STATISTICAL CORRELATION OF PROTEIN AND RIBONUCLEIC ACID COMPOSITION

By George Gamow and Martynas Yčas

DEPARTMENT OF PHYSICS, GEORGE WASHINGTON UNIVERSITY, WASHINGTON, D.C., AND PIONEERING RESEARCH DIVISION, QUARTERMASTER RESEARCH AND DEVELOPMENT CENTER, NATICK, MASSACHUSETTS

Communicated September 29, 1955

Empirical evidence indicates with increasing clearness that ribonucleic acid (RNA) plays a vital role in protein synthesis. It appears rational to assume that the sequence of amino acids characterizing a given protein is uniquely determined by the sequence of nucleotides in the ribonucleic acid molecule.

While RNA is a polymer of four different nucleotides, proteins are polymers of 20 different amino acids. Since it is possible to form 20 kinds of triplets from four different elements, this suggests that each of the 20 amino acids is determined by a triplet of nucleotides, taken without regard to order.[1]

The fact that the internucleotide distances are comparable with the distances of amino acid residues in a protein when both are in the extended form makes it plausible that a given amino acid shares its determining nucleotides with neighboring amino acids. This would necessitate a correlation between neighboring residues in protein sequences, making certain pairs favored and others excluded (1, 2).

However, studies by Gamow, Rich, and Yčas (2) show that there does not appear to be any such interresidue correlation, and all sequences are apparently possible. Thus it appears more probable that the number of determining nucleotides exceeds by a factor of 3 the number of amino acid residues in the synthesized protein, so that neighboring residues do not share determining nucleotides.

It is not possible to test this hypothesis critically against available information on amino acid sequences, since all sequences are permitted. However, the model predicts that the statistical frequency of amino acid residues in proteins will show certain regularities. We have therefore attempted to test whether the distribution of amino acids predicted from the proposed model corresponds with analytically found distributions.

If one arranges the amino acids in a protein in order of abundance, repeats this on a collection of proteins,[2] and takes the mean values (without regard to identity) of the most abundant, second most abundant, third most abundant, etc., one obtains a curve shown in Figure 1. This will be referred to as a "distribution." In

order to see whether this curve corresponds to a random distribution, we compare
it with a mathematical model obtained in the following way. A segment is divided
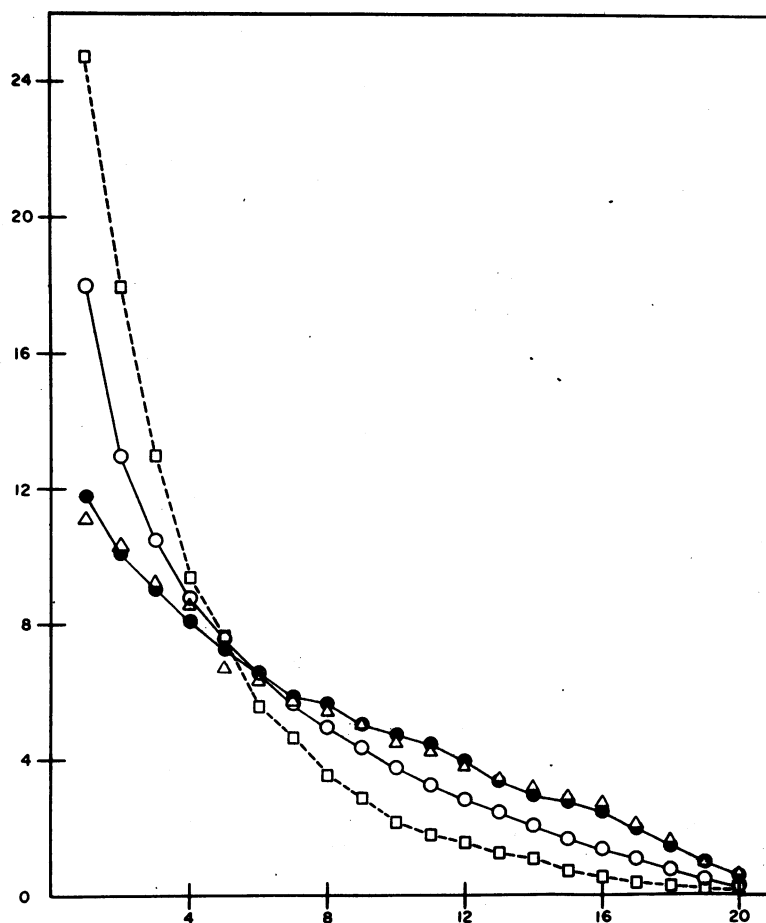into 20 sections at random, and the lengths of the longest, second longest, etc., are

