Eur. J. Biochem. 16 (1970) 1–11

The Diagram, a Method for Comparing Sequences

Its Use with Amino Acid and Nucleotide Sequences

Adrian J. GIBBS

Department of Microbiology, John Curtin School of Medical Research, Australian National University, Canberra

George A. MCINTYRE

Division of Mathematical Statistics, Commonwealth Scientific and Industrial Research Organisation, Canberra

(Received May 4, 1970)

A simple method for detecting similarities in sequences is described. It is used:

1. To provide similarity measures for classifying 25 cytochromes by their amino acid sequences,
2. to detect repetitions in the amino acid sequences of various proteins,
3. to detect regions of possible base-pairing in the nucleotide sequence of a nuclei acid.

Several methods have been used for comparing the sequences of amino acids in different proteins, for estimating the similarity of the sequences, and for using the estimates of similarity to classify the proteins. Some of these methods have been described and compared elsewhere by Gibbs, Kinns, Dale and McKenzie [1], who also describe a way for classifying amino acid sequences using doublet matrices. In this paper we describe another alternative, the "diagonal-match" or diagram method, which, we think has advantages over other methods in that it is basically simple (can be done by hand if necessary), and shows directly all the possible similarities between the sequences. The diagram method is similar in principle to a method for detecting parts missing from sequences suggested to Fitch [2] by Dr. Saul Needleman.

METHOD

The sequences are compared in pairs. For each comparison the two sequences are recorded along adjacent sides of a rectangular matrix, the diagram. With proteins we record the sequences with their N-terminal amino acids in the top left-hand corner of the diagram. Within the body of the diagram every match is recorded; a dot is put wherever a row and column with the same amino acid (recorded at the edge of the diagram) intersect. Similarities of the two sequences are then clearly indicated by diagonal rows of matches (Fig. 1). There are various ways of testing whether the diagonal runs of matches are more than would be expected by chance. We have used two simple methods:

a) The frequencies of lengths of all unbroken diagonal runs of matches are recorded, and are compared with the frequencies of runs expected by chance and calculated directly (Appendix 1).

b) The total number of matches in every diagonal of the diagram is recorded, and again is compared with the numbers expected in each diagonal if the sequences show no similarity.

All these manipulations can be done by hand, but are done more quickly by computer. A composite program, MATCH (written in CDC & ASA FORTRAN IV), for printing the diagram and computing various similarity measures, is available from us.

RESULTS AND DISCUSSION

Amino Acid Sequences—Cytochromes c

To gauge the usefulness of the method, it was used to compare all the cytochromes *c*, whose sequences have been published [3,4], together with cytochrome B5 of cow [3]. Fig. 1 shows three of the diagrams obtained, these are of human cytochrome *c* compared with the cytochromes *c* of rhesus monkey, tuna fish and the bacterium *Rhodospirillum rubrum*. These illustrate the type of diagram obtained with sequences of differing similarity. The monkey and fish cytochromes match with that of man by differing amounts, but do so over their entire length; the principal lines of matches between them lie along the principal diagonals of the diagrams. By contrast the line of matches between the cytochromes of man and *R. rubrum* is, at different parts of the sequences, on three different diagonals, indicating that parts of one sequence are not present in the other. Another feature of the man/monkey and man/fish diagrams is the repetition in the N-terminal part (residues 3–8

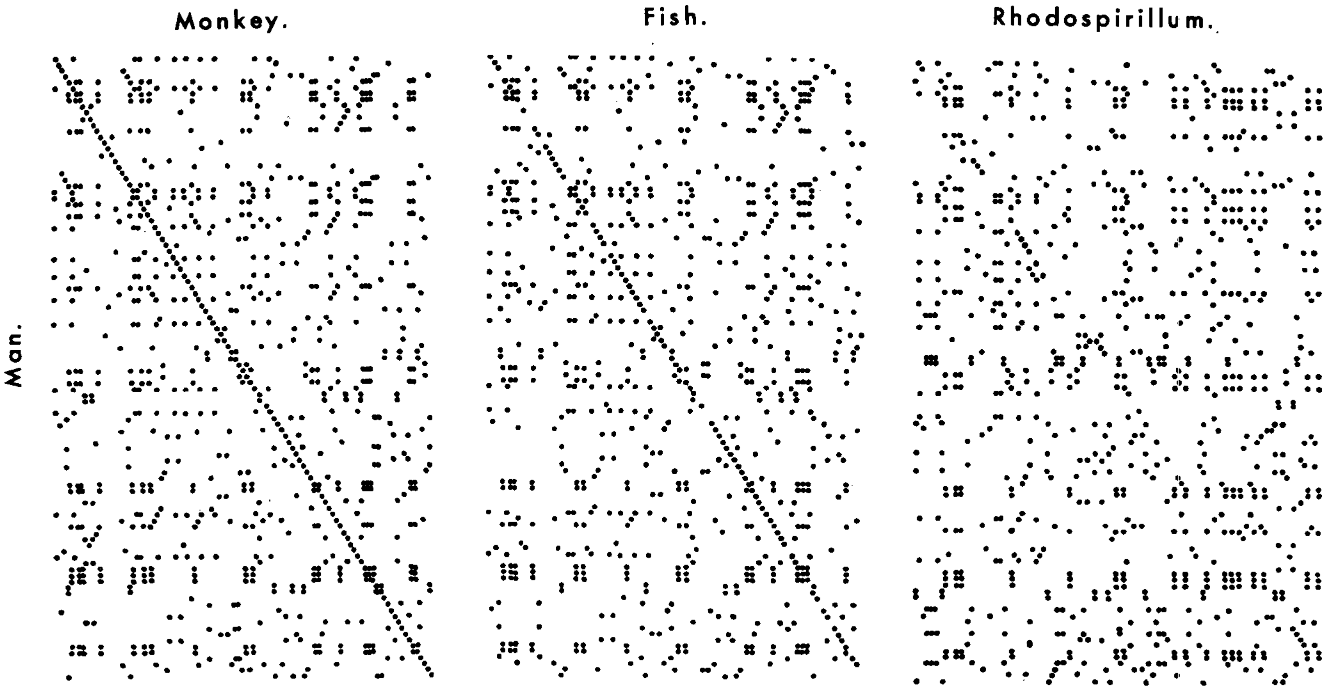

Fig. 1. *The diagrams obtained from comparisons of human cytochrome* c *(left margin of each diagram, N terminus at top) and the cytochromes* c *of monkey, fish and* Rhodospirillum *(upper margin, N termini at left end)*

