Bioinformatics - Recombinant Technology

Instructions, Resources and pre-requisites:

1. You need some basic understanding of bioinformatics – e.g from the computing labs associated with the course

2. You should have attended the lectures on the module recombinant technology and bioinformatics computer labs associated with Advanced Molecular Biology & Bioinformatics (7301BPS) course.

3. You need to have some working knowledge on how to use wEMBOSS. Read the instructions at http://trishul.sci.gu.edu.au/courses/7301BPS/guidewEMBOSS.pdf

4.You should have read the instructions for logging onto wEMBOSS at the URL http://trishul.sci.gu.edu.au/courses/7301BPS/wEMBOSS_instructions.html

5. You should have a password and a login for using http://trishul/sci.gu.edu.au/wEMBOSS

6. Though you can do some of the assignments using online web tools, you are encouraged to use wEMBOSS resources for all your assignments. To assist in your assignments tutorials have been developed.

7. Be prepared to have your web browser open to download sequences and to do searches.

8. The objective for this part is to use and extend what you've learned in your Recombinant Technology module so that you can better connect bioinformatics with biotechnology laboratory experiments. You are required as part of your assignment (refer to the course outline for the submission dates and the assessment criteria for the assignment) to complete the following set tasks:
   * identify restriction sites in a sequence
   * design and evaluate primer pairs and probe sequences
   * locate resources on plasmids

9. A few useful links for restriction sites, DNA sequencing and recombinant DNA can be found at
   * http://www.accessexcellence.org/AE/AEC/CC/restriction.html AND
   * http://www.ultranet.com/~jkimball/BiologyPages/R/RestrictionEnzymes.html

   Extensive resources for primer design, software and tutorials
   * http://www.humgen.nl/primer_design.html

10: Are you confused about primer and target DNA strand orientation?

Primers must be able to anneal to the target DNA in a predictable location (target site) and on a predictable strand (3' or 5') and be able to amplify the region between the primers with DNA Polymerases (eg Taq). Some students may be confused about DNA sequence, strand (3', 5') and primer orientation. The following is a useful guide that may overcome the confusion..

(a) Sequences are always written from 5' to 3'. This includes the sequence of your template DNA (if known), the sequence of the vector DNA into which it is inserted, and the sequence of proposed primers. Don't ever write a primer sequence reversed or you will only confuse yourself and others.

(b) Polymerase always extends the 3' (OH- hydroxyl end) end of the primer, and the sequence you will read will be the same strand (sense or anti-sense) as the primer itself.

(c) Thus, if you choose a primer sequence that you can read in your source sequence (for example, in the vector), the sequence you will obtain will extend from the primer's right (3') end.

(d) Conversely, if you choose a primer from the strand opposite to what your 'source' sequence reads, the resulting sequence will read towards the left.

Summary:
In modern biology, there is always a need to use simple computational applications when working with DNA sequencing, using PCR or digesting DNA with restriction enzymes. Designing primer pairs and locating restriction sites are two examples when such tools have to be called upon. Depending on the project, there may also be a need for selecting appropriate plasmids to meet the intended goals, for example, in gene expression when one has to clone a gene in a plasmid vector.

Typically, primers can often be designed merely by visually inspecting a sequence for short sequences (15-25 bp) which satisfy the following defined criteria:

- the sequence pairs should be of equal length
- have at least 50% G+C content, [Tm= 2(A+T)+4(G+C)], and
- anneal at about the same temperature, ideally between 50-65°C.

Though it is possible to design the primer without resorting to computational software, there are however several primer design programs which can also be used to assist in the design process. Experimental verification of primer function and reaction optimization is always necessary once primers have been designed.

Finding and using genetic markers has long been recognized as being extremely useful in a wide variety of applications. Examples include recognition of disease gene polymorphisms, use in studies of pathogens and epidemiology, selection of plants for desirable agricultural characteristics such as seed yield or height, and analysis of forensic evidence for civil and criminal court cases. The combination of more means of identification of polymorphisms along with the capability of genomic analysis has brought considerable improvements in these areas of inquiry. Additionally as a result of these advances, other areas are undergoing rapid expansion and development. Gene identification and association with function(s) in metabolism, development and cell differentiation, and various response systems are some examples. Means of identifying polymorphisms include restriction fragment length polymorphism (RFLP) and use of probes for specific gene markers, either as part of the gene in question or a closely linked marker to the gene of interest. Designing probes is much like primer design. The object is to select a sequence which is both specific for the target and which has desirable characteristics compatible with the assay application.

