What is a Gene (Part 2)
Shine J & Dalgarno L (1975) Nature 254:34-38

The semester began with a consideration of *What is a Gene?* You considered how the cell determines where the gene begins and where it ends. Begins and ends in what sense? The tour didn't talk about transcription and translation but posed the problem as code-breaking in its pure form. You looked through sequences at the beginning and ends of entities given to be actual genes, and from them you deduced sequences that ***might*** be used by the cell to mark the gene's boundaries. But what were these given genes? We need to superimpose some biology on this problem.

Fig. 1 diagrams how first a gene is transcribed to mRNA and then the mRNA is translated to proteins. The gene is defined by translation, beginning at the start codon for translation and ending with the stop codon. The operon is defined by transcription and may consist of multiple genes. Multigene operons are common in bacteria and archaea, much less common in eukaryotes.[1] In the case of both transcription and translation, initiation of the processes is governed by the binding of a multiprotein complex to a DNA binding site: RNA polymerase to a promoter and a ribosome to a ribosome-binding site.

SQ1. In seeking to define the gene, which process – transcription or translation – was *What is a Gene* concerned with?

SQ2. What insights did you gain from *What is a Gene* as to what determines where ribosomes bind?

SQ3. Perhaps the conclusions you reached in *What is a Gene* are wrong. What are some of the most critical assumptions underlying your conclusions?

A. Our goal in reading Shine and Dalgarno (1975)

Before we go a step further, go get the article that will occupy our thoughts: Shine J, Dalgarno L (1975). Nature 254:34-38. Be certain you have the right article.

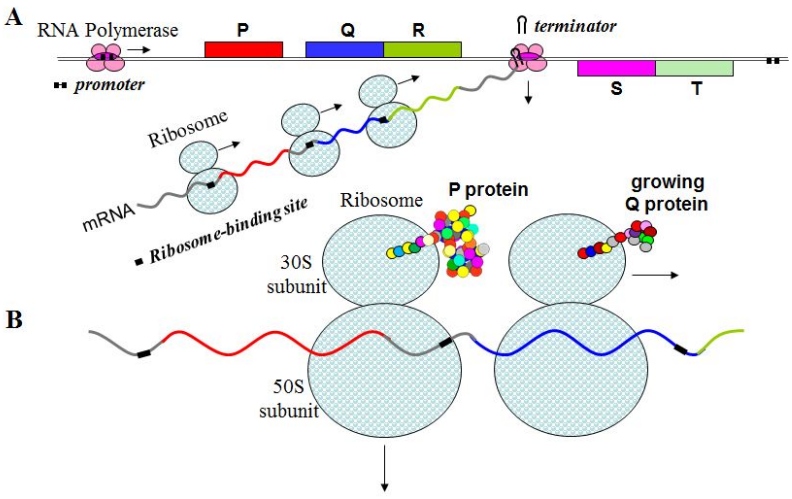

Figure 1. Transcription and translation. (**A**) RNA polymerase (a multi protein complex) binds to the DNA at a promoter well upstream from the operon consisting of three genes, P, Q, and R. It proceeds to transcribe the operon, producing PQR mRNA as it goes. When it transcribes a termination sequence, transcription stops and RNA polymerase falls off the DNA. Meanwhile, ribosomes bind to ribosome-binding sites immediately upstream of genes and proceed to translate the mRNA to proteins. (**B**) A single mRNA may be transcribed from multiple genes, but a ribosome translates only one gene at a time. When it reaches the end of the gene, the ribosome falls off.

[1] ...but not unknown [Blumenthal T (2004) Briefings Funct Genom Proteom 3:199-211].

You got the article (I hope), perhaps because the authority of the printed words directed you to do so, or perhaps because you have of a sliver of faith in my implied belief that you would be better off if you made this small effort. However, neither should be sufficient for you to actually ***read*** the article. For that you should have a substantial reason.