Fig. 1.—Distributions (as defined in text): ●, amino acids, 22 proteins;
△, triplets from 7 nonviral RNA's; □, random triplets, Monte Carlo, 3,000
runs; ○, random distribution, $n = 20$, by von Neumann's formula; *abscissa*,
ordinal rank; *ordinate*, relative frequency in per cent.

averaged over a large number of such divisions. This problem possesses an analyt-
ical solution for which we are indebted to John von Neumann. If the unit length
is divided randomly into n sections and a_1, a_2, a_3, etc., are the mean lengths of the
longest, second longest, etc., sections, then

$$a_j^{(n)} = \frac{1}{n}\left(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{n-j+1}\right).$$

The results given by this formula for $n = 20$ are likewise plotted in Figure 1. It is
obvious that the distribution of amino acid residues in a collection of proteins de-
viates markedly from the random model distribution.

Assuming that the determination of amino acids is by nonoverlapping triplets, this deviation from randomness either originates in the translation procedure operating on a random distribution of nucleotides or is due to a deviation from randomness in the composition of the template itself.

In order to explore the first possibility, it was decided to go through the following Monte Carlo procedure: select four random fractions, normalized to one, and calculate the probabilities of the twenty different triplets. Assuming no bias, the relative frequency of each individual triplet will be given by the product of the frequencies of the component elements. It will be noted that there are three kinds of triplets of four different elements. Four contain three different elements (*abc*), sixteen contain two (*aab*), and four contain one (*aaa*). Since we consider triplets differing only in the order of elements as identical, the relative frequencies must further be multiplied by a weighting factor of 6 for the *abc* type, 3 for the *aab* type, and 1 for the *aaa* type. Repeat this procedure many times and average the amounts of the most abundant, second most abundant, and so on, triplets. This was done for us by Guilio Fermi and Nicholas Metropolis, using the electronic computer MANIAC of the Los Alamos Scientific Laboratory. The result of 3,000 runs is plotted in Figure 1. It will be noted that the curve deviates from the amino acid distribution curve even more than the previous theoretical one. Thus, if the triplet hypothesis is correct, the deviation from randomness of the amino acid distribution must arise from the nonrandom composition of the template.

The distribution, defined in the same way as for proteins, of 7 RNA's[2] is plotted in Figure 2, along with the random distribution expected from Neumann's formula with $n = 4$. The distribution of RNA, like that of protein, does indeed deviate markedly from the random. The attempt was made to see whether this deviation could be due to the fact, recently observed by Elson and Chargaff (3), that the total of adenine plus cytosine tends to equal the sum of guanine plus uracil. If this were the case, then the above-discussed random model should be modified in the following way. A unit length is divided into two halves and each half broken at random into two. According to the result obtained for us by S. Ulam, the means of the longest, second longest, etc., lengths must stand in the ratio $5/12$, $4/12$, $2/12$, and $1/12$. This is also plotted in Figure 2. Although this restriction brings the curve closer to the empirical one, the deviation is still marked. Thus the distribution of RNA does deviate essentially from random and in the same direction as the protein distribution.

Next, using the actual composition of the same 7 samples of RNA, the frequency of individual triplets was calculated for each RNA, the triplets arranged in decreasing order of magnitude, and the average values for the most abundant, second most abundant, etc., calculated. The results obtained (Fig. 1) coincide almost perfectly with the distribution of our sample of 22 proteins.

The same procedure was followed for 3 viruses, where the sample, although smaller, has the advantage of a presumably more direct relation of the RNA and protein. The results are similar to the previous ones, although the fit is less perfect (Figs. 3 and 4).

The above results seem to show that (1) the proportions of amino acids in proteins are not random; (2) this nonrandomness is not due to the application of the triplet translation procedure to a random RNA constitution; (3) the application of

the same translation procedure to the actual RNA composition leads to an excellent agreement with the observed amino acid distribution.

