


Bacterial Genomes

Circular genome,
Polycistronic mRNA,
operons

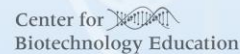


Center for  Biotechnology Education

Hello, I'm Bob Lessick at Johns Hopkins. We will examine the bacterial genome in this short lecture. Be sure to focus on the polycistronic nature of some transcripts.

Objectives

- Describe circular structure of most bacterial genomes
- Define polycistronic
- Define untranslated region
- Describe lactose operon
- Find CDS regions on operon in
 - NCBI sequence record
 - Bacterial gene output



Unlike eukaryotes, most bacteria have a single chromosome and in most but not all cases, that chromosome is circular.

Polycistronic is an important term, meaning that there can be more than one protein-coding region on an mRNA transcript.

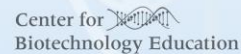
Examining that transcript more closely, you will find two untranslated regions that are not protein coding. Those come at either end of the sequence.

We'll look at the lactose operon, made famous by Jacob and Monod in the 1960s.

And then to bring it back to bioinformatics, we'll look to identify operons on an NCBI sequence and on a gene prediction.

Circular Genome

- Bacterial genomes usually smaller than eukaryotic
 - *Escherichia coli* K12: 4.6 Mb
 - *Saccharomyces cerevisiae*: 12.1 Mb
 - *Homo sapiens*: 3235 Mb (3.2 Gb)
- Bacterial genomes usually circular double-stranded DNA
 - Most eukaryotic chromosomes linear



First. Let's be aware that bacteria tend to have much smaller genomes than eukaryotes. *Escherichia coli* or what we call E coli has a 4.6 megabase genome—that's 4.6 million base pairs.

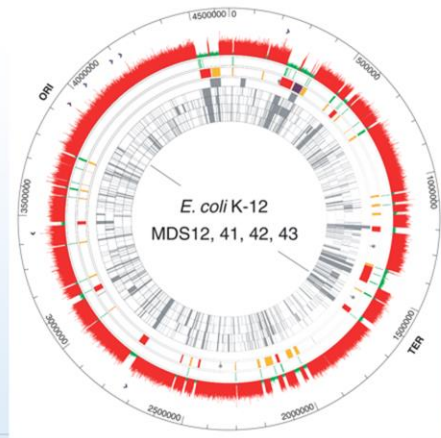
One of the simplest eukaryotes, budding yeast, has a genome of 12.1 megabases, almost three times the size of E coli. And that's a VERY small eukaryotic genome.

The human genome is 3.2 gigabases—that's 3.3 BILLION base pairs! And that's not even the largest genome—some plants are much bigger in genome size than humans.

Bacteria usually have a single circular chromosome—there are exceptions. Many also have extrachromosomal plasmids. Those are small usually circular pieces of additional DNA.

E. coli Genome

- Note position 0 at top of circle
 - Base pairs numbered in clockwise fashion
 - Position 0 represents origin of replication
- + strand genes run clockwise
- - (complementary) strand genes run counterclockwise




Center for  Biotechnology Education

Image from *Science* 312:1044-1046

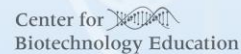
Here's a site map of the *E. coli* genome—look at the top where you see a 0. The 0 position represents what's called the origin of replication. Then the numbers increase clockwise and there are about 4,600,000 or so.

Now remember, there are still two strands! What we call the plus strand is arranged (in this diagram) clockwise.

However, there are genes on the minus or complementary strand and they run in the other direction.

