AN ANALYSIS OF THE TRP OPERON AND HOW TO FIND TRUE, PROKARYOTIC, OPEN READING FRAMES

Saira J. Barlas

BIOINFORMATICS

DT228/4

Denis Manley

Introduction

In this report I'll be discussing the functionality of the five genes in the trp operon as well as the regulation of trp operon genes in the presence of various levels of tryptophan. I'll also be looking into Open Reading Frames (ORF) and how realistic ORFs can be found in a prokaryotic DNA sequence. Lastly I'll be stating as to why "DNA -> AA" is not a sufficient way of determining amino acid sequence for eukaryotic sequences.

1. Describe the functionality of the proteins of the Trp Operon.

The trp operon was first discovered in bacteria, particularly E. coli. It is involved in the biosynthesis of tryptophan and is mostly active in prokaryotes. An operon consists of a set of genes which work together in the production of a protein. The trp operon consists of 5 genes, each of which produce a certain enzyme/protein which aids in the production of tryptophan. Figure 1 shows the 5 genes which are found on the trp operon. They are trpE, trpD, trpC, trpB, and trpA respectively.

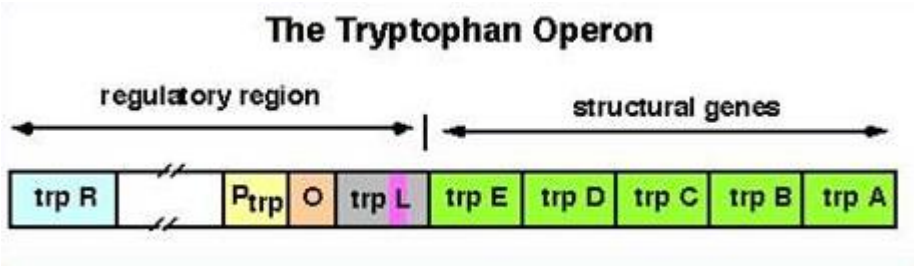

Figure 1. Structure of the Tryptophan Operon

In order for tryptophan biosynthesis to occur, tryptophan requires three enzymes to be active:

1. trpE gene codes for the enzyme anthranilate synthetase which is complexed with phosphoribosylanthranilate transferase which trpD gene is responsible for.
2. trpC gene codes for phosphoribosyl-anthranilate isomerase.
3. trpA and trpB both code for tryptophan synthetase which involves converting chorismate to tryptophan.

As the above 5 genes are located next to each other, they are only served by 1 promoter. A promoter is the region on the operon where the RNA polymerase binds and transcription occurs. Since the genes have only 1 promoter, they are expressed together and require all 3 enzymes to be present in order for tryptophan to be produced. Figure 1 shows 'Ptrp' as the promoter of the operon. Since all these genes are "switched on" or "off" as a group, an operator region on the operon exists which allows RNA polymerase to access the genes. This region is called the operator and is represented in Figure 1 as 'O'.

The whole mechanism controlling the genes is called the trp Operon. RNA polymerase binds to the promoter region and "switches on" the operator in order for transcription of the three enzymes to occur. The genes then produce enzymes in a metabolic pathway which breakdown chorismate to tryptophan amino acid. When the tryptophan levels reach a certain threshold, the operator region is "turned off" and transcription of the enzymes by the 5 genes halts.

2. Explain how the "transcription" of the Trp Operon genes are regulated in the presence of tryptophan and not in the presence of tryptophan.

Attenuation is a way for reducing the function of the trp operon when tryptophan levels are high. It blocks the completion of transcription altogether. RNA polymerase halts the transcribing of the trp operon instantly when detected tryptophan levels are high. Since transcription results in a single mRNA strand being formed from the genes of the trp operon, halting transcription results in a short, stubby mRNA which does not have the ability to encode enzymes to produce tryptophan.

