Differential Expression of Photosynthesis Genes in R. capsulata…
Belasco JG et al (1985). Cell 40:171-181
A tour (Part II)

IV. Determination of the relative steady state concentrations of the *rxcA* messages

IV.A. Elements of the S1-protection experiment

How does Rhodopseudomonas arrange for ***all*** photosynthetic genes to be expressed at the same time, when photosynthesis is desirable, but for ***some*** photosynthetic genes (those encoding proteins of the light-harvesting complex) to be expressed at a much higher level?

We've focused in on a key section of the paper ("*Determination of the relative…*") that seems to promise an answer. Don't get put off by "*steady state*". That just means that they're not concerned with transient conditions, for example the moment after you turn the lights on. They're focusing on conditions where the lights have been on a long time and genes that are supposed to be on have been on a while and have reached cruising altitude. Also, don't be put off by *rxcA*. This refers to the region shown in Figure 1. The region is clearly misnamed (***R***ea***X***ion ***C***enter), because not all of the genes encode proteins of the reaction center.

SQ27. What genes are included in the *rxcA* region and what photosynthesis units are their encoded proteins part of?

The critical experiment described in this section is based on S1-protection analysis, which by now you have a good understanding of (if not, please see SQ24 in Part I). S1-protection experiments require a labeled probe.

SQ28. What question is the S1-protection experiment trying to answer?

S1 experiments require a labeled probe, a piece of either RNA or DNA that has been made so that it is easily detected. One way is to make it radioactive.

SQ29. What DNA was used as the probe in the described S1-protection experiment? Why was that DNA chosen?

You're probably unfamiliar with phage M13. "Phage", short for "bacteriophage" (bacteria eater), is the word used by those in the field to describe viruses that infect bacteria. Take a moment off to find out what is M13 phage.

SQ30. Does M13 phage infect *R. capsulatus*? Is it likely that the DNA sequences of the M13 genome would be found in *R. capsulatus*? If the M13 phage genome were made radioactive would it bind to DNA or mRNA isolated from *R. capsulatus*?

M13 serves as a vector, a carrier, in which to amplify the DNA that will be the significant part of the probe. Phage by their nature replicate themselves, and cloning foreign DNA within its genome causes that foreign DNA to be replicated as well. That way, it is possible to amplify the foreign DNA, to get enough of it to serve as a probe in the experiment.

SQ31. What DNA was cloned into the M13 phage genome? Why that piece?

It might help to take a closer look at the region in question, within the context of the *R capsulatus* genome. Go into CyanoBIKE (go to the BioBIKE portal and CyanoBIKE, VCU mirror). The *R. capsulatus* genome is not part of CyanoBIKE, since the organism is not a cyanobacterium, but you can bring it in by running a special file I've provided. Bring down the

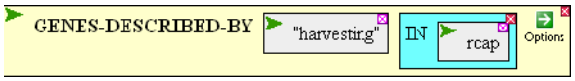
RUN-FILE function from the INPUT-OUTPUT button (or from the alphabetical listing under the ALL button). Enter "rcap.bike" as the file name (including the quotation marks) and specify the SHARED option. The latter says that the file resides in a directory accessible by all. Execute the function.

You should now see a VARIABLES button that has `rcap` in it. That gives you access to the *R. capsulatus* genome. You already know how to see the genome. Bring down the SEQUENCE-OF function, click the *entity* box, and either bring down `rcap` from the VARIABLES button or (equivalently if you spell it right) just type it in. Then execute the function.

You should see the first 6000 nucleotides of the chromosome of *R. capsulatus*. But where are the photosynthesis genes?

SQ32. You're looking for the *rxcA* region. How likely is it that it lies in the first 6000 nt? How big is the genome? How many genes are in it?

