THE DETECTION OF COMPLEMENTATION MAP CLUSTERS BY COMPUTER ANALYSIS

O. J. GILLIE AND R. PETO

National Institute for Medical Research, The Ridgeway, Mill Hill, London, N.W. 7
Medical Research Council Statistical Research Unit, 115 Gower Street, London, W.C. 1

Received October 29, 1968

ALLELIC complementation is now a well known phenomenon which has been found to occur in a wide variety of micro-organisms and in some higher organisms (review, FINCHAM 1966). Allelic complementation may occur when two separate mutations at the same gene locus, but on different homologous chromosomes, are introduced into the same cell. When under these conditions the phenotype is non-mutant, complementation is said to occur. Experimental evidence has shown that this phenomenon can be accounted for in terms of the subunit structure of multimeric protein molecules. In particular it has been shown in alkaline phosphatase of *E. coli* (SCHLESINGER and LEVINTHAL 1963) and in glutamic dehydrogenase of *Neurospora crassa* (CODDINGTON and FINCHAM 1965; CODDINGTON, FINCHAM and SUNDARAM 1966) that inactive mutant enzyme molecules may reaggregate, by virtue of their subunit structure, to form enzymatically active, hybrid protein.

CRICK and ORGEL (1964) proposed a theory of allelic complementation based on the symmetry properties of protein multimers. They concluded that complementation involved the correction of a misfolded mutant polypeptide chain by a wild-type polypeptide chain in regions adjacent to the axes of symmetry of a multimer.

Studies of several complementation maps (GILLIE 1966, 1968) have shown that clustering of mutants occurs on these maps in a way which is consistent with the CRICK-ORGEL theory of complementation. Since no other suggestion has been made to account for this clustering, its existence lends some support to the theory.

Clustering of mutants was originally identified by inspection of the complementation map and matrix. This procedure is not satisfactory because allowance has to be made for any differences between the map and the matrix due to 'exceptional mutants'. For this reason it seemed desirable to look for a less subjective method of analysing complementation data which would identify the clusters by analysing the matrix itself rather than the map. Furthermore, it was considered that the availability of an easy method of analysing complementation data would enable complementation studies to be done on a larger and more routine basis, allowing more detailed analysis and comparison of mutant defects at the functional level. With these aims in mind we have designed a computer program to analyse complementation data.

AIMS AND METHODS OF ANALYSIS

General considerations: Mutants with identical complementation reactions with all other mutants are placed in the same complementation group. These groups are then arranged to form a matrix showing the complementation reactions of the groups with each other. This matrix is then used as the basis for further analysis by our computer method and by map-making.

The computer program works from the matrix to identify a small number of clusters of moderately similar, mutually noncomplementing groups. The clusters of mutants which we identify in this way may be interpreted, according to the CRICK-ORGEL theory, as representing mutations clustered at the particular parts of the polypeptide chain where it intersects the axes of symmetry of the protein multimer. We can only identify a minimum number of these clusters and so infer a minimum number of intersections of the polypeptide chain with the axis of symmetry of the molecule (GILLIE 1968).

In the computer analysis we have imposed the strict requirement that no complementation between members of one cluster should occur. This requirement was sometimes relaxed in the manual analysis (GILLIE 1968) when it was felt that a particular mutant 'naturally' lay in a certain cluster (by virtue of its position in the map), despite having one or two positive reactions with members of that cluster. The computer classifies such mutants either as members of another cluster or on their own as exceptions.

In general, it is our aim to divide the groups up into the minimum number of clusters of mutually noncomplementing complementation groups and a minimum number of exceptions. Sometimes the computer may classify the groups into a larger number of clusters than the minimum possible number. This will be because the requirement that members of one cluster shall be fairly similar in their complementation behaviour has over-ridden the minimisation of the number of clusters.

