Genome Sequence Assembly

Learning Goals:

- Introduce the field of bioinformatics
- Familiarize the student with performing sequence alignments
- Understand the assembly process in genome sequencing

Introduction:

The order of the nucleotides in a gene and the order of the amino acids in a protein are referred to as the sequence of the gene and sequence of the protein respectively. The process in which the sequence of a particular gene is determined is gene sequencing. It is possible to determine the sequence of a protein if the sequence of a gene is known by using the genetic code. If a gene has been sequenced, then its sequence is submitted to the public database at the National Center for Biotechnology Information.

http://www.ncbi.nlm.nih.gov

Determining a protein sequence is more difficult than determining a gene sequence and involves using Mass spectroscopy. Knowing a gene sequence is useful in analyzing the function of a protein as the amino acid composition and order determines the chemical properties of the protein.

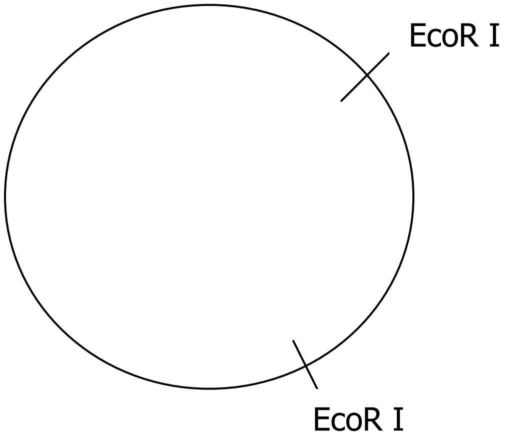
Genome sequencing has become an essential tool in biology. By sequencing the genome of a species, the sequence of every gene is known. Knowing the sequence of a gene enables a biologist to develop primers for PCR and RT-PCR, important laboratory methods for analyzing gene expression and function. Furthermore, the sequence of a gene can provide information as to its function.

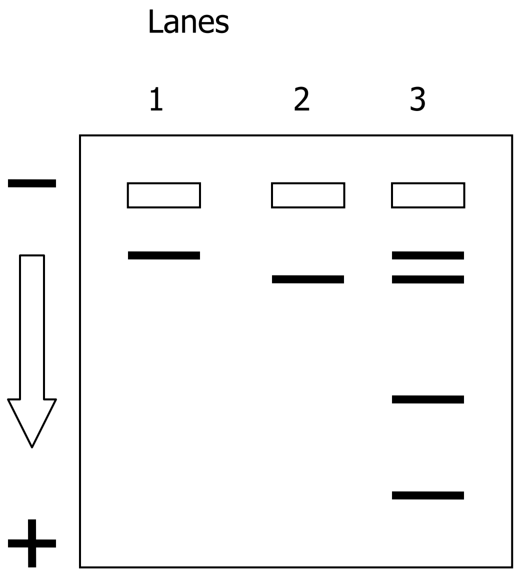
Genome sequencing is an enormous task and requires both laboratory and computational techniques. First, the genome is broken up into many smaller fragments using restriction enzymes. Each fragment is cloned using cloning vectors such as plasmids, yeast artificial chromosomes, or bacterial artificial chromosomes and then sequenced. Ideally a large number of fragments are cloned and sequenced such that there are overlapping regions between the fragments. Once a sufficient number of fragments have been sequenced, the **assembly** process begins. Bioinformatic techniques allow for the production of an entire genome sequence by aligning the fragments and determining their order.

In Class preparatory questions

The answers to the following questions will be reviewed at the beginning of your lab session. It is in your best interest as a student to attempt them prior to Lab, though not required.

1. In the incomplete diagram below (not drawn to scale), Lane 1 contains plasmid DNA that is UNCUT. Lane 2 contains the plasmid DNA cut with the Restriction enzyme EcoR I. Lane 3 contains a DNA ladder, where the bands represent the following sizes: 4,000 bp, 3000 bp, 500 bp, and 250 bp. Draw in the additional band missing from lane 2.





