Biologiske Meddelelser
udgivet af
Det Kongelige Danske Videnskabernes Selskab
Bind **23**, nr. 9

Biol. Medd. Dan. Vid. Selsk. **23**, no. 9 (1958)

CONSTRUCTION AND PROPERTIES OF COMMA-FREE CODES

BY

S. W. GOLOMB, L. R. WELCH, AND M. DELBRÜCK

København 1958
i kommission hos Ejnar Munksgaard

Det Kongelige Danske Videnskabernes Selskab udgiver følgende publikationsrækker:

The Royal Danish Academy of Sciences and Letters *issues the following series of publications:*

	Bibliographical Abbreviation
Oversigt over Selskabets Virksomhed (8°) *(Annual in Danish)*	Overs. Dan. Vid. Selsk.
Historisk-filosofiske Meddelelser (8°)	Hist. Filos. Medd. Dan. Vid. Selsk.
Historisk-filosofiske Skrifter (4°) *(History, Philology, Philosophy, Archeology, Art History)*	Hist. Filos. Skr. Dan. Vid. Selsk.
Matematisk-fysiske Meddelelser (8°)	Mat. Fys. Medd. Dan. Vid. Selsk.
Matematisk-fysiske Skrifter (4°) *(Mathematics, Physics, Chemistry, Astronomy, Geology)*	Mat. Fys. Skr. Dan. Vid. Selsk.
Biologiske Meddelelser (8°)	Biol. Medd. Dan. Vid. Selsk.
Biologiske Skrifter (4°) *(Botany, Zoology, General Biology)*	Biol. Skr. Dan. Vid. Selsk.

Selskabets sekretariat og postadresse: Dantes Plads 5, København V.

The address of the secretariate of the Academy is:

Det Kongelige Danske Videnskabernes Selskab,
Dantes Plads 5, Köbenhavn V, Denmark.

Selskabets kommissionær: Ejnar Munksgaard's Forlag, Nørregade 6, København K.

The publications are sold by the agent of the Academy:

Ejnar Munksgaard, *Publishers,*
6 Nörregade, Köbenhavn K, Denmark.

Biologiske Meddelelser
udgivet af
Det Kongelige Danske Videnskabernes Selskab
Bind **23,** nr. 9

Biol. Medd. Dan. Vid. Selsk. **23,** no. 9 (1958)

CONSTRUCTION AND PROPERTIES OF COMMA-FREE CODES

BY

S. W. GOLOMB, L. R. WELCH, AND M. DELBRÜCK

København 1958
i kommission hos Ejnar Munksgaard

CONTENTS

Synopsis.

The sequence of bases in deoxyribonucleic acids is assumed to represent a coded message, embodying information concerning the sequence of amino acids in proteins. Crick *et al.* [7] suggested that the code might be a "comma-free triplet" code. This means that each amino acid is coded by a triplet of bases, and that the triplets are chosen such that no overlap between any pair of triplets codes for an amino acid. In such a code the triplets do not have to be separated from each other by some kind of comma; they can be run together without causing ambiguities in the message.

This paper concerns the following aspects of comma-free codes:

1) Procedures for the construction of all comma-free triplet codes involving the maximum number (20) of triplets. It is shown that there are five classes of such codes and a total of 408 codes.

2) It is shown that no message written with any of these codes ever contains a fourfold repeat of any base, and that in some of the codes certain threefold repeats are excluded.

3) Certain misprints in the coded message will produce nonsense (the resulting triplet does not code for any amino acid), other misprints will produce missense (the resulting triplet codes for a different amino acid). The codes were studied with respect to missense/nonsense ratio produced by various classes of misprints.

4) DNA has a directional symmetry. The basic structure is such that the message could be read in either direction. The question is posed whether codes could be devised such that if they are read in the wrong direction they give nonsense everywhere, i. e., no triplet or overlap between triplets read in reverse corresponds to any amino acid. Such codes are termed transposable codes. It turns out that a transposable triplet code can code for at most 10 amino acids, which is too few. Therefore quadruplet codes were taken into consideration. These are mathematically more difficult to handle and only a few fragmentary results have been obtained so far.

Printed in Denmark
Bianco Lunos Bogtrykkeri A-S

Part I.

Origin of the Problems, Summary and Discussion of Results.

By M. Delbrück.

The discovery that genetic information in many organisms is transmitted from parent to offspring through desoxyribonucleic acid (*DNA*) and the discovery of the structure of *DNA* by Watson and Crick [1] have raised the problem as to the nature of the code used to carry this information and as to the mechanism by which the code is read. It is believed that one of the intermediate steps of the translation consists in the synthesis of specific proteins and that the essential element of this specificity consists in the sequence of the amino acids in the proteins. Both the *DNA* and the protein are linear polymers. The *DNA* molecule is a duplex of two chains containing principally four bases (adenine = *A*, cytosine = *C*, guanine = *G*, and thymine = *T*). In addition, there are one or more bases occurring in very small proportion which may or may not have any particular significance. The two chains have base sequences which are complementary, *A* always opposite *T*, and *C* always opposite *G*. The duplex as well as the single chain may therefore be looked upon as a message written in a code involving four symbols. In the case of the single chain, the symbols are *A*, *C*, *G*, and *T*; in the case of the duplex, the symbols are the base pairs $A-T$, $T-A$, $C-G$, and $G-C$. The proteins are polypeptide chains of some 20 amino acids and can thus be looked upon as messages written in a code containing some 20 symbols. The problem is thus reduced to one of coding the information contained in a message employing some 20 symbols in a code employing only four symbols and to finding the mechanism for its translation.

The *DNA* as a message container confronts us at once with a peculiar duality feature: it contains *two* complementary messages which are chemically quite different. This is so whether the single chain or the duplex as a whole is the message container. If a single chain is the container, then the complementary chain contains a message which differs from the first one by two operations: (1) reading backward, (2) substitution of the complementary symbol X' for each of the symbols X (A for T, T for A, G for C, and C for G). If the duplex as a whole is the message container, with base pairs as symbols, then we are still dealing with two complementary messages. This is so because the duplex as a whole is symmetric, it can be read in either direction, and the two messages so obtained differ exactly by the same operations as those on the two chains: reversal of direction and substitution of the complementary symbol (base pair $A-T$ for $T-A$, etc.).

By the time the message is translated into protein this duality is apparently gone, as there is no evidence that one piece of genetic material is regularly responsible for two different proteins. Along the path from *DNA* to protein one of the messages is therefore eliminated. The last chapter of this paper arose from a specific suggestion as to the nature of this elimination. To introduce this notion, as well as those which prompted the other mathematical questions and answers dealt with in this paper, it will be necessary to insert a few comments on the biochemical aspects of the problem.

It seems fairly certain that the centers of protein synthesis in the cell are the microsomes, particles which contain no *DNA* but which do contain ribonucleic acid (*RNA*), a linear polynucleotide also containing four bases. It is a likely conjecture that the *RNA* represents an intermediate translation of the code. If so, it would be of the greatest interest to know whether at this point of the translation process the duality of the message has already been eliminated. The fact that the base ratios of *RNA* in some cases deviate from those imposed in *DNA* by the complementarity feature would seem to speak in favour of the idea that the duality has been removed, but neither this fact, nor, if it is a fact, the method by which it is accomplished are clear.

Regarding the actual synthesis of proteins it is now believed [2, 3, 4] that the amino acids are first activated, in two steps: in the first step a complex is formed between amino acid and adenosine monophosphate (*AMP*); while in a second step the amino acid is transferred from *AMP* to a soluble *RNA* fraction. More precisely, there is a specific enzyme for each amino acid which couples only this amino acid to *AMP*. The amino acid is then transferred to a specific site on the soluble *RNA*. At first sight it would seem surprising that for each amino acid there should be a specific enzyme to couple it to *AMP*, since no confusion, i.e., no false synthesis, would occur if this step were unspecific. It is conceivable, however, that in vivo not soluble *RNA* molecules but tri- or tetra-nucleotides play the role of intermediate amino-acid carriers, supplying a specific adaptor to each, this adaptor serving the purpose of fitting it to the code letters in the message. If the same enzyme which couples the amino acid to *AMP* were responsible for the transfer of the amino acid to the specific adaptor then the specificity of the enzyme would make sense. Obviously, the same enzyme could not be charged with the duty of coupling each amino acid to its specific adaptor.

