Molecular Biology Through Discovery
Companion to Crick (1958) *On Protein Synthesis*
Symp Soc Exp Biol 12:138-163

Ordinarily, at this stage I would give a review of the understanding of information flow from DNA to proteins on the eve of the discovery of the genetic code. Happily, there is no need for me to do so, since Francis Crick (**Fig. 1**) performed the task in excellent fashion – and without contaminating knowledge of what was to transpire. I will therefore invite you to download and read his address to the Symposium of the Society for Experimental Biology, September 1957 (published a year later). I'll content myself to offer marginal comments and to suggest which areas would repay your closest reading and which might be skipped.

I. Introduction

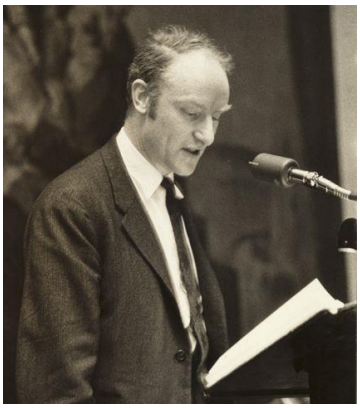


Figure 1: Francis Crick. Lecturing at Cambridge Univ. Courtesy of Stock Photos.

Crick spoke here, as he so often did, to "*…the biologist rather than the biochemist, the general reader rather than the specialist.*" I know of no one who wrote more feature articles for Scientific American (eight between 1954 and 1992) than he did.

The importance of proteins

- You should recognize the main outline of Crick's argument here – it's the same that began our course! Proteins do the work, and the genetic material determines their synthesis.
- (Par.1, line 11): "*…all known enzymes are proteins.*" This was true in 1957, but we now know that certain RNAs (ribozymes) exist that can catalyze reactions. Crick was one of the first to suggest the possibility of catalytic RNA,[1] which was discovered 14 years after his prediction.[2]
- (Par.2, line 1) How the genetic material controls the synthesis of protein makes up a considerable part of molecular biology

SQ1. From your current knowledge of molecular biology, in what ways is the synthesis of proteins controlled by DNA?

- (Par.2, line 2) Crick was certainly correct that there was little evidence at the time for control of protein synthesis by DNA. Nonetheless, he clearly likes the idea. He speaks of "psychological drive", perhaps more important than evidence. What could he mean by that? Perhaps he was thinking about the weight of evidence from the work of Avery, McLeod, and McCarty[3] favoring DNA as the genetic material but not making much headway in the collective consciousness of the time. But when a compelling rationalization for a genetic role of DNA was provided by the Watson-Crick model, then the idea became more palatable, even though there was no more evidence than before.
- (Par.2, line 7) Crick realized that the mechanism of protein synthesis must be the same for all proteins, given that only proteins can make proteins, otherwise it is difficult to escape the requirement for one set of proteins to make the proteins the cell requires, and another set to make the first set, and so on. This problem comes up again later on (p.143). With our present knowledge, you can visualize the process as in **Fig. 2**, with an unchanging machine, the ribosome, acting on a variable source of information, the messenger RNA. The ribosome is a complicated machine, composed of many dozen proteins and large RNAs, but the components are constant and finite.

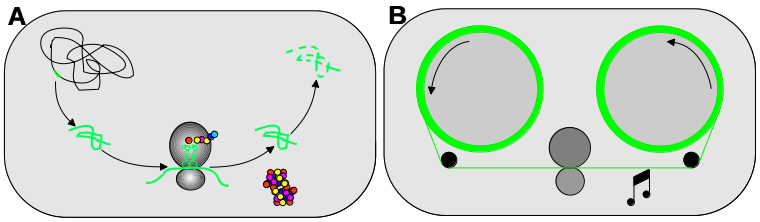


Figure 2: Constancy of mechanism of protein synthesis and variability of message. A. Variable messenger RNA (green) is transcribed, translated to proteins – chains of amino acids (multicolored spheres), by universal ribosomes. **B.** Tape (green) is read by a universal machine that can play any tape, producing music.

