Lecture 12 – Electrophoresis, Sanger DNA Sequencing & Bioinformatics.

Step G. Isolate plasmid DNA from candidate bacteria and digest with restriction endonucleases to verify that the plasmid is correct. Verify fragment sizes using agarose gel electrophoresis (DNA fragments separate based on size).

Electrophoresis – Measurement of the size of DNA fragments:
- Charged DNA molecules are forced through an agarose gel by an electric field. Gel contains crosslinked network of fibers.
- Smaller molecules migrate faster.
- Plot of \( \log(\text{bp}) \) versus distance can be used to determine size.

Application of PCR and Gel Electrophoresis – Identification of Individuals
- Human DNA contains many regions where the DNA sequence is repeated, for example: (AGAAA)(AGAAA). These are found on many chromosomes, as indicated on the right. The sequences flanking (5' and 3' side) the repeats are unique and conserved between individuals.
- A number of diseases are due to an increase in the number of these repeats within protein coding regions.
- The number of tandem repeats varies from individual to individual.
- Therefore the number of tandem repeats can be used to distinguish one individual from another, useful in forensics.

Indicate the PCR primers to amplify the following repeated region:
Sanger DNA Sequencing:

Key Concepts:
- 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 EcoR1 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 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. Usual ratio is 1:100, i.e., most additions will not terminate.

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.
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:

\[
\begin{array}{c}
5' \text{-GG} = \text{Trp}
\end{array}
\]

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

\[
\begin{array}{c}
\text{CCT} \text{CAG ATC}
\end{array}
\]

\[
\text{Pro} \text{ Ile Arg}
\]

Frame 2

\[
\begin{array}{c}
\text{CTC AGA TC}
\end{array}
\]

\[
\text{Leu} \text{ Arg Ser}
\]

Frame 3

\[
\begin{array}{c}
\text{CCT GAT C}
\end{array}
\]

\[
\text{Ser} \text{ Asp}
\]

There is only one correct reading frame. The reading frame from a DNA sequencing experiment can be established by:

- Comparing the protein sequence predicted from the DNA to the most protein sequence determined by chemical methods.

- Prediction based on the length of open reading frame (ORF), most proteins are longer than 50 aa.

Region of HIV DNA Coding for HIV protease.

\[
\begin{array}{c}
5' \text{-ggagccgataqacaagggaaactgtatccttaaatcttcctcagatcactcttttgccaa}^{57}
\end{array}
\]

\[
\begin{array}{c}
\text{Pro} \text{Gln Ile Thr Leu Trp Gln}
\end{array}
\]

\[
\begin{array}{c}
\text{CGACCCCTCGTCACAATAAGATAGGGGGCAACTAAAGGAGCTCTATTAGATACAGGA}^{111}
\end{array}
\]

\[
\begin{array}{c}
\text{Arg Pro Leu Val Thr Ile Lys Ile Gly Gly Gln Leu Ala Leu Leu Leu Asp Thr Gly}
\end{array}
\]

\[
\begin{array}{c}
\text{CGACCATGATACAGTATTAGAAAGGTTTGGCCAGGAAGATGGGAAACCAAAATTGAT}^{179}
\end{array}
\]

\[
\begin{array}{c}
\text{Ala Asp Asp Thr Val Leu Glu Glu Met Ser Leu Pro Gln Arg Trp Lys Pro Lys Pro Lys Met Ile}
\end{array}
\]

\[
\begin{array}{c}
\text{Gly Gly Gly Gly Phe Ile Lys Val Arg Gln Tyr Asp Gln Ile Leu Ile Glu Cys}
\end{array}
\]

\[
\begin{array}{c}
\text{GGACATAAGCTAAGGTACAGTATTAGTTAGGAGCTACACCCTTGCAACCAAAATTGA}^{239}
\end{array}
\]

\[
\begin{array}{c}
\text{Gly His Lys Ala Ile Gly Thr Val Leu Val Gly Pro Thr Pro Val Asn Leu Gly Arg}
\end{array}
\]

\[
\begin{array}{c}
\text{ATCTGTGACCTCAGATTGGTTTGGCAGTTTTAATTTTCCCATAGCCCTATGAGACT}^{354}
\end{array}
\]

\[
\begin{array}{c}
5' \text{-Asn Leu Leu Thr Gin Ile Gly Cys Thr Leu Asp Phe}
\end{array}
\]

No stop codon

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.

a) Locate the change in the DNA sequence in the data, then locate the change in the entire sequence.
b) What is the change in the amino acid sequence (You need to find the correct reading frame)?

**Bioinformatics:**

**HIV Protease Sequence:**

ggagccgagataagacaggaagagtatctcttaacttc
CCTCAGATCAGCTCTTTGGGCAAG
CGACCCCTCTGCATCAAAATAGAGATAGGGGCGGAACTAAGGAAAGGCTCTATTAGATAACAgG
GCAATGATGATCAGTTTTAGAGAATAATGATTTTCCAGGAGATGGAAACCAAAATGATA
GGGGAAAGGAGTTTTTATCAAAAGTAAGACAGCTAGTATGATAGCAGATACTCAGAAATCTGT
GGAGACATATAACCCTATAGTGTTACCTCTATTAGTAGGACCACCACTACTGCTCAACTCGGAAACG
AAATCTGTTGAGATCCAGATTGTTGCACCTTTAAATTTTcccccatgccccatgact\n
**Mutant sequence:** ATACAGTACTAGAAGAAATGAGTTTGCCAGGAAGATGGAAACCAAG

1. Use Blastn to find the location of the sequencing data, using “align two sequences”

```
Query 125 ATACAGTATAGAAGAAATGAGTTTGCCAGGAAGATGGAAACCAAG 169
Sbjct 1 ATACAGTACTAGAAGAAATGAGTTTGCCAGGAAGATGGAAACCAAG 45
```

2. Translate DNA sequencing data in all three reading frames, select the correct reading frame using the known wild-type sequence – you can easily find the correct amino acid sequence since we know that the DNA sequence starts at nucleotide 125 in the entire sequence of the HIV protease. Translation is done using Expasy:

Frame 3 matches the HIV sequence beginning at DNA base 125 (2nd yellow highlight above). Note, we should have ignored frame 1 because it has two stop codons.

3. Once the correct reading frame is known, you could use Blastp to align the two amino acid sequences (complete, and frame 3 from sequence data), showing that this mutation is silent.

```
Query 31 TVLEEMSLPGGRWKP 44
Sbjct 1 TVLEEMSLPGGRWKP 14
```