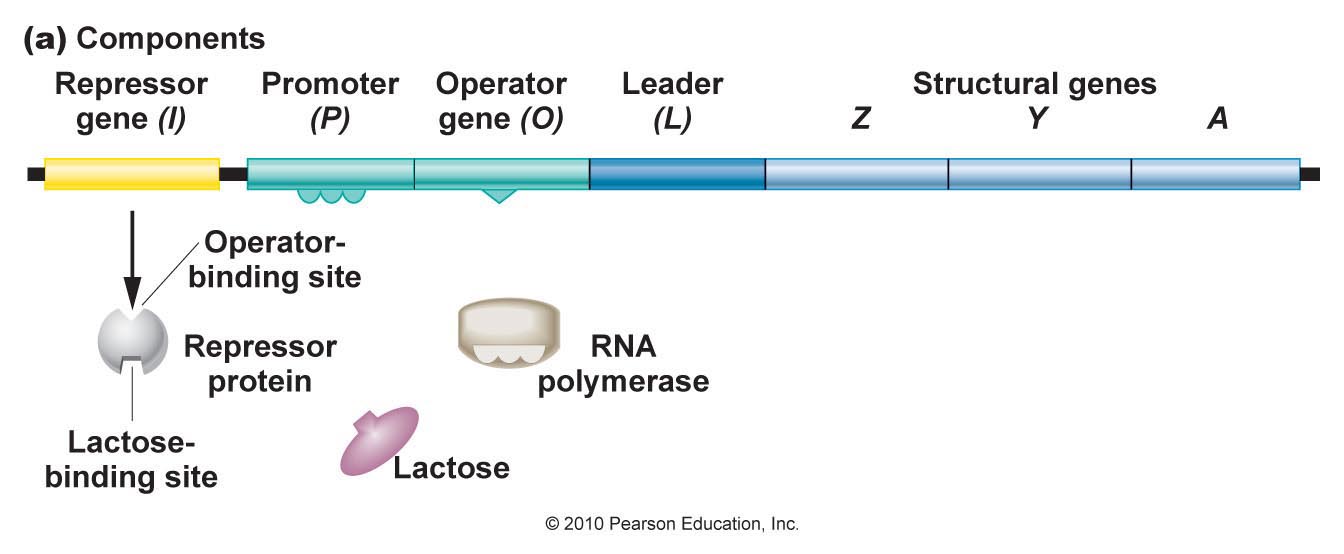
**Bioinformatics Assignment (Part 1) – Jennifer Nolan C16517636**

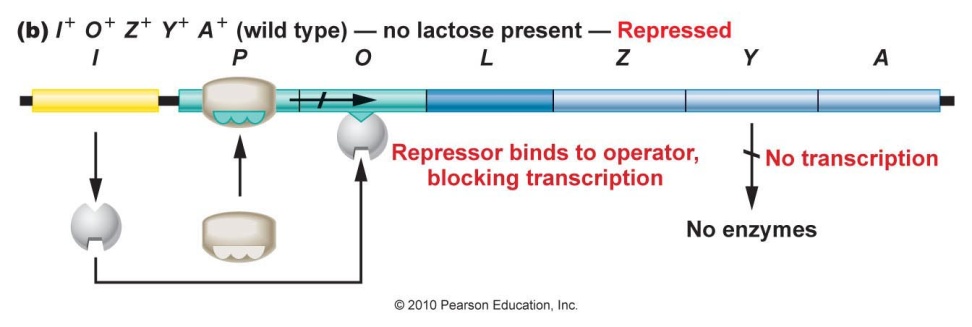
The following report contains an analysis of the lac operon and how to find true, prokaryotic, open reading frames.



*Figure 1: Structure of the Lac operon*

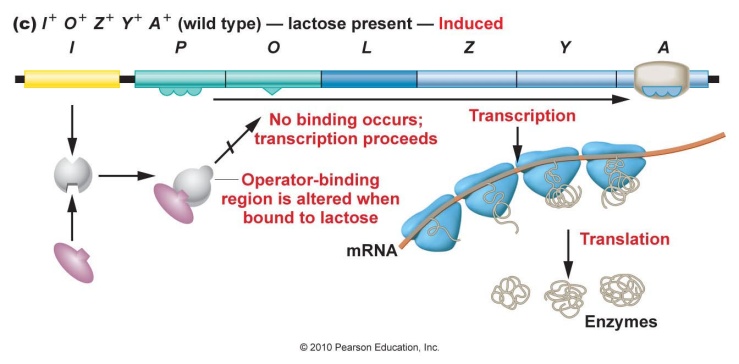
An operon is a gene regulatory system in prokaryotic cells. For E.Coli to metabolise sugar, like lactose, a gene system, the lac operon, needs to produce three enzymes that allow for lactose to be used by the bacteria. The E.Coli lac operon DNA contains the following properties, as shown in Figure 1 above, to regulate the production of the three enzymes: 1 repressor protein that decides if RNA polymerase can perform transcription of the genes or not, a promoter region where the RNA polymerase binds, an operator region that acts as a regulatory region and 3 genes that when transcribed turn into the three enzymes. RNA polymerase binds to the promoter region. The purpose of the repressor protein is to bind to the DNA at the operator region and block the RNA polymerase from moving down the DNA strand and implementing transcription of the genes. Essentially the repressor protein is responsible for whether transcription of the DNA takes places.