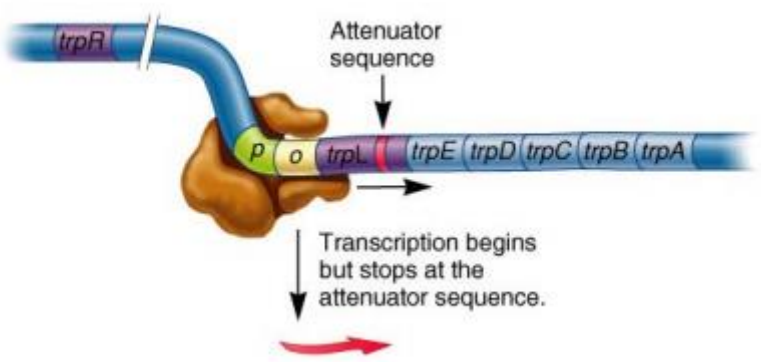

Figure 2. High Tryptophan levels result in attenuation

Attenuation

TrpL, also known as the leader region (as seen in Figure 1 and Figure 2), encodes a short polypeptide and contains an attenuator sequence. When the attenuator is transcribed into mRNA it has self-complimentary sections. Once RNA polymerase starts transcribing the operon, a ribosome can attach itself to the transcript and start translating the leader region. The polypeptide encoded by the leader is short and includes 2 tryptophan residues.

If there is a lot of tryptophan, the ribosome will not have to wait long for tryptophan carrying tRNA and will finish up the leader polypeptide. If the tryptophan levels are low, the ribosome will stall and will be very slow in finishing the polypeptide.

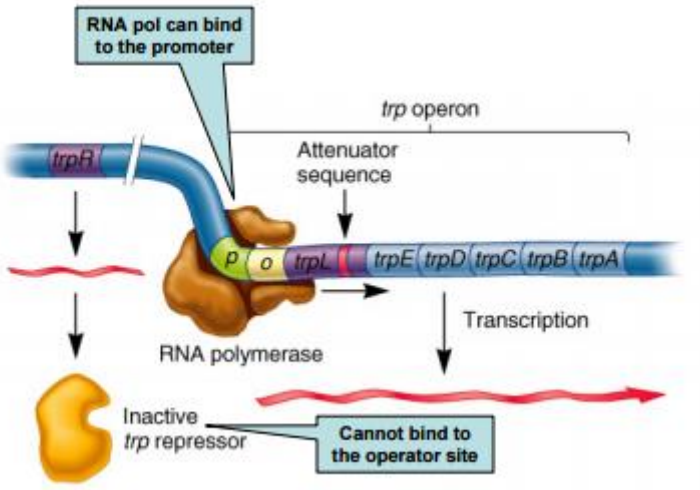

Figure 3. Low Tryptophan levels result in the transcription of the operon

Ribosomes should translate the leader quickly; as the attenuator region comes after the leader, which can be utilised to form different hairpin structures. One structure includes a transcription termination signal and the other does not end termination. If the ribosome translates slowly when there are low tryptophan levels, no termination hairpin will be formed and it will continue transcription. If the ribosome translates quickly when there are high tryptophan levels, it will complete the translation of the leader peptide and will detach itself from the mRNA. This results in the terminator hairpin being formed which causes mRNA to end transcription.

In conclusion, ribosomes will move faster across the leader region when there's a higher level of tryptophan hence forming a terminator hairpin quickly and ending transcription. On the other hand, when tryptophan levels are low, ribosome will stall and this will result in a non-terminator hairpin being formed which will lead to the continuation of transcription.

3. Describe, using examples, how you attempt to find all "potential" realistic Open Reading Frames (ORF) in a prokaryotic DNA sequence such as the trp operon (use the amino acid sequence).

Tryptophan can be found similarly in the following ways:

1. Consider the following sequence:
CGCTACGTCTTACGCTGGAGCTCTCATGGATCGGTTCGGTAGGGCTCGATCACATCGCTAGCCAT

2. Divide the sequence into 6 different reading frames (+1, +2, +3, -1, -2 and -3). The first reading frame is obtained by considering the sequence as codons.

FRAME +1: CGC TAC GTC TTA CGC TGG AGC TCT CAT GGA TCG GTT CGG TAG GGC TCG ATC ACA TCG CTA GCC AT

The second reading frame is formed after leaving the first nucleotide and then grouping the sequence into words of 3 nucleotides.

FRAME +2: C GCT ACG TCT TAC GCT GGA GCT CTC ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG CCA T

The third reading frame is formed after leaving the first 2 nucleotides and then grouping the sequence into words of 3 nucleotides

FRAME +3: CG CTA CGT CTT ACG CTG GAG CTC TCA TGG ATC GGT TCG GTA GGG CTC GAT CAC ATC GCT AGC CAT

The other 3 reading frames can be found only after finding the reverse complement.
Complement:
GCGATGCAGAATGCGACCTCGAGAGTACCTAGCCAAGCCATCCCGAGCTAGTGTAGCGATCGGTA
Reverse complement:
ATGGCTAGCGATGTGATCGAGCCCTACCGAACCGATCCATGAGAGCTCCAGCGTAAGACGTAGCG