You will have the opportunity of using programs for primer design, probe design and restriction enzyme mapping in wEMBOSS, and you are also welcome to explore other software listed under "Useful links", above.

PART 1 Working with wEMBOSS

1.1 Getting sequences, sequence file names, sequence file formats, converting file formats in wEMBOSS

Refer to "A Quick Guide To wEMBOSS" downloadable from the URL http://trishul.sci.gu.edu.au/courses/7301BPS/guidewEMBOSS.pdf during this tutorial.

(a) Search, download & create a mini-databases in wEMBOSS. Download the proteome of *Halobacterium* species from ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Halobacterium_sp/AE004437.faa and copy to your directory in wEMBOSS. For this, create a new file, paste the sequence and save it in your directory as AE004437.faa. This sequence now becomes a mini sequence database which has multiple sequence entries in fasta file format. View this file to check the file format.

(b) Creating a list of sequence names in a file- You may be interested in only a few sequences rather than all the sequences in your newly created database, AE004437. In order for you to work on the sequences of interest you have to first create a list of sequence names. This can be achieved in one of two ways in wEMBOSS as described below:

- Create a file called "my_TFB_NCBI_list" and list the sequence names and / or multiple sequence names: For example, with the following Transcription Initiation Factor IIB (TFB) protein identifiers.

  AE004437.faa:AAG20319.1
  AE004437.faa:AAG19212.1
  AE004437.faa:AAG18850.1
  AE004437.faa:AAG19313.1
  AE004437.faa:AAG18894.1

- Create the following list in the pre-existing file called protlist (= protein list)

  AE004437.faa:AAG20319.1
  AE004437.faa:AAG19212.1
  AE004437.faa:AAG18850.1
  AE004437.faa:AAG19313.1
  AE004437.faa:AAG18894.1
  @my_TFB_NCBI_list

> To create a list of single sequence name and / or multiple sequence names, you will need to understand the syntax for "Uniform Sequence Address –USA" for this-
> * "file"
> * "file:entry"
> * "format::file"
> * "format::file:entry"
> * "database:entry"
> * "database"
> * "@file"

(c) Retrieve the corresponding sequences from the "local fasta database" using seqret

(d). Reformat sequences into other formats eg Swissprot, GenBank etc.

1.2 Creating Files and Databases in wEMBOSS – Assignment 1

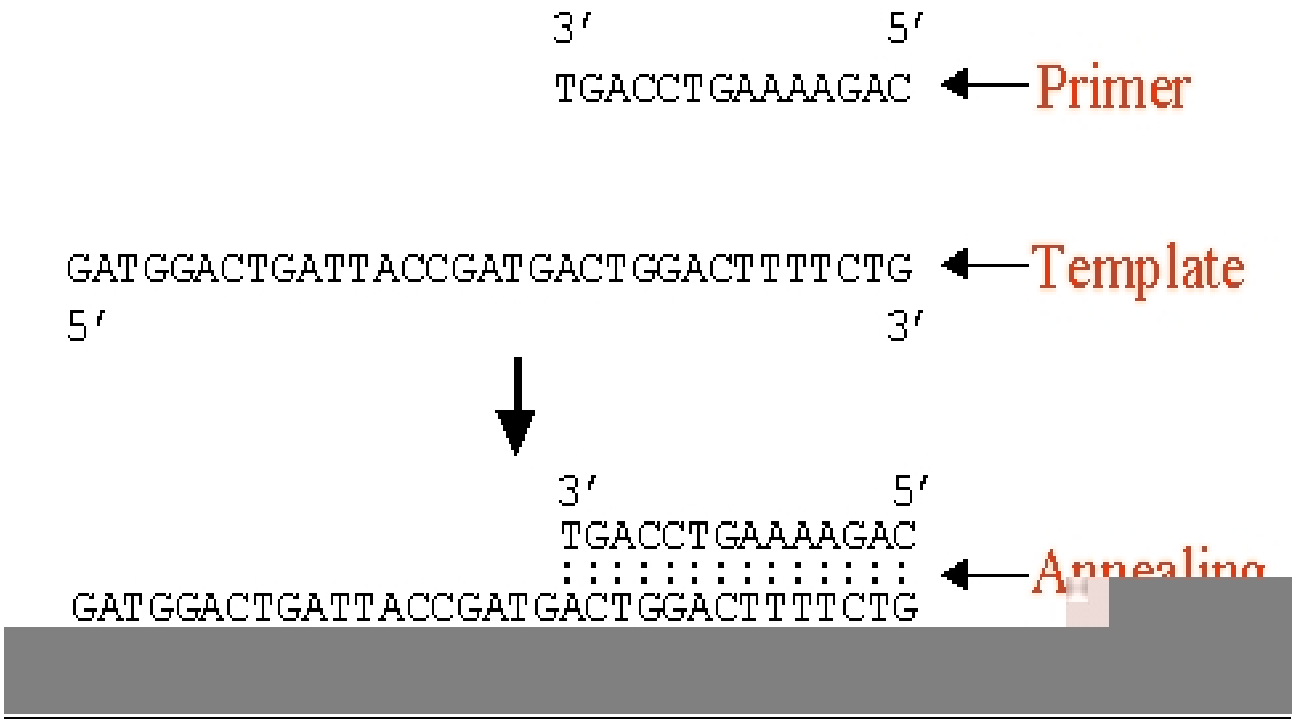
This assignment is based on tutorial 1 and anyone who has understood the tutorial will be able to complete this assignment.