The presence of lactose and glucose regulates the transcription of the lac operon. RNA polymerase binds to the promoter region of the gene depending on the levels of glucose or lactose available.



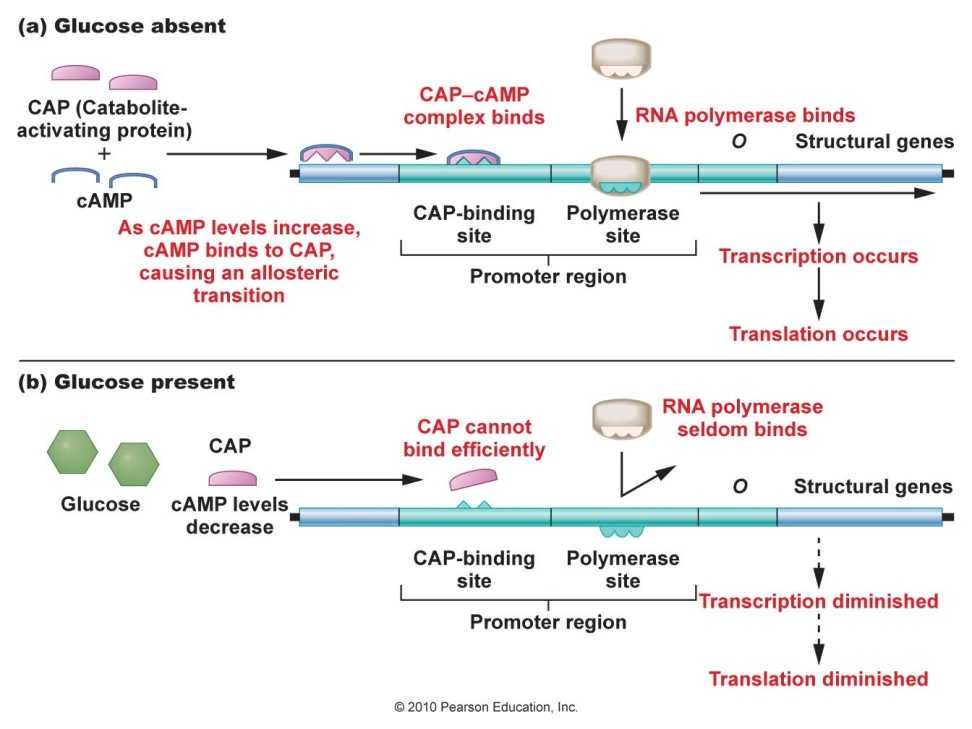
*Figure 2: Diagram showing process for when Lactose is absent*

If lactose is not available, as shown in Figure 2 above, then transcription on the lac operon cannot take place. When the levels of lactose are low, the repressor protein binds to the operator region of the gene. This means that the repressor protein blocks the bound RNA polymerase from moving down the DNA, therefore blocking transcription from taking place.



*Figure 3: Diagram showing process for when Lactose is present*

If lactose is present the repressor protein loses the ability to bind to the gene, as shown Figure 3 above. When lactose is present it binds to the repressor protein and alters the repressor proteins shape. This means that the repressor protein is unable to bind to the lac operon gene any longer. This causes the repressor protein to fall off the operator region and clears the way for RNA polymerase , bound to the promoter region, to transcribe the DNA.



*Figure 4: Diagram showing process for when Glucose is (a). absent and (b). present*

When glucose is present a similar process takes place on the lac operon. When glucose is present, as shown in Figure 4(b) above, the CAP, catabolite activator protein, is altered and therefore cannot bind to the gene efficiently. This prevents RNA polymerase from binding to the promoter region, as RNA polymerase needs CAP to be bound as well for it to transcribe DNA in the presence of glucose. Without the RNA polymerase, transcription of the lac operon cannot take place until the glucose is used up by the bacteria. When glucose is absent, as shown in Figure 4(a) above, the CAP binds to a specific area on the gene, a part of the promoter region known as the CAP binding site. When the CAP is bound, RNA polymerase is also able to bind efficiently to the promoter region of the gene. Once both CAP and RNA polymerase are bound transcription of the gene can begin.

When glucose is scarce other sugars can be used as sources of fuel, for example lactose as explained previously. To use these alternate sugars, alternate genes must be transcribed. This type of process can be seen with the lac operon which is turned on in the presence of lactose and absence of glucose.

