Genome Engineering

Drew Endy (http://mit.edu/endy/)

Questions for Today

A. What can we infer about the design of a genome by considering the physical problems that a biological system solves?

B. How is genetic "real estate" divvied up?

C. What makes a good genetic "part"? What makes a good genome design?

1. Let's check out a time-lapse movie of bacteriophage lambda induction (all the movies in today's class were taken by Francois St-Pierre, a biology graduate student). Note what the bacterial cells are doing. Reproducing and dividing. I.e., they are reproducing machines. This is no small feat. Imagine building a machine that is only one micron long and that can copy itself!!!

2. [Note that sometime this semester, I expect that a research paper should be published that will show for the first time that it is possible to construct a copy of a natural bacterial genome from scratch, and place this newly constructed into an empty cell, causing the cell to start growing and dividing. Thus, while we are considering the design of a bacteriophage genome this in 20.109 this Fall, you should expect to be able to apply the concepts you learn here to design microbial and eukaryotic genomes soon enough].

3. OK, now note that the cells lyse (i.e., are destroyed). In this case, a virus that was latent inside the bacterial cell became active, and destroyed the cells while in the process of making more copies of itself. **What are the various functions that are required to destroy a cell while making progeny phage particles?**

4. Let's check out a second movie. In this case the lambda virus is not yet inside the cell. The virus particles appear green because a protein (green fluorescent protein) has been fused to one of the shell (i.e., capsid) proteins of the lambda particle. As a result, all the particles glow green. A few cells become infected by the phage and eventually produce more green particles, and in the process are destroyed. **What else can we now infer about these "strange" green particles? What new functions must be present?**

5. OK. So, let's take a look at M13, a different phage, the one that we'll be focusing our attention on this term. M13 usually does not destroy its host cell. Instead it maintains some sort of stable existence with E. coli. That is, an E. coli cell infected with phage M13 will continue to grow and divide, but will secrete M13 phage into to the extracellular environment. **Is such a lifestyle more or less sophisticated than what phage lambda does? For example, would you expect that M13 requires more or less functions to infect E.coli, producing more phage but not killing it?**

6. [Note that we do not have a movie of M13 infection. Perhaps somebody would like to make a GFP fusion to an M13 coat protein. We could then make a movie of M13 infection and load it up on YouTube].

7. From the above, make a list of the sorts of biochemical functions (activities, commands, etc) that we expect M13 must have. I.e., the list of things you would expect to see if you "lifted the lid" or "looked under the hood" of a M13 particle. For example:

- particle protein (i.e., coat protein)
- landing gear protein (i.e., tail protein)
- constructing protein (i.e., assembly protein)
- secretory protein (i.e., something to get outside the cell)
- infection protein (i.e., something to inside a new cell)
- copy protein (i.e., something replicate the phage DNA)

8. Could we also infer that there must some information storage mechanism? I.e., a genome that encodes all these functions? [Note that this question seems strange because we know that there is an M13 genome, but your grandparents would not have know this yet].

9. OK, let's get down to business. **How are these functions encoded on the genome?** [Sketch of boxes of different genes arrayed across a genome]. Different functions are encoded as sequences of DNA. For example, a protein will be encoded as a "coding sequence" (aka CDS).

10. Now, are these functions sufficient by themselves? Stated differently, how do we go from a CDS to a biochemical activity? I.e., how are the DNA encoded functions read out inside the cell? [Transcription, translation, and so on]. The key point here is that, in addition to coding sequences specifying different biochemical activities, we also need to think about controlling sequences that determine when and how much of different activities are produced. **In designing a system, control can be as important, or more important than the functions themselves (e.g., think of an furnace with no thermostat, or a plane without any rudder or wing flaps).**

11. So, we can infer that there must be elements for turning ON and OFF transcription, and for turning ON and OFF translation, and DNA replication.