Methods

Determine Order of Fragments

1. Using BLAST you will align two fragments at a time. The output of the program will provide you with the lengths of the fragments and the regions that overlap.

http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi

a. record the length of each fragment

Fragment #	**Length**
1	
2	
3	
4	
5	

b. indicate whether there is an overlap between fragments (Y or N)

	Fragment 1	Fragment 2	Fragment 3	Fragment 4	Fragment 5
Fragment 1					
Fragment 2					
Fragment 3					
Fragment 4					
Fragment 5					

c. Record the regions of overlap (which bases) between the fragments.

d. Draw a diagram of how the fragments align with each other.

e. What is the order of the fragments? ________________________

f. What is the length of the full sequence? _________

```
>Fragment 1
AAAGCAAAACTTGGCAAGCAAACCTTCGATTGATCTCTAAGTTTGATACTGTTGAAGACTTTT
GGGCTCTATACAACCATATCCAGTTGTCTAGTAATTTAATGCCTGGCTGTGACTACTCACTTTTTAAGGA
CGGGATTGAGCCTATGTGGGAAGATGAGAAAAACAAACGAGGAGGACGGTGGCTGATCACACTGAACAAG
CAGCAGAGACGGAGTGACCTCGATCGCTTCTGGCTAGAGACACTGCTGTGCCTTATTGGAGAATCTTTCG
ATGACTACAGTGATGATGTGTGTGGAGCTGTTGTTAATGTTAGAGCTAAAGGTGATAAGATAGCAATATG
GACTACTGAGTGTGAAAACAGAGATGCAGTCACACACATAGGGAGGGTATACAAGGAAAGGTTAGGACTT
CCTCCGAAGATAGTGATTGGTTATCAGTCCCACGCAGACACAGCTACAAAGAGCGGCTCCACCACTAAAA
ATAGGTTTGTTGTTTAAGAAGACACCTTCTGAGTATTCTCACAGGAGACTGCGTCACGCAATCGAGATTG
GGAGCTGAACCAAAGCCT


>Fragment 2
GGAGGCGGAGGGAGCTGGTCCTTAAGGAAGGCACGCGCTTGCTTCTAGATTCCGAAGCGTTTTCAAAGCT
GGTTACAGTCCTTACCACAGCACACCCTTGTGAGGAGCGGTTGTGCGATCAGATCGATCTAAGATGGCGA
C


>Fragment 3
GGAACCGGAAACCACCCCTACCACTAATCCCCCACCTGCAGAAGAGGAAAAAACAGAGTCTAATCA
GGAGGTTGCTAACCCAGAGCACTATATTAAACACCCTCTACAGAACAGGTGGGCACTCTGGTTTTTTAAA
AATGATAAAAGCAAAACTTGGCAAGCAAACCTTCGATTGATCTCTAAGTTTGATACTGTTGAAGACTTTT
GGGCTCTATACAACCATATCCAGTTGTCTAGTAATTTAATGCCTGGCTGTGACTACTCACTTTTTAAGGA
CGGGAT

>Fragment 4

GAGTGACCTCGATCGCTTCTGGCTAGAGACACTGCTGTGCCTTATTGGAGAATCTTTCG
ATGACTACAGTGATGATGTGTGTGGAGCTGTTGTTAATGTTAGAGCTAAAGGTGATAAGATAGCAATATG
GACTACTGAGTGTGAAAACAGAGATGCAGTCACACACATAGGGAGGGTATACAAGGAAAGGTTAGGACTT
CCTCCGAAGATAGTGATTGGTTATCAGTCCCACGCAGACACAGCTACAAAGAGCGGCTCCACCACTAAAA
ATAGGTTTGTTGTTTAAGAAGACACCTTCTGAGTATTCTCACAGGAGACTGCGTCACGCAATCGAGATTG
GGAGCTGAACCAAAGCCTCATCAAAGCAGAGTGGACTGCACTGAAGTTGATTCCATCCAAGTGTTGCTAA
GATATAAGAGAAGTCTCATTCGCCTTTGTCTTGTACTTCTGTGTTCATTCTCCTCCCCCACCCCCAATTT
TTGCTAGTGTGTCCACTATCCCAATCAAAGAATTACAGTATACGTCACCCCAGAACCCGCAGATGTGTTC
CTGGCCCGCTCTGTAACAGCCGGTTAGAATTACCATGACACACACATTTGCCTTTCCACAGTATTCGAAA

>Fragment 5
GCGGTTGTGCGATCAGATCGATCTAAGATGGCGA
CTGTGGAACCGGAAACCACCCCTACCACTAATCCCCCACCTGCAGAAGAGGAAAAAACAGAGTCTAATCA
GGAGGTTGCTAACCCAGAGCACTATATTAAACACCCTCTACAGAACAGGTGGGCACTCTGGTTTTTTAAA
AATGATAAAAGCAAAACTTGGCAAGCAAACCTTCGATTGATCTCTAAGTTTGATACTGTTGAAGACTTTT
GGGCTCTATACAACCATATCCAGTTGTCTAGTAATTT
```