BIOL 213 Genetics (15 September 2000)
The Genetic Code

I'm stressing two things heavily. First, there is a big emphasis on **experiments** and how to evaluate them. It is natural to think that we could go faster if we focused more on the conclusions rather than where they came from. I think this is a short-sighted view. In ten years most of you will have little interest in how many genes it takes to make a protein. All of you, however, will be called upon to make decisions and to evaluate the work of others so you can decide how much you can rely on it. You need to be able to make that decision yourself, learning from expert opinion, but not relying on it. To my mind, the chief characteristic of a good liberal arts institution is the help it gives to students to make the transition to a lifetime of independent thought.

The second thing I'm stressing is **visualization**. Our power of visual intuition is generally much greater than our power of verbal intuition. If you see something in your mind's eye, you're much more apt to avoid the nonsensical inferences that often result from purely verbal reasoning. I want to encourage you to get into the habit of trying to visualize each concept and problem that comes your way in this course. One way of doing so is to simplify problems to a point that you can visualize them. Another is to create analogies that you can visualize and seek insight from them regarding the problem at hand. Scientists -- even quantum physicists -- think visually.

Outline:

I. Review and Introduction to the Coding Problem

II. The Coding Problem - General Considerations (pp.326)
- A. How many bases?
- B. Overlapping vs Nonoverlapping

III. The Decision: What Kind of Code?
- A. Overview of frameshift experiment (Crick et al) (p.152-155)
- B. Experiment: Frameshift mutagenesis of phage T4 *rII* gene

IV. Cracking the Code (pp.328-332)
- A. Experiment: Translation of Homopolymers (Nirenberg & Matthaei)
- B. Experiment: Translation of Random RNAs (pp.328-332)

V. The Genetic Code (pp.326-328)
- A. How to read the code
- B. Important features of the code

I. Review And Introduction To The Coding Problem

Let's summarize the semester up to now:

· What determines the form and function of a cell?
Protein, mostly through their activity as enzymes.

· What is a protein?
A protein is a linear array of amino acids, formed through the interactions of the amino acids into a three-dimensional structure.

· What determines the form and function of a protein?
The order and type of amino acids.

· What determines the order and type of amino acids of a protein?

- The eye transplantation experiments of Beadle and Ephrussi (see problem set question 3.8) and later work with Neurospora mutants (Beadle and Tatum) led to the idea that the primary function of a single gene is to specify a single protein -- the one-gene-one-enzyme hypothesis.
- Of course, we now know that proteins can have quaternary structure, consisting of multiple polypeptides that may be specified by different genes, so a more modern statement of the hypothesis is one-gene- one-polypeptide.
- Other work systematically showed a one-to-one correspondence between mutations in DNA and mutations in protein. The order of mutations in DNA corresponded to the order of mutations in the protein.

• What is DNA?
DNA is a linear array of nucleotides.

THEREFORE, the active information in protein, a linear array of amino acids, is determined by the passive information in DNA, a linear array of nucleotides.

We are now confronted with what was, perhaps the central mystery of genetics: how does the inert genetic information within DNA determine the active information in protein and the form and function of a cell? This problem was partially solved in the early 1960's, constituting one of the high points of scientific achievement.

Q1. What structures of DNA and protein led to the realization that there must be a code that related one to the other?

II. The Coding Problem - General Considerations

II.A. How many bases? (p.326)

By the late 1950's, then, it was clear that we needed to understand how the string of bases or basepairs that comprise a DNA molecule can determine the string of amino acids that comprise a protein (Fig. 1). Once the string of amino acids was synthesized, the protein would fold spontaneously into its mature shape, which would determine its function.