Details of the computer program: The computer selects any complementation group, X , and tries to define the cluster most likely to include X . This is done by examining the remaining groups and selecting that group, X_1 say, which is 'most likely' (the criteria used to define most likely are discussed in detail below) to lie in the same cluster as X . It is also a necessary condition that X_1 should not complement X . The remaining groups are then examined and X_2 selected, where X_2 is the group 'most likely' to lie in the same cluster as both X and X_1 . X_2 must not complement either of them. This process is continued until we have a series of groups $X, X_1, X_2, \dots, X_n$ which cannot be extended further because all the remaining mutants complement one or more of them.

This collection of groups we call the X -derived pseudocluster. A pseudocluster is derived from each of the groups in the matrix in turn and the computer prints these out in a square array, the i th member of the k th row being 1 or 0 according to whether or not the pseudocluster formed by the k th group contains the i th group. Several of the pseudoclusters will be seen to be similar and so we have a rough description of the clustering structure already (see Tables 2 and 3).

The decision that one or another mutant is the 'most likely' extension is effected by computing scores $S_1, \dots, S_n$ for each group not included in the growing pseudocluster. These scores are based on the similarities between the qualitative complementation reactions of the groups being considered as extensions of the collection and groups already included. The score for a group which complements positively with one or more groups already included is zero; other groups have various positive scores depending on the degrees of similarity. The most likely group to be included in the growing pseudocluster is the group with the highest positive score; if none has positive scores the pseudocluster is complete.

Techniques of recursion are available whereby the array of pseudoclusters, P , is used as the starting-point for a better analysis. As well as the scores $S_1, \dots, S_n$ already described the computer calculates scores $r_1, \dots, r_n$ based on P . The r_i are scaled down to percentages and then the computer constructs composite scores, $S_i (r_i - C)$ and chooses the group with the biggest positive composite score to be the next inclusion in the growing pseudocluster. C is a quantity which we specify, called the cut-off point. Groups with $r < C$ are excluded, and C also acts to emphasise r rather than S as C increases towards 100.

Starting with our original P we can do several recursions. As P improves due to recursion, the value of C should be increased to emphasize r (based on P) more and more. Eventually P will become 'self-consistent'—that is to say, for any two groups X_1 and X_2 , if the pseudocluster founded by X_1 includes X_2 then the pseudoclusters founded by X_1 and by X_2 are identical. P now represents a dissection of the complementation groups into clusters and this is the best our method can do with the data.

If too high a value of C was used before P was approximately correct, it was sometimes found that a true cluster might be dissected into subclusters giving more than the minimum number of clusters. To identify the minimum number of clusters, the series of cut-off points 0, 40, 70, 85, 95 would probably be more effective than the series 0, 95.

Our program, written in Atlas algol, together with a description of how to use it is available to anyone who proposes to analyse such data in the near future. We will send copies to any research worker applying to the MRC Statistical Unit for them.

Details of the scoring method used by the computer: (i) The scores $S_1 \dots S_n$. Let E_i be a row of zeroes and ones, the n th element of E_i being 1 if the i th and the n th groups complement positively together and 0 otherwise. If we take the j th and the k th groups X_1 and X_2 and compare E_j and E_k then E_j and E_k will have R positive reactions (ones) in common and N negative reactions (zeroes) in common. We define the *similarity* between X_1 and X_2 to be $(N + R)$. If we are considering extensions of $(X_0, X_1, X_2, \dots, X_n)$ then the score S calculated for the group X is 0 if X is not an admissible extension of $(X_0, X_1, \dots, X_n)$ and the sum of the similarities of X with the X_i otherwise.

A more general criterion was considered based on $R + k \cdot N$. The value of k which made the scores S discriminate most powerfully between the obvious clusters in several different sets of complementation data was found, but it was so close to unity that the extra generality added nothing to the method.

(ii) The scores $r_1 \dots r_n$. Let d_i be the i th column in the array of pseudoclusters P . The score r for the i th group is the sum of the similarities of d_i with the other columns (as above), multiplied by an appropriate factor to make it a percentage of what it would have been if all the things being compared had been identical. When calculating the r_i , putting $k = 0$ and defining the similarity as R is a modification of the method which can be advantageous.