Four years ago Gamow [5] published in these proceedings an important paper in which a first attempt was made to "break the code". The main characteristics of Gamow's attempt were the following: first, it assumed a *direct* translation from *DNA* into protein; second, it assumed an *overlapping* code, the piece of *DNA* determining one amino acid (a diamond shaped structure extending over three base pairs) and the piece of *DNA* determining the next neighboring amino acid overlapping by two thirds of their length; third, it assumed a *degenerate* code, in that several different triplets of base pairs coded for the same amino acid. The reason why it was assumed that a triplet of three pairs code for one amino acid was simple. Two base pairs give only 16 possibilities, which is not enough to code for some 20 amino acids. Three base pairs give 64 possibilities which is more than necessary and therefore permits degeneracy. The overlap feature was introduced for geometrical reasons: with this amount of overlap the spacing from one amino acid to the next would correspond roughly to the spacing from one base pair to the next and this seemed reasonable on structural grounds.

With the increase in our knowledge of amino acid sequences in proteins Gamow's particular scheme, and several others, have been demonstrated to be untenable, and in fact BRENNER has given [6] an elegant proof of the impossibility of all overlapping triplet codes. In 1957 CRICK, GRIFFITH and ORGEL [7] introduced a new idea into the problem. If it is true that groups of base pairs, say triplets, code for one amino acid, and that these triplets are not overlapping, and if the message is formed simply by tacking these triplets end to end, how do we know where one triplet ends and the next begins? Either the message would have to be read strictly in sequence starting at one end, or the triplets might be chosen such that no overlap makes "sense". The triplets which code for amino acids might form a "dictionary" of "words" such that no overlapping triplet in a message written from these words is a word in this dictionary. This is the idea of the comma-free code, and CRICK *et al.* proved that in the case of four symbols and words of length three the maximum size of such a dictionary is 20. They also constructed some of the dictionaries of this size. In such a code, then, the freedom resulting from the fact that there are 64 possible triplets and only 20 amino acids to code for is used to select a comma-free dictionary. GOLOMB, GORDON and WELCH [8] addressed themselves to a mathematically interesting generalization of this problem: what is the maximum size of a comma-free dictionary in the case of an arbitrary number of symbols and an arbitrary length of the words? They were able to obtain a partial solution of this problem, and to develop methods which are useful also for the mathematical developments presented in this paper.

These developments were motivated directly by certain aspects of the *DNA*-protein problem. It is the purpose of this Part I to explain the biological interest of these questions and to summarize and discuss the results. It is hoped that they will stimulate new experimental approaches and point the direction which further mathematical analysis might take, to be of the greatest interest to the biologist.

The overlapping codes considered by previous authors implied no restriction on base neighbours, but did imply certain restrictions, statistical or absolute, regarding amino acid neighbors. It had been hoped that these restrictions would afford clues for breaking the code without any actual knowledge of base sequences in the *DNA*. As it turned out, these restrictions only served to *eliminate* the overlapping codes. The comma-free codes present the opposite situation. They are non-overlapping codes, and as such they contain no restrictions, statistical or absolute, regarding amino acid neighbors, but they do imply restrictions on base sequences. If this is the type of code actually used, then the study of amino acid sequences by itself will be useless for breaking the code. However, information on the neighbor relations among the *bases* may be indicative as to the specific code used, and it becomes of interest to examine these codes closely to see what they imply, statistically or absolute, with respect to base neighbor relations.

A prerequisite for such a study is a method for constructing all possible comma-free codes. This is accomplished for the three letter codes in the first three chapters of Part II of this paper. The key to this construction is the important Theorem 2 which states a surprisingly simple condition which is both necessary and sufficient for a collection of 20 triplets to constitute a comma-free dictionary. With the help of this theorem it can be shown that there are five types of maximal comma-free dictionaries using triplets. Of each type there are a large number of different dictionaries, dictionaries of the same type differing from each other by permutations of letters and by reversals of sections of the dictionary. Chapter 3 is devoted to a study of these permutations and reversals.

Being in possession of these dictionaries we would like to enquire into properties of these dictionaries which might find expression in an experimentally verifiable manner and might thus serve the diagnostic purpose of differentiating between various codes.

The most direct route for breaking the code would of course be available if the conjecture of a specific oligonucleotide adaptor for each amino acid were found to be valid, and if these adaptors should stand in a complementarity or identity relation to the words of the coded message. It would then only be necessary to isolate and characterize the amino acid-adaptor complexes.

It is more likely that a less direct approach will be needed, involving properties of the *DNA* by itself or of the *RNA* by itself, and here a promising attack would lie in the examination of oligonucleotide fragments prepared by hydrolytic procedures with known specificities. Here it is important, before comparing experimental data with the mathematical properties of the codes, to take proper account of the duality feature.

We will illustrate this by examining three rules, proved in Part II of this paper, concerning forbidden symbol combinations.

1) In stochastic messages written from some of the triplet codes one and only one of the triplets of type XXX does not occur.

This is not a property which is verifiable by fragment analysis if the material studied still has the duality feature. Indeed, since the triplet $X'X'X'$ (the complement of XXX) will occur in the *message*, the triplet XXX will occur in the complementary message. If one of the single chains is the message, and does not contain XXX, then its complement will contain it, and its absence from the message proper will not be apparent from a hydrolyzate which does not distinguish the breakdown products of the real message from those of the complement. Similarly, if the duplex, read in one direction, is the real message, and this excludes of the triplets only the triplet XXX (X representing, say, the base pair $A-T$), then one chain excludes the triplet AAA and the other TTT, but each chain will contain the triplet which the other excludes, and the exclusion will not be observable by fragment analysis.

In fact, what is observable is not any rule as expressed in terms of the symbols of the message, but only such a rule after "mixing" it with the complementary rule. Thus, the rule "the triplet XXX does not occur and the triplet $X'X'X'$ does occur" is to be mixed with the complementary rule "triplet XXX does occur and triplet $X'X'X'$ does not occur." The mixed rule then says that both triplets occur.

2) No quadruplet of type $XXXX$ can occur in any message written from any maximal comma-free triplet code.

This rule is not affected by mixing with the complementary rule, since in this case the original and the complementary rule are identical.

3) In two of the five types of triplet code *two* triplets are excluded from the messages. If the two excluded triplets are complementary, then the rule is not changed by mixing, while if they are not complementary, then there are no exclusions after mixing. Whether the two triplets are complementary or not depends on the identification of the symbols with the bases or base pairs.

Up to this point we have been talking about absolute rules: complete exclusions of certain symbol combinations. We now turn to statistical rules. Here the first rule (theorem 1) is that in each maximal comma-free dictionary every letter occurs equally frequently and in the same number of words. Any deviation from equality between the four symbols therefore indicates an unequal frequency of the words, as is indeed obvious from the inequalities between the frequencies of the different amino acids. This inequality is of course insufficient to characterize the type or the specific dictionary. It follows further that it would be very difficult to make predictions concerning the frequencies of permitted digrams and longer combinations. For digrams, for instance, it is easy to calculate the frequencies with which they occur in the different dictionaries, and wide differences will be found between different dictionaries. However, the observable digrams include the overlaps, and their frequencies depend critically on the relative word frequencies which are not known in terms of the words of the code, even though they may be ascertainable, in some cases, in terms of the amino acids. For these reasons it does not seem worth while at this stage to go into a detailed analysis of the statistical relations.

The comma-free codes differ in another important respect from those previously considered: every misprint of necessity alters the message. This is in contrast to degenerate codes, where a large class of misprints produces no alteration at all in the message. The non-degenerate codes are therefore more vulnerable. The errors which occur, may, moreover, be divided into two classes: those which change a word of the dictionary into another word of the dictionary, and therefore one amino acid into another amino acid, one polypeptide into an altered one; and those which change a word of the dictionary into a combination of letters which is not in the dictionary, and therefore one amino acid into no amino acid, one polypeptide of length n into two polypeptides of total length $n-1$. Let us call these two classes of misprints the mis-sense class and the non-sense class. Presumably the non-sense class represents on the average a more severe functional change. It seems likely that the mis-sense to non-sense ratio is an important characteristic of each code and might have played a role in its natural selection. Moreover, it may be that among the individual symbol errors those that interchange similar ones (the purines among themselves and the pyrimidines among themselves) are much more probable than the others.

The mis-sense to non-sense ratios are shown in Tables III and IV for the five types of dictionaries and for various restricted classes of misprints. There exist only slight differences between the five types of dictionaries when the types of misprints are unrestricted. In a maximal comma-free dictionary using four symbols and three-letter words there are sixty letters (20 words with three letters each) and each of them can be misprinted in three different ways, giving a total of 180 different misprints. The class of non-sense misprints ranges in size from 80 (in type I) to 92 (in type V). The differences become more pronounced when restricted classes of errors, of the types alluded to above, are considered. In this case, each letter can be misprinted in only one way (each pyrimidine by the other pyrimidine and each purine by the other purine). The total number of possible misprints is now sixty, and the class of non-sense misprints ranges in size from 16 (for type I) to 32 (for type V).