Table 1. *Length of unbroken runs of matched amino-acids obtained when human cytochrome* c *is compared with other cytochromes* c

Other cytochromes	Run length																							
	0	1				5					10									20		30		50
Monkey	680	36	2	2																			1	1
Rattlesnake	673	43	3	3		1					1						1					1		
Dog	692	36	7	3			1			1			1		1			1			1			
Whale	684	41	5	2			1			1				1				1			1			
Kangaroo	684	45	3	2	1					2				1		1						1		
Chicken	687	33	5	1	1		3													1		1		
Bullfrog	683	41	4	3	1		4			1										1				
Fish	636	44	4	3	3	1				1									1					
Silkworm	641	35	4	1	2	2			1	1				1										
Wheat	651	40	2	1	1	1				1	1	1												
Saccharomyces	672	49	7	1	2	1		1			1													
Rhodospirillum	729	47	6	1	1																			
Pseudomonas	482	21	2	1																				
Random[a]	738	58	5	0.4																				

[a] The runs expected by chance when two random sequences with the composition of human cytochrome *c* are compared.

Table 2. *Sums of matches in the principal and adjacent diagonals of diagrams comparing human cytochrome c with other cytochromes*
The principal diagonal is marked *X*. In each diagram the human cytochrome *c* is along the left margin of the diagram, and the other along the upper margin

															Man											
9	12	4	6	4	3	5	7	7	4	9	7	6	7	9	103	8	7	6	7	8	4	7	7	6	3	Monkey
8	10	4	6	5	5	5	6	8	4	7	6	6	7	8	90	10	10	8	6	8	3	7	8	4	4	Rattlesnake
8	13	5	6	3	3	7	7	8	4	9	9	6	6	7	93	8	5	7	8	9	4	6	8	5	5	Dog
8	12	5	7	4	3	7	7	8	4	9	9	6	6	7	94	9	6	6	7	9	4	6	8	6	4	Whale
8	12	5	7	4	3	7	7	7	4	8	8	6	7	7	94	9	7	5	7	9	4	6	8	5	4	Kangaroo
8	12	5	7	4	3	6	8	7	5	9	7	6	6	7	91	7	6	6	8	9	4	7	8	7	4	Chicken
9	12	6	6	4	3	6	7	8	5	9	7	6	6	7	86	8	6	7	8	10	5	5	9	5	4	Bullfrog
8	12	5	6	4	5	8	6	7	4	6	7	5	6	7	83	7	5	6	6	9	5	7	10	4	2	Fish
8	10	11	3	4	6	5	5	7	4	6	9	4	4	10	39	6	5	5	45	5	6	9	5	4	3	Silkworm
5	3	7	5	7	4	3	6	7	9	4	5	5	1	8	7	11	6	6	6	6	8	7	69	8	8	Wheat
5	6	6	5	7	6	9	6	5	8	4	6	9	11	7	8	5	8	7	11	64	7	6	4	6	7	*Saccharomyces*
8	8	6	8	6	12	8	3	5	4	9	9	9	8	8	21	8	8	12	9	8	10	6	5	6	9	*Rhodospirillum*
4	2	5	5	5	5	5	4	1	8	12	5	7	7	3	4	8	6	6	3	4	5	3	2	6	11	*Pseudomonas*
7·2	7·2	7·3	7·4	7·5	7·6	7·6	7·7	7·8	7·9	8·0	8·0	8·1	8·2	8·3	8·4	8·3	8·2	8·1	8·0	8·0	7·9	7·8	7·7	7·6	7·6	Random[a]

[a] The number of matches in each diagonal expected when two different sequences with the composition of human cytochrome *c* are compared.

and 20—25) of the sequence Val-Glu-Lys(or Asn)-Gly-Gly(or Lys)-Lys.

Table 1 shows the number and length of unbroken runs of matches in certain of the comparisons between human cytochrome *c* and the other cytochromes. The length and number of unbroken runs decreases the more unrelated are the organisms from which the cytochromes were obtained, so that the runs of matches in the man/*Pseudomonas fluorescens* diagram do not differ from what would be expected by chance.

Table 2 shows the total number of matches in the principal and 35 adjacent diagonals in the diagrams of the comparisons listed in Table 1. These again show that the number of matches in the principal diagonal decreases the more unrelated are the organisms from which the cytochromes were obtained. With closely related organisms the greatest number of matches is found in the principal diagonal of each diagram, whereas with very distantly related organisms (*e. g.* man/silkworm, and man/wheat) the line of homology may be in more than one diagonal, indicating again that parts of one sequences are not present in the other. Table 2 also shows the increased number of matches in the diagonal 14th from the principal diagonal caused by the N-terminal repeated sequence mentioned before.

In the Methods section two simple tests were suggested for assessing whether the two sequences in a diagram show similarity. We have used these two tests to provide similarity coefficients, that can be used for classification:

a) A "runs index" of similarity was derived (see Appendix 1) for each diagram from the numbers and lengths of all unbroken diagonal runs of matches and from the runs expected if the sequences had had no similarity (*i. e.* were randomized). A "runs index" was calculated for all pairwise comparisons of the 25 cytochromes. The maximum "runs index" obtained was of 1.9997 between duck and chicken cytochromes *c*, and the minimum —0.0826 between monkey cytochrome *c* and cow cytochrome B5. The "distance" between each pair of cytochromes (*i. e.* the converse of similarity) was calculated as "2.200—runs index", and was used with the CLASS programme (flexible sorting strategy) of Lance and Williams [5, 6] to compute the classification illustrated by the dendrogram in Fig. 2.

b) A "diagonals index" of similarity was derived (see Appendix 2) for each diagram by calculating from the total number of matches in each of the diagonals an observed χ^2 and a maximum χ^2. The latter is defined as the χ^2 obtained when the principal diagonal is filled with matches (*i. e.* complete sequence homology) and all the other matches in the diagram are at random (*i. e.* no repetition within the sequences); it would be obtained by making a diagram of a completely random sequence with itself. The "diagonals index" of similarity was calculated for