How is the information of DNA decoded into the information of protein? The facts of life are these: DNA is composed of FOUR bases while protein is composed of TWENTY amino acids. Two types of codes were considered:

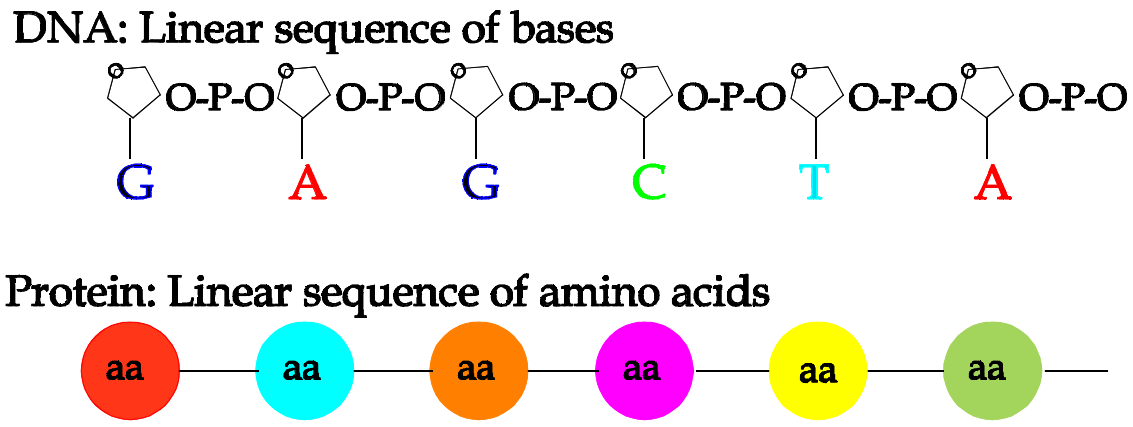


Fig. 1: Problem of how to encode 20 amino acids by 4 bases. Each nucleotide consists of deoxyribose (pentagon) attached to a base (G, A, T, or C). The nucleotides are connected by phosphate to form DNA. Each amino acid residue (colored ball) is connected by a peptide bond to form a protein.

VARIABLE LENGTH CODE

Like written English, where the set of 27 characters (26 letters plus space) encodes thousands of words.

FIXED LENGTH CODE (POSITIONAL)

The unit of information is always a set length. You sometimes see dates written as a fixed code. September 15, 2000 could be written 000915: the first two digits contain the year, the next two the month, and the final two the day.

The first codes imagined in the 1950's were fixed length codes. Suppose the code consisted of single bases specifying single amino acids. Since there are only four bases (A, G, C, T), only four amino acids could be specified. This is too few, so such a code is insufficient to explain how genes specify actual proteins.

Q2. Suppose the code were a doublet code (two adjacent bases specify one amino acid). How many different amino acids could this code specify?

Q3. What is the minimum number of bases required by a code to specify the number of different amino acids that proteins actually contain?

II.B. Overlapping vs. Nonoverlapping

Of fixed length codes, two types can be envisioned, overlapping and nonoverlapping. Take a look at the sequence in Figure 2. Note that the nonoverlapping code considers three bases at a time, then chugs on to the next three, while the overlapping code moves only one base at a time. Each triplet codon adds a single amino acid to the growing polypeptide chain. They differ only in how the RNA is decoded. A machine that read an English nonoverlapping triplet code would interpret the sequence CATEAR as CAT-EAR, but a machine that read an overlapping code would interpret it as CAT-ATE-TEA-EAR. (By the way, it's real hard writing English as an overlapping code. Try it!)

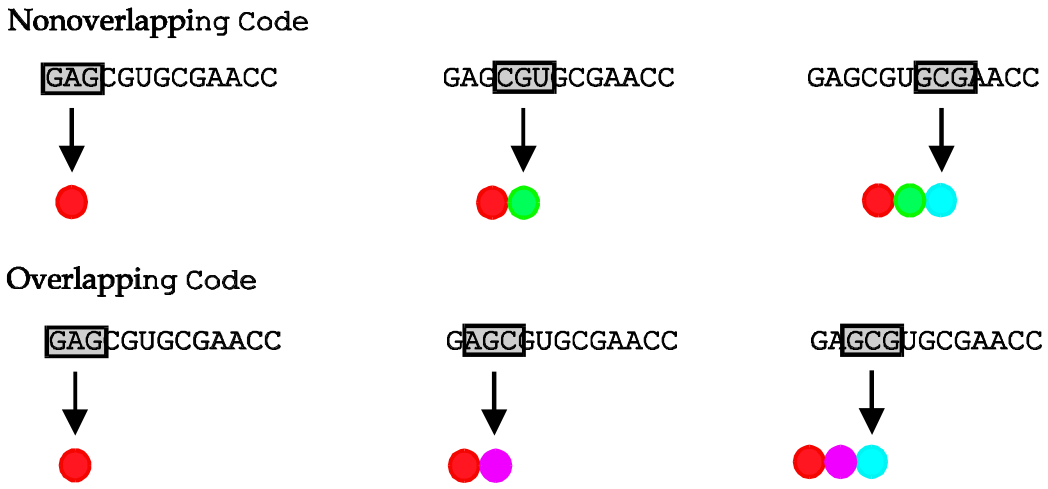


Figure 2: Translation of RNA sequence according to a nonoverlapping or overlapping code.

