Lecture 37: Cloning and DNA Sequencing

Insertion of HIV protease Gene into the Expression Vector

The final product, after PCR.

5' CGGAATTCCCTCAGATC1ACTCTTTGGCAGA...TTAAATTGGGATCCCG 3'
3' GCCCTTAAGCAGGATCTGAGAGAAACCGTATT...AATTAAAGGCCCTGGGCA 5'
EcoRl ProGlnleThrLeuTrpGln---LeuAspPhe BamHl

A detailed view of the expression vector is shown below:

- Promoter for --- Regulatory element --- Ribosome - Start --- Leader --- GAATTC --- GATCC --- Stop --- mRNA ---
- mRNA Syn. --- for mRNA Syn. --- Binding --- Codon --- Peptide --- CTTAGG --- CTTAGG --- Codon --- Termination
  site
  EcoRl
  BamHl

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 --- GATCC --- Stop --- mRNA ---
- mRNA Syn. --- for mRNA Syn. --- Binding --- Codon --- Peptide --- CTTAGG --- Protease --- CTTAGG --- Codon --- Termination

Step A. Cut HIV protease DNA (PCR product) with EcoRl and BamHl

EcoRl
BamHl

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

  site
  EcoRl
  BamHl

Step C. Mix HIV protease fragment and expression vector:


= Start --- Leader --- G-A-T-C-C---Stop---mRNA---
= Codon --- Peptide --- C-T-T-A-G

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

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

EcoRl
BamHl

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.

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.
1. If you want to add EcoR1 AND Bam H1 restriction sites onto the ends of a PCR to make it easier to clone, you can add it to the PCR primers. Which PCR primer set would work for the DNA sequence shown below?

5' CAGCTACGCTCTTCGCAA......TTAATTCGC 3'
3' CCGGATCTAGTAGAAAACGGT......AATTTAAGCC 5'

a. 5'-GAATTCCTT-3' AND 5'-TTCGGATCC-3'
b. 5'-GGAGAATTC-3' AND 5'-CTAGGAAG-3'
c. 5'-GAATTCCTT-3' AND 5'-AAGCTTACG-3'
d. 5'-CTAGGATAC-3' AND 5'-TTCCGGATCC-3'
e. 5'-GAATTCCTT-3' AND 5'-GGATCCGAA-3'

2. If the following DNA fragment was treated with EcoR1 or BamH1, where would these enzymes bind on the DNA and what DNA products would form?

EcoR1: G^AATTC
BamH1: G^GATCC

1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8

2. The following two DNA fragments were digested with EcoR1 and subject to treatment with DNA Ligase (+ATP), what are the possible products?

3. DNA fragments that are produced by a restriction enzyme can always be ligated to each other. Which of the following restriction enzymes produce fragments that could be ligated to fragments produced by a different enzyme?

HindIII (A^AGCTT): -A-A-G-C-T-T- -A A-G-C-T-T-
-T-T-C-G-A-A- -T-T-C-G-A A-

BamHI (G^GATCC): 5'-G-G-A-T-C-C-3' -G G-A-T-C-C-
-3'-C-C-T-A-G-G- 5'-G-C-T-A-G- G-

BglII (C^GATCG): 5'-G-C-T-A-G-G- 3' -C G-A-T-C-G-
-5'-C-C-T-A-G-G- 3'-G-C-T-A-G- C-

-5'-C-T-A-T-G- 3'-C-T-A T-A-G-

HaeIII (GG^CC): 5'-G-G-C-C- 3'-G-G C-C-
-5'-C-C-G-G- 3'-C-C G-G-
DNA Sequencing (dideoxy – Sanger Sequencing)

Key Concept: We will measure the location of each base from the 5’ end of the DNA. Since we know the location of each base, we know the sequence. Obtaining this information is conceptually quite simple. The idea is to generate a collection of all possible DNA chains that:

1. Begin at the same location,
2. End, or terminate, with a known base,
3. Measure the position of termination by measuring the length of the fragment using electrophoresis.

1) All begin at the same location, which is defined by the sequence of the primer. If we choose a primer that anneals on the plasmid adjacent to the place where the DNA was inserted, we can sequence the DNA without knowing anything about the sequence of the inserted DNA.

\[ \text{Primer} \]

\[ 5' - GATTCA - 3' \]

\[ 3' - CTAAGTCGCTATTGCGCTA - 5' \]

DNA sequence to determine - plasmid

2. 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? Can it be incorporated into the growing chain?

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 five additions of a base will terminate. Keep in mind that there are billions of primer-templates in the reaction.

\[ \begin{array}{llll}
5' - GATTCA - 3' & 5' - GATTCA - 3' - dATP & 5' - GATTCA - 3' - dATP & 5' - GATTCA - 3' - dATP \\
2' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT \\
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT \\
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT \\
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT \\
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT \\
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT \\
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT \\
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT \\
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT \\
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT & 3' - CCTAAGTCGCTATT
\end{array} \]

After completion of synthesis the following fragments would be generated.

\[ \begin{array}{llll}
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
5' - GATTCA & 5' - GATTCA & 5' - GATTCA & 5' - GATTCA \\
\ldots
\end{array} \]
3. The position of colored base is measured by electrophoresis – separation by size in an electric field.

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

Converting a DNA sequence to an Amino Acid Sequence (Human translation):

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 =
   
2. The first codon in all genes that encode proteins is ATG (AUG in the RNA), coding for the amino acid methionine in eukaryotic cells and formyl-Met in prokaryotic cells. All internal ATG codons will be translated as Met, e.g. a bacterial mRNA:
   
   5' ---AUG---Met---3' (and AUG
   
   (HIV protease does not start with a Met because it is produced from a longer peptide by proteolysis.)
   
3. Special codons (termination/stop 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.

<table>
<thead>
<tr>
<th>5' Base</th>
<th>Middle Base</th>
<th>3'</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>Phe</td>
<td>Cys</td>
</tr>
<tr>
<td>T</td>
<td>Ser</td>
<td>Tyr</td>
</tr>
<tr>
<td>T</td>
<td>Ser</td>
<td>Cys</td>
</tr>
<tr>
<td>Leu</td>
<td>Ser</td>
<td>Term</td>
</tr>
<tr>
<td>Leu</td>
<td>Ser</td>
<td>Trp</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Arg</td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>Gln</td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>Gln</td>
</tr>
<tr>
<td>A</td>
<td>Ile</td>
<td>Thr</td>
</tr>
<tr>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
</tr>
<tr>
<td>Ile</td>
<td>Thr</td>
<td>Lys</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
</tr>
</tbody>
</table>

There is only one correct reading frame. Signals in the mRNA tell the ribosome what reading frame to use. The reading frame from a DNA sequencing experiment can only be established with certainty by comparing the protein sequence predicted from the DNA to the protein sequence determined by chemical methods (e.g. Edman degradation). In the case of HIV protease, the established reading frame is:

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