(a) Go to the URL http://www.ncbi.nlm.nih.gov/sites/gquery (NCBI-Entrez) and search for "16S rRNA Bacillus".
(b) Note the output from the search.
(c) Select and download approximately 20 sequences (in GenBank format).
(d) Create a mini-database "Bacillus_16SrRNA" and save the sequences
(e) Create a new file with a list of 3 sequence names in nuclist (= nucleotide list)
(f) Create a list of multiple sequence names in nuclist.
(g) Retrieve the corresponding sequences database using seqret
(h) Reformat the sequences in a different format and explain the format

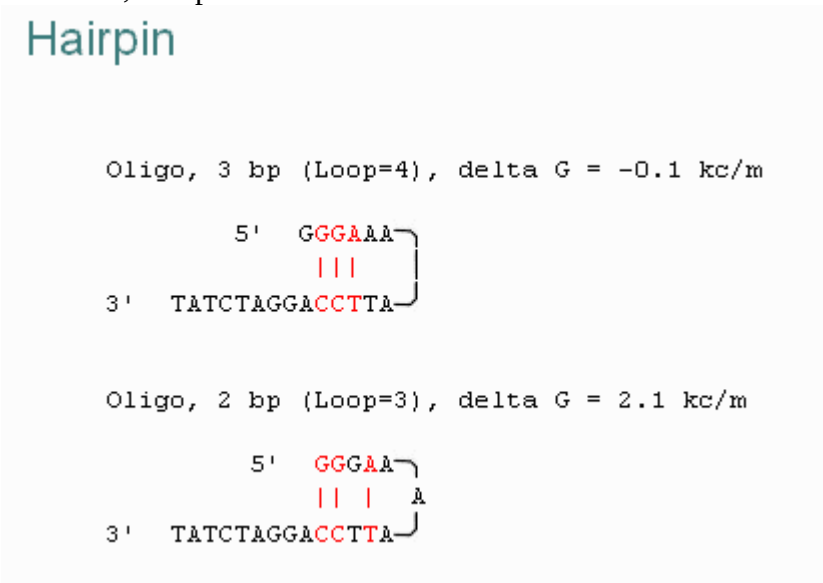
PART 2 Primer and Probe Design

2.1. What is a primer?

A primer is a short synthetic oligonucleotide which is used in many molecular techniques from PCR to DNA sequencing. These primers are designed to have a sequence which is the reverse complement of a region of template or target DNA to which we wish the primer to anneal.