Polycistronic mRNA

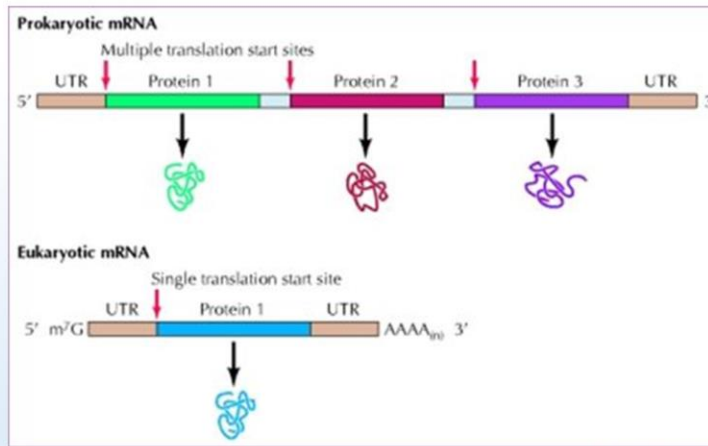
- Eukaryotic transcripts usually have only one coding region
 - One mRNA makes one protein
 - Actually many molecules of same single protein
- Prokaryotic transcripts often have more than one coding region
 - One mRNA make multiple proteins
 - Actually many molecules of each protein



So I have mentioned polycistronic. In eukaryotes, the final spliced mRNA usually only has one protein coding region. That means that only one protein is derived from that mRNA. Actually many copies of that same single protein.

In bacteria, while mRNA splicing usually does not occur, there are often, but not always, more than one protein coding region in a prokaryotic mRNA. If, say, an mRNA has five protein coding regions, then translation makes many molecules of each of those five proteins.

Comparison



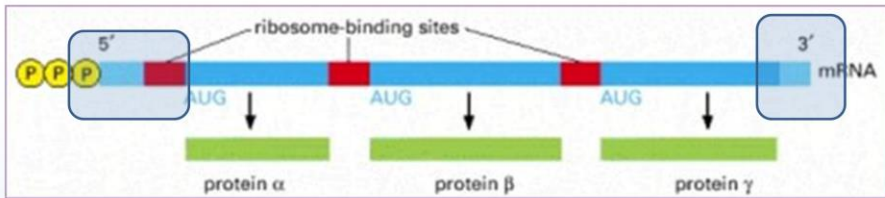
JOHNS HOPKINS UNIVERSITY

Center for Biotechnology Education

--image from Cooper (2000). *The Cell: A Molecular Approach*. Sinauer Associates, Inc.

And here is what it looks like. The diagram is a bit fuzzy but it displays the concept well. The prokaryote is at the top and the eukaryote is at the bottom. Prokaryotes tend to do it this way for efficiency.

UTRs



- Untranslated region before first start codon
– 5' UTR
- Another after last stop codon
– 3' UTR

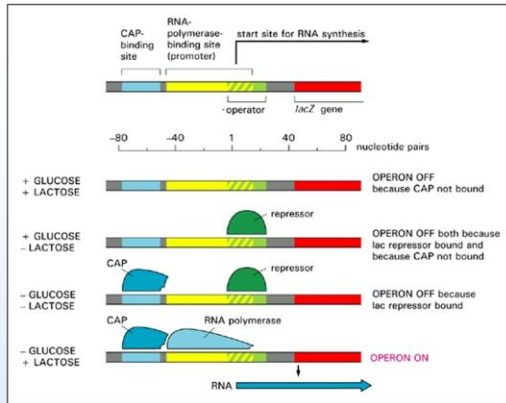
Here's a similar diagram but you can see the untranslated regions or UTRs at either end.

The one at the left of the mRNA is called the 5 prime UTR. The 5' refers to the chemistry of which carbon that phosphate attaches to on the sugar. Or to put it another way, protein sequences go from N-terminus to C-terminus. DNA/RNA sequences go from 5' to 3' and both involve chemistry. The 5' UTR is everything to the left of the first AUG start codon.

The 3' UTR is on the right.

It's everything that is untranslated after the stop codon. Both UTRs are found in eukaryotes as well. So from left to right on a bacterial mRNA, you have a 5' UTR, one or more coding regions, then a 3' UTR.

