Genetic Information Analysis of Bacteriophage Φx 174 *

R. Figueroa

Departamento de Matemáticas, Facultad de Ciencias, Universidad de Chile

and

A. Sepúlveda, M. A. Soto, and J. Tohá

Biofísica, Departamento de Física y Departamento de Matemáticas,
Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile

(Z. Naturforsch. **32 c**, 850–854 [1977]; received June 29/August 1, 1977)

Bacteriophage Φx 174, Genetic Information

The genetic information of Φx 174 genome (genes and intermediate segments) is analyzed in terms of its independent (D_1 index) and dependent information (D_2 and D_3 Markovian indexes), as well as of its ability to generate secondary structure.

Genes B and E, enclosed in A and D respectively, have: 1) values of D_1 and D_3 indexes closer to the theoretical random distribution curves than those of (A–B) and (D–E) gene fractions, and 2) in the ability for secondary structure generation minor differences with genes A and D.

F → G and mRNA start → A intermediate segments differ from randomness in their D_1 and D_2 indexes, but not so much in the D_3 values.

All these data point out the use of code degeneracy for increasing the genetic information density of the virus.

Introduction

F. Sanger *et al.* [1] have recently established the Bacteriophage Φx 174 DNA primary structure, which codes for 9 different protein (A, B, C, D, E, J, F, G and H). These proteins represent a greater information than those expected from the virus genome, indicating a necessary overlapping in DNA information, as has been determined for gen B (enclosed in gene A) and for gene E (enclosed in gene B) [2]. This overlapped information hinders genetic code degeneracy to operate, since gene A triplet's third base corresponds to gene B triplet's first base, and so on, in such a way that the third base remains determined.

For the genetic information analysis of Bacteriophage Φx 174, the following parameters were considered, for each gene sequence, as well as for intermediate segments:

1) Degree of randomness in nucleotide chains.
2) Degree of Markovian dependence of any 2 or 3 successive DNA bases.
3) The ability of primary sequences to generate secondary structure, by successive pairing of complementary bases in some singular zones. (See Methods.)

Requests for reprints should be sent to J. C. Tohá, Facultad de Ciencias Físicas y Matemáticas, Departamento de Física, Universidad de Chile, Casilla 5487, *Santiago*, Chile.

* This work was partially supported by OEA.

Methods

To examine the randomness of Bacteriophage Φx 174 DNA primary structure, the index D_1 [3] was evaluated from the maximal possible informational entropy and the observed one ($D_1 = H_{max} = H_1$), for each gene or intermediate segment. The maximal informational entropy is obtained from equiprobable representation of the four DNA bases, following Shannon's definition:

$$H_{max} = -\sum_{i=1}^{4} p_i \log p_i$$

p_i = frequence of any base in the chain (in this case, $p_i = 1/4$)

and H_1 corresponds to the observed informational entropy (p_i = observed frequency of each base in the polynucleotide analyzed). In consequence, D_1 index measures the tendency of the polynucleotide chain to use some bases more than others.

The D_1 values obtained were compared with those corresponding to the same length randomly generated sequences (Subrutine Randu, IBM 370).

The Markovian dependent information, given by the tendency to use some base pairs or triplets more than others, was evaluated by means of D_2

and D_3 indexes respectively. ($D_2 = H_1 - H_M^1$; $D_3 = H_M^1 - H_M^2$; where $H_M^m = -\sum_{i\sigma} p_{i1}\, p\, i_1\, i_2\, p\, i_1 \ldots i_{m+1} \log p\, i_1 \ldots i_{m+1}$.) All values were calculated in nats. (This is the informational unit when the logarithm is to the base e.)

Finally, the ability to form secondary structure was valorated considering the number of n successive complementarities ($n = 0, 1, 2 \ldots 16$) formed running the nucleotide sequence against itself [4]. As before, the values for each gene or intermediate segment were compared with those corresponding to sequences at random generated.

Results

In this work we have examined the genes (A, B, C, D, E, J, F, G and H) of virus as well as their intermediate segments, in comparison with nucleotide sequences at random generated, ranging from length 19 to 1536.