When designing primers for PCR, sequencing or mutagenesis it is often necessary to make predictions about these primers, for example melting temperature (Tm) and propensity to form dimers with itself or other primers in the reaction. The following program will perform these calculations on any primer sequence or pair- http://www.idtdna.com/html/analysis/Calculator.html The program will calculate both the Tm of the primers, as well as any undesirable pairings of primers. When primers form hairpin loops or dimers, less primer is available for the desired reaction. For example.



Hairpin

Self-Dimer

```
4 bp, delta G = -6.6 kc/m (bad!) (worst= -36.6)
        5' GGGAAAATTCCAGGATCTAT 3'
            ||||  ||||
3' TATCTAGGACCTTAAAAGGG 5'

4 bp, delta G = -5.4 kc/m (bad!) (worst= -36.6)
      5' GGGAAAATTCCAGGATCTAT 3'
              ||||
3' TATCTAGGACCTTAAAAGGG 5'
```

Dimer

2.2 Some thoughts on designing primers (adapted from Innis and Gelfand,1991)

1. Primers should be 17-28 bases in length
2. Base composition should be 50-60% (G+C)
3. Primers should end (3') in a G or C, or CG or GC: this prevents "breathing" of ends and increases efficiency of priming
4. Tm between 55-80 ^{o}C are preferable
5. 3'-ends of primers should not be complementary (ie. base pair), as primer dimers will be synthesised preferentially to any other product
6. Primer self-complementation (ability to form secondary structures eg hairpins) should be avoided
7. Runs of three or more Cs or Gs at the 3'-ends of primers may promote mispriming with G or C-rich sequences (because of stability of annealing), and should be avoided.

Also keep in mind that most oligonucleotide synthesis reactions are only 98% efficient. This means that each time a base is added, only 98% of the oligos will have the correct sequence. Often this is not critical with shorter oligos, but as the oligo length increases, so does the probability that the primer will be incorrectly synthesized (see table below). This is very important in experiments that deal with mutagenesis or cloning. In such case, purification by HPLC or PAGE is recommended.

Table showing primer synthesis efficiency:

Oligonucleotide length	Percent with correct sequence
10 bases	(0.98)10 = 81.7%
20 bases	(0.98)20 = 66.7%
30 bases	(0.98)30 = 54.6%
40 bases	(0.98)40 = 44.6%

2.3 Primer Design - Assignment 2

1. The objective is to design primers 15-25 bp long that will amplify both of the sequences below. Note the sequence format.

>Paddlefish
CCTTGGCCTCTGCCTAATCACACAGATTCTAACAGGATTATTTCTCGCAATACACTACACAGCTGACA
TCTCAACAGCCTTCTCCTCCGTCGCCCACATCTGTCGAGATGTTAACTACGGATGACTAATTCGAAAC
ATTCATGCAAACGGAGCCTCCTTTTTCTTCATCTGCCTCTACCTTCACGTAGCCCGAGGCATATACTA
TGGCTCATACCTCTACAAAGAAACCTGAAACATCGGAGTAGTTCTCCTACTCCTAACTATAATAACCG
CCTTCGTAGGATATGTGCTCCCATGAGGACAGATATCCTTCTGAGGAGCCACCGTAATTACCAACCTT
CTTTCCGCCTTCCCCTACATCGGGGACACCCTAGTACAATGAATCTGAGGTGGTTTCTCAGTAGACAA
CGCCACCCTAACC

>Shovenose Sturgeon
CCTAGGCCTCTGCCTTATTACACAAATCTTAACAGGACTATTTCTTGCAATACACTACACAGCTGACA
TTTCAACAGCCTTCTCCTCCGTCGCCCACATCTGCCGAGACGTAAACTACGGGTGACTAATCCGAAAC
GTCCACGCAAATGGCGCCTCCTTCTTCTTTATCTGCTTGTACCTTCACGTCGCACGAGGTATATACTA
CGGCTCCTACCTCCAAAAAGAAACCTGAAACATCGGAGTAGTCCTCTTACTCCTCACCATAATAACCG
CCTTCGTAGGCTATGTACTGCCCTGAGGACAAATATCATTTTGAGGGGCAACCGTAATCACTAACCTC
CTTTCCGCCTTCCCGTACATCGGCGACACATTAGTGCAATGAATCTGAGGCGGCTTTTCAGTC

2. Copy the two sequences

3. Go to wEMBOSS and log in.
   (a) Open a new file and paste the two sequences into the file.
   (b) Save the file (give the file a name.fasta or a file convention that you can remember)
   (c) Align the two sequences using emma and identify the conserved regions in both the sequence. Make sure that you read the emma manual.

