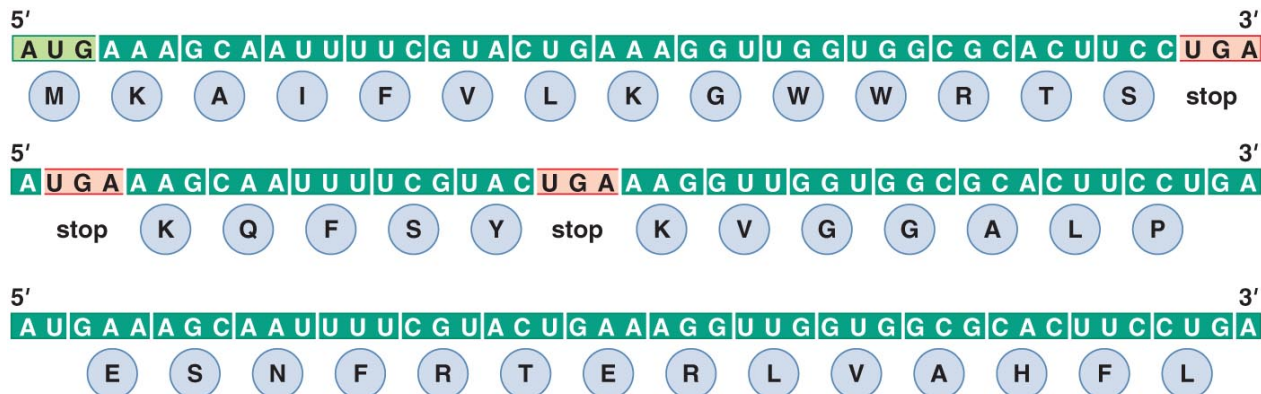


## Prokaryotic Genes, Promoters

***This is supplemental reading for the video lecture on Bacterial Genomes.***

## Reading frames

Single-stranded DNA has three possible translatable amino acid sequences, while dsDNA has six reading frames. The other strand, in the other direction could have genes as well.



Copyright © 2008 Pearson Education, Inc., publishing as Pearson Benjamin Cummings

## Open reading frames (ORFs)

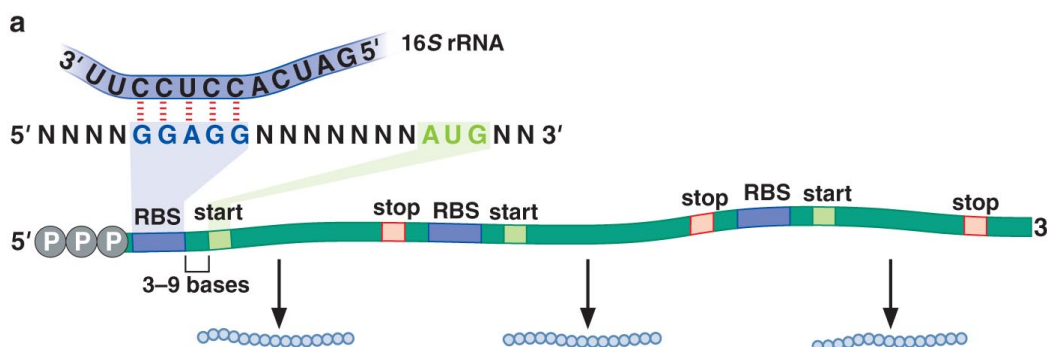
An **open reading frame** is a start codon followed by stop codon, and the start and stop MUST be in the same reading frame. Bacterial mRNA is **polycistronic**. There are more than one ORF per mRNA. In eukaryotes, mRNA is **monocistronic**, meaning one ORF per mRNA. The start and stop codons might NOT be in the same reading frame in eukaryotic pre-mRNA. After splicing, the start and stop codons should be in the same reading frame.

**Discussion topic: What are the advantages and disadvantages of polycistronic or monocistronic mRNA? How does that fit in the context of the organization of the bacterial genome vs. the eukaryotic genome?**

**Discussion topic: Some viral genes rely on frameshifts that alter the stop codon. Try to find examples.**

## Shine-Dalgarno sequence

The Shine-Dalgarno sequence is the bacterial ribosome binding site on the mRNA. The sequence allows for the recruitment of the ribosome to the translation start site. Therefore, the Shine-Dalgarno sequence is just upstream from the start codon.