The curve obtained for $\overline{D_1}$ theoretical mean values of random nucleotides is compared with points corresponding to genes (Fig. 1) and to intermediate segments (Fig. 2). The D_1 values for genes are noticeably over the theoretical curve except for gene E (enclosed in gene D) and gene B (overlapped to gene A), which are significantly closer to $\overline{D_1}$ curve than (D − E) and (A − B), respectively (Fig. 1).

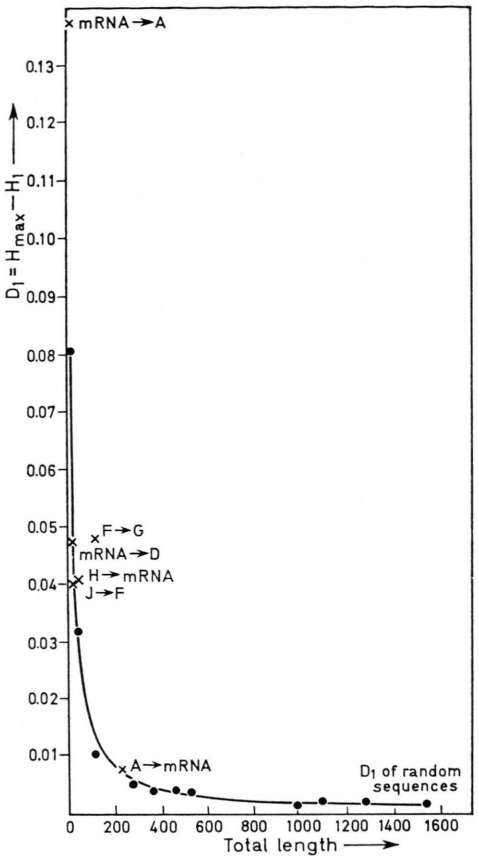

Fig. 2. D_1 values of intermediate segments between Φx 174 genes, in comparison to a D_1 curve of randomly generated sequences.

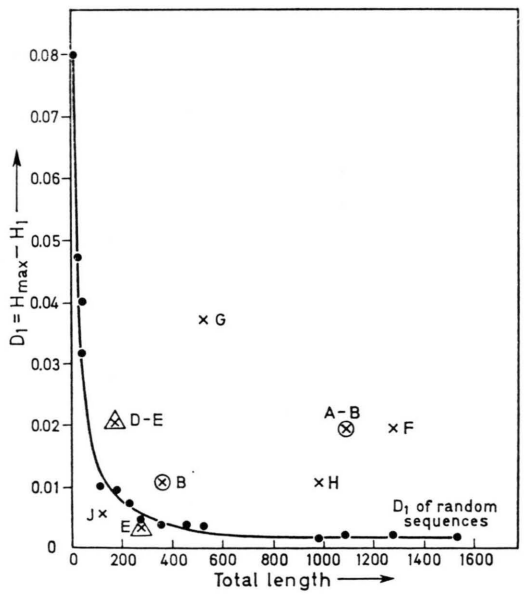

Fig. 1. D_1 values of divergence from equiprobability of Φx 174 genes, in comparison with a $\overline{D_1}$ curve corresponding to randomly generated sequences.

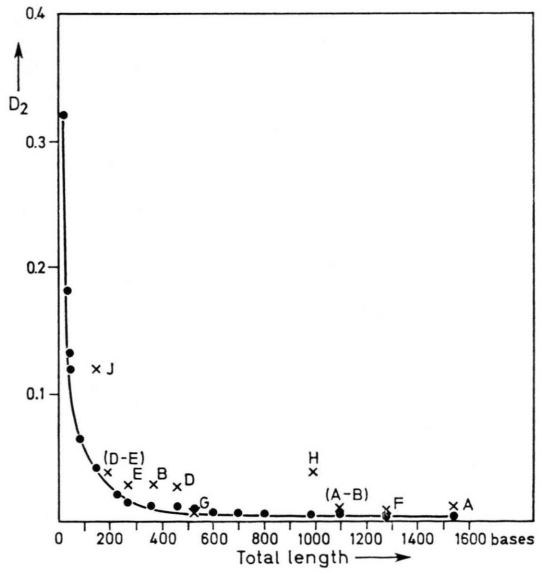

Fig. 3. D_2 values of divergence from independence by pairs of bases in Φx 174 genes, in comparison to a $\overline{D_2}$ curve of randomly generated sequences.