- (Par.3) Are proteins the last word in biological molecules as Crick claimed? You can make a strong argument that he was wrong. In the last fifteen years, the importance of a new class of molecules, small regulatory RNA, has become firmly established.

II. The Problem

Elementary facts about proteins (1. Composition)

- (Par.1, line 3) We are told what to expect in a typical protein.

SQ2. From the numbers given, what is the average molecular weight of an amino acid?

SQ3. Does the given typical number of residues correspond to what you observed in *What is a Gene?*

- (Par.2, line 2) The constancy of mechanism (**Fig. 2**) requires that the operation is also unchanging. Fortunately, the backbone of a protein is the same regardless of which amino acids are in it (recall the notes on *Protein* and the formation of the peptide bond), with no dependence at all on the side groups of the amino acids.
- (Par.2, line 4) S-S links? What are they? You'll recall the cysteine-cysteine cross-links from the Sanger and Tuppy companion (Figures 3 and 4).

SQ4. Put this together with your knowledge of the structure of cysteine (no knowledge? …then look it up). What are the S-S links mentioned by Crick?

- (Par.2, line 6) Crick now claims that proteins are linear, by which he means a linear chain of amino acids.

SQ5. Where did he get that idea? What experiment have you read whose results support this claim?

- (Par.3-4) You may think it obvious that proteins are made from only 20 amino acid. It was definitely not obvious at this time,… because it isn't true. There are many amino acids found in proteins besides the 20 you're familiar with. For example, about 10% of the amino acids in human collagen protein are hydroxyprolines.[4] More than 50% of the amino acids in the egg protein phosvitin are phosphoserine.[5] These unusual amino acids are far from contaminants! Nonetheless, Crick and Watson judged them to be exceptional and are present in a few proteins by some unknown, special mechanism that they would do well to ignore. This illustrates an important point that comes up repeatedly in science. At any given moment, not all observations can be brought under the umbrella of a conceivable hypothesis. Progress depends on seeing which observations can be used to hold together a new explanation and which should be put aside for later consideration.

Elementary facts about proteins (2. Homogeneity)

- Here's another assumption that may seem not very controversial. After all, human insulin is human insulin. But recall the review by Dorothy Wrinch you read a few weeks ago and the important role played in her view by micelles. If enzymes were micelles, as was widely believed in the first half of the century, then proteins would ***not*** be homogeneous, any more than a collection of soap bubbles are homogenous.

Elementary facts about proteins (3. Structure)

- (Par.1, line 1) Crick presents another divergence from the world view illustrated by Wrinch's review. Globular proteins are not extended fibers but rather folded structures, as you have seen in the paper by Perutz et al (1965).[6] Although the first complete protein structure wasn't published until 1960,[7,8] more than two years after Crick's talk, he was well aware of developments in the area. After all, Perutz was Crick's PhD thesis advisor, and both of them worked at the MRC in Cambridge.
- (Par.1, line 7) You're already familiar with proteins denaturing due to heat – recall the question in Problem Set 3 concerning cooked egg whites.

Elementary facts about proteins (Sections 4 and 5) (I'll skip these)

The nature of protein synthesis

- Dounce provides a clear statement of the problem alluded to in the Introduction. What kind of machine can synthesize the astonishing diversity of proteins, including the proteins of the machine itself?

SQ6. With your superior knowledge of molecular biology, how would you respond to Dounce?

The essence of the problem

- (Par. 2) Crick speaks of three flows but focuses on the last – the flow of information – which became the central problem of molecular biology. He encapsulates the problem in a form that will be called (in a few pages) the Sequence Hypothesis, which states:

  > *The structure, and therefore the function, of a protein is determined solely by the order of its amino acids.*

  In a bit that will be followed by:

  > *…and the order of a protein's amino acids is determined solely by the order of nucleotides in the gene that encodes it.*

  How this comes about, both in theory and in practice, motivates a great deal of work in the decades that followed the discovery of the double helix.
- (Par. 2, line 8) He is right in speculating that there are exceptions to the rule that proteins spontaneously assume their structure – those proteins requiring intervention by chaperones.[9]

III. Recent Experimental Work

The role of the nucleic acids

- (Par.1) Crick cites work with which you may be familiar -- the transforming principle of Avery, McLeod, and McCarty[*] -- and recent work by Seymour Benzer on the topology of the rII gene.