Copyright © 2008 Pearson Education, Inc., publishing as Pearson Benjamin Cummings

## Prokaryotic gene structure

We know that many bacterial genes are **polycistronic**. One promoter controls the transcription of an mRNA with multiple open reading frames. What does the promoter sequence look like? What about other regulatory sequences? RNA polymerase binds a region near the transcription start site. Sequences at -10 and -35 are likely to be binding areas for **sigma factor**. The numbers are relative to the transcription start site (TSS), NOT the start codon! The  $\alpha$ -subunit binds an AT-rich region near -40 to -60, called the **UP element**. The **UP element** is found in strong promoters.

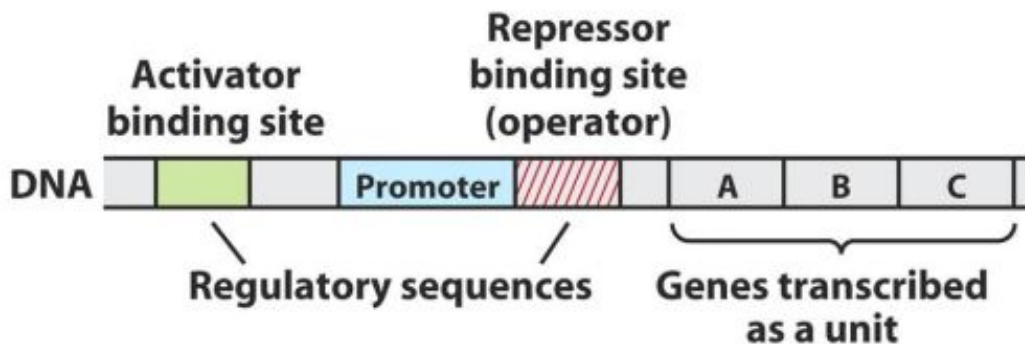
**Discussion topic: How might UP-element presence affect promoter prediction?**

## Sigma factor

The subunit of RNA polymerase that binds the promoter is the sigma subunit or sigma factor. Most of the time, bacteria employ  $\sigma^{70}$ . The sigma factor can change under heat stress. The  $\sigma^{70}$  is part of RNA polymerase under 'normal conditions.' In heat shock conditions,  $\sigma^{32}$  replaces  $\sigma^{70}$ . The  $\sigma^{32}$  subunit binds to heat-shock promoters, and gene expression of 'heat shock proteins' is increased.

## Operons

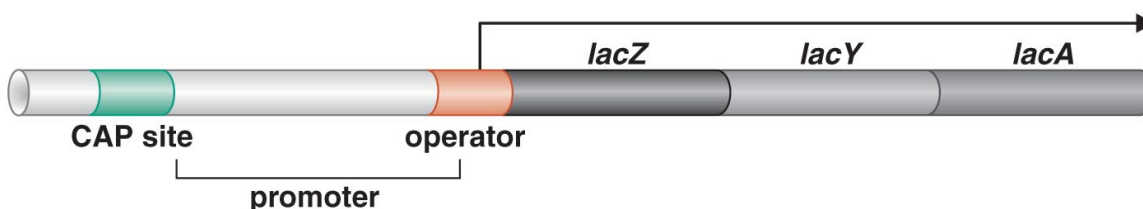
Many bacterial mRNAs are polycistronic. One promoter controls the synthesis of an mRNA. One promoter thus controls the synthesis of every gene on that mRNA. Bacteria tend to group related genes onto one mRNA. For example, lactose digesting genes are on one mRNA. One promoter can produce proteins necessary to digest lactose, and one mRNA can make all the necessary enzymes. A DNA that includes all genes controlled by one promoter is called an **operon**.



--image from *Lehninger Principles of Biochemistry*, Nelson & Cox.

## Negative regulation

The lac operon genes are: lacZ ( $\beta$ -galactosidase), lacY (galactoside permease), lacA (thiogalactoside transacetylase). Lactose (actually a byproduct of lactose) binds the repressor, the repressor dissociates and lactose genes are transcribed. In other words, when lactose is present, you want to turn on the genes to digest lactose. Okay, you could think of that as an on switch (positive regulation), but it's really preventing the off switch from functioning (lactose inhibits negative regulation).



Copyright © 2008 Pearson Education, Inc., publishing as Pearson Benjamin Cummings.

--image from *Molecular Biology of the Gene*, Watson et al.