We turn now to the problem of how the duality of the message in the *DNA* is handled in the translation process. Here we put to ourselves the question whether the dictionary might not be so constructed that it automatically eliminates the duality. This would be accomplished if one of the two complementary sequences of letters contained nowhere letter combinations which occur in the dictionary. This question is examined in the last section of Part II of this paper. It turns out that it is indeed possible to construct dictionaries such that the complement to any message composed of words of the dictionary contains nowhere, neither as the complements of words nor as complements of the overlaps, a word of the dictionary. Such dictionaries are called *transposable* dictionaries and the very strong constraint which they fit does not reduce their size inordinately. A general theorem (theorem 15) about transposable dictionaries is given, establishing an upper bound as to their sizes. In the case of four symbols and three letter words, this maximum size is 10, as compared to 20 in the absence of the constraint. Several such dictionaries are given. Ten words is too few to code for all the amino acids and the simplest way to increase the number of words is by increasing the length of the words to four letters. In this case the upper bound turns out to be 27 and an actual transposable dictionary of size 26 has been constructed. [Added in proof: numerous examples of 27 word dictionaires are now known.]

We wish to emphasize that we consider the postulate of comma-freedom and the postulate of transposability to be almost on the same footing. Indeed the principal virtue of comma-freedom is that any message can be read unambiguously starting at any point, with the proviso, however, that one must know in advance *in which direction to proceed.* Since the equivalence of the two opposite directions in a structural sense seems to be one of the more firmly established features regarding the *DNA* molecule the advance knowledge as to the direction in which to read cannot come from the basic structure. Comma-freedom would therefore seem to be a worthless virtue unless it is coupled with transposability.

Transposable comma-free dictionaries, in contrast to ordinary comma-free dictionaries, are strongly asymmetric with respect to symbol frequencies and symbol combinations. This is due to the fact that in the ordinary comma-free dictionaries every complete equivalence class[1] is represented, whereas in the transposable ones at least one half of these classes are not represented. This may be illustrated by the 26 word dictionary using 4-letter words exhibited in the last chapter. In this dictionary the pairs A and B, and C and D are complementary. The four symbols occur in the dictionary with the frequencies 21, 31, 41, 11. The complementary pairs A, B and C, D occur equally frequently in the dictionary. This is a general rule for transposable dictionaires in which each non-symmetric equivalence class[1] is represented. In the example given each class, with one exception, is, in fact, represented. This exception is the class $ADBC$. Its omission does not lead to a violation of the equality rule for complementary pairs, since it contains an equal number of each pair.

In terms of the bases this rule implies that in the dictionary the pair of bases Adenine-Thymine occurs as frequently as the pair of bases Guanine-Cytosine. For *DNA* molecules this equality in general does not hold. The inequality would have to be due to the unequal frequencies of the words (amino acids) of the dictionary in the coded message (*DNA* molecule).

A strong asymmetry may be inherent in the dictionary with

[1] for definition of this term see Part II, chapter 1.

respect to *purines* vs. *pyrimidines*. This we will illustrate again with the 26 word dictionary. Let us make the identification:

A = adenine
B = thymine
C = cytosine
D = guanine

Thus A and D are the purines (collectively designated as *Pu*), and B and C are the pyrimidines (collectively designated as *Py*). The dictionary is thus seen to contain 32 *Pu* and 72 *Py*. The symbol frequencies in the message will of course depend on the word frequencies, and this may affect the letter frequencies appreciably. However, even for equal word frequency, there would be on one of the chains a large excess of *Py*. Let us call this the *Py* chain. This would be compensated by a corresponding excess of *Pu* on the other chain, the *Pu* chain, and would not be observable as an asymmetry in the total hydrolysate. The asymmetry would be observable, however, in an experiment of the type described by MESELSON and STAHL [9], where N^{15} labelled *DNA* is permitted to duplicate once in the presence of N^{14}, producing "hybrid" *DNA* molecules, half labeled with N^{15}. If the "halves" in these experiments are indeed single polynucleotide chains (of which there is still doubt), then we should obtain two classes of hybrids, those whose *Py* chain is N^{15} labelled and *Pu* chain N^{14} labelled, and vice versa, and these two types of hybrids would differ sufficiently in density (by 0.2 $^0/_0$) to be resolvable by the equilibrium density gradient technique.

The unequal distribution of the purines and pyrimidines over the two chains will lead also to a phenomenon recently reported by SHAPIRO and CHARGAFF [10]. These authors found that the pyrimidines are bunched, i.e., that runs of pyrimidines of length one and two, flanked by purines, occurred less frequently than would be expected for random sequences in which purines and pyrimidines occur equally frequently. Messages constructed from a given dictionary are certainly not to be equated statistically with random letter sequences. The statistical proportion of the symbol relationships in such messages cannot be assessed in detail without a knowledge of the statistics of words. It is clear, however, that a strongly unequal distribution of the pyrimidines over the two chains will introduce a bias in the direction of bunching.

It is perhaps unlikely that the words in the dictionary are of equal length. Indeed if economy with respect to total length of the coded message were the overriding principle in the choice of the code, nature would certainly have selected short words for frequently occurring amino acids and vice versa. We know at present too little about other aspects, particularly those of steric convenience, that might have more decisive influences affecting the selective advantages of various possible codes. The arguments and mathematical results presented in this paper can do no more than to help clarify ideas and direct attention to the types of data on base sequences that would be most helpful.

Part II.

Mathematical Developments.

By S. W. Golomb and L. R. Welch.

1. Definitions and General Theorems.

Let n be a fixed positive integer, and consider an alphabet consisting of the numbers 1, 2 . . . n. With this alphabet form all possible k-letter words $(a_1 a_2 \ldots a_k)$, where k is also fixed. There are evidently n^k such words in all.

Def. A set L of k-letter words is called a *comma-free dictionary* if whenever $(a_1 a_2 \ldots a_k)$ and $(b_1 b_2 \ldots b_k)$ are in L, the "overlaps" $(a_2 a_3 \ldots a_k b_1)$, $(a_3 \ldots a_k b_1 b_2)$, ... , $(a_k b_1 \ldots b_{k-1})$ are not in L.

Let $W_k(n)$ denote the greatest number of words that such a dictionary can contain. In [8] the upper bound

$$W_k(n) \leq \frac{1}{k} \sum_{d/k} \mu(d)\, n^{k/d} \tag{1}$$

was obtained, where $\mu(d)$ is the Möbius function, and the summation is extended over all divisors d of k. It was further shown that the upper bound (1) is actually attained by some dictionary for all values of n when $k = 1, 3, 5, 7, 9, 11, 13, 15$, and this is *conjectured* to extend to all *odd* k. On the other hand, if k is *even*, the upper bound (1) is *not* attained by $W_k(n)$ for any $n > 3^{k/2}$. Specifically, for $k = 2$ it was demonstrated that $W_2(n) = \left[\frac{n^2}{3}\right]$, whereas (1) merely asserts that $W_2(n) \leq \frac{n^2 - n}{2}$. The treatment of the case $k = 2$ actually contained the method for finding *all possible* comma-free dictionaries containing the maximum number of words, $W_2(n) = \left[\frac{n^2}{3}\right]$.

The primary objective here is to provide a constructive method for finding all maximum comma-free dictionaries for $k = 3$. Here the size of such a dictionary is $W_3(n) = \frac{n^3 - n}{3}$.

Before confining attention to the case $k = 3$, certain observations are appropriate concerning all dictionaries for which the upper bound (1) is attained.

Def. Two k-letter words are in the same *equivalence class* if they differ only by a cyclic permutation of the letters. (Thus *ABC, BCA, CAB* are all in the same equivalence class.) An equivalence class is *complete* if it contains k distinct members. (The example *ABC, BCA, CAB* is complete. The equivalence class of *AAA* is *degenerate,* containing only one member.)

Degenerate equivalence classes are never represented, and complete equivalence classes are represented at most once, in any comma-free dictionary, as shown in [7] and [8].

Theorem 0. The upper bound (1) is attained by a comma-free dictionary if and only if every complete equivalence class has *exactly* one representative in the dictionary.

Theorem 1. In any comma-free dictionary for which $W_k(n)$, the size of the dictionary, attains the upper bound (1), all n letters of the alphabet occur equally often among the $kW_k(n)$ letters in the dictionary. Also, each letter occurs in the same number of different words.

