## Lecture and Dialog on Synthetic Biology

Drew Endy (endy@stanford.edu)

### *Goals for Today*

**A. Coding in DNA**

**B. DNA refinement**

**C. New DNA codes?**

**Announcements **

i. Last chance to send in preferred office hours / tutorial. We'll start next week. Email endy@stanford.edu

**Today**

0. (From last time). We can change the function of a DNA element (such as a transcription promoter) by changing individual bases of DNA (e.g., from a T7 to a T3 RNA polymerase promoter). Such changes might impact the *strength* or *specificity* of the DNA element. We can also change the architecture of DNA, rearranging natural genetic elements (e.g., all the genes of yeast in alphabetical order) or their sequence-specific architecture (e.g., unstuffing overlaping genes).

*A. Coding in DNA*

1. Given that DNA is code, we might imagine that it would be possible to store messages or otherwise inside DNA. Turns out that this is true!

2. In order to encode messages in DNA you first need to decide on a code. The code that most people tend to start with (although this has issues, as we'll see) is the "genetic code," which is the mapping from the 64 triplet codons to the 20 amino acids.

| First Position (5´ end) | | Second Position | | | | | | | | | Third Position (3´ end) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | U | | C | | A | | G | | | |
| | U | UUU | Phe | UCU | Ser | UAU | Tyr | UGU | Cys | U | |
| | | UUC | | UCC | | UAC | | UGC | | C | |
| | | UUA | Leu | UCA | | UAA | *Stop* | UGA | *Stop* | A | |
| | | UUG | | UCG | | UAG | *Stop* | UGG | Trp | G | |
| | C | CUU | Leu | CCU | Pro | CAU | His | CGU | Arg | U | |
| | | CUC | | CCC | | CAC | | CGC | | C | |
| | | CUA | | CCA | | CAA | Gln | CGA | | A | |
| | | CUG | | CCG | | CAG | | CGG | | G | |
| | A | AUU | Ile | ACU | Thr | AAU | Asn | AGU | Ser | U | |
| | | AUC | | ACC | | AAC | | AGC | | C | |
| | | AUA | | ACA | | AAA | Lys | AGA | Arg | A | |
| | | AUG | Met | ACG | | AAG | | AGG | | G | |
| | G | GUU | Val | GCU | Ala | GAU | Asp | GGU | Gly | U | |
| | | GUC | | GCC | | GAC | | GGC | | C | |
| | | GUA | | GCA | | GAA | Glu | GGA | | A | |
| | | GUG | | GCG | | GAG | | GGG | | G | |

3. The reason many folks start with this code is that the 20 amino acids have single English letter abbreviations, which means that if you can spell out a word as a string of amino acids, you can then reverse translate (i.e., lookup) the DNA sequence that would encode those amino acids.

# The Single-Letter Amino Acid Code

---

- **G** - Glycine (Gly)
- **P** - Proline (Pro)
- **A** - Alanine (Ala)
- **V** - Valine (Val)
- **L** - Leucine (Leu)
- **I** - Isoleucine (Ile)
- **M** - Methionine (Met)
- **C** - Cysteine (Cys)
- **F** - Phenylalanine (Phe)
- **Y** - Tyrosine (Tyr)
- **W** - Tryptophan (Trp)
- **H** - Histidine (His)
- **K** - Lysine (Lys)
- **R** - Arginine (Arg)
- **Q** - Glutamine (Gln)
- **N** - Asparagine (Asn)
- **E** - Glutamic Acid (Glu)
- **D** - Aspartic Acid (Asp)
- **S** - Serine (Ser)
- **T** - Threonine (Thr)

4.

For example, Ham Smith is a beloved Nobel Laureate who discovered restriction enzymes thereby helping to give birth to biotechnology as we know it. Ham's name translates pretty easily into the single letter amino acid code: HAMSMITH, where H stands for histidine, A stands for alanine, and so on.

5. A ha! Once you have this mapping, you can then lookup the DNA sequence that would code for HAMSMITH. Since the process of reading from nucleic acids to proteins is called translation, this new process of going backwards from amino acid sequences to nucleic acid sequence is called "reverse translation" (i.e., going back from amino acids to nucleic acids). There are many tools online that will perform reverse translation for you. Look around.

6. By hand or using these tools, you can find out that catgcgatgagcatgattacccat is the DNA sequence that will code for HAMSMITH.

7. Now, you might ask a few questions at this point. What if there is a mutation? Or, what else could I write? Or, aren't there better codes?

8. As one example, a single point mutation to Ham's signature, cctgcgatgagcatgattacccat,gives the DNA sequence that will code for PAMSMITH, a well known illustrator or Tarot cards from the early 20th century, according to Wikipedia.

9. Hmm. I'll let you think about the other questions (i.e., better codes). Also, can you think of other messages people might want to encode in DNA?

*B. <u>DNA Refinement</u>*

10. If we hop back to the standard genetic code, we can pay more attention to the fact that there is a "degeneracy" in the code. That is, often times more than one codon will code for a single amino acid. For example, there are six codons that code for serine. As a second example, there are three stop codons. Why? And, does it matter which codon is used (either in a natural piece of DNA, or in something that we engineer)?

11. When there is a choice of codon (for a particular amino acid) researchers have found that there is a preferred frequency of codon usage. Different organisms often tend to prefer different codons. You can analyze a natural genome and determine that organisms preferred codon usage. See this website for one interesting online resource: http://www.kazusa.or.jp/codon/

12. Codon usage can be related to tRNA abundances inside cells. tRNAs are the molecules that carry the amino acids into the ribosome, as the ribosome is translating a messenger RNA into a protein. If a codon is used frequently, then likely the tRNA levels inside the cell should be higher. However, cells can regulate the process and rates of translation by using rare codons, or limiting tRNA levels, et cetera.

13. Nevertheless, the fact that the genetic code is degenerate (for protein coding sequences, at least) gives us some options as would-be engineers of biology. For example, if we need to make a minor change to the coding sequence of a protein, we can likely change the codons in order to erase or add a feature. If we are lucky and good engineers, the codon changes will be "silent," meaning that they won't lead to

a change in the amino acid sequence of the protein. This will become important when we go to standardize DNA parts (later this term).

14. As a quick example, if you found a sequence of DNA that contained GAATTC, this would be a site recognized by a restriction enzyme called EcoRI. If this sequence was in a protein it would could for glutamic acid and phenylalanine. If you wanted to get rid of the EcoRI site you might consider changing the sequence to GAGTTC, which would still code for the same two proteins, but would no longer be recognized by EcoRI!

*C. <u>New DNA Codes</u>*

15. Look again at the genetic code. Do you think that the code itself is optimized for something in particular? That is, has evolution selected the standard genetic code to be good at anything in particular?

16. What is you wanted to design a new genetic code that was optimized for different properties?

17. Could you make a genetic code in which every individual point mutation (i.e., every single letter genetic change) would be selected against?