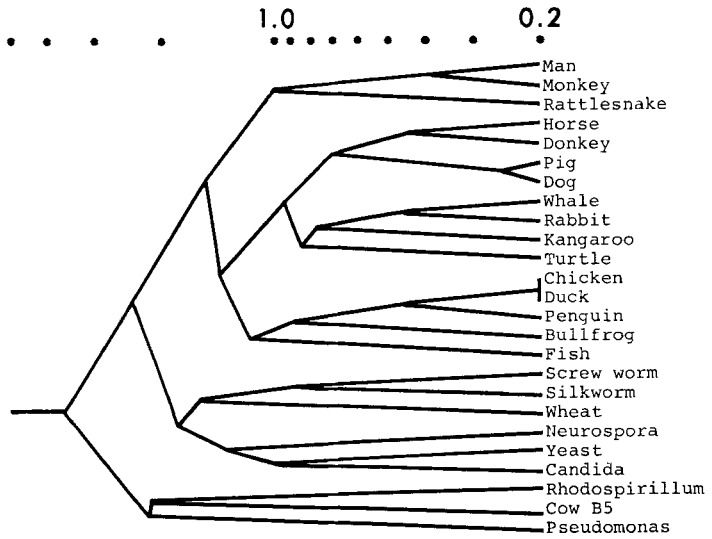

Fig. 2. *A dendrogram illustrating a "runs index" classification of 24 cytochromes* c *and cow cytochrome B5.* The similarity between the sequences of every pair of cytochromes was estimated by the "runs index" (see text), and the dendrogram was computed from these similarities [5,6]

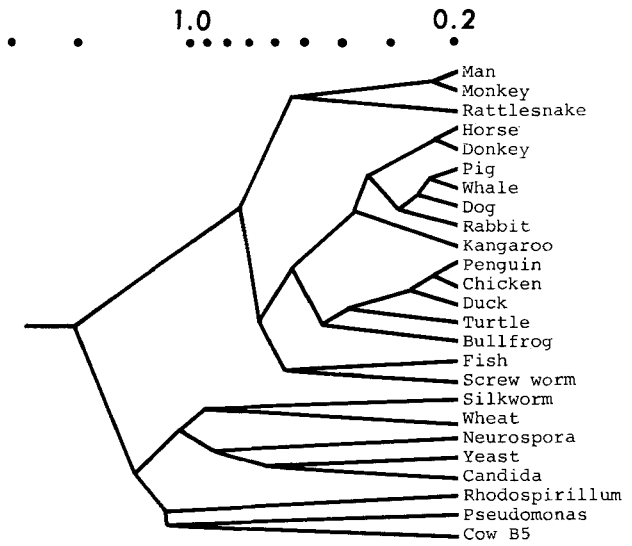

Fig. 3. *A dendrogram illustrating a "diagonal index" classification of 24 cytochromes* c *and cow cytochrome B5.* The similarity between the sequences was estimated by the "diagonals index" (see text), and the dendrogram was computed from these similarities [5,6]

all pair-wise comparisons of the 25 cytochromes. The maximum "diagonals index" obtained was 0.9692 between chicken and penguin cytochromes *c*, and the minimum — 0.0254 between chicken and *P. fluorescens* cytochromes *c*. The "distance" between each pair of cytochromes was calculated as "1.2000-diagonals index" and was also used with the CLASS programme to compute a classification (Fig. 3).

The two classifications are similar, resemble closely classifications of the cytochromes derived in other ways [1], and approximate to current ideas on the relationships of the organisms from which the cytochromes were obtained. However, there are some obvious differences:

a) Both classifications show the close similarity between rattlesnake and primate cytochromes *c*. This similarity was noted by Smith [7], but not found by Fitch and Margoliash [8], and was found, but discounted, by Dayhoff and Eck [3]. A similarity like this must make one seriously question the value of interpreting classifications of this sort strictly in terms of phylogeny [9,10].

b) The relationships between tuna fish, silkworm and screw-worm cytochromes *c* are different in the two classifications. In the "diagonals index" classification screw-worm classifies with fish and the vertebrates, while silkworm classifies with the other non-vertebrates, whereas in the "runs index" classification the two insect cytochromes classify together and group with the non-vertebrates.

These differences reflect the difference between the two similarity indices. A greater "diagonals index" is obtained when the matches occur in one diagonal, as in the fish/screw-worm diagram, than when they are divided between two diagonals as in the fish/silkworm and silkworm/screw-worm diagrams. By comparison the "runs index" is dependent on the length of unbroken runs of matches, so that the insect cytochromes, which have longer runs of matching amino acids [runs of 29, 20, 7($\times$3), 6 and 5], give a greater "runs index" than the fish/silkworm (runs of 15, 10, 8 and 7) and fish/screw-worm [runs of 19, 9($\times$2), 8 and 7] comparisons.

The relationships between chicken, duck and penguin cytochromes are also different in the two classifications for the same reasons.

The two similarity indices we have used could perhaps be combined to give a single similarity index for use in classification; however using the indices separately has some advantages for, as shown above, the differences between the classifications give added information. Obviously more subtle measures of similarity could be used with the diagram method to give improved classifications, however, we doubt the value of using the "minimum mutational distance" measure used by Fitch [2,10] because of the great number of untested and untestable assumptions in this particular measure of similarity.

The probability with which the observed χ^2 (calculated for the diagonals index) could be obtained by chance can also be estimated. For example the diagram (Fig. 4), obtained with bovine α-lactalbumen and chicken lysozyme, whose similarity has been reported

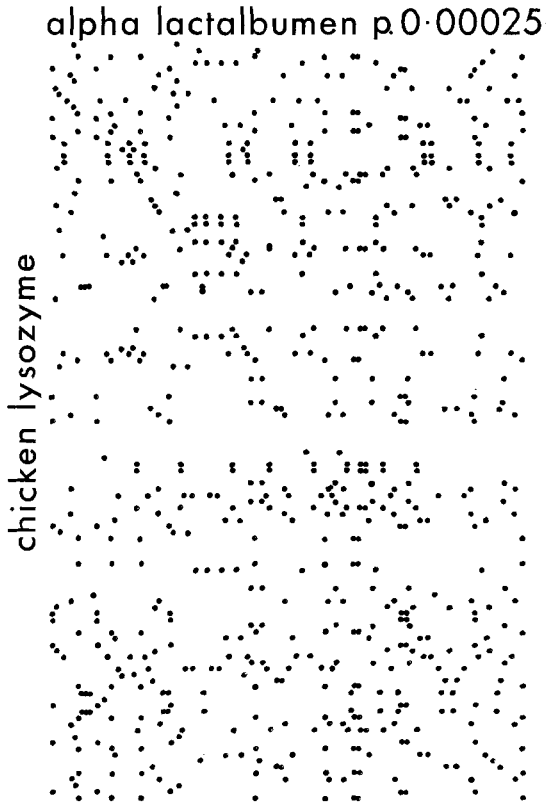

Fig.4. *The diagram obtained from a comparison of chicken lysozyme (left margin, N terminus at top) and bovine alpha lactalbumen (upper margin, N terminus at left end)*