4. There are many other alignment programs which can used. Search the web for such programs and list at least three together with their URLs.

5. The conserved regions that have noted in 3(c) could be good targets for PCR primers to bind to both sequences. Design primers for these target sites. Note that the two primers need to bind to the target DNA such that the free 3' ends of each primer point towards each other. You may wish to review the rules used to for primer design.

6. After you have identified the sequence of your primers, check the primers with the programs used to calculate melting temperature (Tm) and the formation of primer dimers. If the Tm is less than 55°C or the primer has the potential to form bad hairpins or dimers, try another region of the sequence.

2.4 Primer Design - Assignment 3

The objective of the assignment is to database such as GenBank, DDBJ, EMBL to search for and download sequences, to transfer and write these sequences using wEMBOSS and to design PCR primers to amplify specific regions of the target DNA.

1. Go to Genbank (or DDBJ or EMBL) and search for the following entries. Note the regions of the DNA these sequences correspond to. Which species?
   AF016978 OVU12869 BBU12864

2. Download (or copy) the sequence entries in fasta format.

3. Go to wEMBOSS and log in.
   (a) Open a new file and paste all the three sequences into the file.
   (b) Save the file (give the file a name.fasta or a file convention that you can remember)

4. Use "emma" (clustalw) and align the three sequences.

5. Design PCR primers which will amplify all three DNAs. The Tm of the primers should be >50°C.

6. Now design primers such that the PCR products (amplicons) from two different species which will be of different sizes. This will allow for rapid identification of the species.

7. List the Tm of primers?

8. Check the primers for potential to form hairpin structures, dimers etc at the URL http://www.idtdna.com/analyzer/Applications/OligoAnalyzer/

2.5 Primer Design - Assignment 4

1. Check out the following sites to become familiar with available tools for primer design - http://www.chemie.uni-marburg.de/~becker/ Other good resources, for restriction analysis can also be found at the same URL or at the following URL http://www.hgmp.mrc.ac.uk/GenomeWeb/nuc-primer.html Browse a few of listed sites to get a feel for what they have to offer. Among the available freeware, GeneWalker, Primer3, and PrimerDesign are frequently cited and recommended. You can use Primer3 in wEMBOSS.

2. The following is the complete mRNA sequence (1090 nucleotides) for equine ubiquitin C-terminal hydrolase. The protein is of interest as it is found in the synovial fluid at above-normal concentrations in horses suffering from osteoarthritis.

   >gi|10336505|dbj|AB049188.1|AB049188

```
CTGTTTTTCCTACTCCCTGGCTTCTCCTCCTTCTCGCTCTTCGCGAAGATGCAGCTCAAACCGATGGAGA
TTAACCCCGAGATGCTGAACAAAGTGCTGGCCAGGCTGGGGGTCGCCGGCCAGTGGCGCTTCGTGGACGT
GCTGGGGCTGGAGGAGGAGACTCTGGGCTCGGTGCCAGCGCCTGCCTGCGCCTTGCTGCTGCTGTTTCCC
CTCACGGCCCAGCATGAGAACTTCAGGAAAAAACAGATTGAAGAACTGAAGGGACAAGAAGTCAGTCCTA
AGGTGTACTTCATGAAGCAGACCATTGGGAACTCCTGCGGTACCATCGGACTTATCCACGCCGTGGCCAA
TAACCAGGACAAACTGGAGTTTGAGGATGGATCGGTCCTGAAACAATTTCTTTCTGAAACGGAGAAGTTA
TCCCCTGAAGACAGAGCCAAATGCTTTGAAAAGAATGAGGCCATTCAGGCAGCCCATGATGCTGTGGCAC
AGGAAGGCCAATGTCGGGTAGATGACAAAGTGAACTTTCATTTTATTCTGTTTAACAACGTGGATGGCCA
CCTCTATGAACTTGATGGGCGGATGCCTTTCCCGGTGAACCATGGCACCAGTTCAGAGGACCTGCTGCTG
CAGGACGCCGCCAAGGTCTGCAGAGAATTCACTGAGCGTGAGCAAGGCGAAGTCCGCTTTTCTGCTGTGG
CGCTCTGCAAGGCAGCCTAATGCCCTGTAAGAGGGACTTGGCTTTTTTCCTCTCTCCCCTTCAACGTGAA
ATATATCCTGACCGATGCAGTCTAAGATGCTTCCCTACTTGTAGAACACAGCTGTTCTCCTTTGGTTCTG
CAGGCCTGCTCCTCCCCTCCGCCACACCCAAGCACTAGCAGAGCTCAGCTGTCGATCGAGCAAAGTTTGG
TGTAAGCTTCAGGTGGCGAAGCATTTCCCCCAGTGTATGTCTTGTATCTCAATATCTAATGCTTTAAATG
GCTACTTTGGTTTGTGTCTGTAAGTTAAGGCCTTGGATGTGGTTTAATTGTTTGTCCTTAAAAGGAATAA
AACTTTTCTGCTGATAAGAAAAAAAAAAAAAAAAAAAAAA
```