## CAP, Lac repressor

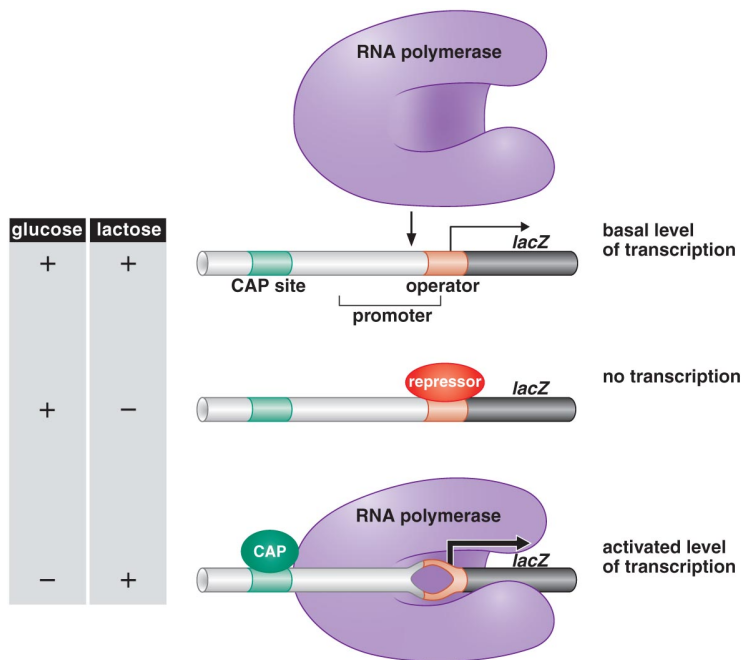
The CAP (catabolite activator protein) is sometimes called CRP (cAMP receptor protein). It is the activator that binds CAP site. The lac repressor binds the operator and blocks RNA polymerase binding.

### Lac operon summary

**Absence of lactose:** Lac operon genes off (repressor bound to operator)

**Presence of lactose NOT glucose:** Lac operon genes on (repressor gone, CAP bound to activator)

**Presence of lactose AND glucose:** Basal level - (repressor gone but no CAP bound to activator). Glucose is the preferred energy source

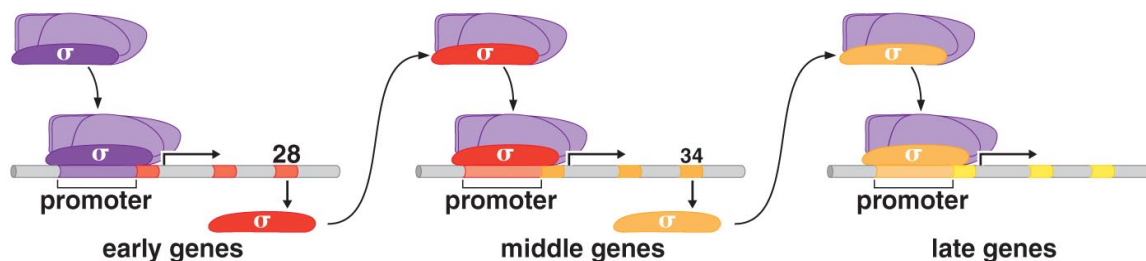


Copyright © 2008 Pearson Education, Inc., publishing as Pearson Benjamin Cummings.

--image from Molecular Biology of the Gene, Watson et al.

## Sigma factors

$\sigma^{70}$  binds most promoters.  $\sigma^{32}$  binds heat shock promoters.  $\sigma^{54}$  binds nitrogen metabolism promoters. Some viruses make their own sigma factor(s), including phage SPO1 (*Bacillus subtilis*). Note how the virus hijacks the sigma 70 polymerase for the early genes. The early gene transcription/translation includes a sigma factor that turns on the middle genes. Then another viral sigma factor turns on the late genes. The genes are controlled by time of transcription.



Copyright © 2008 Pearson Education, Inc., publishing as Pearson Benjamin Cummings.

--image from Molecular Biology of the Gene, Watson et al.

## Finding promoters

In designing a program to find promoters, often training sets of promoters known to bind  $\sigma^{70}$  are used to

allow for the finding of unknown DNA that may also bind  $\sigma^{70}$ , presumably locating promoters. Other training sets may be used to find promoters that bind alternate sigma factors.

**Discussion topic: Why are false positives more of a problem for promoter finding than for gene finding?**

### Structures of promoters

Not all promoters fit the standard model. A 2014 paper by Meysman et al. (*PLoS One* 9:e88717) describes the varieties of structures in bacterial promoters. They found distinct classes of promoter structure (GC content, curvature). They hypothesized that the seven different sigma factors of *E. coli* would match well with the different promoter structures, however, those results proved mostly insignificant.