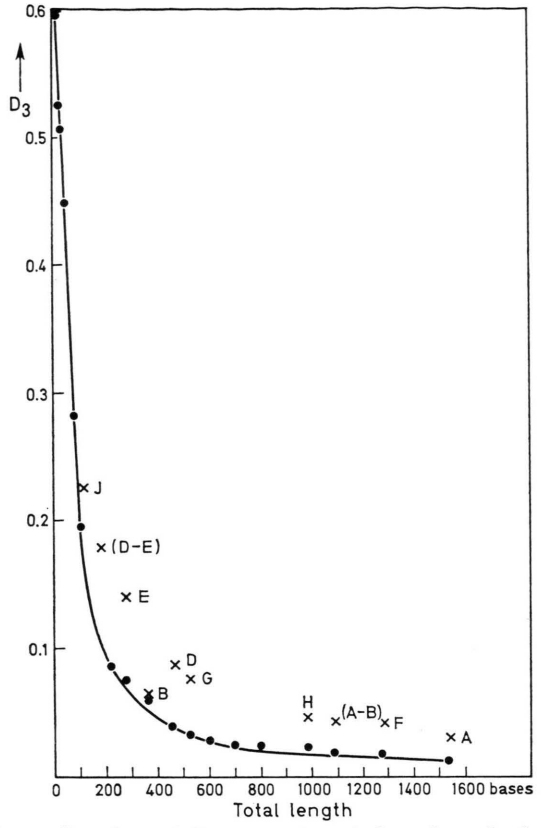

Fig. 4. D_3 values of divergence from independency by base triplets in Φx 174 genes, in comparison to a $\overline{D}_3$ curve of randomly generated sequences.

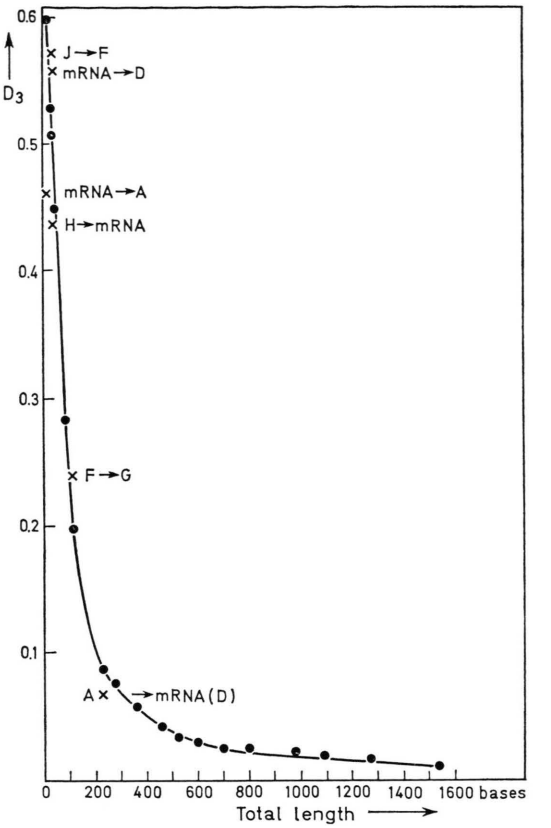

Fig. 6. D_3 values of intermediate segments between Φx 174 genes, in comparison to a $\overline{D}_3$ curve of randomly generated sequences.

