BROAD
INSTITUTE

CODES WITHOUT COMMAS

By F. H. C. Crick, J. S. Griffith, and L. E. Orgel

MEDICAL RESEARCH COUNCIL UNIT, CAVENDISH LABORATORY, AND DEPARTMENT OF THEORETICAL CHEMISTRY, CAMBRIDGE, ENGLAND

Communicated by G. Gamow, February 11, 1957

This paper deals with a mathematical problem which arose in connection with protein synthesis. We present the solution here because it gives the "magic number" 20, so that our answer may perhaps be of biological significance. To make this clear, we sketch in the biochemical background first.

It is assumed in one of the more popular theories of protein synthesis that amino acids are ordered on a nucleic acid strand (see, for example, Dounce[1]) and that the order of the amino acids is determined by the order of the nucleotides of the nucleic acid. There are some twenty naturally occurring amino acids commonly found in proteins, but (usually) only four different nucleotides. The problem of how a sequence of four things (nucleotides) can determine a sequence of twenty things (amino acids) is known as the "coding" problem.

This problem is a formal one. In essence, it is not concerned with either the chemical steps or the details of the stereochemistry. It is not even essential to specify whether RNA or DNA is the nucleic acid being considered. Naturally, all these points are of the greatest interest, but they are only indirectly involved in the formal problem of coding.

The first definite proposal was made by Gamow.[2] His code, which was suggested by the structure of DNA, was of the "overlapping" type. The meaning of this is illustrated in Figure 1. Gamow's code was also "degenerate"—that is, several sets of three letters (picked in a special way) stood for a particular amino acid. However, all the 64 ($4 \times 4 \times 4$) possible sets of three letters stood for one amino acid or another, so that any sequence whatever of the four letters stood for a definite sequence of amino acids.

It is easy to see that codes of the overlapping type impose severe restrictions on the allowed amino acid sequences. Unfortunately, no such restrictions have been found, although considerable (unpublished) efforts have been made, by a number of workers, to find them. Part of this work has been reviewed by Gamow, Rich,

```
(ns commas
  "CODES WITHOUT COMMAS -- with apologies to F.H.C. Crick et al")

(comment "http://www.pnas.org/content/43/5/416" is the paper.
         The year is 1957. What do we know now?)

(def nucleotides ["Adenine" "Cytosine" "Guanine" "Thymine"])
nucleotides ; => ["Adenine" "Cytosine" "Guanine" "Thymine"]
(count nucleotides)                                ; => 4

(def essential   {:F "phenylalanine"
                  :H "histidine"
                  :I "isoleucine"
                  :K "lysine"
                  :L "leucine"
                  :M "methionine"
                  :T "threonine"
                  :V "valine"
                  :W "tryptophan"})

(def conditional {:C "cysteine"
                  :G "glycine"
                  :P "proline"
                  :Q "glutamine"
                  :R "arginine"
                  :Y "tyrosine"})

(def dispensable {:A "alanine"
                  :D "aspartic acid"
                  :E "glutamic acid"
                  :N "asparagine"
                  :S "serine"})

(def amino-acids (merge essential conditional dispensable))
(count amino-acids)                                ; => 20

"The problem of how a sequence of four things (nucleotides)
 can determine a sequence of twenty things (amino acids)
 is known as the 'coding' problem."

(comment Now ... given that. What can we find out?)
```

Read the => in this comment as "evaluates to".