### Strategies

#### Comparative genomics

By comparing intergenic regions of related species, promoters can be isolated based on the assumption that base substitutions have no constraint outside the promoter, but substitutions are limited within the promoter, which must maintain its ability to bind sigma factor. The challenge is to find genomes at the appropriate evolutionary distance. If species are too closely related, then the nonpromoter regions have not fully diverged and false positives are a problem. If the species are too far apart on the evolutionary scale, then only well-conserved promoters are detected. The Footy program was described in 2007 by Grech et al. (*Res. Microbiol.* 158:685-693). It uses a position weight matrix approach to analyze noncoding regions of related genomes. This approach has proven limited, however.

|                                     |                      |                                   |                       |                      |
|-------------------------------------|----------------------|-----------------------------------|-----------------------|----------------------|
| <i>C. glutamicum</i> SigA consensus | TTGNCA               |                                   | GNTANANTNG            |                      |
| <i>C. glutamicum</i> SigH consensus | KGGAA <sup>W</sup>   | YG <sup>T</sup> TR <sup>W</sup> W |                       |                      |
|                                     | -35 (P2 $\sigma^H$ ) | -35 (P1 $\sigma^A$ )              | -10 (P2 $\sigma^H$ )  | -10 (P1 $\sigma^A$ ) |
| <i>C. glutamicum</i>                | TGGAACAAC            | TTTGTGGCAT                        | TTACCGTTGCTATAT       | ATGTAAGC             |
| <i>C. efficiens</i>                 | GGGAATAAC            | CTTGTGGTTAAGCT                    | CGTTGC                | AAAAGGTGTAAGGTTGAGT  |
| <i>C. diphtheriae</i>               | GGGAACAAC            | TTTGGGGTCTGTGAGT                  | TACTACTAGTGT          | CAGGTTGAGT           |
| <i>C. pseudotuberculosis</i>        | GGGAACAAC            | TTTGGCGTGAGATGT                   | CGTTATGGATAGTGT       | CAGGTTGAGT           |
| <i>C. ulcerans</i>                  | GGGAACAAC            | TTTGTGGGTAGTGC                    | CGTTATGTATAGTGT       | CAGGTTGAGT           |
|                                     | :****:***:***.*:..   | :***..                            | ....:***.***:*****:.. | ..:*****:            |

--image from Grech et al. (2007), *Res. Microbiol.* 158:685-693

#### Finding sequences unique to promoters

Because promoters have consensus sequences, promoters can be found by the statistical overrepresentation of specific short sequences in the intergenic regions. Another method looks at the weaker stability of the DNA double helix associated with promoters. Recall that one feature of polymerase binding is promoter melting. Both methods have limitations. A training set approach is good at finding promoters very similar to what is in the training set, but may miss "unusual" promoters and give lots of false positives. Jacques et al. (*BMC Bioinformatics*, 2006, 7:423) used matrices that represent the genomic distribution of hexanucleotide pairs. However, false positives were a problem--they looked heavily at -35 and -10 sequences that may not be relevant for all promoters.

#### Incorporating microarray/RNA-seq data

Meng and Vigron, in [a 2014 paper](#), describe a prediction method that incorporates transcriptome data. They describe a "guilt-by-association" methodology, looking for co-expressed genes. The thinking is that co-expressed genes share common regulators. Their [CTSP Tool](#) has a web-based interface. It does require uploading expression data.

Another tool had a similar approach. BioProspector is a C program that uses Markov modeling to examine upstream regions of co-expressed genes. It can be [downloaded from the Liu lab](#) at Stanford.

**Discussion topic: What are microarrays? How might microarray data be used to find eukaryotic promoters?**

### **Alternative sigma factors**

Yu et al. (*J Bacteriol.* 2006. 188:23-8206-8212) use weight matrices to find promoters that are associated with  $\sigma^{28}$ , an alternative sigma factor in *Chlamydia*. The  $\sigma^{28}$  factor is also known as the motility sigma factor because it is involved in flagella biosynthesis. Song et al. (*FEMS Microbiol Lett.* 2007. 271:222-229) used similarity to *E. coli* followed by a weight matrix to identify  $\sigma^{28}$  promoters in multiple species.

Rhodium et al. (*Nucleic Acids Res.* 2012 40:2907-2924) developed detailed full-length promoter models for  $\sigma^E$ , an alternative sigma factor in *E. coli*. Their goal is to use the models to accurately predict promoter strength. Those findings could be useful for engineering promoters of particular strengths.

### **Programs**

[BPROM](#) is the SoftBerry program for finding bacterial promoters. It uses Linear Discrimination Function to find potential  $\sigma^{70}$  promoters with a claimed accuracy of 80%. Here is a help page and sample output. A [Neural Network Promoter Prediction](#) program may be found at the Berkeley Drosophila Genome Project page, but it can be set for prokaryotes.