An alternative method of analysis: In the beginning a more complicated method of analysis was used. In constructing a pseudocluster we do not want our next inclusion to be from a different cluster. If this happens, many of the groups that complement positively with either cluster will be excluded, whereas if we had chosen a group from the correct cluster it is likely that fewer restrictions on the extensibility of the pseudocluster would have been imposed. This suggested the definition of the S_i which we used in the beginning.

Several groups will already be formally excluded from the growing pseudocluster because they complement positively with one or more of the groups already included in it, and any extension of the cluster will in general increase the number of formally excluded groups. The group that is 'most likely' by this criterion to lie in that cluster is defined as that group whose inclusion would cause the least number of extra exclusions.

This criterion (extensibility) is computationally less convenient than the similarity criterion and gave practically identical results wherever it was used.

RESULTS

The first tests of the computer programme, which used the method outlined above were made using the data of GROSS for the *leu-2* locus of *Neurospora crassa* (GROSS 1962). Two recursions on the first pseudocustering matrix with C taking the values 0 and then 70 were needed to produce a self-consistent matrix. This self-consistent matrix was found to have classified the groups into the same two clusters and two exceptional groups which have been described previously (GIL-

The complementation matrix for the ad 5/7 data of COSTELLO and BEVAN (1964)

The groups have been labelled 1 to 48 for the purposes of ready comparison with Tables 5 and 6. Actual mutant isolation numbers corresponding to the groups are given in Table 4.

The pseudocustering matrix derived after one recursion from the ad 5/7 data of COSTELLO and BEVAN (1964)

The recursion was performed using a cut-off point $C = 85$. The matrix is self-consistent and shows five clusters. Columns occurring three or more times are listed beside the matrix. All the groups represented by a one in any of the five listed columns fall into the same cluster. Groups not represented in any of the five columns are exceptions.

LIE 1966, 1968). Since clustering at the *leu-2* locus is comparatively straight forward, this was not considered to be a very rigorous test of the method.

The method has also been used to analyse four other sets of data: these are described below.

The ad-5/7 locus in Saccharomyces cerevisiae: data of COSTELLO and BEVAN (1964): The method of analysis is best illustrated using the results from this locus. The results for other loci will be presented in summary form only.

The original complementation matrix for the *ad-5/7* data is presented in Table 1. The first analysis gave a non-self-consistent pseudoclustering matrix. A recursion on this first pseudoclustering matrix, using cut-off point $C = 85$, produced a second pseudoclustering matrix (Table 2) which was self-consistent and in which all except six groups were placed in one of five clusters. However, a different result was obtained when another recursion on the first pseudoclustering matrix using $C = 0$, then $C = 70$ produced a self-consistent matrix (Table 3) with three clusters and four exceptions. In Tables 2 and 3 clusters may be identified by classifying the columns of the matrix; if a column occurs more than once or twice, then the groups which are represented by those columns fall into the same cluster.

At first sight, the computer analysis of this locus has produced two different results showing that the matrix consists of three or five clusters. Further analysis shows that both these results are reasonable and consistent with each other.

The results of the three and five cluster analyses have been compared group by group in Table 4. This shows (last two columns of Table 4) that clusters *B* and *C* of the three cluster analysis have been sub-divided into 'sub-clusters' B_1 , B_2 , and C_1 , C_2 in the five cluster analysis. It is more clear how this can occur when we examine the interaction of clusters one with another.

Table 5 shows what we have called the cluster interaction matrix. Each entry in this matrix represents the estimated probability, calculated directly from the complementation matrix, that complementation will occur between any two mutants selected at random from the corresponding clusters. The cluster interaction matrix must have a row of zeroes down the diagonal, since mutants in the same cluster show no complementation amongst themselves. Zeroes which appear in other parts of the matrix show that these clusters are similar and may be fused. In Table 5b the cluster interaction matrix of the five cluster analysis shows zeroes which are not on the diagonal. In fact the matrix in Table 5b can be reduced by fusing clusters (or 'subclusters') B_1 and B_2 and C_1 with C_2 to give the three cluster matrix as in Table 5a. The matrix in 5b can also, however, be reduced by fusing subclusters B_2 and C_1 to give a four cluster matrix. We can see, therefore, that the analysis of the matrix into five clusters is not by means of an arbitrary division of clusters *B* and *C* into subclusters, but rather the division of these clusters into subclusters on the basis of their interaction with each other.