Proof. The dictionary contains one representative from each complete equivalence class. *Which* representative is inconsequential insofar as the letters comprising the individual words and the entire dictionary are concerned. Since any permutation of the alphabet leaves the class decomposition of the k-letter words invariant, permutation of the alphabet does not affect the distribution of letters in the dictionary, nor the frequency with which letters appear in distinct words. Therefore, these distributions and frequencies must be the same for all n letters of the alphabet.

Corollary. If all words containing a given letter are dropped from a maximum comma-free dictionary L attaining the upper bound (1) for $W_k(n)$, the remaining words form a maximal comma-free dictionary L' for the $(n-1)$-letter alphabet. The number of different words of L in which a given letter appears is thus $W_k(n) - W_k(n-1)$.

Proof. The deletion of certain words from a comma-free dictionary cannot destroy the comma-free property. Thus expurgation of all words of L containing the letter X leaves a comma-free dictionary L' involving only an $(n-1)$-letter alphabet. Moreover, every complete equivalence class for these $n-1$ letters had a unique representative in L, which has not been dropped during the expurgation process. Thus L' is maximal by Theorem 0.

2. Classes of Dictionaries for $k = 3$.

The fundamental result concerning the structure of maximal comma-free dictionaries for $k = 3$ is expressed in the following theorem. Unless otherwise specified it will henceforth be assumed that $k = 3$.

Def. The first two letters of a word are an *initial digram*; the last two letters are a *final digram.*

Theorem 2. For $n > 2$, the necessary and sufficient condition that a collection of $(n^3 - n)/3$ words constitute a maximal comma-free dictionary is that no initial digram ever occurs as a final digram.

Proof. The condition is *sufficient* for maximal comma-freedom, since $(n^3 - n)/3$ is the correct dictionary size, and violation of comma-freedom from $(ABC)(DEF)$ could only occur if either BCD or CDE were in the dictionary along with ABC and DEF. But BCD begins with the final digram of ABC, and CDE ends with the initial digram of DEF.

To show *necessity,* suppose that ABC and BCD were both in the maximum comma-free dictionary L. To prevent conflicts, A must never end a word, nor D begin a word, and in particular, $A \neq D$. Therefore, the equivalence class $\{AAX, AXA, XAA\}$ must be represented by AAX, and the class $\{YDD, DYD, DDY\}$ must be represented by YDD, for all $X \neq A$ and $Y \neq D$. For any Z distinct from both A and D, consider the class $K = \{ADZ, DZA, ZAD\}$. (Here the hypothesis $n > 2$ is required.) Observe that ADZ is contained in the overlap of $(AAX)(YDD)$ when $X = D$, $Y = Z$; while ZAD is in the overlap of $(AAX)(YDD)$ when $X = Z$,

$Y = A$. Also, DZA ends in A. Thus L can contain no representative of the class K, and by Theorem 0, L is not maximal. This contradiction completes the necessity proof.

Notes. 1) The case $n = 2$ is truly exceptional. Thus $\{110, 100\}$ is a maximal dictionary, although 10 occurs both initially and finally.

2) The proof of Theorem 2 can be used to show that a comma-free dictionary in which a digram occurs both initially and finally falls short of the maximum by at least $n-2$, corresponding to the $n-2$ choices of Z in (ADZ).

3) The method given in [8] to obtain maximum comma-free dictionaries for $k = 3$ uses all words XYZ satisfying $X < Y \geq Z$, where $<$ and $\geq$ refer to alphabetical ordering. It is clear that no initial digram could then be a final digram; and since the number of words satisfying the inequality is $(n^3-n)/3$, the dictionary is maximum comma-free by Theorem 2.

The following result may be remembered as a "law of the excluded middle".

Theorem 3. 1) In a maximum comma-free dictionary, every letter with one possible exception occurs in the middle position of some word.

2) If every letter is sometimes a middle letter, then every digram occurs, either initially or finally.

3) If A is the excluded middle letter, then every digram except AA occurs either initially or finally.

Proof. 1) If neither A nor B occurs as a middle letter, then the complete equivalence class AAB, ABA, BAA is unrepresented in the dictionary, contradicting maximality.

2) Suppose every letter can occur in the middle position, and CD is neither an initial nor a final digram. Let XCY be a word with C in the middle. Then XCD could be added into the dictionary, because XC is a legitimate initial digram, and CD, having never occurred initially, is a legitimate final digram. But this contradicts the assumption that the dictionary was already maximal.

Note. The corresponding result concerning the absence of initial and final letters is as follows:

In a maximum comma-free dictionary, every letter occurs both initially and finally, with at most one exception. If D is exceptional, it occurs *either* initially *or* finally, and also occurs $n(n-1)$ times as a middle letter.

The proof of this assertion is inherent in the proof given subsequently for Theorem 9.

3) Suppose A never occurs in the middle position. Then the digram AA cannot occur initially or finally. However, any digram $CD \neq AA$ has one member (or both) unequal to A, and that member can be made to occur in the middle of a word, exactly as in 2) above.

Def. A *section* of a comma-free dictionary is the set of all words having a given middle letter.

Theorem 4. In a maximum comma-free dictionary, the following conditions hold for the sections.

1) The number s of sections is either n or $n-1$.
2) The number of words in each section has the form ij, with $i \leq n$, $j \leq n-1$.
3) $\sum_{\sigma} ij = (n^3-n)/3$, where σ runs through all sections.
4) $\sum_{\sigma} (i+j-1) = n^2-n$.

Proof.

1) This restates Theorem 3, part 1.
2) The words having a given middle letter D involve certain initial digrams ending with D, and certain final digrams beginning with D. By Theorem 2, all linkages of such initial and final digrams are consistent with comma-freedom, hence must occur in the interest of maximality. The size of the section is thus the number of initial digrams ending in D times the number of final digrams beginning with D. Each of these factors is at most n; and they cannot both equal n, since DD cannot occur both initially and finally.
3) The sum of the sizes of the parts equals the size of the whole.
4) In the section of size ij, the number of digrams used either initially or finally is $i+j$. By Theorem 3,

$$\sum_{\sigma} (i+j) = \begin{cases} n^2 & \text{if } s = n \\ n^2-1 & \text{if } s = n-1, \end{cases}$$

since there are n^2 possible digrams.

Therefore

$$\sum_{\sigma}(i+j-1) = n^2-n.$$

Corollary 4.1. To obtain all maximum comma-free dictionaries for $k = 3$ and fixed n, it suffices to consider the dictionary size $(n^3-n)/3$ partitioned into n or $n-1$ sections, each of size ij, with $i \leq n$, $j < n$, where the sum of the *weights* $(i+j-1)$ of the individual sections equals n^2-n.

Corollary 4.2.

$$\sum_{\sigma}(i-1)(j-1) = 2\binom{n}{3}.$$

Example. When $n = 4$, the dictionary size is $(n^3—n)/3 = 20$, and the sum of the weights of the sections is $n^2-n = 12$. The only partitions of 20 into three or four sections of total weight 12 are:

I.	$20 = 12+6+2,$	$12 = 6+4+2.$
II.	$20 = 12+6+1+1,$	$12 = 6+4+1+1.$
III.	$20 = 12+4+4,$	$12 = 6+3+3.$
IV.	$20 = 9+9+2,$	$12 = 5+5+2.$
V.	$20 = 9+9+1+1,$	$12 = 5+5+1+1.$

Dictionaries of all five types exist, as shown (by sections and digrams) in Table I. [Added in proof: in a paper just published (Proc. Acad. Sci. Amsterdam, Series A, **61:** 253–258) H. Freudenthal obtained the same five types without use of our Theorem 2.]

Table I.

The five types of maximum comma-free dictionary for $k = 3$, $n = 4$. Braces indicate the reversible subcollections of Theorem 7.

I	II	III	IV	V
AD DA *BD DB* *CD DC* *DD*	*AD DA* *BD DB* *CD DC* *DD*	*AD DA* *BD DB* *CD DC* *DD*	*AD DA* *BD DB* *CD DD*	*AD DA* *BD DB* *CD DD*
AC CA *BC CB* *CC*	*AC CA* *BC CB* *CC*	*AC CA* *BC CC*	*AC CA* *BC CB* *DC CC*	*AC CA* *BC CB* *DC CC*
AB BA *BB*	*BB BA*	*AB BA* *CB BB*	*AB BA* *BB*	*BB BA*
	AA AB			*AA AB*

3. Reversible Portions of Dictionaries.

Once the basic types of maximum comma-free dictionaries for $k = 3$ and fixed n are constructed, it is important to determine those transformations which preserve maximality, comma freedom, and type. Two obviously permissible transformations are permutation of the alphabet, and reading all words of the dictionary backwards. Crick *et al.* [7] observed that certain *portions* of a dictionary may sometimes be read backwards. The aim of this chapter is a precise mathematical formulation of the circumstances under which portions of the dictionary may be so reversed, and a criterion for the number of *different* dictionaries which result from a given dictionary by the reversal process. The basic result appears as Theorem 6.