A further prediction from the model may be noted. Because of different weighting factors for triplets of the *abc*, *aab*, and *aaa* types, individual amino acids would be expected to be consistently abundant, rare, or of intermediate frequency within
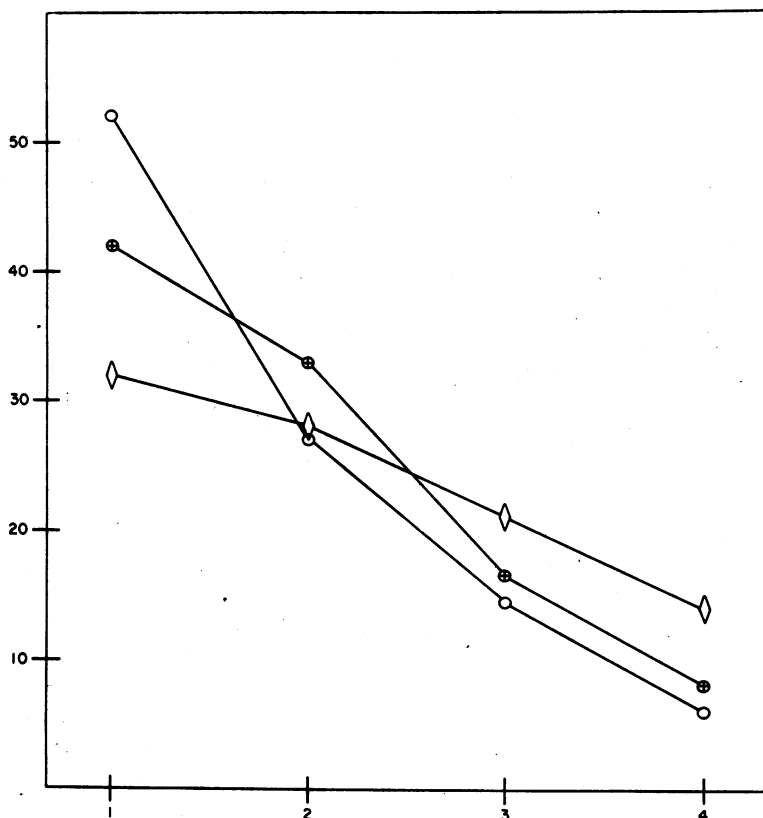
FIG. 2.—Distributions (as defined in text): ◇, 7 nonviral RNA's; ○, random distribution, $n = 4$, by von Neumann's formula; ⊕, random distribution if $A + B = C + D$ (after S. Ulan); *abscissa*, ordinal rank; *ordinate*, relative frequency in per cent.

wide limits of variation in composition of the RNA template. Tristram (4) has indeed observed this to be the case, each amino acid tending to be normally distributed about its characteristic frequency.

We consider that these results speak strongly in favor of the original hypothesis that amino acid residues in proteins are selected by independent triplets of nucleotides taken without regard to order.

FIG. 4.—Same procedure as in Fig. 3, for 3 plant virus RNA's.

FIG. 3.—Same procedure as in Fig. 1, for 3 plant virus proteins.

1015

APPENDIX I

RANDOM DIVISION OF A UNIT LENGTH INTO n PARTS
(AFTER J. VON NEUMANN)

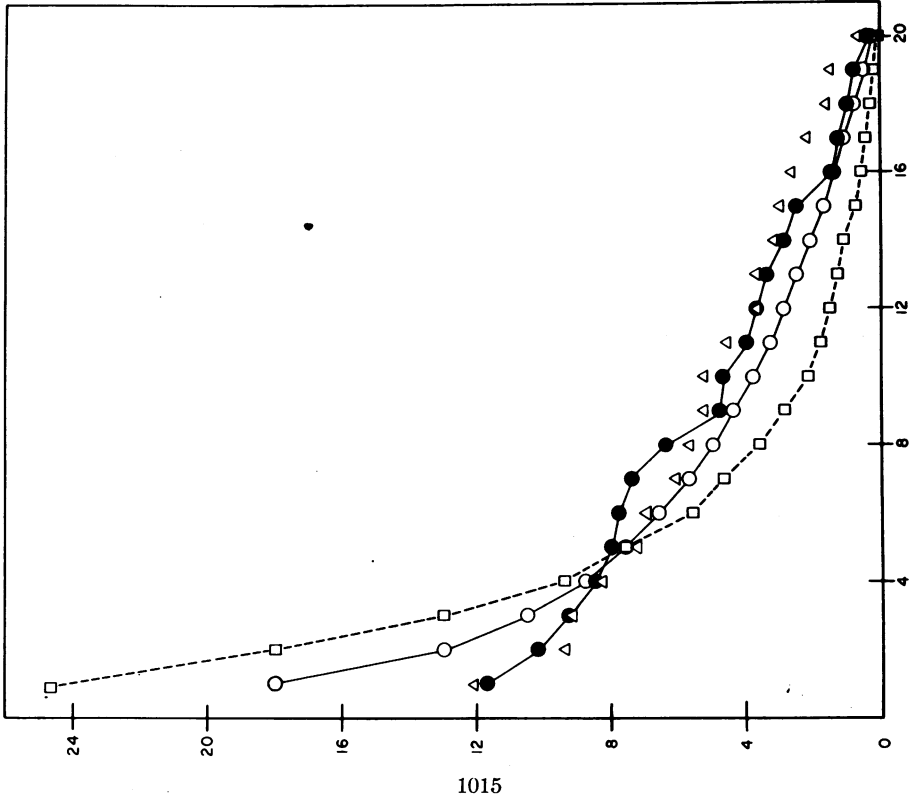Consider a unit length randomly divided into n parts so that the lengths of individual sections, as they follow from left to right, are $x_1, x_2, x_3, \ldots, x_n$. The values of x_i are subjected to the conditions