Comparison of the cluster analysis made by computer with that made previously by inspection of the map (GILLIE 1968) shows that the two analyses are in substantial agreement, i.e. they are about 90 percent mutually consistent.

The ad-5/7 locus in Saccharomyces cerevisiae: data of DORFMAN (1964): Complementation at the *ad-5/7* locus in *Saccharomyces cerevisiae* was also investi-

TABLE 4

*Comparison of cluster classification by the inspection method and by the computer method
for ad 5/7 (COSTELLO and BEVAN 1964)*

Number of complementation group	Representative mutant isolation numbers	Number of mutants in group	Cluster classification by map inspection (GILLIE 1968)	Cluster classification by computer (from Tables 5+6)	
1	631	1	a	A	A
2	653	1	a	A	A
3	E9	1	a	A	A
4	E19	1	a	A	A
5	E17	1	a	A	A
6	75	1	a	A	A
7	200	1	a,c	ex	ex
8	491	1	b,c	B	B ₁
9	518	1	b	B	B ₁
10	E5	1	b	B	B ₂
11	E3	1	b	B	B ₂
12	399	1	b	B	B ₂
13	332	1	b	B	B ₁
14	344	2	b	B	B ₂
15	441	1	b	B	B ₂
16	367	1	b	B	B ₁
17	335	1	b	B	B ₁
18	353	1	b	B	B ₁
19	346	1	b	B	B ₂
20	401	1	b	B	B ₂
21	375	1	b	B	B ₂
22	385	1	b	B	B ₂
23	205	2	b	B	B ₁
24	7	1	b,c	ex	ex
25	66	2	b	B	B ₁
26	E8	1	b	B	B ₁
27	427	1	b	B	B ₁
28	141	1	b	B	B ₁
29	434	1	b	B	B ₁
30	594	1	b	ex	ex
31	334	1	b	ex	ex
32	669	1	b	ex	ex
33	E22	1	c	ex	ex
34	E4	1	c	C	C ₁
35	630	1	c	C	C ₁
36	E30	1	c	C	C ₁
37	35	1	c	C	C ₁
38	86, 28	2	c	C	C ₁
39	676	1	c	C	C ₂
40	6	1	c	C	C ₂
41	125	1	c	C	C ₂
42	E2	2	c	C	C ₂
43	E26	1	c	C	C ₂
44	677	1	c	C	C ₂
45	E28	1	c	C	C ₁
46	13	2	c	C	C ₂
47	E34	1	c	C	C ₂
48	26	1	c	C	C ₁

TABLE 5

Cluster interaction matrices for the ad 5/7 data of COSTELLO and BEVAN (1964)

(a)	A	B	C		
A	0	.76	1.0		
B	.76	0	.48		
C	1.0	.48	0		
(b)	A	B ₁	B ₂	C ₁	C ₂
A	0	.81	.70	1.0	1.0
B ₁	.81	0	0	.38	.78
B ₂	.70	0	0	0	.67
C ₁	1.0	.38	0	0	0
C ₂	1.0	.78	.67	0	0

Two cluster interaction matrices are given. One for the three cluster analysis (Table 2) and another for the five cluster analysis (Table 3). Each entry in the matrix represents the estimated probability that complementation will occur between any two mutants selected at random from the corresponding clusters.

gated by DORFMAN using independently induced mutants. Our analysis into clusters using the computer method was 85 percent consistent with results obtained previously by inspection of the map (see Table 6). Four clusters and

TABLE 6

Comparison of cluster classification by the inspection method and by the computer method for ad 5/7 (DORFMAN 1964)