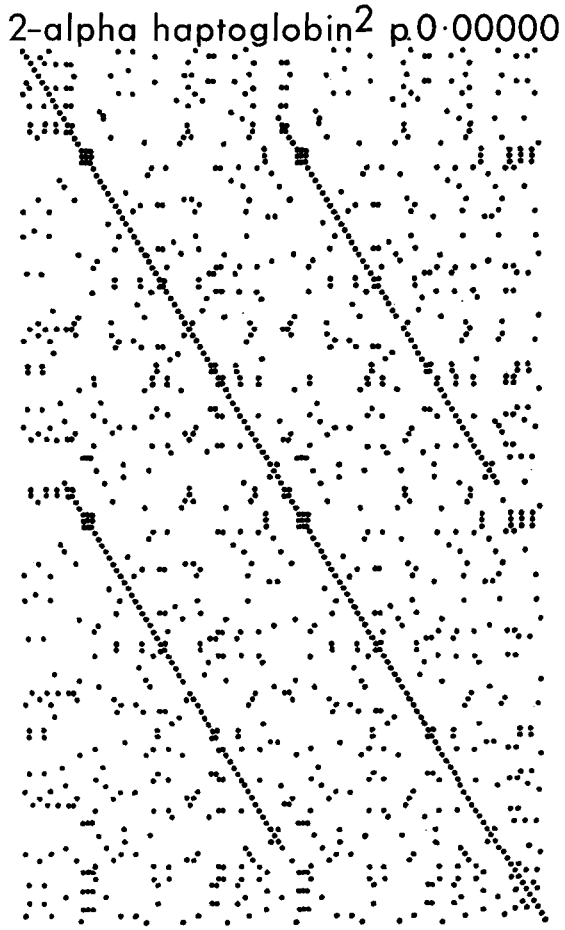

Fig.5. *The diagram obtained by comparing 2-alpha haptoglobin with itself (same sequence along left and upper margin, N termini of both in upper left corner)*

by Brew, Vanaman and Hill [11], gave a runs index of 0.139, and a diagonals index of 0.032 with a probability of 0.00025 of occurring by chance.

Repetitions within a Sequence: Internal Homology

Since Smithies, Connell and Dixon [12] showed that human 2α-haptoglobin has a sequence, the two halves of which are almost identical to one another (Fig.5) and also to the sequence of 1α-haptoglobin, it has become fashionable to examine the amino acid sequences of proteins for repeated sequences, and there are now many reports of such repetitions [13]. However, few of the claimed repetitions have been tested statistitically, and many rely on the addition of gaps to parts of the sequence, and this has inherent dangers [14].

The diagram method may be used to compare a sequence with itself, and will thereby detect repetitions for these will show as runs of matches on diagonals other than the principal diagonal. Further more repetitions will be indicated by the diagonals index being greater than 1, as the observed χ^2 will be greater than the maximum χ^2. For example 2α-haptoglobin compared with itself (Fig.5) gives a diagonals-index of 1.6095, which has a probability of occurring by chance of less then 0.000001. The programme MATCH includes an option for use when comparing sequences with themselves; as before the programme calculates the diagonals and runs indices and in addition calculates χ^2 and its associated probability for one half of the diagram excluding the principal diagonal. Table 3 shows the results obtained using this option with various proteins reported to show internal repetitions, and also several for which no such claims have been made.

The diagram method confirms the reported repetitions in the bacterial ferredoxins [15,16] and one of the bacterial subtilisins, and, in addition, suggests that there may be repetitions in lamprey globin. The diagram method did not detect the internal repetitions reported in the heavy and light chains of immunoglobulin [18,19], clupeine z [20], human α-haemoglobin [17], the cytochromes *c* of *Neurospora crassa* [21] and *Rhodospirillum rubrum* [4], nor did it

Table 3. *Results of comparing different proteins with themselves by the diagram method to detect sequence repetitions*

Protein	Diagonals index	Probability of occurring by change
Clupeine Z	0.2532	0.958
Cytochrome *c* — *Neurospora*	0.991	0.644
Cytochrome *c* — *Rhodospirillum*	0.992	0.620
Ferredoxin — *Closteridium pasteurianum*	1.186	0.000005
Ferredoxin — *Closteridium butyricum*	1.136	0.00019
Ferredoxin — *Micrococcus aerogenes*	1.148	0.000094
Haemoglobin — human α	0.996	0.547
Haemoglobin — human β	1.021	0.110
Haptoglobin — human 2α	1.609	0.000000
Immunoglobulin — heavy chain	0.998	0.558
Immunoglobulin — light chain	0,999	0.515
Substilisin — BPN	1.016	0.153
Subtilisin — Carlsberg	1.041	0.005
Other proteins		
Azurin — *Bordetella*	1.004	0.375
Fd-phage	0.989	0.591
Ferredoxin — *Leucana glauca*	0.991	0.653
Globin — carp α	0.992	0.648
Globin — lamprey	1.033	0.017
Haptoglobin — human 1α	1.022	0.135
Lipotropin	0.986	0.699
Lysozyme — T4 phage	0.982	0.898
Nuclease — *Staphylococcus aureus*	0.987	0.773
Penicillinase	1.012	0.195
Ribonuclease — T1	0.998	0.497
Rubredoxin — *Micrococcus aerogenes*	1.029	0.115
Tobacco mosaic virus	1.004	0.369
Trypsin inhibitor — cow	0.979	0.793
Tryptophan synthetase — *Escherichia coli*	0.986	0.864

detect repetitions in the sequences of 16 other different proteins.

The diagrams obtained by comparing with themselves the *Neurospora* and *Rhodospirillum* cytochromes, and human α-haemoglobin show that the parts of the sequences claimed to show homology give no more obvious diagonal runs of matches than other parts of the sequences. The repetitions in clupeine z are claimed [20] to show that this protein of 31 amino acid residues (20 of them arginine) is a repetition of a 5 residue sequence, however the diagram method shows that the observed repetitions could be derived by chance from a protein of the same unusual composition.