OK, so you're going to have to look for the region, and going through the genome bit by bit will be way too slow. How about letting the computer do the walking? There are many ways to do the search. Here's one. Bring down from the GENES-PROTEINS button the GENES-DESCRIBED-BY function. In the *query* box, type "harvesting", since this word is pretty unlikely to be in the name of any gene except those encoding proteins of the light-harvesting complex. Choose the IN option from the Options Icon menu, and type `rcap` in the *value* box. Then execute the function, producing a popup window with the results.



SQ33. Which of the genes are the ones you're looking for? Why?

Copy the name of one of the genes you're looking for and paste it into the **Go to:** box and click **Go**.

SQ34. Were you brought to a region of the genome that is familiar to some extent? What genes do you see there?

SQ35. The legend to Figure 1 states "*Although the L and M genes overlap by 5 nucleotides…*". Is this true? If so, what are those 5 nucleotides? If not, why not?

It would be nice to have closer match with the diagram of Figure 1, which shows a significant number of nucleotides preceding the B870β gene. I'll estimate that the number of nucleotides to the left of the gene is about the same as the length of gene **L**.

SQ36. How long (in nucleotides) is gene L?

SQ37. What is the coordinate of the nucleotide before the B870β gene by an amount equal to the length of gene L?

Type that coordinate into the **Go to:** box and click **Go**.

Now we have some space. The article claims that the S1-protection probe consists of the M13 phage genome (which bears no similarity to the *R. capsulatus* genome, so we don't care about it) into which an *rxcA* fragment (which we ***do*** care about) has been cloned. That fragment is described as extending from the EcoRI site to the PstI site. These are restriction sites, DNA sequences recognized and cut by specific restriction enzymes. If you already know something

about restriction enzymes, fine. If not, just leave it at this: restriction enzymes bind to DNA and cut it at specific sequences. The restriction enzyme EcoRI recognizes the sequence GAATTC, and the enzyme PstI recognizes the sequence CTGCAG. You can find specific DNA sequences within a genome by clicking the **Display only?** box, typing the sequence into the **Search:** box, and pressing Enter.

SQ38. Where does the EcoRI site of interest lie in the genome? What is its coordinate?

SQ39. Where does the PstI site of interest lie in the genome? [*Don't forget to click the Display only box!*] What is its coordinate?

SQ40. What is the length of the EcoRI-PstI fragment? How does this compare to the length given in the article?

SQ41. If we believe Figure 1, about where does the end of the short *rxcA* messages lie in the genome? About where does the long message lie?

IV.B. Results of the S1-protection experiment

Of course, we don't believe Figure 1, at least not until we're given a reason to do so. Figure 2 may be the place to find some reasons. So take a look at the right panel of Figure 2. It shows total RNA (including mRNA) from *R. capsulatus*, hybridized with the radioactive probe, digested with S1 nuclease, and run out on a gel to separate surviving RNA-DNA hybrids by size.

SQ42. What size fragments are detected in the experiment? Why do you say so? [Understand that nothing the authors say should be given any weight unless evidence is provided. And if evidence is provided, then you don't need what the authors say, except to guide your own observations of the evidence.]

SQ43. Compare the sizes of these fragments to the sizes of the messages shown in Figure 1. Why do you not see a 2.7 kb (2700 nt) fragment in Figure 2?

SQ44. Consider the longest fragment. What to you predict to be the coordinate of its 5' end (its beginning)? What assumptions are you making? Where does that coordinate lie in the genome?

SQ45. Presuming that the short fragments begin at the same point (which remains to be shown in other experiments), at what approximate coordinates do they end?

SQ46. Which is the most abundant fragment? Which is the least? Which is the most abundant message in Figure 1? Which is the least?

IV.C. Implications of the S1-protection experiment

This observation poses somewhat of a problem. Recall the sentence from the Introduction, "*If these photosynthesis gene products are in fact encoded by a single transcript, then control of gene expression must occur at a level beyond the simple modulation of transcription initiation.*" Modulation of transcription initiation was at the time (and largely remains) the mechanism we know the most about regarding the control of gene expression. The usual picture is to turn on the water spigot and let the hose worry about itself.

