ADVANCES IN GENETICS

VOLUME 12

Edited by

E. W. CASPARI

Biological Laboratories
The University of Rochester
Rochester, New York

and

J. M. THODAY

Department of Genetics
University of Cambridge
Cambridge, England

1964

ACADEMIC PRESS • NEW YORK AND LONDON

ACADEMIC PRESS INC.
111 FIFTH AVENUE
NEW YORK, NEW YORK 10003

United Kingdom Edition
PUBLISHED BY
ACADEMIC PRESS INC, (LONDON) LTD.
BERKELEY SQUARE HOUSE, LONDON W.1

Library of Congress Catalog Card Number: 47-30313

PRINTED IN THE UNITED STATES OF AMERICA

ADVANCES IN GENETICS

VOLUME 12

CONTRIBUTORS TO VOLUME 12

ERIC H. DAVIDSON, *The Rockefeller Institute, New York, New York*

VERNE GRANT, *Rancho Santa Ana Botanic Garden, Claremont, California*

FRANK LANNI, *Department of Microbiology, Emory University, Atlanta, Georgia*

FRIEDRICH OEHLKERS, *Institute of Botany, University of Freiburg i.Br., Germany*

ADVANCES IN GENETICS

VOLUME 12

CONTENTS

Cytoplasmic Inheritance in the Genus *Streptocarpus*

FRIEDRICH OEHLKERS

THE BIOLOGICAL CODING PROBLEM*†

Frank Lanni

Department of Microbiology, Emory University, Atlanta, Georgia

* Publication 592 from the Division of Basic Health Sciences. The main body of the review is based on literature and other communications received by mid-January 1963. The Addendum is based on later papers received by mid-August.

† Abbrevations: mRNA = messenger RNA. sRNA = soluble (adaptor) RNA. Hb = hemoglobin. TMV = tobacco mosaic virus. A, C, G, U, T = adenine, cytosine, guanine, uracil, thymine, respectively, or the corresponding nucleosides and nucleotides. I = inosinic acid (used twice). Forms such as ApUpUp. show nucleoside units linked by 3′,5′-phosphodiester groups. Other abbreviations are defined in context.

I. Introduction

Suddenly, in 1961, with the discovery by Nirenberg and Matthaei that synthetic ribonucleotide polymers can function as messengers in polypeptide synthesis, the biological coding problem was brought to bay after almost a decade of frustrated pursuit.

This is not to say that all aspects of the problem have been settled. But the results quickly obtained in the laboratories of Nirenberg (Matthaei *et al.*, 1962; Nirenberg *et al.*, 1963) and Ochoa (Speyer *et al.*, 1962a; Ochoa, 1963) practically assure the validity of a revolutionary idea apparently first stated by Dounce (1952) and independently by Gamow (1954a, b). Having in mind the topologically linear character of polynucleotide and polypeptide chains, and assuming that nucleic acids govern somehow the structure of proteins, Dounce and Gamow proposed that a specific correspondence ("coding relation") exists between particular sets of nucleotides ("code words") and particular amino acids. [Caldwell and Hinshelwood (1950), whose paper recently came to the reviewer's attention, made a similar proposal and should be counted among the coding pioneers.]

The efforts of early cryptographers, led mainly by Gamow, Crick, and their colleagues, to discover general properties of the postulated coding relation and to "break" the code have been reviewed enough (Gamow *et al.*, 1956; Crick, 1958, 1959; Yčas, 1958, 1962; Levinthal, 1959; Chantrenne, 1961; Tavlitzki, 1962; Zubay and Quastler, 1962) and need only brief comment here. The available facts for decoding were very limited: a few amino acid sequence formulas (often incomplete and often, it now appears, erroneous), a few amino acid replacements in homologous proteins from different species, the gross composition of a few nucleic acids and proteins. From such shards, working with much persistence, imagination, and resourcefulness, unsure even that a code existed, the pre-Nirenberg workers devised, tested, and rejected scheme after scheme. The three explicit codes emerging from this work are listed in Table 1. Of these, one (Gamow *et al.*, 1956) was quickly abandoned (Yčas, 1958). The two most recently under consideration (Yčas, 1961;

Woese, 1961a) seem to have been abandoned (Yčas, 1962; Woese, 1962) with the advent of synthetic polymers. (The code of Zubay and Quastler [1962] depends partly on synthetic-polymer data and is listed in Table 1 for convenience. This code will be discussed below.)