[*] He cites a more recent article by Rollin Hotchkiss' group. Hotchkiss worked with Avery and continued the study of transformation after Avery's retirement.

I'm going to skip now all the way to p.152, omitting the discussion on what was known about the biochemistry of protein synthesis.

IV. Ideas about protein synthesis

The Sequence Hypothesis

- The main virtue of the hypothesis was how it simplified the problem. Don't worry about the bewildering variety of protein structures and functions. Just get the DNA sequence right and the rest will take care of itself.

The Central Dogma

- The Central Dogma is more familiar to most, probably because of the catchy name. Some have misrepresented it as the flow of information: DNA → RNA → protein. If that were how it worked, then the discovery of RNA viruses such as HIV that reverse transcribe their RNA to DNA to replicate themselves would refute the model.

SQ6. Make up a scheme in which the sequence hypothesis is false but DNA still determines the structure of protein.

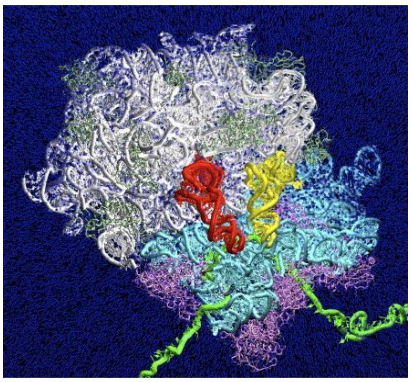
SQ7. Imagine an exception to the Central Dogma.

Some ideas on the cytoplasmic protein synthesis

- (Par.1, line 1). The idea that there must be RNA templates in the cytoplasm stems from the observation that protein synthesis takes place in the cytoplasm, with little or none in the nucleus (reviewed in the pages I skipped over) and that RNA contains the information necessary to encode the protein (the Sequence Hypothesis). Again, "microsomal particles" refers to what we would now call "ribosomes".

- (Par.2) Read this section carefully, as it represents well the views held at the time and the resulting confusion regarding the role of RNA in protein synthesis. But before we get into that, it might help to go through a short review of ribosomes and protein synthesis.

  Ribosomes are composed of two assemblages of proteins, the large and small subunits, each consisting of dozens of proteins. Each subunit is assembled on an RNA scaffold, the large subunit containing 23S RNA (28S RNA in eukaryotes) and the small subunit containing 16S RNA (18S in eukaryotes).[†] The full ribosome reads messenger RNA (mRNA) that binds to it, using transfer RNAs (tRNAs) to connect the triplet codons on the mRNA to the amino acid to be added to the growing peptide chain. **Fig. 3** shows the ribosome at work, both in a view (**A**) that emphasizes the various RNA components, and a more abstract view (**B**) that emphasizes the physiological role of the ribosome.



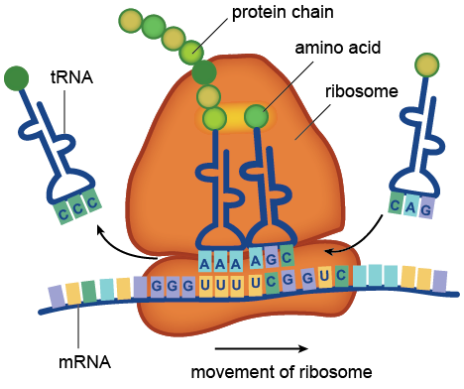


Figure 3. The ribosome. (**A**) View showing the RNA components: white 23S rRNA, blue 16S rRNA, green mRNA, and two tRNAs (red and yellow).[14] (**B**) View showing the functioning of the ribosome, with an amino acid (green ball) attached to a tRNA about to be linked to the growing protein chain. (From www.shmoop.com/biology-cells/structure-function.html)

[†] The large subunit also contains smaller RNAs. 'S' stands for 'Svedberg unit', which is a measure of how fast the molecule moves when centrifuged.

SQ8. Provide labels in Fig. 3B for all the RNA, indicating 5' and 3' ends.