$$\sum_{i=1}^{n} x_i = 1 \quad \text{and} \quad 0 < x_i < 1.$$

Now let us define n numbers y_j as

$$y_1 = \text{smallest of all } x\text{'s},$$
$$y_2 = 2^d \text{ smallest of all } x\text{'s},$$
$$\ldots \ldots \ldots \ldots \ldots \ldots$$
$$y_n = \text{largest of all } x\text{'s}.$$

The values of y_j are subject to the conditions

$$\sum_{j=1}^{n} y_j = 1 \quad \text{and } 0 < y_1 < y_2 < \ldots < y_n < 1.$$

Considering the problem in n-dimensional space, we can use the statistical weight

$$d\pi = dy_1\, dy_2 \ldots dy_{n-3}\, dy_{n-2}\, dy_{n-1}$$
$$= dy_1\, dy_2 \ldots dy_{n-3}\, dy_{n-2}\, dy_n$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$
$$= dy_1\, dy_3\, dy_4 \ldots dy_{n-2}\, dy_{n-1}\, dy_n$$
$$= dy_2\, dy_3\, dy_4 \ldots dy_{n-2}\, dy_{n-1}\, dy_n.$$

The problem is to find the mean values of $y_1, y_2, \ldots, y_n$ for all possible divisions of the unit length. Put

$$z_j = y_j - y_{j-1} \quad (j = 1, 2, \ldots, n).$$

Then, clearly,

$$y_j = \sum_{k=1}^{k=j} z_k,$$

and the restricting conditions on z_k become

$$nz_1 + (n-1)z_2 + \ldots + z_n = 1; \quad z_k > 0.$$

The weight $d\lambda$ will now be proportional to

$$d\lambda = dz_1\, dz_2 \ldots dz_{n-3}\, dz_{n-2}\, dz_{n-1}$$
$$= dz_1\, dz_2 \ldots dz_{n-3}\, dz_{n-2}\, dz_n$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$
$$= dz_1\, dz_3\, dz_4 \ldots dz_{n-1}\, dz_n$$
$$= dz_2\, dz_3\, dz_4 \ldots dz_{n-1}\, dz_n.$$

Put

$$\omega_k = (n + 1 - k)z_k \quad (k = 1, 2, \ldots, n).$$

Then the restrictions become

$$\sum_{k=1}^{n} \omega_k = 1, \qquad\qquad \omega_k > 0,$$

and the statistical weight is proportional to

$$
\begin{aligned}
d\mu &= d\omega_1 \, d\omega_2 \ldots d\omega_{n-3} \, d\omega_{n-2} \, d\omega_{n-1} \\
&= d\omega_1 \, d\omega_2 \ldots d\omega_{n-3} \, d\omega_{n-2} \, d\mu_n \\
&\quad .\;\;.\;\;.\;\;.\;\;.\;\;.\;\;.\;\;.\;\;.\;\;.\;\;.\;\;.\;\;.\;\;. \\
&= d\omega_1 \, d\omega_3 \, d\omega_4 \ldots d\omega_{n-1} \, d\omega_n \\
&= d\omega_2 \, d\omega_3 \, d\omega_4 \ldots d\omega_{n-1} \, d\omega_n.
\end{aligned}
$$

because of the symmetry of restricting conditions in ω-space,

$$\bar{\omega}_1 = \bar{\omega}_2 = \ldots = \bar{\omega}_n,$$

and, since

$$\sum_{k=1}^{k=n} \omega_k = 1,$$

then

$$\bar{\omega}_1 = \bar{\omega}_2 = \ldots = \bar{\omega}_n = \frac{1}{n}.$$

Therefore,

$$\bar{z}_k = \frac{\bar{\omega}_k}{n+1-k} = \frac{1}{n} \cdot \frac{1}{n+1-k}$$

and

$$\bar{y}_j = \sum_{k=1}^{k=j} \bar{z}_k = \sum_{k=1}^{k=j} \frac{1}{n} \cdot \frac{1}{n+1-k},$$

$$\bar{y}_j = \frac{1}{n}\left(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{n-j+1} \right),$$

as given in the text.