Yet there can be no question that the early workers made many sig-

TABLE 1
Early RNA Codes in Comparison with the Primary (U-Rich) Codes of Nirenberg and Ochoa*

Amino acid	Nirenberg-Ochoa	Gamow *et al.* (1956)	Yčas (1961)	Woese (1961a)	Zubay-Quastler (1962)
ala	(UCG)	{AAC}	G	UAG	UCG
arg	(UCG)	{AGG}	G	AGG	UGC
asn	(UAA), (UAC)	{AGU}	U	GAU	UCA
asp	(UAG)	{AGU}	U	GAU	UCA
cys	(UUG)	—	C	UCC	?CG
gln	—	{GGC}	A	UAU	UUA
glu	(UAG)	{AUU}	A	UAU	UUA
gly	(UGG)	{CUU}	A	GAG	UUG
his	(UAC)	—	U	UCU	UGU
ilu	(UUA)	{ACC}	U	CAU	UAC
leu	(UUA), (UUC), (UUG)	{AGC}	A	UCG	UCU
lys	(UAA)	{GCC}	C	CCG	UGA
met	(UAG)	—	C	CUU	UAU
phe	UUU	{GUU}	A	UUG	UUU
pro	(UCC)	{CCU}	C	CCC	UCC
ser	(UUC), (UCG)	{GCU}	U	AAG	UGG
thr	(UAC), (UCC)	{ACU}	C	CAC	UAG
try	(UGG)	—	A	UUC	UAA
tyr	(UUA)	—	G	UUU	?AU
val	(UUG)	{AAU}	G	CAG	UUC

* The amino acid abbreviations are from Yčas (1961); *cys* = cysteine or half-cystine. All the codes refer to messenger RNA (mRNA) or its presumptive equivalent, not to soluble RNA (sRNA). Forms such as (UCG) mean that the internal order is not specified but that only one of the possible permutations is applicable. Forms such as {AAC} mean that all the possible permutations are applicable, i.e., the internal order is irrelevant. Plain forms such as UAG mean that the internal order is fixed relative to a suitable reference within the same set of codes; permutation of the reference code, itself arbitrarily ordered, would permute the entire set.

Gamow *et al.* deduced their codes from the over-all composition of the protein and RNA of two RNA viruses. Yčas used six RNA viruses. Woese deduced a different set of code compositions from the same six viruses and, going a step further, deduced relative internal orders from amino acid replacements in miscellaneous proteins. Zubay and Quastler used the Nirenberg-Ochoa codes UUU for *phe* and (UCC) for *pro* and deduced additional codes and relative permutations from miscellaneous amino acid replacements, leaning heavily on the expected changes A → G and C → U for nitrous acid-induced replacements in tobacco mosaic virus.

nificant and durable contributions. They transformed the vague notions that preceded them into precise and testable ideas, many of which (discussed later) still serve to guide thinking and experimentation. Two of their main decoding techniques—using compositional correlations between nucleic acids and proteins to deduce the composition of codes, and then using amino acid interchanges to deduce something about internal structure—are among the important techniques employed today, but with greatly increased power. Through their sometimes evangelistic zeal, the early workers generated widespread interest and optimism. They recruited everywhere, and soon a diverse band of *aficionados*—cosmologists, plainer physicists, mathematicians, geneticists, biochemists, molecular biologists, etc—was at work (and still is, with the possible exception of cosmologists) on *the* problem. By 1961, when Nirenberg and Matthaei announced their discovery that polyuridylic acid specifically stimulates the incorporation of phenylalanine into peptide linkage in cell-free protein-synthesizing systems, and when they inferred that one or more residues of uridylic acid constitute the code for phenylalanine, the *aficionados* had made all of us ready. We joined in the *olé* heard round the world.