SQ9. Provide labels in Fig. 3B indicating carboxy and amino termini.

Ok that's enough current events. Back to Crick.

SQ10. Compare Crick's view of the role of RNA with your current, presumably superior view.

- (Par.2, line 10) The idea that ribosomes are like RNA viruses, i.e. protein shells housing RNA (and that viruses are renegade ribosomes) shackled Crick's mind during the 1950's and, as we shall see later, impeded the discovery of mRNA.
- (Pars.5-6) A study of the structure of DNA by Watson and Crick paid big dividends, and expecting a similar reward, Watson spent a good deal of his time after leaving Cambridge trying to elucidate the structure of RNA. The results were disappointing,[10] because except for small RNAs (like tRNA) and small regions where RNA basepairs with either itself or another RNA, it doesn't have a regular structure.

The adaptor hypothesis

- (Par.1, line 3) Crick refers to a specific class of theories that explain how the order of nucleotides in RNA determine the order of amino acids, the most famous of which being the diamond code of George Gamow.[11] This is the subject of a problem on a problem set.
- (Par.1, line 8) He can't imagine any way that amino acids can bind directly to RNA, hence no way that the RNA can directly determine the order of amino acids. However, it takes little imagination to see how RNA can bind to RNA, by the same basepairing that joins the two DNA strands of the double helix. Therefore, he postulates the existence of a small adaptor RNA that can bind to the information-laden RNA on one side and to the appropriate amino acid on the other.
- (Par.3) But the concept of an adapter merely kicks the core question down the road. How is the appropriate amino acid attached to the adapter with the appropriate RNA? Here Crick imagines that each of the 20 amino acids must use a different enzyme that is clever enough to recognize both the amino acid and the specific adapter. His imagination was right on the mark, as we now know that specific aminoacyl-tRNA synthetases do precisely what Crick saw to be necessary.

SQ11. Of all the components of translation mentioned thus far, which one *knows* the genetic code? That is which one is analogous to a code book where you look up a word in the secret code and find its meaning?

The soluble RNA

- 'Soluble RNA' (which we would now call tRNA) stands in contrast to 'microsomal RNA' (which we would now call rRNA). The latter is not soluble because it's bound to ribosomes. Let's take a step back and look at each type of RNA, according to our present day knowledge.

Table 1 shows the characteristics of different classes of RNA. Take a look at it.

SQ12. What fraction of all RNA is composed of what Crick called microsomal RNA? What fraction is soluble?

It is perhaps understandable that in the 1950's RNA was considered to be in one of these two classes. mRNA, a small fraction of the whole, was unknown at the time.

SQ13. Why is it that almost all of the RNA in the cell is rRNA or tRNA? Draw on an analogy between a ribosome and a factory that makes vintage cars to order.

Back to Crick.

Table 1: Comparison of classes of RNA in a typical bacterium (*E. coli*)

Class	Types	Fraction	Sizes	Stability	Function
Ribosomal RNA (rRNA)	3	80%	2904 bp, 1542 bp, 120 bp	Stable	Scaffold for ribosomal proteins. Participates in translation.
Transfer RNA (tRNA)	37	12%	76 bp to 91 bp	Stable	Connects codon to amino acid in translation.
Messenger RNA (mRNA)	1000's	5%	~300 bp to ~10,000 bp	Unstable	Carries information on amino acid sequence of specific protein.
Other RNA	???	???	<400 bp	variable	Other structural RNA; regulatory RNA

- (Pars.1-2) Crick couldn't fathom that adaptor RNA could be longer than three nucleotides.

  SQ14. Why did he like three? What was his problem with longer than three?

Subsequent steps and ***Two types of RNA***

- It may be difficult for you to grasp what Crick is proposing since it is so utterly different from what we believe today to be true. I certainly have a difficult time with it. Here's what I think is his model.
  - A microsomal particle (ribosome) contains the RNA needed to encode a specific protein
  - That RNA is replicated (just as viral RNA is somehow replicated)
  - Some of the RNA is broken down into triplet codons, and each triplet is attached to the appropriate amino acid, forming adaptors. You therefore get the adaptors you need in the right proportions.
  - The adaptors turn around and bind to the RNA, thereby ordering the attached amino acids in the proper order.