APPENDIX II

Random Division of a Unit Length into Four Parts under the Condition that the First Division Is Made in the Middle
(After S. Ulam)

Consider a unit length broken into two halves, I and II, and each of these again broken randomly into two. Let x be the longest part of I and y the longest part of II. Then the distributions of both x and y are uniform in the interval $1/4$ to $1/2$. We plot x and y in a two-dimensional diagram (Fig. 5) and pick at random a point P within this square. If the point is above the diagonal (as shown in the diagram), we take its y co-ordinate (because it is larger than x), and if it is below the diagonal, we take its x co-ordinate (which is, in this case, largest). The center of gravity of the triangle ULA has the co-ordinate $y = 1/4 + 2/3 \cdot 1/4 = 5/12$, and the same figure gives the x co-ordinate of the triangle AMU. This gives us the mean length of the longest piece.

The second longest piece will be the shorter of the two x or y. This will correspond to the x co-ordinate of the center of gravity of ULA, or the y co-ordinate of the center of gravity of AMU, and is equal to $^1/_4 + ^1/_3 \cdot ^1/_4 = ^4/_{12}$.
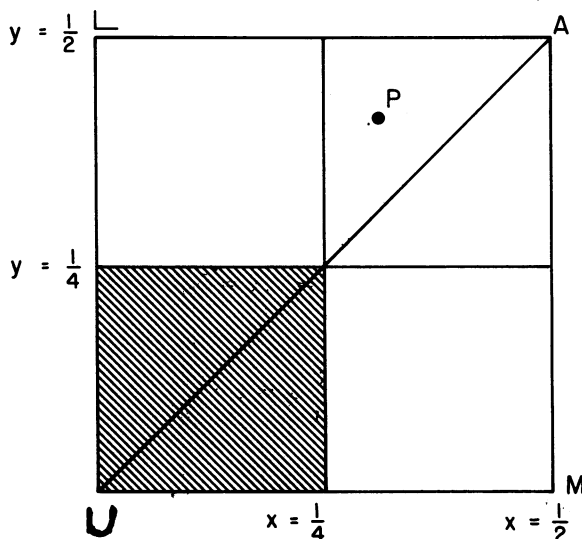

FIG. 5—Graphical solution of restricted division problem.

The shortest of all pieces will correspond to a point in the shaded area of Figure 5. If the point in this area is above the diagonal, we take its x co-ordinate, and if it is below, its y co-ordinate. In each case it is equal to $^1/_3 \cdot ^1/_4 = ^1/_{12}$, which gives the mean length of the shortest piece. The length of the next shortest is, of course, $1 - ^5/_{12} - ^1/_{12} = ^2/_{12}$. Thus the four mean lengths will be

$$^5/_{12}; \ ^4/_{12}; \ ^2/_{12}; \ ^1/_{12}$$

as given in the text.

[1] One of us has previously suggested (1) that the coincidence of the number 20, which is the number of combinations of four different things taken three at a time and the number, also 20, of the different kinds of amino acids occurring in proteins is not accidental. Our belief that this is the case is strengthened by certain biochemical findings. Besides the 20 commonly occurring amino acids, certain others are found as components of one or a few proteins (2). In three cases something is known of their mode of formation. The present evidence strongly indicates that thyroxin (5), hydroxyproline (6), and phosphoserine (7) are formed from tyrosine, proline, and serine residues already incorporated into protein. This is in contrast with the 20 commonly occurring amino acids, which are incorporated as such. Thus the evidence to date does not make it necessary to assume that template configurations for selecting more than 20 amino acids exist.