Def. Reversal of a word means writing it backward. *Reversal of a set* means writing every word in the set backward.

Def. A palindromic set is a subset of the dictionary which is setwise invariant under reversal. A *palindromic word* (or *palindrome*) is a word invariant under reversal.

Examples. The reverse of ABC is CBA.
The set (ABC, CBA) is palindromic.
The word ABA is a palindrome.

Def. A *reversible set* W is a subset of a maximum comma-free dictionary L such that simultaneous reversal of all members of W leaves maximality and comma-freedom intact.

Lemma A. Given a dictionary L,

1) Every palindromic set is reversible.
2) If ABC and CBA are distinct words in L, a reversible set containing either must contain both.
3) The intersection of a reversible set and a palindromic set is palindromic.

Proof.

1) The reversal of palindromic sets leaves L invariant.
2) If ABC is reversed and CBA is not, the dictionary size is reduced by one word, violating maximality.
3) Let M be palindromic and let W be reversible.

If ABC is in $M \cap W$, then CBA is in M by palindromicity, and in W by 2) above, so that CBA is in $M \cap W$.

Def. Two reversible sets are called *congruent* if they differ only by a palindromic set. (The effect on L of reversing the one is the same as reversing the other.)

Def. A subset T of L is *digram-closed* if for any digram AB occurring (either initially or finally) in some word of T, the digrams AB and BA never occur outside of T.

Example. In the dictionary,

$L = \{010, 011, 020, 021, 022, 120, 121, 122\}$, the subset

$T_1 = \{010, 011\}$ is digram-closed, since none of the digrams 01, 10, 11 occur outside of T_1. However, the subset $T_2 = \{120, 121, 122\}$ is not digram-closed, because 20 occurs in T_2 while 20 and 02 occur outside of T_2.

Def. For any subset S of L, $\bar{S}$ will denote the complement of S in L. The largest palindromic set in L will be denoted by Q. (It is shown in Theorem 8 that $Q \neq L$)

Lemma. If S is reversible, if $S_1 = S \cap \bar{Q}$, and if T is the union of S_1 with the set of all palindromes and palindromic pairs having a digram in common with a word of S_1, then T is a digram-closed subset of L which is congruent to S.

Proof. Since T is congruent to S_1 and S_1 is congruent to S, T is congruent to S.

To prove that T is digram-closed, we first establish: *i*) If the digrams AB and BA (not necessarily distinct) are both initial or both final in L, then either no word of L or all words of L containing such digrams are in T.

Without loss of generality, suppose AB and BA are both initial, and ABC is in T. Then CBA is in the new dictionary Γ, so that BA is final in Γ. Therefore, every word of L beginning with BA belongs to T by Theorem 2. Since BA is assumed initial in L, some word BAD does exist in L, hence belongs to T. By repetition of the foregoing argument, using BAD instead of ABC, every word of L beginning with AB belongs to T, and *i*) is proved.

The lemma is next proved for the following case: *ii*) Let AB occur in a word of S_1, without loss of generality *initially*. Thus ABC is in S_1, for some C. Suppose that BA occurs as a digram in some word of $\bar{T}$. (The alternative is that BA occurs only in words of T, which is treated later.) By i), the word of $\bar{T}$ in which BA occurs can be written XBA.

Since AB is initial and BA is final, maximality requires ABA to be in L. Whether or not ABA is reversed, AB is initial in Γ. By the construction of S_1, ABC in S_1 implies that CBA is not in L. But BA is final in L. Therefore, CB is not initial in L, and by Theorem 3 must be final. By i), every word containing BC or CB must be reversed. In particular, XB is initial and BC is final, so that XBC is in L, and must be reversed. Therefore BX is final in Γ.

Since Γ is maximal, with BX final and AB initial, it follows that ABX is in Γ. Since XBA was assumed to remain unreversed in the transition from L to Γ, ABX is in $\bar{T}$. But XBA and ABX cannot belong simultaneously to $\bar{T}$, by the definition of T. Therefore, every word containing BA is in T.

If $A = B$, case ii) is concluded, while if $A \neq B$ then BA occurs at least once, and repetition of the argument proves that every occurrence of AB is in T.

iii) It remains to consider ABC in $T \cap \bar{S}_1$ and neither AB nor BA is a digram in any word of S_1. Then by construction, CBA is in $T \cap \bar{S}_1$ and BC or CB occurs as a digram in some word of S_1. By ii), every occurrence of BC or CB is reversed in the transition from L to Γ. By Theorem 3, BB is a digram of L. Since CB is initial and BC is final, either BBC or CBB is in L, and is reversed. Because BB cannot be both initial and final, neither BBC nor CBB is in Q, and the one reversed is therefore in S_1. Case ii) now implies that every occurrence of BB is reversed, which covers both ABB and BBA. But not both of these are in L, so that one or the other is in S_1. This contradicts the assumptions of iii).

Theorem 5. A subcollection S of L is reversible if and only if S is congruent to a digram-closed subset T of L.

Proof. If S is reversible, it is equivalent to a digram-closed subset T of L by the foregoing Lemma.

Conversely, if T is digram-closed, reversing T will preserve the dictionary size, and will keep initial and final digrams separate. By Theorem 2, comma-freedom is thereby preserved. Thus if S is congruent to T, S is reversible.

Def. The palindromic sets in a dictionary L, and their complements relative to L, constitute the *palindromic algebra* P in L.

Theorem 6. The reversible sets of a dictionary L form a Boolean algebra R of subsets. The palindromic algebra P is a Boolean subalgebra of R which is in fact an ideal. The distinct dictionaries obtainable from L by reversing reversigle sets are in two-to-one correspondence with the elements of the quotient algebra R/P.

Proof. Any collection of subsets of a set L, closed under $\cup$ and $\cap$, containing L and $\emptyset$ and the complement of every set in the collection, is a Boolean algebra [11]. That both R and P contain L, $\emptyset$, and complements is clear.

The closure of P under unions and intersections is also obvious (cf. Lemma A, part 3). The closure of R under unions and intersections follows from Theorem 5. The fact that reversible sets are essentially closed with respect to digrams is preserved under union and intersection operations. Finally P is an ideal of R, because it is clearly a subring, and moreover if $M \in P$ and $W \in R$, them $M \cap W \in P$. The two-to-one correspondence arises because L itself may be reversed at will.

Theorem 7. To preserve comma-freedom, maximality, section sizes and weights under reversal of subsections of the dictionary L, it suffices to operate on the unmatched digrams in the individual sections of L.

Proof. This is a direct corollary of Theorems 4 and 5.

Example. See Table I for the reversible subsections corresponding to $k = 3$, $n = 4$.

Note. The group of reversals of Theorem 7, times the group of permutations on the n-letter alphabet, generate the complete set of comma-free dictionaries from the representative ones obtained in Corollary 4.1. For example, the five representative dictionaries for $n = 4$, shown in Table I, correspond to 408 distinct maximum comma-free dictionaries in all, as shown in Table II.

Theorem 8. For $k = 3$, no maximum comma-free dictionary is palindromic.

Proof. In a palindromic dictionary, if the digram XX occurs initially it must also occur finally. By Theorem 2, this is impossible for $n > 2$. Hence none of the digrams XX can occur, precluding maximality by Theorem 3. The case $n = 2$ is settled separately by observing that the only palindromic possibility, $\{ABA, BAB\}$, is not comma-free.

Table II.

Numbers of distinguishable maximum comma-free dictionaries for $k = 3$, $n = 4$.

	Basic Dictionary	Permutations of Alphabet	Reversals	Product
	I	24	8	192
	II	12	8	96
	III	12	4	48
	IV	12	4	48
	V	6	4	24
Totals:	5	66	28	408

4. Characteristics of Messages.

For the case $k = 3$, $n = 4$, it is important, for the biological applications, to obtain results concerning the possible messages written with words selected stochastically from the maximum comma-free dictionaries. The investigation here will assume $k = 3$, but n will remain free.