  SQ15. Anything missing from this scenario? Other problems you see with it?

The coding problem

- (Par.1) Length How can a set of four nucleotides (A, C, G, T) be used to specify a set of 20 amino acids (Ala, Cys,…Tyr, Val)? This sounds like a theoretical problem, and for the 1950's it was.

  SQ16. The simplest code is that one nucleotide specifies one amino acid. If this code is used, how many amino acids can be specified?

  SQ17. Suppose that two nucleotides can be used to specify one amino acid. Now how many amino acids can be specified? Three nucleotides?

  SQ18. The Morse Code uses only two symbols (short, long), yet it can specify each of the 26 letters of the English alphabet. How many symbols do you predict must be used to represent one letter?

  SQ19. In fact, no letter is represented by more than four short/long symbols. How is this possible? (Hint: I lied about the number of symbols)

- (Par.2, line 5) Crick makes what sounds like an random quip how cosmologists have no compunction about making theories without facts. This was a joke, a reference to George Gamow (Fig. 4), who constructed the first proposed genetic code. Gamow is known for many things in his action-packed life, but one of them is fleshing out the Big Bang hypothesis to explain how nuclei heavier than hydrogen arose at the birth of the universe.

Figure 4. George Gamow. Theoretical physicist, cosmologist, popular science writer, inventor of first genetic code

- (Par.3 and Fig. 1) Overlapping: Crick's Fig. 1 shows how a sequence of letters might be interpreted differently depending on whether a triplet code is overlapping, partially overlapping, or non-overlapping.

  SQ20. Try it out with less arbitrary English letters. How would you interpret T-E-A-T-E-N-D as a triplet code that is either overlapping, partially overlapping, or non-overlapping?

- (Par.3) Degeneracy: More than one codon determining the same amino acid…

  SQ21. Of course the genetic code we're familiar with was unknown when Crick wrote this article. But look at it (e.g. on the course web site, Resources and Links). How would you characterize it with respect to length, overlapping, and degeneracy?

  SQ22. How could you modify our genetic code so that it is the same as in reality, except that it is not degenerate?

- (Par.4, line 3) "*It is quite easy to disprove Gamow's code…*" And you will do it, in Problem Set 5.

- (Par.5, line 1) Why is Crick concerned with discovering whether there are restrictions in allowed amino acid sequences? By allowed sequences, he considers the possibility that, say, the amino acids phenylalanine and cysteine never occur next to each other in that order. This is important, because overlapping codes set severe restrictions on what codons can lie next to each other, and this leads to restrictions on amino acid adjacency. If you can show that in nature any amino acid can lie next to any other amino acid, then triplet overlapping codes immediately are off the table.

  SQ23. Consider the codon AGC in an overlapping triplet code. How many possible codons can follow it? If AGC encoded proline, how many amino acids could follow proline?

  SQ24. How many amino acid-amino acid pairs are possible in proteins?

- (Par.6, line 6) The focus on the coding ratio is symptomatic of an idea that influenced thought in the 1950's. In 1946, in the first major biology meeting after World War II (hence widely attended) William Astbury pointed the curious coincidence revealed by X-ray crystallography that the spacing between nucleotides in DNA and the spacing of amino acids in extended protein is nearly the same. You'll recall (Notes for *DNA Structure*) that the spacing between nucleotides is 3.4 Å. **Fig. 5** shows in graphical terms the structure that gave rise to Asbury's figure for amino acid spacing.