SQ47. Are the photosynthesis gene products of the *rxcA* region in fact encoded by a single transcript?

SQ48. Does it follow that control of gene expression may still be explained by the simple modulation of transcription initiation?

Suppose you are an RNA message in the process of being made. RNA, like DNA, is synthesized 5' to 3', shown in Figure 1 as left to right. Suppose your synthesis has proceeded up to the middle of the B870β gene. Tell me, are you a long message (one that will extend through the reaction center **L** and **M** genes) or a short message? How do you know?

If the long and short messages in fact begin at the same place (crudely suggested by Figure 2, more precisely measured in a later experiment), then why do we have messages of different sizes? The authors conceived of two possibilities:

(1) *Differential termination*: Most messages stop just after the B870α gene, some extend to beyond gene X.

(2) *Differential degradation*: All messages extend to beyond gene X. Most are degraded back to just after the B870α gene.

Much of the rest of the article tries to distinguish between these two possibilities, but we'll leave the story at this point.

V. The critical sequence element

Whether termination or degradation is important, there's something very interesting about the region just after the B870α gene. What's so special? In examining this region, the authors found the sequence shown in Figure 4. Let's confine our attention to the upper sequence.

SQ49. Find that sequence in the displayed *R. capsulatus* genome. Where is it relative to the genes of *rxcA*?

SQ50. If you were to search for `UAAGCGUU`… you would no doubt be unsuccessful. Why? What could you search for that would give you a match?

SQ51. Figure 4 shows this sequence folded back on itself. Can any random sequence do this? Why this one?

SQ52. Write out the DNA sequence corresponding to the beginning of the hair pin to the end (just the part of the sequence that's folded back, from GUUGC… to GCGAC), typing T instead of U. Label one end 5' and the other 3'. Write below it the second DNA strand, also labeled with 5' and 3'. Read the bottom strand (of course 5' to 3') and compare it to the upper strand. What do you observe, and what do you call such sequences?

SQ53. This isn't the first time in this tour you've encountered such a sequence. Recall the EcoRI site (GAATTC) and the PstI site (CTGCAG). Do the same is in the previous question, writing out the second strand for these sequences.

What's going on? Why are palindromes everywhere? Any time you're struck by something remarkable – something that seems to arise too frequently or too infrequently – you need to protect yourself with the following sort of question:

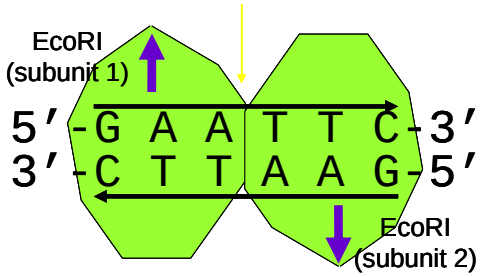
SQ54. Maybe you're being fooled. Maybe palindromes are very common, just by chance. What is the probability that a random 6-nt piece of DNA will be a palindrome?

Palindromes in natural DNA frequently turn up in regions important in regulation. This is because they have two properties, both of which you have now seen:

(1) Palindromes when expressed as single-stranded RNA can fold back on itself

(2) Palindromes read the same forwards (on the upper strand) and backwards (on the lower strand). This means that a protein that recognizes the upper strand will also recognize the lower strand.

This last property explains why so many DNA-binding proteins (like EcoRI) bind to palindromic sequences. They do so because it is easier to evolve a 3-nt binding specificity, and binding twice, as shown to the right, gives twice as much specificty for the same number of distinct proteins. Since DNA-binding proteins often control the initiation of gene expression, unusual palindromes in DNA are often indicative of regulatory sites.



Representation of EcoRI binding to its recognition site. Note that two copies of the protein subunit binds in opposite orientations. Note also that if you flip the figure upside-down, you get what you started with.

SQ55. Give a specific example for each of the two ways (I know of) that palindromes may be significant in nature.