The rapid pace of discoveries with synthetic polymers affects the strategy not only of decoding but also of reviews. By its somewhat central position, the coding problem draws information and ideas from a bewildering array of sources: the molecular structure of nucleic acids and proteins, the biosynthesis of these macromolecules, fine-structure genetics, biochemical genetics, infective heredity, molecular disease, evolution, virology, etc. Limitations of space and of this reviewer prohibit a comprehensive survey of all these aspects of the subject. We shall therefore use the synthetic polymers as a focal point of discussion, leading up to them with a suitable background and proceeding from them along a few selected lines. Fortunately, excellent reviews of collateral topics have appeared recently and no doubt will continue to appear. Wherever suitable reviews are available, we shall cite these in preference to original articles.

II. Fundamental Concepts of Coding

Recent as they are, the Nirenberg-Ochoa codes have already engendered controversy and confusion. Much of the difficulty appears to be semantic. Certain terms (e.g., *coding ratio, code letter, degeneracy, overlapping*) are used in different senses by different workers. Others (e.g., *codon*) have crept into the literature without adequate definition. The more serious trouble comes, however, when various workers urge that the "codes" are triplets and not doublets, or doublets and not

triplets, or a mixture of doublets and triplets, or a mixture of unambiguous doublets (each coding just one amino acid) and ambiguous doublets (each coding more than one). Here, too, analysis suggests that *code* takes a variety of meanings and that the implications of certain proposals are not always fully appreciated.

To prepare for a discussion of these issues, we go back to the remarkably prescient ideas of Dounce and Gamow as a way of introducing the fundamental concepts of coding.

A. Postulates of the Dounce-Gamow Schemes

Dounce (1952) and Gamow (1954a, b) proposed very similar schemes that we shall combine from the two sources and dissect into the following postulates, which were either stated or implied:

1. *One nucleic acid, one polypeptide:* To each "distinct" polypeptide made by a cell there corresponds a "distinct" nucleic acid [section, we might now say, of a nucleic acid megamolecule (Thomas, 1963)] that constitutes the genetic determinant (structural gene, cistron) of the polypeptide.

The term "distinct" needs comment. Both Dounce and Gamow stressed primarily the *sequence* of residues in polypeptides and polynucleotides; and two polypeptides, or two polynucleotides, could be regarded as distinct if they differed in sequence. Then Postulate 1 would imply perfect genetic control of amino acid sequence; i.e., the set of polypeptide chains made under the control of a given allele of a structural gene would be sequentially homogeneous. It seems best to state this as a separate postulate (Postulate 2). Therefore, we shall relax the meaning of "distinct" and group as "one polypeptide" all the chains made under the control of a given allele of a structural gene, regardless of the homogeneity of the polypeptide products. Similarly, "one nucleic acid" may include chains (or duplexes) that differ in sequence but correspond to identical polypeptide chains. Thus, a mutation affecting the sequence of a determinant nucleic acid need not be reflected in the corresponding polypeptide (see Postulate 12, degeneracy). It remains an experimental matter to decide whether all of the genetic information for a polypeptide resides in a single, structurally cohesive unit (the structural gene or cistron; see below).

2. *Sequential homogeneity of polypeptides:* The polypeptide chains produced under the control of a given allele of a structural gene are uniform in amino acid sequence; i.e., the selection of an amino acid residue is completely determinate at each site in the sequence (or the indeterminacy s so small as to be negligible).

3. *Sequence hypothesis:* The specificity of a determinant nucleic acid

resides in its nucleotide sequence, and it is the nucleotide sequence that specifies the amino acid sequence of the corresponding polypeptide.

According to an extreme form of the sequence hypothesis (Crick, 1958) no extra information is needed to guide the specific folding of a finished polypeptide chain.

4. *Translation on templates:* The specificity of nucleic acids is translated into that of proteins (polypeptides) by a process in which nucleic acids function as templates in protein synthesis.