Number of complementation group	Representative mutant isolation numbers	Number of mutants in group	Cluster classification by map inspection (GILLIE 1968)	Cluster classification by computer (from Table 3)
1	105	1	a	ex
2	146	1	a	ex
3	31	1	a	ex
4	117	1	b	B
5	166	1	b	B
6	E290	1	b	B
7	E323	1	b	B
8	135	1	b	B
9	7	2	b	B
10	123	1	b	B
11	E336	1	b	B
12	26	1	b	B
13	20	1	e	ex
14	1	1	e	E
15	33	1	e	E
16	22	1	e	E
17	E222	1	e	E
18	156	1	e	ex
19	147	1	e	E

TABLE 6—(Continued)

Number of complementation group	Representative mutant isolation numbers	Number of mutants in group	Cluster classification by map inspection (GILLIE 1968)	Cluster classification by computer (from Table 3)
20	E143	1	b,e	B
21	E194	1	b,c,e	C
22	111	1	b,c,e	C
23	E188	1	b,c,e	C
24	E346	1	b,c,e	C
25	104	1	d	ex
26	138	1	d	D
27	132	1	d	D
28	142	1	d	D
29	155	1	d	D
30	19	1	d	D
31	133	1	d	D
32	119	1	d	D
33	137	1	d	D
34	106	3	d	D
35	102	2	d	D
36	17	3	d	D
37	127	1	d	D
38	2	3	d	D
39	10	1	d	D
40	38	2	d	D
41	158	1	d	C
42	34	3	d	D
43	39	1	d	C
44	25	2	d	D
45	3	1	d	C
46	118	1	c	ex
47	124	1	c	C
48	9	1	c	C
49	110	1	c	ex
50	6	2	c	C
51	150	1	c	C
52	16	1	c	ex
53	15	1	c	ex
54	149	1	c	ex
55	F27	1	c	C
56	120	1	c	C
57	107	1	c	C
58	23	1	b,c,d	C
59	165	2	b,c,d	C
60	11	2	b,c,d	C
61	145	1	b,c,d,e	C
62	13	10	b,c,d,e	C
63	125	1	b,c,d	C
64	164	1	b,c,d,e	C
65	4	11	b,c,d,e	C
66	42	2	b,c,d,e	C
67	E303	1	a,c,d,e	C

TABLE 7

Cluster interaction matrix for the ad 5/7 data (DORFMAN 1964)

	B	C	D	E
B	0	.31	.99	.89
C	.31	0	.24	.37
D	.99	.24	0	.96
E	.89	.37	.96	0

Each entry in the matrix represents the estimated probability that complementation will occur between any two mutants selected at random from the corresponding clusters.

eleven exceptions were identified. DORFMAN's original matrix contained some blanks (unknown results); these were excluded by the computer when similarities for the groups concerned were calculated, but otherwise did not affect the method of analysis.

The computer analysis divided the matrix into four clusters, which were very similar to clusters *c*, *d*, *b*, *e* identified in the previous analysis made by inspection of the map. One of the clusters identified previously by inspection (*a*), which formed the 'tail' of the DORFMAN map was represented by three exceptional groups in this analysis. One of these three groups complements the other two and so according to our strict rules cannot be included in a cluster with the other two. The cluster interaction matrix for this locus is shown in Table 7.

The His-B locus in Salmonella typhimurium: data of LOPER et al. (1964): Unfortunately, the original complementation matrix was not available to work from for this locus and we had to reconstruct the matrix from the map and notes published by LOPER *et al.* (1964). On analysis we found four clusters with eight exceptions (see Table 8) in the data for this locus. The division into clusters produced by this analysis was essentially the same as the division into four 'basic complementation groups' made by LOPER *et al.* (1964), and identical with the division into clusters made by inspection (GILLIE 1968). The cluster interaction matrix for this locus is given in Table 9.