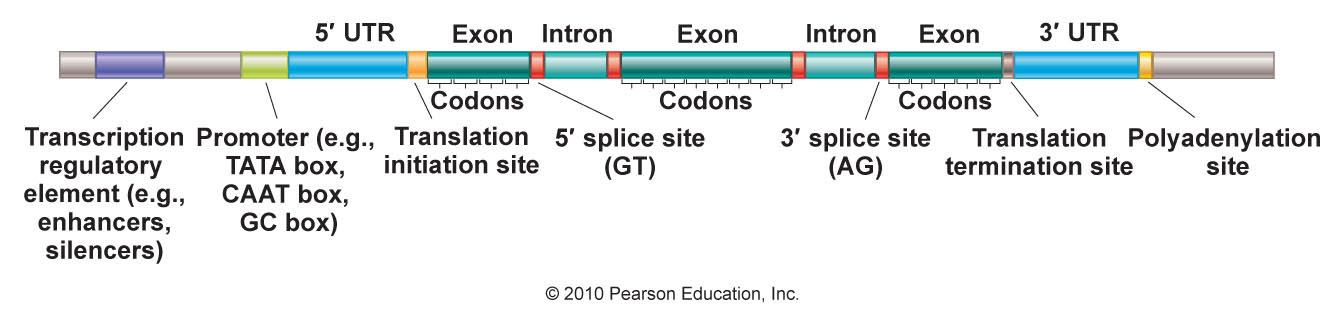
The lac operon, as described previously, contains Open Reading Frames (ORFs) within its DNA sequence. The following process is used to find all potential Open Reading Frames (ORF) in a prokaryotic DNA sequence, like the lac operon.

An Open Reading Frame (ORF) is a sequence of DNA that could contain the potential coding sequence (CDS) of a gene. An ORF in prokaryotes can be directly translated into an amino acid sequence. An ORF must contain a start codon eventually followed by a stop codon. With the lac operon there are 3 genes or CDS’s all in close proximity, making the protein CDS contiguous.

The following process can be used to find all the potential reading frames in a DNA sequence. Firstly, the document, usually a fasta file, containing the DNA sequence is opened. The descriptor line of the file is ignored. Next, the DNA sequence in the file is converted into a contiguous DNA sequence, by removing the line separators. The sequence is then split into three reading frames, base 1, 2 and 3. The DNA sequence then needs to be translated into an amino acid sequence using the “DNA<->AA” translation table. Convert the first reading frame from a DNA sequence into an amino acid sequence using the translation table. Then shift one position to the right in the DNA sequence, for the 2nd reading frame, and translate it into an amino acid sequence. Repeat the previous process for the third reading frame. Get the reverse compliment of the original DNA sequence, split it into three reading frames and convert into the amino acid sequence using the same process as mentioned previously. Mark all the stop codons, denoted with a ‘TAA’, ‘TGA’ or ‘TAG’, and start codons, denoted with an ‘ATG’, in the DNA sequence. Also, mark all the stop amino acids, denoted with a ‘\*’ and start amino acids, denoted with an ‘M’, in the amino acid sequence. Lastly, look for all potential ORFs, a sequence beginning with a start codon or amino acid followed by a stop codon or amino acid. If there is no ORFs found, then there are no ORFs in that reading frame.

If there is a potential ORF it must start with a start codon or amino acid and stop with a stop codon or amino acid. However, there can be false positives with ORFs. To avoid these the following method can be used. Determine the length of the ORF and if it is less than 20, for an amino acid, then eliminate it as a false positive. Also, something to remember about ORFs is that the DNA codon “ATG” or the amino acid “M” can also exist as part of the gene sequence, meaning that is doesn’t have to represent the start codon.

As mentioned above, the amino acid sequence of the prokaryotic gene can be derived directly using the “DNA<->AA” translation table. However, the same technique cannot be used for determining the amino acid sequence for eukaryotic gene coding sequences (CDS).



*Figure 5: Diagram showing the structure of a Eukaryotic gene*

The ORF in prokaryotes can be manually translated directly into an amino acid sequence, using the “DNA<->AA” translation table, because the CDS in a prokaryotic cell is contiguous. The same technique cannot be used for eukaryotic cells. This is because of the following reasons. Firstly, eukaryotic genes contain more than one promoter region, with the promoter closest to the coding sequence being referred to as the basal promoter. Secondly, as shown above in Figure 5, the protein coding sequence of eukaryotic DNA consists of exons, regions that are translatable, and introns, regions that are not translatable. The introns would need to be removed to determine the true amino acid sequence. Unlike prokaryotic DNA where the coding sequence is contiguous, eukaryotic gene regulation is more complex. Therefore, as the eukaryotic DNA coding sequence is not contiguous, the “DNA<->AA” translation table cannot be used to directly translate the DNA sequences of eukaryotic cells. This means that a different technique is required when converting eukaryotic genes to amino acids.