12. Let's take a look at a few key types of bacterial genetic elements:

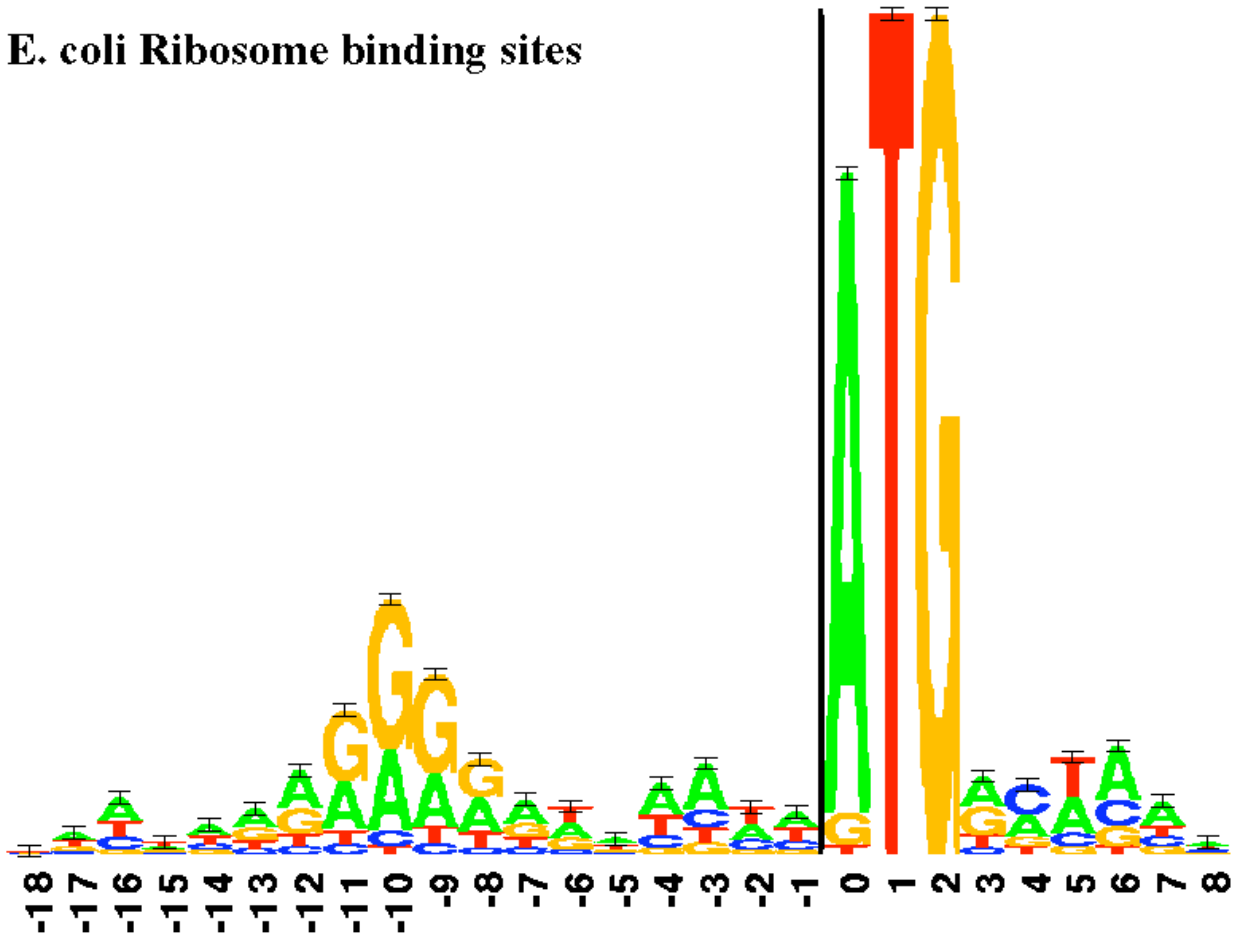
12a. The Promoter, champion of transcription initiation. Standard promoters that operate inside E.coli will have what's called a -35 region, and a -10 region. These are the regions that interact with RNA polymerase, the enzyme that carries out transcription (by polymerizing an RNA molecule, copying the underlying DNA template). The numbers -35 and -10 refer to the position on the DNA that the RNA polymerase binds prior to initiating transcription at the +1 site. I.e., for such a promoter, the E.coli RNA polymerase binds upstream of where transcription starts (i.e., the transcription start site). The exact sequence of a promote can vary, and as a result will act as a stronger or weaker promoter (i.e., initiate more or less messenger RNA synthesis). A typical E.coli promoter will look like:

-- TTGACA -- (15-19 base pairs) --- TATAAT -- (a few more base pairs) -- START

(-35) (-10)

12b. The Ribosome Binding Site (RBS), master of translation initiation. Standard RBSs that operate inside E.coli will be about 7 base pairs long and will be positioned about 6 base pairs upstream of the start of a CDS encoding a protein. The exact sequence of an RNA can vary (as w/ promoters, sequence differences will produce different rates of protein synthesis initiation). For example, a very strong RBS might have the sequence <u>AGGAGG</u>(nnnnnnn)ATG, where the underlined sequence in the RBS, and the seven n's are any space sequence separating the RBS from the ATG start codon (what's an ATG start codon? hang on). Note that you do not always need a "formal" RBS to start protein synthesis. There are natural examples of proteins that are synthesized from mRNA

that do not appear to contain any RBSs!!! [see the RBS "sequence logo" depiction below for a depiction of sequence variation at E.coli RBSs].



[above image c/o T. Schneider via http://www.lecb.ncifcrf.gov/~toms/gallery/ribo.logo.gif]

12c. The Coding Sequence (CDS), encoder of proteins! Coding sequences will be made up of codons, triplets of DNA that define the genetic code for a particular organism. 64 codons can be defined from the 4 bases. Some codons specify what amino acid should be incorporated into a protein next. Other codons define START and STOP sequences that control protein synthesis (i.e., translation). For example, ATG and GTG are two "allowable" start codons in E. coli. And, UAA, UAG, and UGA are three "allowable" stop codons. A start codon will direct the initiation of protein synthesis (usually in partnership w/ a RBS). **A stop codon will halt protein synthesis. A CDS is defined as the sequence of DNA between a start and stop codon.**

12c-bonus. But, note that the start and stop codons need to be "in frame" in order to function as a pair. In frame refers to the fact that codons are three base pairs long, so any strain of DNA contains three distinct reading frames, depending on which base you decide to start reading from. And, if you consider the complementary strand of double stranded DNA, a sequence of DNA contains up to six (!!!) reading frames. Stay tuned.

13 . Enough. Let's take a look at an example DNA sequence. What can you find?

gc**[**atg**{**cgcaa**agggagg**cgac**ATG**gcaggttacggcgc**taa]**agga**}**atccgaaat**TGA}**aaa

14. What parts can you find in the above sequence?

15. What can you say about the architecture of genetic elements?

16. Why might this particular architecture be good or bad?