Q4. Suppose that the third base (G) of the RNA sequence shown in Fig. 2 were mutated. How many amino acids would be affected if the code were nonoverlapping triplet? How many if the code were overlapping triplet?

III. The Decision: What Kind Of Code?

III.A. Overview of frameshift experiment (Crick et al)

Variable vs fixed? Overlapping vs nonoverlapping? Triplet vs who-knows-what? In one remarkable experiment executed by Francis Crick, Leslie Barnet, Sydney Brenner, and RJ Watts-Tobin, these questions were largely resolved. This is a confusing experiment but a very important one. If you can grasp what they did, how they did it, and what they found, then you've gone a long way in understanding the heart of the coding problem and its solution. They had available to them the following tools:

- BACTERIOPHAGE T4: A virus (also called phage) that infects *E. coli.*
- *rII*: A region of T4 DNA consisting of two genes that are essential for the infection of *E. coli* by T4 under certain conditions.
- PROFLAVIN: a mutagen, i.e. something that alters DNA to cause mutations.

Let's examine each tool in turn.

III.A.1. Bacteriophage T4 (pp.152-155)

You have already seen in Lab 1 how a single *E. coli* cell can be detected by its growth into a colony on a plate. In a similar way, you can detect a single virus by its ability to propagate within a population of bacteria, reproducing itself along the way. Fig. 6-13 illustrates this process. A single phage infects a single cell and kills it. That event would be undetectable – how could you find the one dead cell amongst the billion live ones? However, before the cell died, the phage hijacked its metabolism to create several tens of progeny phage. These phage that are released upon the death of the cell infect adjacent cells, and the cycle is repeated.

Within hours, millions of phage have been produced, killing millions of cells. This can be seen, as a clearing on a plate that was covered with *E. coli*.

III.A.2. The *rII* region of phage T4 (p.155)

As you know, plasmids often confer on their hosts some selective advantage, e.g. resistance to certain antibiotics. Some viruses also make themselves enticing. The bacteriophage λ (lambda) does not always kill *E. coli* but sometimes integrates into *E. coli* DNA and lives there within its host. *E. coli* carrying integrated λ are resistant to infection by some phage. Phage T4 has learned how to overcome this defense. We don't understand how λ protects *E. coli* nor how T4 circumvents the problem, but two T4 genes, *rIIA* and *rIIB* are required for T4 to infect cells already occupied by λ. T4 phage that have suffered mutations in the *rII* region are no longer able to infect *E. coli* (λ), but they have no problem infecting *E. coli* lacking phage λ.

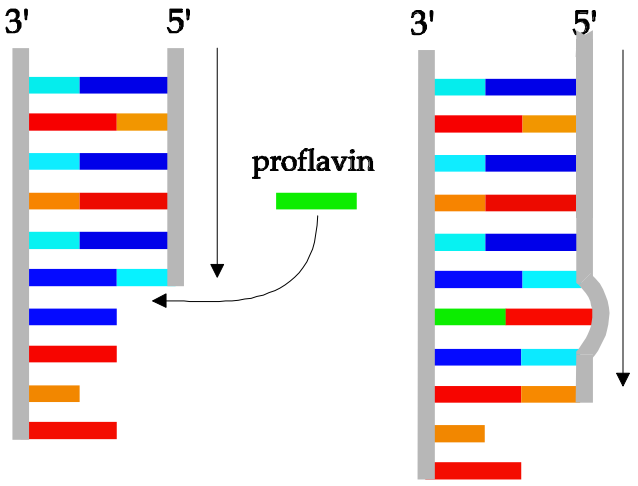
Mutations in the *rII* region has a second effect. For an unknown reason, T4 phage with a mutant *rII* region kill *E. coli* more rapidly than wild-type (normal) T4. In fact, the "*r*" in "*r*II" stands for *r*apid lysis (or killing). The characteristics of T4 with respect to the *rII* region are shown in Fig. 6-14 and summarized in Table I below.