The ad-2 locus in Saccharomyces cerevisiae: data of WOODS and BEVAN (1966): The *ad-2* data of WOODS and BEVAN (1966) presented difficulties which we did not encounter in analysing data at the other loci. Five analyses were made using various procedures and were all in agreement in identifying at least five main clusters, but there was a good deal of variation with respect to certain groups which could be assigned to one cluster or another in a fairly arbitrary way. The *ad-2* data were also difficult to analyse by map construction, there being numerous exceptional interactions which could not easily be fitted into the map. However, comparison of the previous manual analysis into clusters based on the map (GILLIE 1968) with one of the computer analyses showed a 70 percent agreement.

The arrangement of groups in clusters in two analyses is shown in Table 10 and the cluster interaction matrices for these analyses are shown in Table 11. Analysis A was performed on the complete matrix and analysis B on a modified matrix in which weak complementation reactions (noted in the original data of

TABLE 8

*List of clusters identified at the His-B locus of Salmonella typhimurium
(data of LOPER et al. 1964)*

A	B	C	D
380	143 (4)	612	641
167	65 (3)	902	53
488	241	391 (2)	923
138	289	20	456 (2)
669 (2)	542	482	206
243	480	569	374 (3)
234	578	118	61
262	40 (4)	12 (3)	590
562	257	79	...
136	217	369	...
286	...	...	...
865	...	...	...
355	...	...	...
59	...	...	...
328	...	...	...
656	...	...	...
812	...	...	...
429	...	...	...
Exceptions: 353, 116, 56, 824, 662, 425, 573, 470			

The complementation groups comprising each cluster are designated by the isolation number of one of the mutants comprising the group. The number in parentheses represents the number of mutants per group when this exceeds one. Exceptional mutants which cannot be placed in clusters are listed separately.

TABLE 9

Cluster interaction matrix for the His-B locus of Salmonella typhimurium (LOPER et al. 1964)

	A	B	C	D
A	0	.62	.84	.76
B	.62	0	.9	1.0
C	.84	.9	0	.92
D	.76	1.0	.92	0

Each entry in the matrix represents the estimated probability that complementation will occur between any two mutants selected at random from the corresponding clusters.

TABLE 10

*Cluster membership assigned to the ad-2 mutants of Saccharomyces cerevisiae
(WOODS and BEVAN 1966) in two analyses*

Complementation group (Representative mutant isolation number)	Analysis A	Analysis B
1	1	ex
110	1	1
67	1	1
14	1	1
113	1	2
12	ex	2
181	2	2
161	2	2
117	2	2
85	2	2
134	2	2
2.7	3	3
2.5	3	8
189	3	8
2.8	3	3
2.10	3	3
2.13	3	3
2.19	3	3
2.14	3	3
25	9	3
6	9	3
172	ex	8
27	ex	ex
184	ex	ex
65	7	ex
4	7	ex
2.15	4	ex
178	4	4
105	4	4
88	4	4
68	4	4
3	4	4
103	4	4
2.4	4	4
2.6	4	4
106	4	4
75	4	4
11	4	4
195	4	4
92	4	4
77	4	4
203	4	4
87	4	4

TABLE 10—Continued

Complementation group (Representative mutant isolation number)	Analysis A	Analysis B
101	4	4
102	4	4
116	4	4
71	4	4
69	4	4
29	4	4
100	6	4
36	6	4
167	6	4
26	6	4
170	5	4
2.21	5	4
82	6	ex
162	6	5
7	5	5
2.1	5	5
80	5	5
138	5	5

Analysis A was performed taking all complementation reactions (weak and strong) as positive and analysis B was performed taking strong complementation reactions as positive. Mutants having the same number in analyses A or B fell in corresponding clusters.

WOODS and BEVAN 1966) were considered as negative reactions. These two analyses are in 70 percent agreement (43 agreements and 18 disagreements) over the assignment of groups to clusters.

Comparing these two analyses in more detail, it can be seen that the ten mutants involved in cluster 3 produce four of the disagreements. These ten mutants react similarly with all other mutants and are mutually noncomplementing with the exception of two pairs of mutants (pair 2.5 and 189, and pair 25 and 6) which do show complementation between themselves. These two pairs cannot lie in the same cluster and either pair may be arbitrarily excluded by the computer. This may be a case where strict adoption of the rule that no complementation may occur between members of a cluster is unrealistic.