Let me suggest this one. One theme running throughout this semester is how to recognize a gene. Soon you will be applying your insights to genes for which the answer is not already known, and the world will rely on your answers. It is therefore important to understand (as much as possible) how the cell recognizes a gene so that you may as well. Your first go at it, through *What is a Gene*, you relied on the accuracy of the gene definitions in BioBIKE. You learned how to assess the significance of patterns occurring at the beginnings and ends of genes... so long as the gene definitions are correct.

But ***are*** they correct? Any truth they contain can be traced back to laboratory experiments someone did on those or similar genes. Shine and Dalgarno started with the sequences of some bacteriophage genes whose beginnings were already known by experiment.

SQ4. Given how a gene is defined, what kinds of experiment could determine the beginning of a specific gene?

Their goal was the same as ours in *What is a Gene* – to determine the sequences that may be used by a cell to mark the beginning of a gene. So let's consider that question again, through their eyes.

B. Initial reading of Shine and Dalgarno (1975)

With that question in mind, go through the article quickly, noting parts of it that relate to what you want to find out and therefore might warrant closer attention. Go on... I'll wait.

SQ5. Which sections and tables did you mark as most worthy of your attention? Which least worthy? (recall that "worthy" means relevance to the question you brought to the article)

OK. Now we can compare scorecards. Here's how I marked the sections:

Abstract: Encouraging, and I gather that 16S RNA[2] is a major player in this article (but now that I'm encouraged, I see no need to spend any more time on the abstract, since the article should explain everything in the abstract but in greater depth)

Introduction (section not labeled): Seems to be saying things I'm already familiar with. I'll probably read this more carefully to get some gratification.

Cistron specificity: [NOTE: "cistron" is an old word for gene] This section seems to be talking about how general is the signal to start of translation (hence the start of the gene). This is interesting, but right now I want to know what the signal is. So, I'll leave this section for later or maybe never.

Base sequence or secondary structure? The first sentence is very provocative (don't know what "30S" means? See Fig. 1). It poses two possibilities: one that was considered in *What is a Gene* and one that wasn't. I mark this section high priority.

---

[2] If you're uncertain about "16S RNA" you might want to pay another visit to *Flow of Information*, on the topic page, Introduction to Molecular Biology.

3'-terminal sequences: This looks like a description of method, which I'll avoid until I have no choice. Ditto with Table 1.

Complementarity to coliphage ribosome-binding site sequences: This section describes the main table in the article (Table 3), which is certainly a point in its favor. I'm not sure at this point whether I'm interested in Table 2. The section as a whole talks about ribosome binding... I think I'll understand the section if I understood what Table 3 meant. I'll mark Table 3 as high priority and the rest as maybe.

Initiation factors: Seems pretty obscure. Maybe I can get away without learning what initiation factors are.

Concluding paragraph: (note that the last paragraph is nominally part of "Initiation factors" but really functions as a summary). No doubt about it. They really think that 16S ribosomal RNA is important in determining the beginning of genes. New goal for article: Figure out what the connection is between the two.

I have arrived at my prioritized list. First I want to go more carefully through the section "Base sequence or secondary structure?". Then I want to understand Table 3 (the section "Complementarity to coliphage..." maybe helpful in this goal. Through all this, I hope to understand what evidence is there that 16S ribosomal RNA determines the beginnings of genes.

C. Base sequence or secondary structure?

Paragraph 1: The last sentence summarizes their views on secondary structure. Maybe they're right or maybe they're wrong. I won't know unless I look inside the many articles they cite. But Shine and Dalgarno don't add their own data to this discussion, so I'm going to skip over this issue to get to the one for which they ***do*** present evidence.

Paragraph 2: Sequences of ribosome-binding sites... this paragraph looks like it's going to give some answers! They present seemingly pertinent evidence, then at the end draw a conclusion that uses some very good words (e.g. "16S RNA", "initiation of protein synthesis"). Time to go through the paragraph in some detail.