Table 1: Growth of Phage T4 on *E. coli*

Host Strain[a]	Wildtype T4	*rII*⁻ T4
E. coli B	normal plaques	large plaques
E. coli K(λ)	normal plaques	no plaques

[a]*E. coli* B and K are two strains, whose differences need not concern us. Crick et al used *E. coli* B as a strain that *rII*⁻ T4 could infect (because it did not carry an integrated λ phage). The *E. coli* K strain they used happened to carry an integrated λ phage, so could not be infected by *rII*⁻ T4.

III.A.3. Proflavin



Proflavin and similar mutagens work by mimicking the shape of base pairs. It is the same length as a G-C or A-T base pair, and like them, proflavin is flat. It can therefore insert itself between the bases of DNA, thereby causing single base insertions or deletions. Still confused?

Proflavin put between two bases confuses DNA polymerase, and as a result, a single extra nucleotide is inserted. It can also cause deletions if it intercalates between bases of the newly synthesized strand. Both kinds of mutations -- one-base insertions and one1-base deletions would cause a shift in the reading frame if the code were nonoverlapping, causing all succeeding triplets to be misread. This is explained further below.

Q6. Reconsider the nature of overlapping and nonoverlapping codes, as shown in Fig. 2. How would the amino acid sequence change if proflavin caused an insertion of a base in a sequence read by an overlapping code? How would it change if the sequence were read by a nonoverlapping code?

III.B. Experiment: Frameshift mutagenesis of phage T4 *rII* gene

In this remarkable experiment, Crick et al showed that an insertion or deletion of one base had a lethal effect on the *rII* region, consistent with a nonoverlapping code and inconsistent with an overlapping code. They showed further that while two insertions or deletions were just as bad as one, THREE insertions or deletions brought the gene back to normal. How can three wrongs make a right?

The only explanation that made sense was that the genetic code was a nonoverlapping triplet (or multiple of three). Their results are represented in Fig. 3.

Figure 3. English analogy of frameshift experiment of Crick et al. Original sequence, read as nonoverlapping triplets, produces a good rII protein, enabling phage T4 to infect *E. coli* K(λ). An insertion of a base (R) into the sequence by the action of proflavin ruins not only the triplet codon it is part of but all succeeding codons, by shifting the reading frame. As a result, only the first part of the protein is normal. Following the point of mutation, the amino acids encoded by the triplets are more or less random. A second nearby insertion (another R) does not help: the message past the mutation is still nonsense. A third insertion, however, (an O), restores the reading frame to normal. There are still some mistaken amino acids, but the large majority of amino acids are as in the wild-type protein. Phage T4 is again able to infect *E. coli* K(λ).

Note that the suppression does not produce complete sense, but it's a whole lot better than the alternative.

Q7. Suppression by secondary proflavin-induced mutations does not generally produce a wild-type protein (one with all the original amino acids intact). What considerations govern whether such a protein will be good enough? Think about the fraction of the protein that is changed (how big is a typical protein?) and the position of the altered amino acids within the protein.

Q8. What would have been the results of the frameshift experiment if the genetic code were not based on fixed length triplet codons but instead on variable length codons? Imagine that cytosines represent spaces that separate codons.

We won't go through the experiment in detail, but it is useful to see how Crick et al could have looked for the rare triple insertion mutants. Mutants with a defective *rII* region by reason of an insertion or deletion are not able to infect *E. coli* K(λ) productively. Triple insertion mutants with adequate *rII* region ARE able to infect *E. coli* K(λ) productively.

Q9. Crick et al added proflavin to a phage stock that already had two insertions close to each other in the *rII* region. Suppose that once every 100 million phage, the proflavin causes an insertion close by to the other two, while in the remaining phage, proflavin causes a mutation elsewhere or, more likely has no effect at all. A billion of the treated phage are used to infect *E. coli* K(l) and the mixture is spread on a plate (see Fig. 6-14). What will you see on the resulting plate?