Lac operon



- Glucose preferred energy source
- Lactose used in absence of glucose
- +1 = transcription start site

– -40 = 40 bases “upstream”

– +80 = 80 bases “downstream”



Center for Biotechnology Education

Here is the well studied lactose operon. Bacteria prefer glucose as an energy source.

But if there is no glucose to be found, they’ll deal with lactose. So they have to find a way to turn on the genes involved in lactose synthesis. This describes some complex regulation and it shows what is called the regulatory region of the gene, or in this case, the operon. That’s the term for the multiple genes on a single mRNA.

Look primarily at the numbers below the first graphic. The +1 position is the transcription start site. That is NOT the start codon---it is the start of the mRNA—transcription, not translation! The mRNA is to the right of that +1 position. Or think of it this way---+1 refers to the beginning of the 5’UTR of the mRNA. The region of DNA to the left of +1 is called the regulatory region. Here is where proteins bind that help control the level of transcription. A need for lactose digestion ultimately determines whether transcription of the lactose enzymes is on or off. The direction to the left, or to the 5’ side is frequently called upstream. The -35 position is 35 bases upstream of the transcription start site.

And to the right or to the 3’ side is called downstream.

Gene prediction output

Four mRNAs: First has three CDSs

```

Prediction of potential genes in microbial genomes
Time: Tue Jan 1 00:00:00 2005
Seq name: gb|AE016830.1|:30001-38800 Enterococcus faecalis V583, complete genome
Length of sequence - 8800 bp
Number of predicted genes - 7
Number of transcription units - 4, operons - 2

```

N	Tu/Op	Conserved pairs (N/Pv)	S	Start	End	Score
1	Op 1	.	+	CDS	3 - 101	147
2	Op 2	.	+	CDS	125 - 1330	852
3	Op 3	.	+	CDS	1430 - 1810	449
4	Tu 1	.	+	CDS	2001 - 4592	2396
5	Tu 1	.	-	CDS	4629 - 7244	2114
6	Op 1	.	-	CDS	7788 - 8297	670
7	Op 2	.	-	CDS	8305 - 8772	525


```

Predicted protein(s):
>GENE 1 3 101 147 32 aa, chain +
ALGVVHLDAGNVQVIIGTKVTVTRNQLEMILG
>GENE 2 125 - 1330 852 401 aa, chain +
VCPDFAIIPPCTVCTQADPEFDEPCADITDEPTIDPDPNADPAVTATYCPDTPMURVQCV

```

Next two are single transcription units

Final mRNA has 2 CDSs

Center for  Biotechnology Education

Here is an output from FGENESB at softberry.com. This is explained in the separate video on Bacterial Gene Prediction. Note that there are four transcription units meaning that four mRNAs come out of this region of genomic DNA. Of those four transcription units, two are operons, which each have more than one CDS on that mRNA.

The first mRNA has three CDS regions...Op1, Op2 and Op3.


Next are two single-CDS transcription units. Tu1 and Tu1. Why not Tu2? That would imply a second CDS on an mRNA...that's how the numbering works.

The last mRNA has 2 CDS regions, Op1 and Op2. Hopefully that explained the notation.

Summary

- Bacterial genomes usually circular
- Prokaryotic genes are somewhat simpler than eukaryotic genes
 - No splicing of mRNA
- Prokaryotic genes can be polycistronic
 - More than one CDS per mRNA
- Operons allow similar genes to be transcribed on same mRNA



Center for  Biotechnology Education

To summarize, bacteria usually have circular genomes. Eukaryotes tend to have linear chromosomes—and more than one!

With prokaryotic genes, you don't have to deal with the splicing issue—usually.

However, one complication is that multiple coding regions often occur in a bacterial mRNA transcript.

Finally, why do it this way. Bacteria waste little space in their genome – efficiency allows for genes in a certain pathway to be transcribed in one mRNA molecule so that all necessary enzymes can be translated in one shot. That's it for now!