1

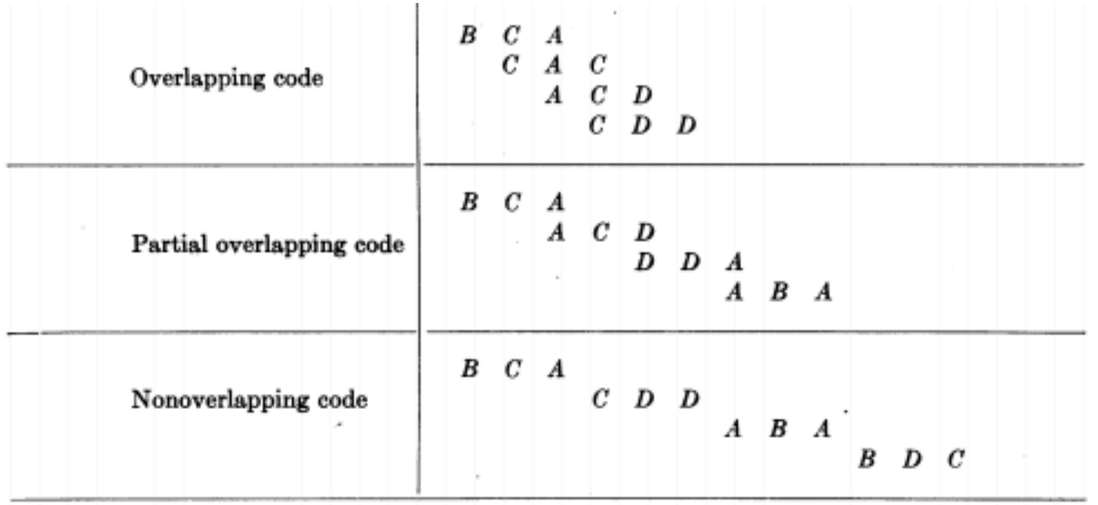

FIG. 1.—The letters *A*, *B*, *C*, and *D* stand for the four bases of the four common nucleotides. The top row of letters represents an imaginary sequence of them. In the codes illustrated here each set of three letters represents an amino acid. The diagram shows how the first four amino acids of a sequence are coded in the three classes of codes.

If each amino acid were coded by *two* bases (rather than the three shown in Fig. 1), we should only be able to code $4 \times 4 = 16$ amino acids. It is natural, therefore, to consider nonoverlapping codes in which *three* bases code each amino acid. This confronts us with two difficulties: (1) Since there are $4 \times 4 \times 4 = 64$ different triplets of four nucleotides, why are there not 64 kinds of amino acids? (2) In reading the code, how does one know how to choose the groups of three? This difficulty is illustrated in Figure 2. The second difficulty could be overcome by reading off from one end of the string of letters, but for reasons we shall explain later we consider an alternative method here.

. . . , B C A, C D D, A B A, B D C, . . .

or

. . . . B, C A C, D D A, B A B, D C . . .

FIG. 2.—The commas divide the string of letters into groups of three, each representing one amino acid. If the ends of the string of letters are not available, this can be done in more than one way, as illustrated. The problem is how to read the code if the commas are rubbed out, i.e., a comma-less code.

```
"Begin digression: Model base pairing with a simple Clojure map."

(str "Model" \space "base" \space "pairing" \.) ; => "Model base pairing."

(def sentence ["Model" \space "base" \space "pairing" \.])
sentence ; => ["Model" \space "base" \space "pairing" \.]

(apply str sentence)                            ; => "Model base pairing."
((partial apply str) sentence)                  ; => "Model base pairing."

(def string (partial apply str))
(string sentence)                               ; => "Model base pairing."

nucleotides            ; => ["Adenine" "Cytosine" "Guanine" "Thymine"]

(first nucleotides)                       ; => "Adenine"
(first (second nucleotides))              ; => \C
(map first nucleotides)                   ; => (\A \C \G \T)

(def ACGT (string (map first nucleotides)))
ACGT                                      ; => "ACGT"
(string (reverse ACGT))                   ; => "TGCA"

(def pair (zipmap ACGT (reverse ACGT)))
pair                                      ; => {\A \T \C \G \G \C \T \A}

(map pair ACGT)                           ; => (\T \G \C \A)

[(rand-nth ACGT) (rand-nth ACGT)]         ; => [\G \A]

(def strand (repeatedly (fn [] (rand-nth ACGT))))
(string (take 23 strand))                 ; => "TCTGTCCCCGTAGACAAGACGTT"
(string (take 23 (map pair strand)))      ; => "AGACAGGGGCATCTGTTCTGCAA"

"End digression: Now where were we?" "We were counting things."

(count ACGT)                                    ; =>  4
(count amino-acids)                             ; => 20
(count (for [x ACGT]                [x]))       ; =>  4
(count (for [x ACGT y ACGT]         [x y]))     ; => 16
(count (for [x ACGT y ACGT z ACGT] [x y z])) ; => 64
```