SQ6. List the evidence in the paragraph that justifies the last sentence.

SQ7. What is the specific sequence in the ribosomal RNA they said is important?

SQ8. What do they claim is the specific sequence of ribosome binding sites.

SQ9. How are the two related to each other?

Shine and Dalgarno make some bold claims. Let's test some of them. They make a claim about 16S RNA in *E. coli*. We should be able to confirm or refute this by pulling up the sequence and looking for ourselves. *E. coli* is not a cyanobacterium, so you won't find the sequence in CyanoBIKE. But PhAnToMe/BioBIKE has a large variety of bacterial genomes. That would be the logical place to look.

Go to PhAnToMe/BioBIKE

- Go to the BioBIKE Portal

- Click PhAnToMe/BioBIKE public site (NOT CyanoBIKE)

Now that you're in PhAnToMe/BioBIKE (and if you're not, there's no point in going any further), we have two immediate tasks: (1) Find the *E. coli* genome, and (2) Find the 16S RNA gene within that genome.

To find a bacterial genome, mouse over the ORGANISMS button and click Bacteria. After a few seconds, the status line (near the green arrows) will tell you that the bacteria menu has been loaded and invite you to use it. Now the ORGANISMS button will have an item called Bacteria (alpha). Mouse over that to get an alphabetical list of bacteria. Find *Escherichia coli*... Whoops! There are over a dozen strains! I believe they're all pathogenic *E. coli* (none used by Shine and Dalgarno or any sane person studying basic molecular biology), except for the laboratory strain *E. coli* K12. Click that strain to bring it into the workspace.

To find the 16S RNA gene... let's hope that the organism is well annotated. If it is, then the gene should be described as some variant on "16S RNA". Go to the GENES-PROTEINS button and bring down the GENES-DESCRIBED-BY function. Fill in the *query* box with the term you want to search for (in quotes, since you want to search for the specific letters, not the contents of some nonexistent variable by that name). If you executed the function at this point, BioBIKE would search every gene of every bacterium and every phage. It could take a while. Better to tell it where you want to find the gene. So mouse over the Options icon and click IN and then Apply. Drag in the *E. coli* box you got in the previous paragraph into the *value* box of IN. Now execute the function.

If you tried searching for "16S RNA", then you failed, but in doing so you learned that different people have different ideas what to call genes. Shine and Dalgarno call the gene 16S RNA, but the annotators of *E. coli* had a different idea. If your first try was too specific to find a match, try something less specific, perhaps just "16S".

SQ10. If you searched for "16S", then you probably got two matches. Consider each and decide the prospects of each for getting you what you want.

If you were looking for chocolate, then surely something labeled "chocolate-eating monster" would be low on your list of candidates. Whatever else 16S RNA is, it isn't a protein. The other match is much more encouraging. If you could find part of the gene, you'd probably be able to find the whole thing. And finding genes is already in your repertoire, almost. Copy the name of the gene and then bring up the sequence of *E. coli*, using SEQUENCE-OF as usual. Paste the gene name into the Go To box of the sequence viewer, and click Go. You will be brought to the position in the genome where the gene lives, and there you'll find a second gene on top of it. What is that gene? Click on it to find out. Under "Description" you'll see how the annotators of *E. coli* chose to describe the gene. What does the term mean? Google is your friend.

SQ11. Once you've found the gene, determine whether Shine and Dalgarno's claim regarding 16S RNA is correct, at least in the case of *E. coli*. If you found an appropriate sequence, what are its coordinates?

Shine and Dalgarno also make a significant claim about "...all coliphage RNA ribosome-binding sites examined to date...". However, without visiting the articles cited in the previous paragraph, we're not in a position to evaluate that claim. We'll put this one on hold until we look at Table 3.

Paragraph 3: No new information here but a useful summary (I like the way these two write!).