Theorem 9. In a message written from a maximum comma-free dictionary, every triple XYZ is capable of occurring unless $X = Y = Z$. The triple XXX *fails* to occur if and only if either

1) X is the excluded middle letter or
2) the section with X in the middle contains $n(n-1)$ words. (Thus all n^3 triples are capable of appearing in the message with at most two exceptions.)

Proof. Unless $X = Y = Z$, the equivalence class XYZ, YZX, ZXY is nondegenerate, hence represented in the dictionary. In a message, the representative word can appear consecutively, so that the other two members of its class appear in the overlaps.

If A is the excluded middle letter, it is clearly impossible to find AAA in the message. Suppose D is not an excluded middle letter, but DDD never occurs in messages. The digram DD occurs, without loss of generality, initially. Should any word end in D, the message could contain DDD. Thus every digram XD occurs initially.

Next, every digram DY occurs finally. For consider the effect of expurgating from the dictionary all words containing D. The residue is a maximum comma-free dictionary for $n-1$ letters (see

Corollary to Theorem 1). It contains $n(n-1)$ fewer words than the original dictionary. Adding the D-words back, there are $n(n-1-\alpha)$ words in the section for D, where α is the number of digrams DY not occurring finally. These occur initially in other sections, and contribute at most $\alpha(n-1)$ words. But $n(n-1-\alpha)+\alpha(n-1)<n(n-1)$ for $\alpha>0$. Thus the section for D must contain the full $n(n-1)$ words, and there is at most one such section.

Examples. In the five dictionaries of Table I, all possible situations are illustrated. Thus patterns I and III exclude both AAA and DDD; pattern II excludes only DDD; pattern IV excludes only AAA; and in pattern V no triples are excluded.

Theorem 10. In a message written from a maximum comma-free dictionary, the sequences $XXXX$ will never appear.

Proof. In order for $XXXX$ to appear in the message, either XXX is a word in the dictionary, which is impossible, or XX ends one word and begins another, which is likewise impossible.

Theorem 11. Let L contain I initial digrams, F final digrams, and $N=\dfrac{n^3-n}{3}$ words in all. Of the n^4 possible quadruples of letters from the n-letter alphabet, exactly $IF+(i+f)N$ quadruples are capable of occurring in messages written from L. (Here i and f are the number of initial and final letters in L, respectively.)

Proof. An overlap quadruple $VXYZ$ breaks either as $(VXY)(Z\cdot\cdot)$ or $(..VX)(YZ..)$ or $(..V)(XYZ)$. These possibilities are mutually exclusive by Theorem 2. The second case leads to IF quadruples. The first and third cases combine to yield $(i+f)\,N$ quadruples.

Corollary. Let $Q(n)$ denote the possible quadruples. Then

$$\frac{11\,n^4-8\,n^2}{12}\geqq Q(n)\geqq\frac{11\,n^4-4\,n^3-17\,n^2+10\,n}{12}.$$

Moreover, $Q(n)/n^4\sim 11/12$ as $n\to\infty$.

Proof. As in Theorem 9, there is at most one letter which fails to occur both initially and finally.

Thus

$$(2\,n-1)\left(\frac{n^3-n}{3}\right)\leqq(i+f)\,N\leqq 2\,n\left(\frac{n^3-n}{3}\right).$$

Also, $I+F$ is either n^2 or n^2-1, and it is easily seen that

$$\left(\frac{n^2+n-2}{2}\right)\left(\frac{n^2-n}{2}\right)\leqq IF\leqq\left(\frac{n^2}{2}\right)^2.$$

Adding inequalities,

$$\frac{11\,n^4 - 4\,n^3 - 17\,n^2 + 10\,n}{12} \leq Q\,(n) \leq \frac{11\,n^4 - 8\,n^2}{12}.$$

The asymptotic relation follows immediately.

Example. When $n = 4$, $194 \leqq Q(4) \leqq 224$.

Both extremes are attained among the 408 dictionaries of Table II.

5. Missense and Nonsense.

Def. If one letter A of a word W is misread as B, the resulting word W^1 may be either in or out of the dictionary. In the former case the word W^1 is called *missense*, and in the latter case, *nonsense*.

Def. If for each individual occurrence of the letter A in a dictionary L, A is misread as B, the missense-to-nonsense ratio $\frac{M}{N}(AB)$ is the number of times $M(AB)$ that missense results, divided by the number of times $N(AB)$ that nonsense results.

Theorem 12. In any comma-free dictionary L which attains the size (1),

1) $\frac{M}{N}(AB) = \frac{M}{N}(BA)$,

2) $M(AB) + N(AB) = \frac{k}{n} W_k\,(n)$,

3) $M(AX) = M(AY)$ whenever X and Y occur symmetrically in L.

Proof. By Theorem 1, A occurs a total of $\frac{k}{n} W_k\,(n)$ times in L. Since misreading A as B produces either missense or nonseense disjunctively,

$M(AB) + N(AB) = \frac{k}{n} W_k\,(n)$, which is 2). By the symmetry of missense,

$M(AB) = M(BA)$ independent of the comma-free hypothesis. By 2), $M(AB) + N(AB) = M(BA) + N(BA)$. Combining these relations yields 1). Statement 3) is self-evident.

Example. For the five dictionaries of Table I, the missense-to-nonsense structure is shown in Table III.

Table III.

Missense/Nonsense for the Five Patterns with $k = 3$, $n = 4$.

Dictio- nary	Transitions AB	AC	AD	BA	BC	BD	CA	CB	CD	DA	DB	DC
I	13/2	9/6	3/12	13/2	11/4	5/10	9/6	11/4	9/6	3/12	5/10	9/6
II	12/3	10/5	4/11	12/3	10/5	4/11	10/5	10/5	9/6	4/11	4/11	9/6
III	11/4	11/4	3/12	11/4	8/7	7/8	11/4	8/7	7/8	3/12	7/8	7/8
IV	13/2	6/9	6/9	13/2	8/7	8/7	6/9	8/7	4/11	6/9	8/7	4/11
V	12/3	7/8	7/8	12/3	7/8	7/8	7/8	7/8	4/11	7/8	7/8	4/11

These figures can be summed over certain sets of errors, corresponding to probable misprints. The results of four such summations are shown in Table IV.

Table IV.

Missense/Nonsense Summed over Classes of Transitions.

Dictionary	Unrestricted	Transitions $(AB)(CD)$	$(AC)(BD)$	$(AD)(BC)$
I	100 / 80	44 / 16	28 / 32	28 / 32
II	98 / 82	42 / 18	28 / 32	28 / 32
III	94 / 86	36 / 24	36 / 24	22 / 38
IV	90 / 90	34 / 26	28 / 32	28 / 32
V	88 / 92	32 / 28	28 / 32	28 / 32

Remarks. Reversals of the type in Theorem 7, as well as permutations of the alphabet, leave the total missense-to-nonsense ratio of the dictionary unchanged. The numbers shown in Table IV are invariant under the permutation-and-reversal group which generates 408 dictionaries from the five patterns in Table I.

6. Extensions to Larger Values of *k*.

Theorem 2 furnishes the key to the structure of maximum comma-free dictionaries when $k = 3$. Some partial analogues to Theorem 2 for the case $k = 5$ and beyond will now be presented.

Theorem 13. A necessary condition that L be a maximum comma-free dictionary for $k = 5$ is that no initial tetragram be also a final tetragram. (This holds for *all* values of n).

Proof. Suppose $A\gamma$ and γB are both in L, where γ is a tetragram. Then as in Theorem 2, no word can end in A nor begin in B, and $A \neq B$. Thus the words $AABBB$, $AAABB$, and $AAAAB$ must be in L as the representatives of their classes. The equivalence class $ABABB$, $BABBA$, $ABBAB$, $BBABA$, $BABAB$ is then unrepresented, because $BABBA$, $BBABA$, and $BABAB$ begin in B, while $ABABB$ is found in the overlap of $(AAAAB)$ $(ABBBB)$, and $ABBAB$ is found in the overlap of $(AAABB)$ $(ABBBB)$. Thus L could not be maximal.

Note. As in Theorem 2, the constraint that no initial digram be a final digram, nor any initial tetragram a final tetragram, suffices for comma-freedom. It is not likely that the digram condition is *necessary* when $k = 5$. However, the upper-bound dictionary for $k = 5$ given in [8] consisted of all words $ABCDE$ from the ordered n-letter alphabet 1, 2, . . . , n satisfying either

$$A < B \geqq C,\ D \geqq E \text{ or } A < B < C < D \geqq E,$$

wherein initial digrams and initial tetragrams cannot occur finally.