BROAD INSTITUTE

CODES WITHOUT COMMAS

By F. H. C. Crick, J. S. Griffith, and L. E. Orgel

MEDICAL RESEARCH COUNCIL UNIT, CAVENDISH LABORATORY, AND DEPARTMENT OF THEORETICAL CHEMISTRY, CAMBRIDGE, ENGLAND

Communicated by G. Gamow, February 11, 1957

This paper deals with a mathematical problem which arose in connection with protein synthesis. We present the solution here because it gives the "magic number" 20, so that our answer may perhaps be of biological significance. To make this clear, we sketch in the biochemical background first.

It is assumed in one of the more popular theories of protein synthesis that amino acids are ordered on a nucleic acid strand (see, for example, Dounce[1]) and that the order of the amino acids is determined by the order of the nucleotides of the nucleic acid. There are some twenty naturally occurring amino acids commonly found in proteins, but (usually) only four different nucleotides. The problem of how a sequence of four things (nucleotides) can determine a sequence of twenty things (amino acids) is known as the "coding" problem.

This problem is a formal one. In essence, it is not concerned with either the chemical steps or the details of the stereochemistry. It is not even essential to specify whether RNA or DNA is the nucleic acid being considered. Naturally, all these points are of the greatest interest, but they are only indirectly involved in the formal problem of coding.

The first definite proposal was made by Gamow.[2] His code, which was suggested by the structure of DNA, was of the "overlapping" type. The meaning of this is illustrated in Figure 1. Gamow's code was also "degenerate"—that is, several sets of three letters (picked in a special way) stood for a particular amino acid. However, all the 64 ($4 \times 4 \times 4$) possible sets of three letters stood for one amino acid or another, so that any sequence whatever of the four letters stood for a definite sequence of amino acids.

It is easy to see that codes of the overlapping type impose severe restrictions on the allowed amino acid sequences. Unfortunately, no such restrictions have been found, although considerable (unpublished) efforts have been made, by a number of workers, to find them. Part of this work has been reviewed by Gamow, Rich,


```
(ns commas
  "CODES WITHOUT COMMAS -- with apologies to F.H.C. Crick et al")

(comment "http://www.pnas.org/content/43/5/416" is the paper.
         The year is 1957. What do we know now?)

(def nucleotides ["Adenine" "Cytosine" "Guanine" "Thymine"])
nucleotides ; => ["Adenine" "Cytosine" "Guanine" "Thymine"]
(count nucleotides)                              ; => 4

(def essential   {:F "phenylalanine"
                  :H "histidine"
                  :I "isoleucine"
                  :K "lysine"
                  :L "leucine"
                  :M "methionine"
                  :T "threonine"
                  :V "valine"
                  :W "tryptophan"})

(def conditional {:C "cysteine"
                  :G "glycine"
                  :P "proline"
                  :Q "glutamine"
                  :R "arginine"
                  :Y "tyrosine"})

(def dispensable {:A "alanine"
                  :D "aspartic acid"
                  :E "glutamic acid"
                  :N "asparagine"
                  :S "serine"})

(def amino-acids (merge essential conditional dispensable))
(count amino-acids)                              ; => 20

"The problem of how a sequence of four things (nucleotides)
 can determine a sequence of twenty things (amino acids)
 is known as the 'coding' problem."

(comment Now ... given that. What can we find out?)
```

Read the => in this comment as "evaluates to".