SQ12. Do you now see why Shine and Dalgarno are so enthusiastic about sequencing the 3' ends of 16S RNA from different bacteria?

D. **Complementarity to coliphage ribosome-binding site sequences**

I can now understand why Shine and Dalgarno want to test their idea with different bacteria, but it makes some sense to confirm what they claim with respect to the known case of *E. coli*. In the first sentence they provide us with three test cases, three genes from a phage that infects *E. coli*. Best of all, they tell us where to find the relevant sequences for these genes.

SQ13. What three genes did they provide? Where can you find the genes' ribosome-binding sites? What specific sequences are claimed to be those ribosome-binding sites?

SQ14. Are Shine and Dalgarno right in their claim of base-pairing?

Now we have specific genes to work with. Again we have two tasks: (1) Find the phage genome, and (2) Find the three genes within the genome. You can try to find the phage genome in the same way you found *E. coli*, but using the Viruses menu. Once you've activated the menu, you can click Viruses (alpha) and search for... This doesn't look right. None of the viruses have names like "R17". They all begin with names of bacteria, which in fact are their hosts. OK. Try scanning the Escherichia phages for R17. Any luck? No need to panic. Maybe the name is weird. Go to the GENOME menu, mouse over GENOME-DESCRIPTION, and click ORGANISM/S-NAMED. Since you're interested in finding only phage (not bacteria), select the PHAGE-ONLY option. Finally, type "R17" into the name box and execute the function. …Now you can panic.

The result might seem like very bad news, but if you look at the first footnote of Table 3, you'll find a ray of hope. Its last sentence points to two other phages that will do just as well as R17.

SQ15. Using the tools you've gotten experience with, try to find these two phages in BioBIKE.

I think we have one reasonable candidate, and how reasonable it is depends on the relationship of "Enterobacteria" to "Escherichia coli". There are many ways of resolving this issue. I'll let you find one.

SQ16. Bring up the sequence of MS2, using SEQUENCE-OF. How big is the phage? How many genes does it have?

SQ17. Use the search box within the sequence viewer to see if you can find the sequences Shine and Dalgarno claim should be there in the three ribosome binding sites shown in Table 3. Where are they situated relative to their associated genes? What is "A-protein"?

SQ18. At least in the cases you've now observed, how distant is the ribosome binding site from the start codon of a gene?

SQ19. Evaluate Shine and Dalgarno's claim (at least with respect to *E. coli* and phage MS2) that there is a complementary relationship between the 3' end of 16S ribosomal RNA and ribosomal binding sites in front of genes.

This looks pretty good (though it is only three protein – not a large sample), but you may sure that bacteria didn't devise its protein synthesis machinery just for the benefit of phages. You'd think the same mechanism works in the bacteria as well. Is this true?

SQ20. Define a variable containing the upstream regions from the genes of *E. coli* -- recall that you did something very similar to this in *What is a Gene*. Display the sequences. Sensitized by your reading of Shine and Dalgarno (1975), do you see any common features amongst many of the upstream regions?

SQ21. Use another trick from *What is a Gene* – filtering – to collect all upstream sequences that have a plausible ribosome-binding site according to Shine and Dalgarno. I suggest that you choose a pattern that matches all the binding sites of minimal length. Note that you can provide FILTER with a pattern that contains alternative patterns. "abc|def|ghi" is a pattern that demands that the sequence matches "abc" OR "def" OR "ghi".

SQ22. What fraction of coding genes in *E. coli* is preceded by ribosome-binding sites as you defined them in SQ20?

SQ23. Recall that in *What is a Gene*, Part III, you plotted the position-specific distribution of nucleotides upstream of all the genes of *Anabaena variabilis*. How would you now interpret that plot?

SQ24. Repeat that plot with the upstream regions of *E. coli*. Repeat it again with those upstream regions that did NOT match the pattern you specified in SQ20. What conclusion do you draw from these plots?