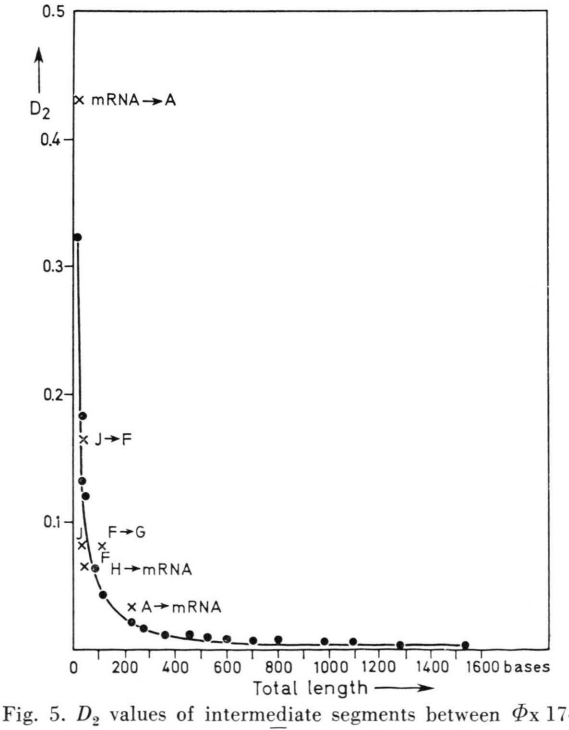

Fig. 5. D_2 values of intermediate segments between Φx 174 genes, in comparison to a $\overline{D}_2$ curve of randomly generated sequences.

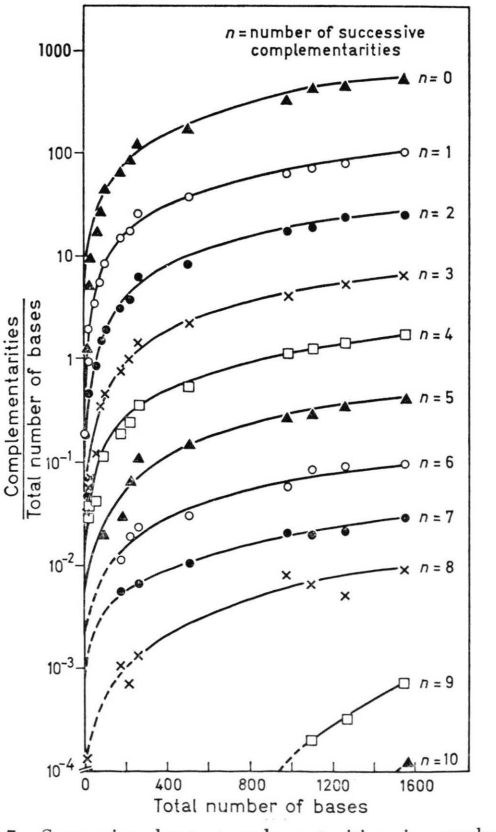

Fig. 7. Successive base complementarities in randomly generated sequences.

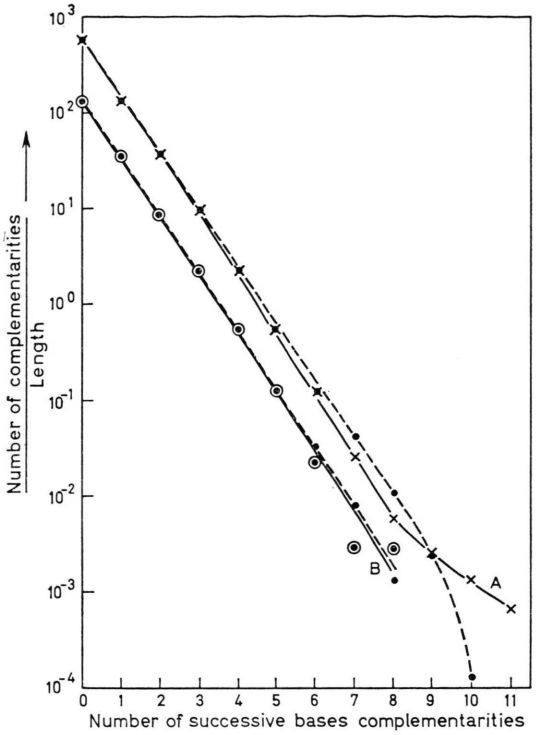

Fig. 8. Theoretical successive base self-complementarities found in genes A and B, and in randomly generated sequences of equal length. (A—B values, not shown in Fig., are similar to A values.)