Gamow chose DNA for the template. Dounce, luckier or wiser, chose RNA. As discussed in Section IV,E, the transfer of information (specificity) from DNA to polypeptide is now believed to occur in two steps: first, *transcription* into a special *messenger* RNA (mRNA) synthesized on a DNA template; second, *translation* into polypeptide on a template of mRNA. To simplify the following discussion, we shall assume that mRNA, which is believed to be single-stranded, copies only one strand of a DNA duplex. Accordingly, the term *nucleotide* will refer either to a nucleotide in mRNA or to the corresponding nucleotide in the "active" DNA strand. Because of the complementarity of the Watson-Crick DNA duplex, it would be possible in most references to DNA to substitute *nucleotide pair* without change of meaning.

5. *Selective sites, selective nucleotides, and selective packets:* With respect to amino acid selection at a particular site (*target site*) in a polypeptide, nucleotides at certain sites in the determinant nucleic acid can be replaced freely by others without effect on the selection; nucleotides at certain other sites cannot be replaced freely. To be freely replaceable, a nucleotide must be replaceable without selective effect regardless of concurrent nucleotide replacements at other sites, i.e., regardless of the actual sequence of the molecule, the length, however, remaining fixed. The sites at which nucleotides cannot be replaced freely are the *selective sites* for the polypeptide target site in question. The particular nucleotides that occupy the selective sites in a given nucleic acid comprise the *packet of selective nucleotides,* or *selective packet,* for the target site.

Note that the number and location of selective sites are, by definition, invariant under nucleotide replacement and, hence, invariant under amino acid replacement at the target site. The *size* of a selective packet (number of selective nucleotides) and its *shape* (spacing of selective nucleotides) are therefore also invariant under replacement. It will be important to recall these features when *selective packet* becomes *coding unit* (= *code word*). Packets for different target sites need not, however, have the same size or shape under the present postulate and definitions.

6. *One selective packet, one polypeptide target site:* Two distinct

target sites in a polypeptide require two selective packets, each of which includes at least one selective nucleotide not included in the other.

7. *Colinearity hypothesis:* The order of selective packets in a determinant nucleic acid is the same as the order of their respective target sites in the corresponding polypeptide. This assumes that the packets have a shape which makes it meaningful to speak of their order.

It is now generally assumed (Section IV,E) that the order of information along polynucleotide sequences remains invariant during transcription from DNA to RNA, as well as during replication. Hence, the colinearity hypothesis relates polypeptides to both types of nucleic acid.

Postulates 8–12, referring to special features of selective packets, are stated separately since they are subject to independent verification.

8. *Uniformity of size:* All selective packets contain the same number of selective nucleotides (or nucleotide pairs, counted in a duplex).

9. *Actual size:* Triplets. Gamow's packets ("diamonds") include two nucleotides and a nucleotide pair but are formally equivalent to triplets.

10. *Shape:* The selective nucleotides comprising a packet occupy adjacent sites in the nucleic acid.

11. *Overlap:* Certain packets, appropriately placed in the chain, share one or more selective nucleotides.

Both Dounce and Gamow proposed extreme overlap; e.g., in the chain . . . ABCDCAD . . . each of the successive triplets ABC, BCD, CDC, etc., would have a corresponding target site in the polypeptide.

12. *Degeneracy:* Certain selective packets may be replaced by certain others without effect on amino acid selection.

The postulate needed to convert a merely deterministic scheme into a coding scheme has yet to be stated. A triplet or other packet of selective nucleotides has size, shape, composition, internal order, and one other important property: position in the chain. For all we have said so far, two packets of the same kind, but located at different positions, could correspond to different amino acids, or sets of amino acids, at the respective target sites. In the extreme case, the correspondence at one position need bear no relation whatsoever to that for the same kind of packet at a second position. Suppose, for example, that the packet ABC corresponds at one position to amino acid X and at a second position to amino acid Y. By definition (Postulate 5), the two packets already contain *all* of the selective nucleotides bearing on amino acid selection at the two target sites. Hence, we are not allowed to overcome the difficulty simply by expanding the packets. Since position can be measured from a suitable reference point independently of the *kinds* of intervening nucleotides, one way out is to use position as an extra deterministic fea-

ture. But this fails to give the simplest possible correlations implied by the notion of coding.

The fundamental Dounce-Gamow postulate, one that abstracts a selective packet away from its context and thus rejects at the level of molecular fine-structure something that has long been dear to geneticists, is then:

13. *Lack of position effect:* The correspondence between selective packets and amino acids is, for packets of the same kind, invariant with respect to position of the packets in the polynucleotide.