The inability of the diagram method to detect repetitions within the light and heavy chain of immunoglobulin is more interesting. The homologies within these sequences were found by Hill *et al.* [19] and by Edelman *et al.* [8] by comparing the light chain with the heavy, and when this is done by the diagram method there is clear evidence of sequence similarities (runs index 0.081, diagonals index 0.033 with a probability of 0.0036 of occurring by chance), yet when the light chain is compared with itself no sequence repetitions are indicated (runs index -0.069, diagonals index 0.989, half diagram diagonals probability 0.702), and similarly there is no evidence of repetitions when the heavy chain is compared with itself (runs index 0.032, diagonals index 0.998, half diagram diagonals probability 0.558). The diagram made with the light and heavy chain (Fig. 6A and B) clearly shows the sequence similarities reported by Edelman *et al.* [18] between the C-terminal half of the light chain and three regions of the C-terminal part of the heavy chain. When these and the other regions of the two chains are compared individually by the diagram method, it can be seen (Table 4) that: a) the C-terminal half of the light chain shows sequence similarities with the three C-terminal quarters of the heavy chain, b) the three C-terminal quarters of the heavy chain show sequence similarities with one another, c) the N-terminal half of the light chain has some similarity with the C-terminal half of the light chain, and, surprisingly, a greater similarity with the C-terminal quarter of the heavy chain, and d) the N-terminal quarter of the heavy chain shows no similarity with other parts of the sequences.

Thus the heavy chain has clearly resulted from a replication of the light chain, but it is impossible

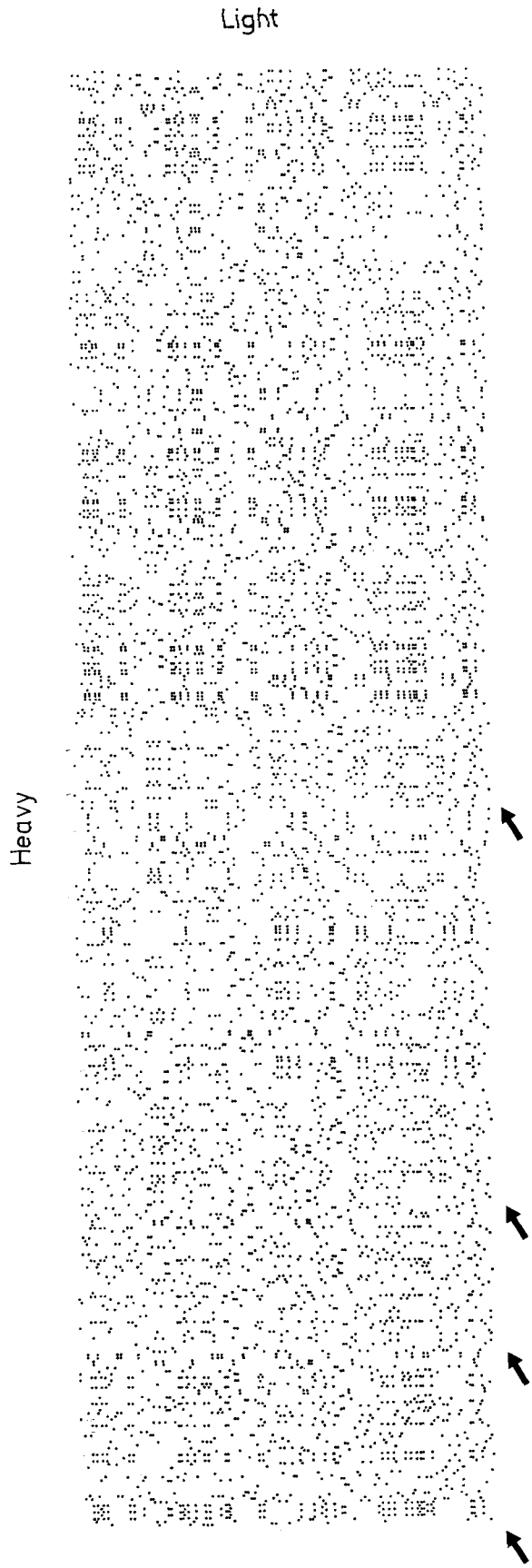

Fig. 6A. *The diagram obtained from a comparison of the heavy chain of immunoglobulin (left margin, N terminus at top) and the light chain of immunoglobulin (upper margin, N-terminus at left).* Arrows indicate the most noticeable diagonal runs of matches between the C-terminal half of the light chain and the three C-terminal quarters of the heavy chain. These runs can be seen most clearly by looking along the line of each arrow at a shallow angle to the page

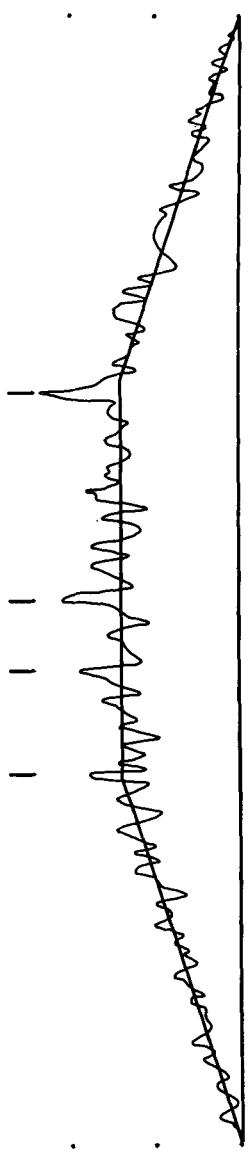

Fig. 6B. *A graph showing the sums of matches on each diagonal of Fig. 6A; running sums of five adjacent diagonals are plotted at the position of the central diagonal of the five.* The left end of the graph corresponds to the diagonal at the C-terminal end of the heavy chain, and the right end to the C-terminal end of the light chain. The straight line indicates the number of matches expected if the matches were uniformly distributed over the diagram, and the arrows show the diagonals arrowed in Fig. 6A

Table 4. *Comparisons by the diagram method of parts of the light and heavy chains of immunoglobulin*
Residues are numbered from the N-terminus of the proteins. The numbers represent the probability that the χ^2 calculated from the sums of matches in the diagonals of the diagram could occur by chance

						Selfed
Light 1—110	.198	.550	.664	.727	.039	.702
	light 111—216	.425	.203	.159	.221	.402
		heavy 1—110	.783	.386	.739	.535
			heavy 111—220	.208	.237	.962
				heavy 221—340	.131	.433
					heavy 341—446	.298

to be sure whether this is the result of a doubling of the entire light chain [19] or a triplication of just the C-terminal part of the light chain.