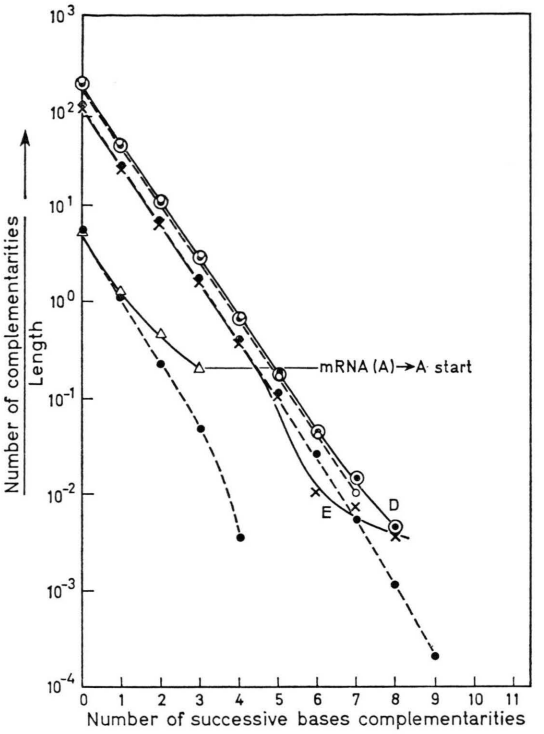

Fig. 9. Theoretical successive base self-complementarities found in genes D and E and intermediate segment mRNA (A) → A start, and in randomly generated sequences of equal length. (D—E values, not shown in Fig., are similar to *D* values.)

In Fig. 2, intermediate segments between F and G genes and between gene A mRNA start and A, appear randomless.

Figs 3 and 4 show the values of D_2 and D_3 for genes, referred to random theoretical distribution curves $\overline{D}_2$ and $\overline{D}_3$. The points for genes are above the curves, being specially remarkable the D_3 values. However, values for genes E and B are more approximate to $\overline{D}_3$ curve than (D − E) and (A − B) D_3 points respectively.

In the case of intermediate segments, the values of D_2 (Fig. 5) displaying a greater difference with $\overline{D}_2$ curve, are those corresponding to segments between F and G, and between mRNA start (A) and A. These differences diminish when referring to D_3 values (Fig. 6).

In Fig. 7, groups of n successive base complementarities formed by nucleotide sequences at random generated, are plotted as a function of the length of the sequence.

The ability for secondary structure generation of genes and intermediate segments is analyzed in Figs 8 and 9, where no noticeable differences with random curves were found.

Discussion

When we analyze Φx 174 genetic information, a significant pattern of non-randomness appears, looking specially to D_1 and D_3 values. This fact could support the hypothesis of a restriction in the basic information of the micro-organism favouring a greater fidelity of message.

Genes B and E, included in A and D respectively, display D_1 and D_3 values significantly closer to $\overline{D}_1$ and $\overline{D}_3$ theoretical curves than segments (A − B) and (D − E), indicating a greater randomness, possibly promoted by the third base invariancy.

In the case of intermediate segments: F → G and mRNA (A) → A, the values of D_1 and D_2 are significantly different from theoretical values, being

this difference lessened in the D_3 comparison, results in accordance with the information of these segments not translated as codons.

Finally, the minor ability to form secondary structure found in genes B and E, relatives to A and D is not relevant.

[1] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III, P. M. Slocombe, and M. Smith, Nature **265,** 687 [1977].
[2] B. G. Barrell, G. M. Air, and C. A. Hutchison III, Nature **264,** 34 [1976].
[3] L. L. Gatlin, Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (L. M. Le Cam, J. Neyman, and E. L. Scott, ed.), p. 277–296, Univ. California Press 1972.
[4] R. Figueroa, A. Soto, G. González, M. Pieber, C. Romero, and J. Tohá, J. Theor. Biol. **36,** 321 [1972].