[2] We have used data on the amino acid composition of 22 proteins: whale myoglobin, horse hemoglobin, aldolase, triosephosphate dehydrogenase, phosphorylase, zein, ovalbumin, β-lactoglobulin, α-casein, conalbumin, fibrinogen, bovine serum albumin (4); prothrombin (8); carboxypeptidase (9); papain (10); ribonuclease (11); ACTH (12); insulin (13); tropomyosin, actin (14); barley β-globulin (15); and lysozyme (16). Virus proteins: tobacco mosaic (17), turnip yellow (18), and tomato bushy stunt (19). Since analysis of a chemical hydrolyzate of a protein does not distinguish glutamine from glutamic acid, or asparagine from aspartic acid, we have estimated the abundance of these amino acids as follows: the published amide content of each protein was

assigned to glutamine and asparagine in the ratio in which glutamic and aspartic acid occur in the chemical hydrolyzate. This procedure was unnecessary for insulin, ACTH, lysozyme (20), and zein, where the actual content of the four acids is known. The data on the amide content of tobacco mosaic virus protein are from Schramm (21). No data are available on the amide content of turnip yellow virus protein, and we estimate the content to be the mean of that of tobacco mosaic and tomato bushy stunt viruses. This value is obviously very uncertain. The RNA composition data are taken from the compilation of Magasanic and refer to the composition of the RNA of calf liver, calf spleen, carp nucleotropomyosin, cat brain, sea urchin eggs, yeast, and *Escherichia coli* (22). Virus RNA data: tobacco mosaic (23), turnip yellow (24), and tomato bushy stunt (19). The nucleoide and amino acid compositions of RNA and protein are expressed in moles per cent.

REFERENCES

1. GAMOW, G. *Nature*, **173**, 318, 1954; *Kgl. Danske Videnskab. Selskab Biol. Medd.*, **22**, 3, 1954.
2. GAMOW, G., RICH, A., and YČAS, M. *Advances in Biol. and Med. Physics*, Vol. **4**, New York: Academic Press, 1955 (in press).
3. ELSON, D., and CHARGAFF, E. *Nature*, **173**, 1037, 1954; *Biochem. et Biophys. Acta*, **17**, 365, 1955.
4. TRISTRAM, G. R., in *The Proteins*, ed. H. NEURATH and K. BAILEY **1a**, 181. New York: Academic Press, 1953.
5. ROCHE, J., and MICHEL, R. *Ann. Rev. Biochem.*, **23**, 481, 1954.
6. STETTEN, M. R. *J. Biol. Chem.*, **181**, 31, 1949.
7. BURNETT, G., and KENNEDY, E. P. *J. Biol. Chem.*, **211**, 969, 1954.
8. LAKI, K., KOMINTZ, D. R., SYMONDS, P., LORAND, L., and SEEGERS, W. H. *Arch. Biochem. and Biophys.*, **49**, 276, 1954.
9. SMITH, E. L., and STOCKELL, A. *J. Biol. Chem.*, **207**, 501, 1954.
10. SMITH, E. L., STOCKELL, A., and KIMMELL, J. R. *J. Biol. Chem.*, **207**, 551, 1954.
11. HIRS, C. H. W., STEIN, W. H., and MOORE, S. *J. Biol. Chem.*, **211**, 941, 1954.
12. BELL, P. H. *J. Am. Chem. Soc.*, **76**, 5565, 1954.
13. HARFENIST, E. J. *J. Am. Chem. Soc.*, **75**, 5528, 1953.
14. KOMINTZ, D. R., HOUGH, A., SYMONDS, P., and LAKI, K. *Arch. Biochem. and Biophys.*, **50**, 148, 1954.
15. BROHULT, S., and SANDEGREN, E., in *The Proteins*, ed. H. NEURATH and K. BAILEY, **2a**, 487. New York: Academic Press, 1954.
16. FEVOLD, H. L. *Advances in Protein Chem.*, **6**, 187, 1951.
17. KNIGHT, C. A. *J. Biol. Chem.*, **171**, 297, 1947.
18. ROBERTS, E., and RAMASARMA, G. B. *Proc. Soc. Exptl. Biol. Med.*, **80**, 101, 1952.
19. DEFREMERY, D., and KNIGHT, C. A. *J. Biol. Chem.*, **214**, 559, 1955.
20. OHNO, K. *J. Biochem. (Japan)*, **41**, 345, 1954.
21. SCHRAMM, G. *Advances in Enzymol.*, **15**, 449, 1954.
22. MAGASANIC, B., in *The Nucleic Acids*, ed. E. CHARGAFF and J. N. DAVIDSON **1**, 373. New York: Academic Press, 1955.
23. KNIGHT, C. A. *J. Biol. Chem.*, **197**, 241, 1954.
24. MARKHAM, R., and SMITH, J. D. *Biochem. J.*, **49**, 401, 1951.