Theorem 14. For all $k \geqq 5$, a necessary condition for L to be a comma-free dictionary which attains the upper bound (1) is that no initial $(k-1)$-gram occur finally.

Proof. Let L be maximum comma-free, and suppose γ is a $(k-1)$-gram such that both $A\gamma$ and γB are in L. Then as in Theorem 13, all "duplex" words of the type $ABB \cdots BB$, $AAB \cdots BB$, $\cdots$, $AAA \cdots AB$ must be in L. Consider the equivalence class containing $ABABB \cdots B$. The only cyclic permutations not beginning with B, and hence possible candidates for inclusion in L, are $ABABB \cdots B$ itself, and $ABB \cdots BAB$. Each of these, however, occurs in the overlaps of successive "duplex" words.

Speculation. All dictionaries in [8] which attained the upper bound (1) for odd k have the property that no initial i-gram is a final i-gram for any even $i < k$. To what extent this characterizes the maximum comma-free dictionaries is not known. Theorem 2 shows that for $k = 3$, this property is characteristic, *except* when $n = 2$. It is reasonable to hypothesize the possibility of finding maximum dictionaries satisfying the condition of non-ambiguity of even-grams for *all* odd k. The conjecture that for odd $k > 3$ these are the *only* maximum comma-free dictionaries is less likely.

7. Transposable Dictionaries.

The chemical structure of deoxyribonucleic acid (*DNA*) is known to comprise two oppositely polarized strands, here viewed as sequences whose terms are the four bases (here labelled *A*, *B*, *C*, *D*) in seemingly random order, except that *A* in one strand is always opposite *B* in the other, and *C* in one strand is always opposite *D* in the other. (This situation is illustrated in Figure 1.)

It is quite possible that a 180° rotation of this configuration would give rise to new ambiguities, so that in addition to the comma-free condition, an auxiliary restriction should be imposed on the dictionary structure. These considerations motivate the definitions and theorems which follow.

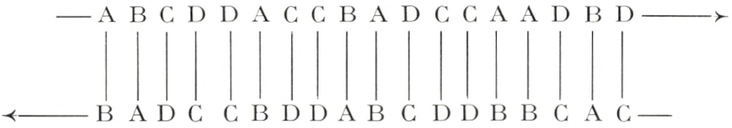
A B C D D A C C B A D C C A A D B D
B A D C C B D D A B C D D B B C A C

Fig. 1. The two opposititely polarized strands of *DNA*.

Def. In an alphabet in which there is an even number n of letters, to each letter X can be assigned a *transpose* letter X' such that $X \neq X'$ and $(X')' = X$. There are then $\frac{n}{2}$ *pairs* of transpose letters.

Def. Given the word or letter sequence $AB \cdots G$, the *transpose* is defined to be $(AB \cdots G)' = G' \cdots B'A'$. A word or collection of words transformed back into itself under transposition is called *symmetric*.

Def. Given a comma-free dictionary L, the *transpose dictionary* L' contains the transposes of the words of L. Further, L is called *transposable* provided that no word of L can appear as a word of L' nor as an overlap of two words of L'.

Theorem 15. The number of words in a comma-free transposable dictionary L, containing k-letter words from an n-letter alphabet, does not exceed one-half the number of non-symmetric complete equivalence classes.

Proof. Every complete class $(AB \cdots FG,\ B \cdots FGA, \cdots, GAB \cdots F)$ has a transpose class $(G'F' \cdots B'A',\ A'G'F' \cdots B', \cdots, F' \cdots B'A'G')$. If $AB \cdots FG$ is in L, then $G'F' \cdots B'A'$ is in L', and looking at the overlaps of $(G'F'\cdots B'A')\,(G'F'\cdots B'A')$ the entire transpose class is seen to occur. Thus, representing one class in L precludes the representation of its transpose class. If the class is symmetric, it is its own transpose class, and cannot be represented at all. Otherwise, either the original class, the transpose class, or neither, but not both, may be represented in L. Hence the theorem.

Corollary. For odd k, the size $S_k(n)$ of a comma-free transposable dictionary satisfies

$$S_k(n) \leqq \frac{1}{2\,k}\sum_{d/k} \mu(d)\, n^{k/d}. \tag{2}$$

For even k, (2) holds with strict inequality.

Proof. The statement for odd k simply combines the inequality (1) with Theorem 15. The strict inequality for even k devolves from the fact that there are always symmetric, complete classes—e.g. the class of $AA\cdots AA'A'\cdots A'$, which is clearly symmetric, and complete because $A \neq A'$.

Note. A careful count of symmetric classes shows

$S_2(n) \leqq (n^2 - 2\,n)/4$ and $S_4(n) \leqq (n^4 - 3\,n^2 + 2\,n)/8$. [Added in proof: The general result is

$$S_k(n) \leqq \frac{1}{2\,k}\sum_{d/k} \mu(d)\, n^{k/d} - \frac{1}{4}\sum_{2d/k} \mu(d)\, n^{k/2\,d}.]$$

Theorem 16. For $k = 3$ and even n, there exist comma-free transposable dictionaries which attain the size $\frac{n^3 - n}{6}$, which corresponds to the upper bound (2) when $k = 3$.

Proof. In each transpose pair X, X', designate one member as *primary* and the other as *secondary*. (This designation can be performed in any of $2^{n/2}$ ways.) With the $\frac{n}{2}$ primary letters, form a maximum comma-free dictionary of

$$\frac{\left(\frac{n}{2}\right)^3 - \frac{n}{2}}{3} = \frac{n^3}{24} - \frac{n}{6}$$

three-letter words. This is Part I of the new dictionary. Adjoin Part II, consisting of all sequences of primary-secondary-primary letters. There are $\frac{n}{2}$ choices for each of the three positions, hence $\frac{n^3}{8}$ words in Part II. Part I and Part II are clearly disjoint, and their union is a dictionary L containing $\frac{n^3-n}{6}$ words.

No initial digram of Part II can occur finally in L, because no word of L ends in a secondary letter. No final digram of Part II occurs initially in L, because no word of L begins with a secondary letter. Thus the only violations of comma-freedom in L are within Part I, but Part I was comma-free by hypothesis.

It remains to show that L is transposable. Transposes of words from Part I consist entirely of secondary letters. Transposes of words from Part II have the pattern secondary-primary-secondary. Hence any three consecutive letters in a message written using words of the transpose dictionary L' will include at least two secondary letters. Since words of L contain at most one secondary letter each, L is transposable.

Caution. If $n = 4$, there are only 2 primary letters. A comma-free dictionary using these two letters and not satisfying the digram condition of Theorem 2 must not be used as Part I of a transposable dictionary. Not only does the proof of Theorem 16 break down in such cases, but the conclusion itself is false.

Examples. For the four-letter alphabet A, B, C, D, with $A' = B$, $C' = D$, suppose that A and C are designated as primary. The

Table V.

The three patterns for transposable comma-free dictionaries, for the case $k = 3$, $n = 4$.

Part I	1. *A A C*	2. *A A C*	3. *A A C*
	C A C	*C C A*	*C A C*
Part II	*A B A*	*A B A*	*A B B*
	A B C	*A B C*	*A B C*
	C B A	*C B A*	*C B B*
	C B C	*C B C*	*C B C*
	A D A	*A D A*	*A D B*
	A D C	*A D C*	*A D C*
	C D A	*C D A*	*C D B*
	C D C	*C D C*	*C D C*

first two dictionaries of Table V are formed in accordance with Theorem 16, while the third dictionary is an independently discovered example.

Given any transposable comma-free dictionary in Table V, a symmetry group H containing sixteen operators can be applied to obtain new dictionaries. This group is generated by the three alphabetic permutations (AB), (CD), and $(AC)(BD)$, and by the operation of reading all words backward. The eight alphabetical permutations form a group isomorphic to the dihedral group of the square with consecutive vertices $ACBD$. Under the operations of H, dictionaries 1. and 3. in Table V lead to sixteen dictionaries each. However, dictionary 2. leads to only eight dictionaries, because the permutation $(AC)(BD)$ leaves dictionary 2. unchanged. Thus, Table V yields a total of forty distinct transposable comma-free dictionaries for the case of $k = 3$, $n = 4$, and these forty are believed to exhaust the possibilities.

All forty of the ten-word dictionaries just described satisfy the digram condition of Theorem 2. Moreover, each is embeddable in one or more of the 408 twenty-word dictionaries obtainable from Table I. The extent to which these digram and embeddability properties are fortuitous is not known.