The diagram also suggests that there is some similarity of sequence between the C-terminal half of the light chain and the part of the heavy chain between the two C-terminal quarters.

The discrepancy between the results obtained by the diagram method when the light and heavy chains are compared with each other rather than with themselves may have two causes, either the diagonals index is not sufficiently sensitive, or somewhat different parts of the C-terminal half of the light chain are represented in the three homologous regions of the heavy chain.

The latter possibility is interesting for it suggests that after the light chain gene replicated to give the heavy chain, there has been selection against having the same amino acid sequence in different parts of the heavy chain. This selection may be common in proteins and have some definite evolutionary advantage, for several proteins, when compared with themselves by the diagram method, gave no evidence of internal repetitions and have a diagonals index of less than 1 (Table 3). Thus the χ^2 obtained by the diagram method, when a sequence is compared with itself, may be the product of two opposing features, internal repetitions, which may be the residue of gene replication, and which will increase the χ^2, and selection against repetition within the sequence which will decrease the χ^2.

The reported sequence repetitions within proteins have almost invariably been interpreted as showing that the gene responsible for the particular protein has arisen by the duplication or replication of part or all of a smaller gene, and although other interpretations are possible for many of these proteins, they seem not to have been considered. For example, it is likely that the iron-containing moiety of bacterial ferredoxin is symmetrical, and hence the protein moiety must have specific, repeated and perhaps symmetrical amino acid sequences to bind to the iron-containing moiety. Similarly, Fitch [17] reported that a repetition, separated by 66 amino acid residues, in human haemoglobin molecules was evidence of gene duplication in the origin of this molecule. Fitch detected this similarity by calculating the minimum number of single nucleotide changes theoretically needed to interconvert parts of the sequences; however this measure will treat as similar not only sequences with the same amino acids but also sequences with amino acids with similar properties [22]. Thus Fitch's method may have detected regions in the haemoglobins with groups of polar or non-polar amino acids, and it is noteworthy that those amino acids in the haemoglobin protein subunits that are involved in the intersubunit binding sites (but not in the haem binding sites) occur mostly in two parts of the sequence about 60 residues apart [23].

Nucleotide Sequences—R17 Bacteriophage Fragment

There is an increasing number of reports of the sequence of nucleotides in small nucleic acid molecules, and there has been much speculation on the possible secondary structure of these molecules resulting from G-C and A-U(T) base pairing between different parts of the molecule. The diagram method is a convenient way for showing all the possible base pairings between different parts of a nucleotide sequence. Fig. 7 shows the diagram obtained with the sequence recently reported by Adams, Jeppesen, Sanger and Barrell [24], of a large fragment of the nucleic acid of the R17 bacteriophage. The sequence is recorded along the left hand edge of the diagram, and along the top of the diagram is recorded the complementary sequence, with the 5′ terminus of each sequence next to the top left hand corner of the matrix. Diagonal rows of matches indicate sequences in common, and this complementary diagram is symmetrical about the bottom left to top right principal diagonal. Marked on the diagram in Fig. 7 is the well defined line of possible base-pairings noted by Adams *et al.* [24] and this indicates how the diagram may be interpreted. For example, the run

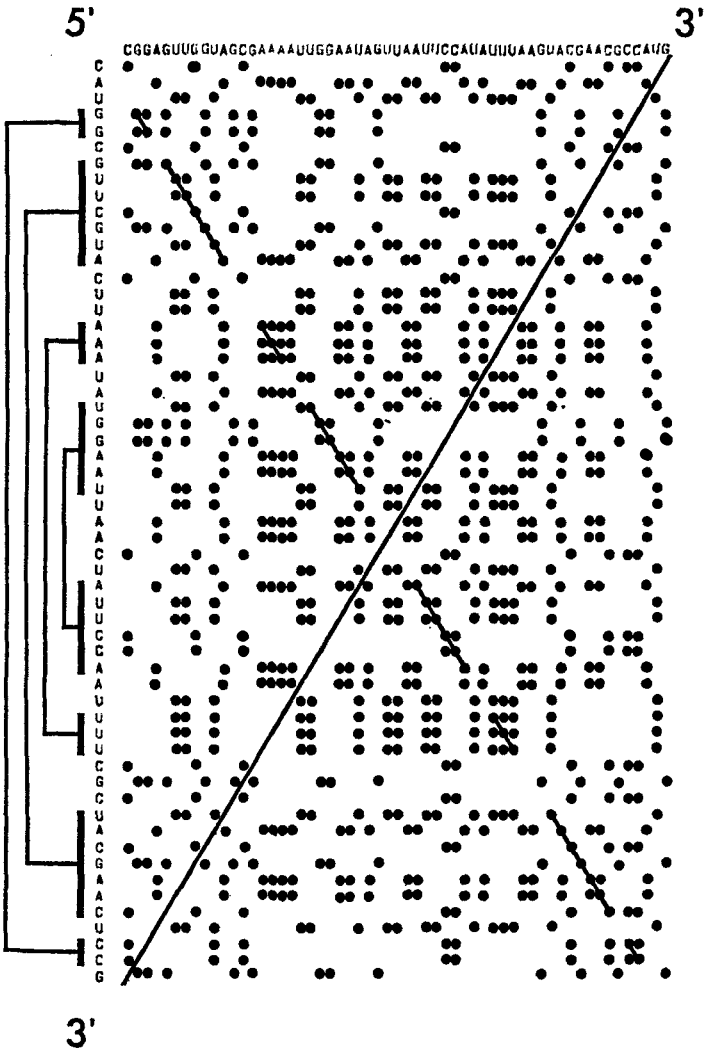

Fig. 7. *The diagram obtained by comparing the nucleotide sequence of a fragment of the nucleic acid of the R17 bacteriophage (left margin, 5′ terminus at top) with its complementary sequence (G = C, A = U) (upper margin, 5′ terminus at top).* For explanation see text

of nucleotides on the marked line and nearest to the 5′ terminus is a pair of nucleotides, -G-G-. In the mirror position in the other half of the diagram (at the 3′ terminal end of the sequences) is a pair of matches opposite the complementary sequence -C-C-. The next run of nucleotides marked is opposite the sequence -^{5}G-U-U-C-G-U-A^{3}, and in the mirror position is the run of matches opposite the complementary sequence -^{3}C-A-A-G-C-A-U^{5}. These and the other complementary sequences noted by Adams *et al.* [24] are indicated by bracketing on the left hand side of the diagram and it is clear that although base-pairings are possible between other parts of sequence, the marked line is, by far, the most complete and is also symmetrical, thus allowing a helical structure.