Now same process as that of +1, +2 and +3 strands is repeated for -1, -2 and -3 strands with reverse complement sequence
FRAME -1: ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA GAG CTC CAG CGT AAG ACG TAG CG
FRAME -2: A TGG CTA GCG ATG TGA TCG AGC CCT ACC GAA CCG ATC CAT GAG AGC TCC AGC GTA AGA CGT AGC G
FRAME -3: AT GGC TAG CGA TGT GAT CGA GCC CTA CCG AAC CGA TCC ATG AGA GCT CCA GCG TAA GAC GTA GCG

3. Now mark the start codon and stop codons in the reading frames

FRAME +1: CGC TAC GTC TTA CGC TGG AGC TCT CAT GGA TCG GTT CGG TAG GGC TCG ATC ACA TCG CTA GCC AT
FRAME +2: C GCT ACG TCT TAC GCT GGA GCT CTC ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG CCA T

FRAME +3: CG CTA CGT CTT ACG CTG GAG CTC TCA TGG ATC GGT TCG GTA GGG CTC GAT CAC ATC GCT AGC CAT
FRAME -1: ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA GAG CTC CAG CGT AAG ACG TAG CG
FRAME -2: A TGG CTA GCG ATG TGA TCG AGC CCT ACC GAA CCG ATC CAT GAG AGC TCC AGC GTA AGA CGT AGC G
FRAME -3: AT GGC TAG CGA TGT GAT CGA GCC CTA CCG AAC CGA TCC ATG AGA GCT CCA GCG TAA GAC GTA GCG

4. Identify the open reading frame (ORF) - sequence stretch beginning with a start codon and ending in a stop codon.

FRAME +2: ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG
FRAME -1: ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA
FRAME -3: ATG AGA GCT CCA GCG TAA

5. Based on the amino acid table the peptide sequence is found

First Nucleotide		Second Nucleotide: U code	U Amino acid	C code	C Amino acid	A code	A Amino acid	G code	G Amino acid	Third Nucleotide
	U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
		UUC		UCC		UAC		UGC		C
		UUA	leu	UCA		UAA	STOP	UGA	STOP	A
		UUG		UCG		UAG	STOP	UGG	trp	G
	C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
		CUC		CCC		CAA		CGC		C
		CUA		CCA		CAC	gln	CGA		A
		CUG		CCG		CAG		CGG		G
	A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
		AUC		ACC		AAC		AGC		C
		AUA		ACA		AAA	lys	AGA	arg	A
		AUG	met	ACG		AAG		AGG		G
	G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
		GUC		GCC		GAC		GGC		C
		GUA		GCA		GAA	glu	GGA		A
		GUG		GCG		GAG		GGG		G

Figure 4. Amino Acid Table

FRAME +2: ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG
met asp arg phe gly arg ala arg ser his arg stop

FRAME -1: ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA
met ala ser asp val ile glu pro tyr arg thr asp pro stop

FRAME -3: ATG AGA GCT CCA GCG TAA
met arg ala pro ala stop

By analysing the ORF we can predict the possible amino acids that are produced during the translation process.

4. The sequence of a prokaryotic gene can be translated manually by utilising the "DNA -> AA" translation table. Discuss, using suitable examples, why this is not an efficient technique in determining the amino acid sequence for eukaryotic gene coding sequences.

The region of the nucleotide sequences from the start codon (ATG) to the stop codon is called the Open Reading frame.

Gene finding in organism specifically prokaryotes starts form searching for an open reading frame (ORF). An ORF is a sequence of DNA that usually starts with the start codon "ATG" and ends with any of the three termination codons (TAA, TAG, TGA). Depending on the starting point, there are six possible ways (three on forward strand and three on complementary strand) of translating any nucleotide sequence into amino acid sequence according to the genetic code. These are called reading frames.

The eukaryotic gene finding is altogether a different task as the eukaryotic genes are not continuous and interrupted by intervening noncoding sequences called 'introns'. In addition, the organisation of genetic information in eukaryotes and prokaryotes is different.