For $k = 4$, considerably less is known. For $k = 4$, $n = 4$, the note following Theorem 15 yields 27 as an upper bound to the size of a transposable comma-free dictionary. No such example has been constructed. In Table VI, however, a 26-word dictionary for this case is exhibited. Part I of this dictionary was obtained from an 18-word maximal comma-free dictionary using only A, B and C, from which two words were omitted because of the transposability constraint. In Part II, representatives of ten of the remaining eleven complete non-symmetric classes appear. They were chosen to represent their classes according to the following criteria:

1) D never begins or ends a word,
2) D occurs as the second letter,
3) In case of competing candidates, the one with the fewest D's is chosen.

The 27th class, which contains $A\ D\ B\ C$, cannot be added consistently to the dictionary already formed.

Table VI.

A comma-free transposable dictionary for $k = 4$, $n = 4$, containing 26 words. (One non-symmetric complete class is unrepresented.)

Part I		Part II	
A B B B	*A C C B*	*C D C B*	*B D C C*
A C B A	*A C C C*	*C D B C*	*C D D B*
A C B B	*B A C A*	*C D C A*	*B D B C*
A C B C	*B B C A*	*C D C C*	*C D B B*
A A C A	*B B C B*	*A D C C*	*B D C B*
A B C A	*B C C A*		
A B C B	*B C C B*		
A C C A	*B C C C*		

In contrast to the equal frequency of letters (Theorem 1) in maximum comma-free dictionaries, the dictionary of Table VI contains *A*, *B*, *C*, *D*, with the respective frequencies 21, 31, 41, 11. The dictionaries of Table V are also seen to deviate significantly from equal letter frequency. The proper analog of Theorem 1 for the case of transposable dictionaries asserts that every *transpose pair* of letters is represented equally often in a dictionary containing words from all the non-symmetric complete equivalence classes. The sharing of representation by members of the same pair is less rigidly constrained.

California Institute of Technology, Pasadena, U. S. A.

References.

1. Watson, J. D., and Crick, F. H. C.: "A Structure for Deoxyribose Nucleic Acid." Nature **171**: 737 (1953).
2. Hoagland, M. B., Stephenson, M. L., Scott, J. F., Hecht, L. I., and Zamecnik, P. C.: "A Soluble Ribonucleic Acid Intermediate in Protein Synthesis." J. Biol. Chem. **231**: 241–257 (1958).
3. Berg, P., and Ofengand, E. J.: "An Enzymatic Mechanism for Linking Amino Acids to RNA." Proc. Natl. Ac. Scie. (U.S.A.) **44**: 78–86 (1958).
4. Schweet, R. S., Bovard, F. C., Allen, E., and Glassman, E.: "The Incorporation of Amino Acids into Ribonucleic Acids." Proc. Natl. Ac. Scie. (U.S.A.) **44**: 173–177 (1958).
5. Gamow, G.: "Possible Mathematical Relation Between Deoxyribonucleic Acid and Proteins." Biol. Medd. Dan. Vid. Selsk. **22**: no. 2 (1954).
6. Brenner, S.: "On the Impossibility of All Overlapping Triplet Codes in Information Transfer from Nucleic Acids to Proteins." Proc. Natl. Ac. Scie. (U.S.A.) **43**: 687–694 (1957).
7. Crick, F. H. C., Griffith, J. S., and Orgel, L. E.: "Codes Without Commas." Proc. Natl. Ac. Scie. (U.S.A.) **43**: 416–421 (1957).
8. Golomb, S. W., Gordon, B., and Welch, L. R.: "Comma-free Codes." Can. J. Math. **10**: 202–209 (1958).
9. Meselson, M., and Stahl, F. W.: "The Replication of DNA in *Escherichia coli*." Proc. Natl. Ac. Scie. (U.S.A.) **44**: nr. 7 (1958).
10. Shapiro, H. S., and Chargaff, E.: "Studies on the Nucleotide Arrangement in Deoxyribonucleic Acids. II. Differential Analysis of Pyrimidine Nucleotide Distribution as a Method of Characterization." Biochem. Biophys. Acta **26**: 609–623 (1957).
11. Jacobson, N.: "Lectures in Abstract Algebra." D. van Nostrand Co., Inc., 1951, Vol. I, p. 207.

Indleveret til selskabet den 9. juni 1958.
Færdig fra trykkeriet den 7. november 1958.

Det Kongelige Danske Videnskabernes Selskab

Biologiske Meddelelser

(Biol. Medd. Dan. Vid. Selsk.)

Bind **22** (kr. 65,00)

		kr. ø.
1.	Jensen, P. Boysen: Untersuchungen über Determination und Differenzierung. 2. Über die Wachstumsvorgänge in der Spitze der Wurzelhaare von *Phleum*. With an English Summary. 1954	3,50
2.	Böving, Adam G.: Mature Larvae of the Beetle-Family Anobiidae. 1954	35,00
3.	Gamow, G.: Possible Mathematical Relation between Deoxyribonucleic Acid and Proteins. 1954	2,00
4.	Børgesen, F.: Some Marine Algae from Mauritius. Additions to the Parts Previously Published. VI. 1954	8,00
5.	Jensen, P. Boysen: Untersuchungen über Determination und Differenzierung. 3. Über die Wirkungsweise des determinierenden Faktors, der bei der Bildung der Wurzelhaare von *Lepidium, Sinapis* und *Phleum* tätig ist. With an English Summary. 1955	4,50
6.	Salomonsen, Finn: The Evolutionary Significance of Bird-Migration. 1955	6,00
7.	Munch-Petersen, Agnete, Kalckar, Herman M., and Smith, Evelyn E. B.: Uridyl Transferases, their Occurrence and Physiological Role. 1955	3,00
8.	Gamow, G.: On Information Transfer from Nucleic Acids to Proteins. 1955	1,00
9.	Hevesy, G.: Conservation of Skeletal Calcium Atoms through Life. 1955	2,00

Bind **23**

(*uafsluttet/in preparation*)

1.	Rasmussen, Erik: Faunistic and Biological Notes on Marine Invertebrates III. The Reproduction and Larval Development of some Polychaetes from the Isefjord, with some Faunistic Notes. 1956	11,00
2.	Petersen, Johs. Boye, and Hansen, J. Benth: On the Scales of some Synura Species. 1956	7,00
3.	Brøndsted, H. V.: Experiments on the Time-Graded Regeneration Field in Planarians. With a Discussion of its Morphogenetic Significance. 1956	7,00
4.	Børgesen, F.: Some Marine Algae from Mauritius. Final Part. Edited by Tyge Christensen. 1957	5,00

		kr. ø.
5.	JENSEN, P. BOYSEN: Untersuchungen über Determination und Differenzierung. 4. Über den Aufbau des Zellwandgerüstes der Pflanzen und die Determination desselben. With an English Summary. 1957	6,00
6.	LARSEN, KAI: Cytological and Experimental Studies on the Genus *Erodium* with Special References to the Collective Species *E. Cicutarium* (L) L'Her. 1958	4,00
7.	PETERSEN, JOHS. BOYE, and HANSEN, J. BENTH: On the Scales of some Synura Species. II. 1958	3,50
8.	JENSEN, P. BOYSEN: Untersuchungen über Determination und Differenzierung. 5. Über die Wirkungsweise des Wuchsstoffes in dem Epikotyl von *Phaseolus* (Die Brückentheorie der Wuchsstoffwirkung). With an English Summary. 1958	7,00
9.	GOLOMB, S. W., WELCH, L. R., and DELBRÜCK, M.: Construction and Properties of Comma-Free Codes. 1958	5,00

On direct application to the agent of the Academy, EJNAR MUNKSGAARD, Publishers, 6 Nörregade, Köbenhavn K., a subscription may be taken out for the series of *Biologiske Meddelelser.* This subscription automatically includes the *Biologiske Skrifter* in 4to as well, since the *Meddelelser* and the *Skrifter* differ only in size, not in subject matter. Papers with large formulae, tables, plates etc., will as a rule be published in the *Skrifter,* in 4to.

For subscribers 'or others who wish to receive only those publications which deal with a single group of subjects, a special arrangement may be made with the agent of the Academy to obtain the published papers included under one or more of the following heads: *Botany, Zoology, General Biology.*

In order to simplify library cataloguing and reference work, these publications will appear without any special designation as to subject. On the cover of each, however, there will appear a list of the most recent papers dealing with the same subject.

The last published numbers of *Biologiske Meddelelser* within the group of **General Biology** are the following:
Vol. **22**, nos. 3, 7—9. — Vol. **23**, no. 9.

Printed in Denmark.
Bianco Lunos Bogtrykkeri A/S