3. Find 3-4 primer pairs which could be used to help fully sequence the gene. You may use Primer3 in wEMBOSS or another program of your choosing.

4. Briefly outline how you would test the functionality of the designed primers.

5. Briefly outline how you would test the specificity of the designed primers.

2.6 Probe Design and Selection -Assignment 5

1. Molecular probes are used in a variety of applications, such as Southern and Northern blots and microarrays. For a sequence to be useful as a probe, it needs to be specific for the target sequence. It must not bind to anything else which might be present in the sample being screened. Designing probes is similar to designing primer pairs. In fact, one way to begin is to use some of the same primer design programs. Other approaches may work as well.

2. Tissue typing provides an excellent example of an activity which has been accomplished using a variety of methodologies over the years, including the development and use of probes. Screening for human major histocompatibility complex [MHC] antigens is required for tissue typing of possible donors for organ transplants and of the recipients. It is also useful in genetic linkage analysis for a variety of disease associations. Originally, screening was done using a cytotoxicity assay system for some of the loci, and a mixed lymphocyte response [MLR] assay for other loci. Use of RFLPs was introduced in the late 1980's followed by introduction of PCR methods in the 1990's. Microarrays are now being developed and introduced for HLA screening.

3. HLA and tissue typing references:

   Tutorial on different HLA screening methods

   http://www.umds.ac.uk/tissue/what1.html

   http://www.anthonynolan.org.uk/HIG

Publications :

Feolo M, Fuller TC, Taylor M, Zone JJ, Neuhausen SL (2001). A strategy for high throughput HLA-DQ typing. J Immunol Methods 258:65-71

Cai H, White PS, Torney D, Deshpande A, Wang Z, Keller RA, Marrone B, Nolan JP.(2000). Flow cytometry-based minisequencing: a new platform for high-throughput single-nucleotide polymorphism scoring. Genomics 66:135-43. Erratum in: Genomics 2000 69:395

4. The Assignment Task: The HLA allele with the strongest association with a disease is HLA B27. It has a strong association frequency with ankylosing spondylitis and one type of arthritis. Therefore it is of interest to screen for the presence of this particular allele. The complete mRNA sequence for B27 is in GenBank: gi|187657|gb|M12678.1|HUMMHB27A.Using a method of your choosing, design a probe of 20-30 nucleotides which will detect the presence of B27 in DNA samples. (See Summary question 2 below.)

   a. Outline your approach for designing a probe specific for HLA B27

   b. Give the sequence you designed.

   c. Test your sequence for specificity to B27 and verify that it recognizes only B27 and no other B allele, nor any other gene, in the human genome.

   d. As a follow up on HLA B27's association in disease, you may be interested in reading the following paper which combines two computational methods to identify peptide sequences from *Chlamydia trachomatis* predicted to be involved in binding B27 and which may be involved in the pathogenesis of B27 associated disease.