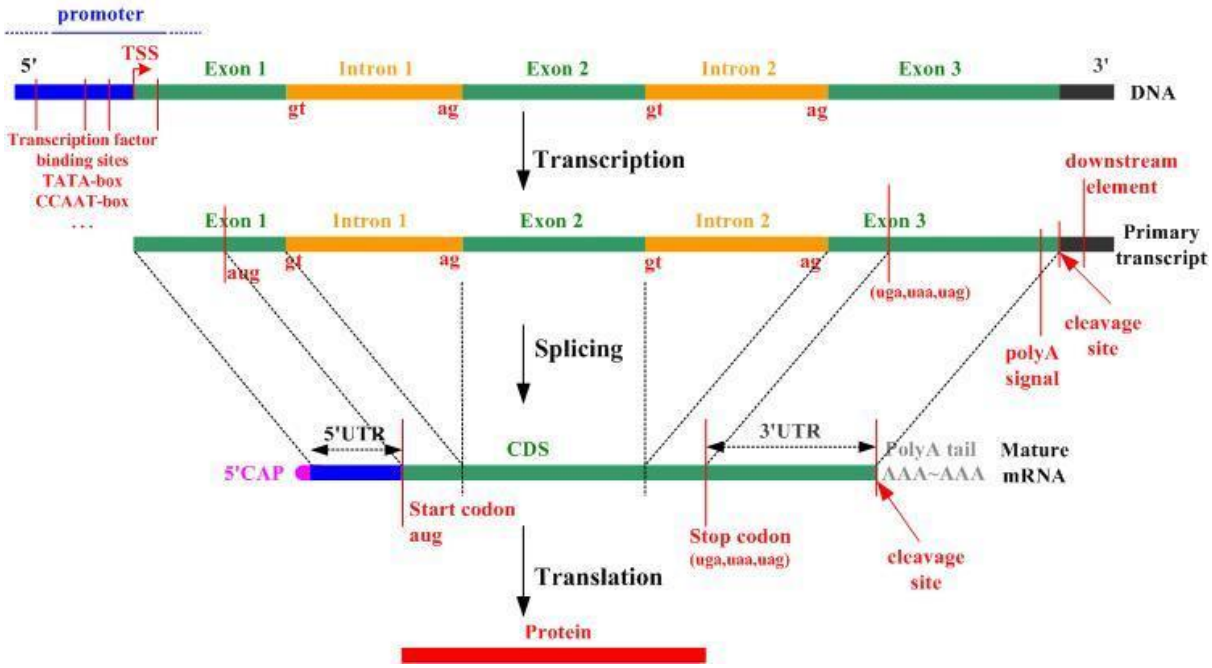

Figure 5. Exons and Introns in Eukaryotes

The Coding Sequence (CDS) is the actual region of DNA that is translated to form proteins. While the ORF may contain introns as well, the CDS refers to those nucleotides (concatenated exons) that can be divided into codons which are actually translated into amino acids by the ribosomal translation machinery.

Mainly, CDS means only that the sequence is known to be transcribed and, therefore, it is coding for something, neither gene nor protein has to be known. Any full mRNA sequence (obtained from cDNA sequencing) will have a full coding sequence. ORF is usually predicted based on DNA sequence and not proven to be transcribed.

Conclusion

Tryptophan is an amino acid widely produced in prokaryotes and can be easily regulated using attenuation. It contains 5 genes which produce 3 enzymes to break down chorismate into tryptophan. Attenuation is a technique which regulates transcription depending on the tryptophan levels present. ORF gives us the ability to find open reading frames in prokaryotic cells yet due to the presence of introns and exons in eukaryotic cells, ORF would not be the correct method.

References

1. Trp operon - mirror site (no date) Available at: https://www.csun.edu/~hcbio027/biotechnology/lec4a/trp.html
2. Biology and Edition, the 5 (no date) The Operon. Available at: http://www.hammiverse.com/lectures/18/4.html
3. Tryptophan Operon (no date) Available at: http://mol-biol4masters.masters.grkraj.org/html/Gene_Expression_I5B-Tryptophan_Operon.htm
4. Finding ORF of a given sequence (theory): Bioinformatics virtual lab I: Biotechnology and biomedical engineering: Amrita Vishwa Vidyapeetham virtual lab (no date) Available at: http://vlab.amrita.edu/?sub=3&brch=273&sim=1432&cnt=1
5. What's the difference between Cds and Orf? (no date) Available at: https://www.biostars.org/p/47022/
6. The Trp Operon, Khan academy (no date) Available at: https://www.khanacademy.org/science/biology/gene-regulation/gene-regulation-in-bacteria/a/the-trp-operon
7. Gene Expression and Regulation (no date) Available at: http://www.blackwellpublishing.com/trun/pdfs/chapter12.pdf
8. Prokaryotic Gene Expression (no date) Available at: http://www.csus.edu/indiv/p/peavyt/184/9lect184%20f07.pdf