IV. Cracking The Code

IV.A. Experiment: Translation of Homopolymers (Nirenberg & Matthaei)

The frameshift experiment gave the overall nature of the code but did nothing to tell us which specific triplet codons coded for which specific amino acids. Indeed, Crick despaired of cracking the code, knowing of no way to proceed. Ironically, the same year the frameshift experiment was completed, little known Marshall Nirenberg and Heinrich Matthaei presented their work at one of the first international meetings devoted to molecular biology, work that served as the basis for the ultimate cracking of the code. Because they were outside the inner circle (and because Nirenberg chose a hopelessly boring title for his talk), few attended the talk, but Matthew Meselson (the graduate student half of the Meselson-Stahl experiment) did show up and upon realizing the import of what he had heard, prevailed upon the meeting organizers to let Nirenberg deliver his talk a second time, this time to a packed house.

The key element of Nirenberg and Matthaei's work was that they found unusual conditions in vitro (in a test tube) under which ribosomes would translate artificial RNA. (By the way, it is important to realize that the experiment worked only because Nirenberg and Matthaei inadvertently tortured the ribosomes. In cells, ribosomes will not translate RNA unless it contains an appropriate start signal.) Synthetic RNA composed solely of uracils (UUUUUUU...) was translated in their system to a polypeptide composed solely of phenylalanines (phe-phe-phe-...). Combining this result with that of the frameshift experiment led to the conclusion that UUU encoded phe, the first codon assignment.

Q10. What modification of this experiment immediately suggests itself to find the assignments of three other codons?

IV.B. Experiment: Translation of Random RNAs (pp.328-332)

You can only go so far with homopolymeric RNA (i.e., RNA composed of just one kind of nucleotide). To go further, two labs (Nirenberg's and Severo Ochoa's) looked at the translation of several synthetic polymers that had random sequences but known RATIOS of bases. For example, one polymer consisted solely of G's and U's in a ration of G:U = 70%:30%. It might look like GGUGGGGUUGGGUGGGUGUG.... Translation of such an RNA would not give a defined polypeptide, but the RATIOS of the amino acids in the polypeptides would be informative.

The table shown on p.329 shows more precisely the number of each kind of triplet you would expect from a GU polymer with 70% G's.

Q11. How many possible triplets are there composed solely of G's and U's? Why this number?

Q12. Would you expect that all triplets occur at the same frequency? Why (not)?

Q13. Why is the probability of the occurrence of GGU given as 0.7 x 0.7 x 0.3?

Q14. Some triplets have the same calculated frequencies some others. Why?

Q15. If a synthetic RNA is made from A and C in the ratio of 9:1, what's the probability of finding an ACC codon?

The experiment in broad outline is simple: Translate a random GU polymer with G:U = 70%:30% and measure the frequencies of the amino acids in the protein product. You can find a description of Nirenberg and Matthaie's experiment on page 330. The protocol had to address several problems:

a. How to translate the artificial RNA polymer?
The polymer couldn't be translated within a cell as no way was known of introducing artifical RNA (or any kind of RNA). The mechanism of translation was unknown, so it was not possible to use characterized components. The solution was to break open cells and rely on them to provide the unknown requirements for translation.

b. How to ensure that only the artificial RNA polymer is translated?
Breaking the cell released not only ribosomes but also the cell's own RNA. Fortunately, RNA is short-lived. Nirenberg and Matthaei simply waited long enough after breaking the cell for the RNA to degrade. To ensure that cellular RNA was not replenished after breaking the cell, they destroyed the cellular DNA.

c. How to focus on one amino acid at a time?
They wanted to know not merely the total number of amino acids incorporated into protein by translation of the artificial RNA. More important than that was the amount of each of the twenty separate amino acids incorporated. They focused on each of the twenty in turn by doing the experiment twenty times, each time radioactively labeling one amino acid.

d. How to measure the amount of amino acid incorporation into protein?
Nirenberg and Matthaei took advantage of the fact that proteins are so big that they get caught up in filter paper, while amino acids are small enough to flow through. Therefore, the radioactive amino acid that was not incorporated into protein could be washed away and the radioactive amino acid in protein retained on the paper.