2.7. Restriction mapping - Assignment 6

1. Check out the following sites to get a feel for some of what is available.
   http://www.accessexcellence.org/AE/AEC/CC/restriction.html
   http://www.ultranet.com/~jkimball/BiologyPages/R/RestrictionEnzymes.html
   http://internalmed.wustl.edu/divisions/enzymes/INDEX.HTM

2. The following is the complete mRNA sequence (1090 nucleotides) for equine ubiquitin C-terminal hydrolase.

   >gi|10336505|dbj|AB049188.1|AB049188

```
CTGTTTTTCCTACTCCCTGGCTTCTCCTCCTTCTCGCTCTTCGCGAAGATGCAGCTCAAACCGATGGAGA
TTAACCCCGAGATGCTGAACAAAGTGCTGGCCAGGCTGGGGGTCGCCGGCCAGTGGCGCTTCGTGGACGT
GCTGGGGCTGGAGGAGGAGACTCTGGGCTCGGTGCCAGCGCCTGCCTGCGCCTTGCTGCTGCTGTTTCCC
CTCACGGCCCAGCATGAGAACTTCAGGAAAAAACAGATTGAAGAACTGAAGGGACAAGAAGTCAGTCCTA
AGGTGTACTTCATGAAGCAGACCATTGGGAACTCCTGCGGTACCATCGGACTTATCCACGCCGTGGCCAA
TAACCAGGACAAACTGGAGTTTGAGGATGGATCGGTCCTGAAACAATTTCTTTCTGAAACGGAGAAGTTA
TCCCCTGAAGACAGAGCCAAATGCTTTGAAAAGAATGAGGCCATTCAGGCAGCCCATGATGCTGTGGCAC
AGGAAGGCCAATGTCGGGTAGATGACAAAGTGAACTTTCATTTTATTCTGTTTAACAACGTGGATGGCCA
CCTCTATGAACTTGATGGGCGGATGCCTTTCCCGGTGAACCATGGCACCAGTTCAGAGGACCTGCTGCTG
CAGGACGCCGCCAAGGTCTGCAGAGAATTCACTGAGCGTGAGCAAGGCGAAGTCCGCTTTTCTGCTGTGG
CGCTCTGCAAGGCAGCCTAATGCCCTGTAAGAGGGACTTGGCTTTTTTCCTCTCTCCCCTTCAACGTGAA
ATATATCCTGACCGATGCAGTCTAAGATGCTTCCCTACTTGTAGAACACAGCTGTTCTCCTTTGGTTCTG
CAGGCCTGCTCCTCCCCTCCGCCACACCCAAGCACTAGCAGAGCTCAGCTGTCGATCGAGCAAAGTTTGG
TGTAAGCTTCAGGTGGCGAAGCATTTCCCCCAGTGTATGTCTTGTATCTCAATATCTAATGCTTTAAATG
GCTACTTTGGTTTGTGTCTGTAAGTTAAGGCCTTGGATGTGGTTTAATTGTTTGTCCTTAAAAGGAATAA
AACTTTTCTGCTGATAAGAAAAAAAAAAAAAAAAAAAAAA
```

3. Create a restriction map showing the cut sites for 2 enzymes of your choosing. For this, do the following:
   (a) Search wEMBOSS and list a few programs that can be used for this purpose.

(b) Search programs on the web that be used for this purpose. List these together with their URL

2.8. Selecting Plasmids - Assignment 7

1. It is beyond the scope of the tutorial to provide you with all the current information on vector and gene expression technologies. Decisions on selection of the best plasmid is task dependent eg cloning of PCR products vs gene expression, Check out the following sites for basic information
   http://www.dur.ac.uk/~dbl0www/Bioinformatics/DNA_corner.htm (search vectors)
   http://www.carolina.com/biotech/plasmid_problems/plasmid_guide.asp

2. Answer the following questions:
   a. How can having a restriction map on an mRNA or cDNA sequence be useful in helping to select a plasmid for cloning?

   b. Why do some plasmids have several different restriction sites in specific regions, especially within an antibiotic gene or an enzyme gene?

   c. What factors dictate restriction enzyme choice?

2.9 Primer Design, Plasmid Selection and Gene Expression:

The gene of interest usually has to be amplified from genomic or vector DNA by PCR (polymerase chain reaction) before it can be cloned into an expression vector. The first step is the design of the necessary primers. An example of primer design to restriction enzyme and cloning vector selection at the URL **http://www-nmr.cabm.rutgers.edu/bioinformatics/Primer_Primer/**