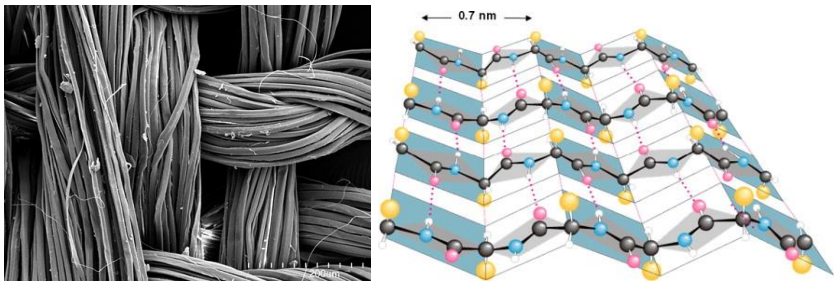


Figure 5. Fibrous protein. (**A**) Fibers of woven silk (scanning electron micrograph from vartest.com/date/2011/09/). (**B**) beta pleated sheets structure typical of many fibrous proteins (Irving Geis[15]). Note that the 0.7 nm (= 7 Å) is the distance between the R groups of *two* amino acid residues, so the amino acid spacing is half that number.

  SQ25. From the equivalence in spacing, it would seem that the coding ratio should be one. What's the argument in favor of this ratio? Why would the ratio seem to favor overlapping codes? How did Crick dispense with the argument?

- (Par.7 to end of section) Crick poses two problems: (1) Nonoverlapping triplet codes have too many codons for the number of amino acids, (2) How do you know where to begin translating the message?

  SQ26. Take the first problem. Why does Crick consider the number of triplet codes to be too many?

SQ27. Take the second problem. How in fact is this problem solved?

But this is 1957, and Crick doesn't know all the molecular biology you know, so instead he comes up with an ingenious solution,[12] by postulating that the sequence of codons establish where the gene begins and ends, by ensuring that if you start at the wrong place, you'll read nonsense, i.e. codons that have not been assigned any amino acid. In such a system, AAA could not be a codon, because then two such codons side by side, AAAAAA, would be ambiguous: which AAA should you read? If you read the wrong one, then the next codon may not be next to the AAA codon.

SQ28. If AGA is a legal codon, then what two codons become illegal? [Hint: try the same trick of putting two identical codons next to each other]

SQ29. Is BCCBDDABBACA a legal message in the example code Crick shows in the middle of p.160? How about ACBDDABCCABB?

SQ30. Does Crick's code carry with it restrictions on what amino acids may be next to each other?

SQ31. Many were blown away by the elegance of this code.[13] What characteristics made it so appealing?

References

1. Crick FHC (1968). The origin of the genetic code. J Mol Biol 38:367-379.
2. Gilbert W (1986). Origin of life: The RNA world. Nature 319:618.
3. Avery OT, MacLeod CM, McCarty M (1944). Studies on the chemical nature of the substance inducing transformation of Pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. J Exp Med 79:137-158. (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2135445/)
4. Pine EK, Holland JF (1965). Heterogeneity in the composition of human collagen. Arch Biochem Biophys 115:95-101.
5. Rosenstein RW, Taborsky G (1970). Nonphosphorylated sereine residues in phosvitin. Biochem 9:658-659.
6. Perutz MF, Kendrew JC, Watson HC (1965). Structure and function of hemoglobin. II. Some relations between polypeptide chain configuration and amino acid sequence. J Mol Biol 13:669-678.
7. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North ACT (1960). Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5 Å resolution. Nature 185:416-422.
8. Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Davies DR, Phillips DC, Shore VC (1960). Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å resolution. Nature 185:422-427.
9. Hartl FU, Bracher A, Hayer-Hartl M (2011). Molecular chaperones in protein folding and proteostasis. Nature 475:324-332. (http://dx.doi.org/10.1038/nature10317)
10. Rich A, Watson JD (1954). Some relations between DNA and RNA. Proc Natl Acad Sci USA 40:759-764.
11. Gamow G (1954). Possible relation between deoxyribonucleic acid and protein structures. Nature 173:318.
12. Crick FHC, Griffith JS, Orgel LE (1957). Codes without commas. Proc Natl Acad Sci USA 43:416-421.
13. Judson HF (1996). *The Eighth Day of Creation*. Cold Spring Harbor Laboratory Press. p.318.
14. Sanbonmatsu KY, Tung CS (2007). High performance computing in biology: Multimillion atom simulations of nanoscale systems. J Struct Biol 157:470-180.
15 Chen P-Y, McKittrick J, Meyers MA (2012). Biological materials: Functional adaptations and bioinspired designs. Prog Materials Sci 57:1492-1704.