The idealized results of the experiment are shown on page 331. The actual results were not as readily interpreted, as you would expect in an experiment of this type. Note that there were only six reported amino acids that were observed to be incorporated into protein, even though there are eight possible triplet codons composed of G's and U's. The text interprets the results using prior knowledge of the genetic code, information that was of course not available to Nirenberg and Matthaei. Let's look at the data from their perspective. Comparing the tables on p.329 and p.331, you can see straight away that GGG could encode only glycine, the most frequently incorporated amino acid. Likewise, UUU is likely to encode phenylalanine. What can we say about cysteine? It's 6% frequency of incorporation corresponds to that expected from either UUG, UGU, or GUU. It must be encoded by precisely one of these triplets. This kind of experiment is not capable by itself of deducing the genetic code, but by combining the results of experiments with several artificial RNA polymers, much progress was made.

Q16. From the information given in Tables 13-3 and 13-4 and the results of the UGUGUGU... dinucleotide experiment, what can you infer about how leucine is encoded?

V. The Genetic Code (pp.326-328)

From these and other experiments, it was possible to make definitive assignments for 61 out of the 64 possible triplets, and these are shown in Table 12-2. This table is more than a climax to the intense work of an emerging field of biology. It's more like Newton's laws, something that is used everyday to predict and understand the world. It is therefore essential that you understand how to USE the information in the genetic code.

V.A. How to read the code

First, you need to know how to read it. If you want to find the amino acid encoded by the triplet codon AGC, look at the left side for A, the top for G, and that gets you to the box where you can scan for AGC. You should find the amino acid to be serine (ser). A list of the 20 amino acids and their three- and one-letter abbreviations was provided on p.9 of the notes on Protein (first day of class).

Second, note three special codons for which there are no amino acids given. UAG, UAA, and UGA are the STOP codons, marking the end of an encoded polypeptide chain. There is another special codon, AUG. Most but not all genes begin with AUG. It is sometimes called the START codon, though there are genes that start with GUG or UUG. Besides marking the beginning of the protein, AUG is also used to encode the amino acid methionine. In this respect, it is just a normal codon. AUG may be found not only at the beginning but also in the middle of genes. Ribosomes don't start new proteins there because a specific binding site is required for ribosomes, sites that are found only before AUG's that begin genes.

Q17. Starting from the first conventional start codon, translate the RNA strand given below:

GAAGCAUGUCCGAGCAAUGAGCCGA

V.B. Important features of the code

1. The genetic code is degenerate: There are 64 possible triplet codons but only 20 possible codons. Nonetheless, 61 of the 64 codons are assigned amino acids. Most amino acids are encoded by more than one codon. This many-to-one relationship is what defines "degeneracy".

Q18. The three amino acids most commonly found in human protein are leucine, glycine, and serine. The three amino acids least commonly found in human protein are tryptophan, methionine, and histidine. Draw a conclusion about how degeneracy relates to the natural frequencies of amino acids.

2. Not all amino acid changes are possible from a single basepair mutation: Virtually all mutations found in nature are single events: single basepair changes or single insertions or deletions. This fact places a strong limitation on what amino acid changes are observed. For example, leucine, encoded by CUA, can mutate in one of three positions: C in the first position to U, A, or G; U in the second position to C, A, or G; and in the third position to U, C, or G. In principle, then, a one-base change in the codon can lead to any one of nine other codons and any one of nine amino acids. Immediately, one sees that it is impossible with a single base change to get to triplets encoding all 20 amino acids.

Actually, the situation is even more extreme. Of the nine codons listed above, five of them also encode leucine. Changes to one of these are called silent mutations, mutations that do not effect the amino acid sequence of the protein. Two other of the nine codons encode isoleucine or valine, amino acids very similar to leucine. These are called conservative changes. Changes to these amino acids very well may not affect the structure of the protein. Only two of the nine possible changes lead to a hydrophobic-to- hydrophilic amino acid change. The failure of many mutations to produce functional changes in protein is no accident. The genetic code appears to be built with that in mind.

Q19. List the changes that can be produced by a single basepair mutation in the AGA codon encoding arginine and label each silent, conservative, hydrophobic-to-hydrophilic, hydrophilic-to-hydrophobic, or other.