Similar diagrams made from the published sequences of transfer and 5S-ribosomal ribonucleic acids are more complex, and although those of the transfer ribonucleic acids clearly confirm the possibility of the clover-leaf structure proposed for these molecules, they show that other more complex base-pairing arrangements are also possible.

This use of the diagram method is, we believe, a useful complement to the rather more complex methods for determining the possible position of base-pairing in nucleic acids suggested by Tumanyan *et al.* [25] and Richards [26].

APPENDIX 1. RUNS INDEX

Matching Sequences of Two Different Proteins

If there are k different amino acids present in one or both of the two proteins and f_{1i}, f_{2i} ($i = 1, k$) are the respective frequencies of the amino acids present

in the first and second protein then the total numbers of amino acids present are $\sum_{i=1}^{k} f_{1i} = N_1$ and $\sum_{i=1}^{k} f_{2i} = N_2$, the number of matches NMATCH $= \sum_{i=1}^{k} f_{1i}, f_{2i}$, and the total number of cells is N_1N_2. The prior probability of a match occurring in a particular cell is $p = N\text{MATCH}/N_1N_2$.

If the sequences in the two proteins are not related, *i. e.* in random order relative to one another, the probability that a match will not have another match contiguous to it on the diagonal is $p\,q^2$ for those matches not on the boundary, $p\,q$ for those on the boundary and p for the two cells on the corners of the opposing diagonal direction where $q = 1 - p$. The expected number of such single matches is

$$p\,[q^2\,(N_1 - 2)\,(N_2 - 2) + 2q\,(N_1 + N_2 - 3) + 2].$$

Similarly the expected number for S contiguous matches on a diagonal is

$$p^s\,[q^2\,(N_1 - S - 1)\,(N_2 - S - 1) + 2q\,(N_1 + N_2 - 2S - 1) + 2].$$

A run of $S - 1$ is defined as S contiguous matches on a diagonal. We have used as a comparitive index the log ratio of the sum of observed lengths of run squared to the sum of expected lengths of run squared, *i. e.* $\log_{10}\left(\sum_{i=1}^{N} X_i^2 f_i\ (\text{obs}) \Big/ \sum_{i=1}^{N} X_i^2 f_i^1\ (\text{exp})\right)$ where x_i is the length of run, f_i the frequency of this length of run, f_i^1 the corresponding expected frequency and N the smaller of N_1, N_2. The index will be sensitive to the number of breaks in the sequence of matches on the main diagonal but with few breaks it will be perhaps unduly sensitive to the position of these breaks. With a single break the index will be least when the break occurs at or near the middle of the diagonal and greatest when it occurs near either end.

Matches in Diagonals of Diagrams Comparing a Protein with Itself

In this instance there is perfect matching on the principal diagonal. We can examine the distribution of length of runs in the diagonals to one side of the main diagonal. The prior probability of a match occurring in a particular cell in this region is $(N\text{MATCH-}N)/N^2$.

If there is random order in the amino acid sequence in a protein the probability of a match not having another match contiguous to it on the same diagonal is pq^2 for those matches not on the boundary of the triangular domain, pq for those on the boundary and p for the cell in the corner, where $q = 1 - p$. The expected number of such single matches is

$$p\,[q^2\,(N - 2)\,(N - 3)/2 + 2q\,(N - 2) + 1].$$

Similarly the expected number of S contiguous matches is

$$p^s[q^2\,(N - S - 1)\,(N - S - 2)/2 + 2q\,(N - S - 1) + 1].$$

APPENDIX 2. DIAGONALS INDEX

Matches of Diagonals of Diagrams Comparing Two Different Proteins

The criterion adopted here is to calculate χ^2 for the total matches on each of the $N_1 + N_2 - 1$ diagonals. This is related on the one hand to the expected value of χ^2 with no association in the ordering of amino acids in the two proteins and to its value when there is matching in all cells on a main diagonal and random matching elsewhere. The choice of this last reference is very arbitrary but is realistic as in the very great majority of pairings the departure from randomness comes almost wholly from the sequence of matches on the main diagonal or a diagonal close by. It will be assumed for the purpose of presentation that if N_1 and N_2 are not equal, N_1 is the smaller.

With all N_1 cells on a main diagonal matched there are (NMATCH-N_1) other matches in $N_1\,(N_2 - 1)$ other cells lying on $N_1 + N_2 - 2$ diagonals. Define $\alpha = N\text{MATCH}/N_1\,N_2$ and $\beta = (N\text{MATCH-}N_1)/N_1\,(N_2 - 1)$. If X_j is the number of cells on any diagonal j the expected number of matches is βX_j and the contribution to χ^2 for the table excluding the main diagonal is

$$(\text{obs} - \beta X_j)^2 \left\{ \frac{1}{\beta\, x_j} + \frac{1}{(1-\beta)\, x_j} \right\} = (\text{obs} - \beta X_j)^2 / X_j \beta\,(1 - \beta).$$

If random this has an expectation of $(N_1 + N_2 - 3)/(N_1 + N_2 - 2) = K$ *i. e.* $(\text{obs} - \beta X_j)^2$ has an expectation of $K\,\beta\,(1 - \beta)\,X_j$ or $\text{obs} = \beta\,X_j + \varepsilon$ where ε has an expectation of 0 and variance of

$$K\,\beta\,(1 - \beta)\,X_j.$$

If we now consider the full table and with N_1 matches on the main diagonal the contribution to χ^2 from this main diagonal is $[N_1\,(1 - \alpha)]^2/N_1\,\alpha\,(1 - \alpha) = N_1\,(1 - \alpha)/\alpha$. For a diagonal with X_j cells the calculated expected number of matches is $X_j\alpha$ so that $\text{Obs} - X_j\alpha$ is $X_j\,(\beta - \alpha) + \varepsilon$ and the expected value of $(\text{Obs} - X_j\alpha)$ is $X_j^2\,(\beta - \alpha)^2 + K\beta\,(1 - \beta)\,X_j$.