B. Some Coding Terms

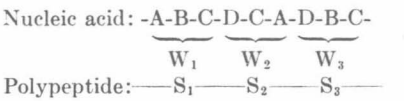
Postulate 13 is what makes it possible to conceive of a simple table of correspondences between selective packets and amino acids. Such a table will be called a *dictionary.* A selective packet will be called a *coding unit, code word,* or simply *word;* we leave it to the context to make clear whether the reference is to one or more packets located at definite positions in a nucleic acid, or to the set of all such packets. (Equivalent or nearly equivalent terms in current use are *codon, nucleotide configuration, letter.*) The terms *code* and *genetic code* may be used to refer to part of the dictionary ("the code for alanine") or to all of it. An individual selective nucleotide will be called a *coding subunit* or *letter. Size* (= number of letters in a word) and *shape* (= spacing of letters in a word) have the same meanings as before. Words that have the same shape, and therefore also the same size, are said to be *congruent.*

The *coding ratio,* often confounded with *word size,* is defined as follows (from Crick, 1959): "If B consecutive bases (nucleotides) are required to code A consecutive amino acids, the coding ratio is the number B/A, when B and A are large." In the Dounce-Gamow codes the coding ratio is unity, whereas the word size is three.

A *sense word* is any word appearing in the dictionary of a given organism. A *nonsense word* is one that does not appear in the dictionary but can be derived from a sense word by one or more nucleotide replacements. *Co-degenerate words* are sense words that are mutually replaceable without effect on amino acid selection. A *nonsense mutation* is the replacement of a sense word (in a nucleic acid) by a nonsense word. A *missense mutation* is the replacement of one sense word by another that is not co-degenerate with it.

A sense word or nonsense word situated in a nucleic acid is said to be *targeted* if it has a target site in the corresponding polypeptide, and *untargeted* if it does not. In the Dounce-Gamow schemes, all the words that can be read in a nucleic acid are targeted. In certain other schemes,

not all the words that can be read are targeted. To illustrate, consider a non-overlapping triplet scheme in which the postulated association between words (W) in a certain stretch of nucleic acid and target sites (S) in the corresponding polypeptide is:



The triplets ABC, DCA, DBC are targeted; BCD, CDC, CAD, ADB are not. Two main ways have been suggested for rendering the unwanted words inoperative in amino acid selection. One is to make all unwanted words nonsense words ("commaless codes"; Crick *et al.*, 1957); inserting appropriate nucleotides as punctuation ("commas") between adjacent targeted words amounts to the same thing, since any word made up partly or entirely of punctuating nucleotides must be a nonsense word. The alternative, presently in favor (Crick *et al.*, 1961), is to distinguish targeted and untargeted words by their placement in the nucleic acid. Any device (processes, structures, etc.) used to make this distinction, and to select a target site for each targeted word, may be called a *reading frame*. A (systematic) change in the assignment of target sites to words constitutes a *shift of the reading frame* (Crick *et al.*, 1961). Note that the concept of reading frame does not contravene Postulate 13, since the reading frame assigns targets but is not imagined to affect the translation of targeted words. From the viewpoint of the polypeptide, however, the reading frame does participate in the determination of the sequence.

How are we to define *universality* of the genetic code? For the code to be universal, it seems legitimate, necessary, and sufficient to require that all organisms have exactly the same dictionary, i.e., that both the distinction between sense and nonsense, and the meaning of sense words, be invariant. It seems undesirable to demand further that each organism contain at least one representative of each word in its nucleic acid; this condition might better be called *ubiquity*. Accordingly, we shall say that the code is universal if each word evokes the same response, in terms of amino acid selection, in each organism (or suitable cell-free extract) in which the word occurs or into which it is introduced. As Benzer and Weisblum (1961) point out, universality may be difficult to prove, whereas a single discrepancy would suffice to prove non-universality.

Dictionaries represent, of course, an extreme extrapolation of the notion of templates. The language of coding differs fundamentally, therefore, from that of classical genetics. To illustrate, assume a dictionary