly Arg Trp Met Ile} \]

**G**

\[ \text{Gly His Asp Ala Ile Met Val Val Glu Gly Thr Val Leu Glu Gly Glu Met} \]

\[ \text{AAT CTT GGA GGG CAC GGT TTC AAT TTA GAA TTT TAA TCC CTT GAA CAG} \]

\[ 3' - \text{Asn Leu Leu Leu Glu Glu Cys Thr Leu Asp Phe} \]

**Example 1** – Sequencing primer was **ACAGGA** (1st yellow highlight). Find the 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?