The expected contribution of this diagonal to χ^2 is

$$(\chi_j^2\,(\beta - \alpha)^2 + K\beta\,(1 - \beta)\,X_j)/X_j\alpha\,(1 - \alpha).$$

Pooled over the diagonals other than the main diagonal the sum is

$$[(\beta - \alpha)^2 N_1 (N_2 - 1) + (N_1 + N_2 - 3)\beta (1 - \beta)]/\alpha (1 - \alpha)$$

so that the expected χ^2 under these special circumstances is

$$\text{CHIMAX} = N_1 (1 - \alpha)/\alpha + [(\beta - \alpha)^2 N_1 (N_2 - 1) + (N_1 + N_2 - 3)\beta (1 - \beta)]/\alpha (1 - \alpha).$$

It is desirable that the range of the index should be approximately 1 to 0. We have therefore used as the Diagonals Index

(observed χ^2-degrees of freedom)/(CHIMAX-degree of freedom) where the degrees of freedom $= N_1 + N_2 - 2$.

The numerator can be negative but the ratio would depart little from zero. Also the observed χ^2 can potentially be greater than CHIMAX if long sequences of amino acids are repeated in both sequences so as to give long runs on diagonals other than the main diagonal. With most amino acid sequences the Diagonals Index lies between .99 and −.01.

Matches on Diagonals of a Diagram Comparing a Protein with Itself

In this situation the expected value of CHIMAX changes slightly because of the symmetry on the two sides of the main diagonal. With $N = N_1 = N_2$ and $K = 2(N - 2)$ the estimate for CHIMAX is

$$N (1 - \alpha)/\alpha + [(\beta - \alpha)^2 N (N - 1) + (2N - 4)\beta (1 - \beta)]/[\alpha (1 - \alpha)].$$

A more satisfactory measure of internal homologies (repetitions in the sequence) than the departure of χ^2 from CHIMAX is given by examining departures of observed from expected frequencies for the $N - 1$ diagonals above the main diagonal. The expected value [27] of $\chi^2 = \Sigma (\text{obs-exp})^2/\text{exp}$ is $N (N - 1)/(N + 1)$ or to a close approximation, $N - 2$. For a set of 500 random permutations of 80 residues of an artificial protein with 5 amino acids equally represented, the sample estimates of the population mean and variance were 78.54 and 168.13. These compare with exact values of 78.03 and 153.45 respectively. The program gives probabilities based on the χ^2 distribution with $N - 2$ degrees of freedom, through normalisation using the Wilson-Hilferty approximation.

Note added in proof (29th July 1970). Haber and Koshland [*J. Mol. Biol.* 50 (1970) 617], using a different method, have also been unable to detect the internal repetition found in haemoglobin by Fitch [17], and give alternative reasons for doubting the value of using Fitch's method to assess the relatedness of dissimilar sequences.

REFERENCES

1. Gibbs, A. J., Kinns, H. R., Dale, M. B., and MacKenzie, H. G., *Syst. Zool.* (1970) submitted for publication.
2. Fitch, W. M., *Biochem. Genet.* 3 (1969) 99.
3. Dayhoff, M., and Eck, R. V., *Atlas of Protein Sequence and Structure 1967—68* Nat. Biomed. Res. Found, Maryland 1968.
4. Dus, K., Stetten, K., and Kamen, M. D., *J. Biol. Chem.* 234 (1968) 5507.
5. Lance, G. N., and Williams, W. T., *Comput. J.* 9 (1967) 373.
6. Lance, G. N., and Williams, W. T., *Nature (London)*, 212 (1966) 218.
7. Smith, E. L., *Harvey Lect.* 62 (1966—67) 231.
8. Fitch, W. M., and Margoliash, E., *Science*, 155 (1967) 279.
9. Fitch, W. M., and Margoliash, E., *Brookhaven Symp. Biol.* 21 (1969) 217.
10. Fitch, W. M., *J. Mol. Biol.* 16 (1966) 9.
11. Brew, K., Vanaman, T. C., and Hill, R. L., *J. Biol. Chem.* 242 (1967) 3747.
12. Smithies, O., Connell, G. E., and Dixon, G. H., *Nature (London)*, 196 (1962) 232.
13. Nolan, C., and Margoliash, E., *Ann. Rev. Biochem.* 37 (1968) 727.
14. Cantor, C. R., *Biochem. Biophys. Res. Commun.* 31 (1968) 410.
15. Tanaka, M., Nakashima, T., Benson, A., Mower, H., and Yasunobu, K. T., *Biochemistry*, 5 (1966) 1666.
16. Benson, A. M., Mower, H. F., and Yasunobu, K. T., *Arch. Biochem. Biophys.* 121 (1967) 563.
17. Fitch, W. M., *J. Mol. Biol.* 16 (1966) 17.
18. Edelman, G. M., Cunningham, B. A., Einar Gall, W., Gottlieb, P. D., Rutishauser, V., and Waxdal, M. J., *Proc. Nat. Acad. Sci. U. S. A.* 63 (1969) 78.
19. Hill, R. L., Delaney, R., Fellows, R. E., and Lebovitz, H. E., *Proc. Nat. Acad. Sci. U. S. A.* 56 (1966) 1762.
20. Black, J. A., and Dixon, G. H., *Nature (London)*, 216 (1967) 152.
21. Cantor, C. R., and Jukes, T. H., *Proc. Nat. Acad. Sci. U. S. A.* 56 (1966) 177.
22. Volkenstein, M. V., *Nature (London)*, 207 (1965) 294.
23. Perutz, M. F., *J. Mol. Biol.* 13 (1965) 646.
24. Adams, J. M., Jeppesen, P. G. N., Sanger, F., and Barrell, B. G., *Nature (London)*, 223 (1969) 1009.
25. Tumanyan, V. G., Sotnikova, L. Y., and Kholopov, A. V., *Dokl. Akad. Nauk. SSSR*, 166 (1966) 1465.
26. Richards, E. G., *Eur. J. Biochem.* 10 (1969) 36.
27. Haldane, J. B. S., *Biometrika*, 31 (1939) 346.

A. J. Gibbs' present address:
Rothamsted Experimental Station
Harpenden (Herts.), Great Britain

G. A. McIntyre
Division of Mathematical Statistics, C. S. I. R. O.
Canberra, Australia