Both analyses (Table 10) agree in forming a large cluster 4, but analysis B includes some extra groups in cluster 4 which analysis A places in clusters 5 and 6. Eight disagreements between the two analyses are a result of this. Distinct clusters would not always be expected to occur on the CRICK-ORGEL hypothesis. If for example an 'intersection' of the polypeptide with the axis of symmetry is rather large, as would be the case if the polypeptide chain ran almost parallel to the axis of symmetry for some distance, then mutants at one end of the 'intersection' could complement positively with mutants in the middle or at the far end and precise clusters would not be expected. Clusters 4, 5 and 6 might well represent a phenomenon of this kind. The low interaction probabilities of clusters 4, 5 and 6

TABLE 11

Cluster interaction matrices for the ad-2 data (WOODS and BEVAN 1966)

A.		1	2	3	4	5	6
	1	0	0.6	0.95	0.58	1.0	1.0
	2	0.6	0	0.87	0.3	0.97	1.0
	3	0.95	0.87	0	0.58	1.00	0.98
	4	0.58	0.3	0.58	0	0.30	0.17
	5	1.0	0.97	1.00	0.30	0	0.36
	6	1.0	1.0	0.98	0.17	0.36	0
B.		1	2	3	4	5	8
	1	0	0.24	0.79	0.50	0.73	1.00
	2	0.24	0	0.77	0.44	0.97	1.00
	3	0.79	0.77	0	0.38	0.97	0.29
	4	0.50	0.44	0.38	0	0.16	0.89
	5	0.73	0.97	0.97	0.16	0	1.00
	8	1.00	1.00	0.29	0.89	1.00	0

Matrix A is derived from analysis of the complete data (see Table 10, column A) and matrix B is derived from an analysis in which weak interactions are omitted (see Table 10, column B). Each entry in the matrix represents the estimated probability that complementation will occur between any two mutants selected at random from the corresponding clusters.

TABLE 12

Analysis of three different randomised matrices of the same dimensions as ad-2 (61 groups) containing the same frequency of positive and negative test results

	Number of clusters	Number of groups in each cluster	Number of exceptions
Randomised matrix a	12	5,4,3,7,2,5,3,8,6,6,4,2	6
Randomised matrix b	12	8,6,8,5,4,2,3,2,8,2,4,2	7
Randomised matrix c	12	6,7,5,7,4,4,5,3,2,3,3,2	10
<i>ad-2</i> *	8	5,5,8,23,6,6,2,2	4
<i>ad-2</i> †	6	3,7,3,8,28,5	7

* Using all complementation reactions, results from Table 10A.

† Using good complementation response only, results from 10B.

(Table 11) and the fact that cluster 6 is absorbed into clusters 4 and 5 when weak interactions are ignored, is consistent with this hypothesis.

Since the *ad-2* map has a comparatively large number of not-so-well defined clusters the question may well be asked whether this amount of clustering might not be obtained if an absolutely random matrix of the same size and with the same frequency of positive and negative tests were subjected to our procedure. We have done this (see Table 12) and found that in three different trials of this kind, 12 clusters were always found with from 6 to 10 'exceptional groups' which could not be placed in clusters. This result is quite different from the result obtained for the *ad-2* data proper (see Table 12). However, a further difference was revealed when the estimated probabilities in the cluster interaction matrices were examined. These are the probabilities that complementation will occur between any two mutants selected at random from specified clusters. The frequency distribution for the probabilities from the *ad-2* data proper was skewed with more than half of the probabilities greater than 0.8, whereas the probabilities for the random data appeared symmetrically distributed about the mode which was between 0.5 and 0.6 (see Figure 1). This asymmetrical distribution of probabilities, together with the smaller number of clusters than are found in analyses of randomised matrices show that we have detected real structure in the *ad-2* matrix.

DISCUSSION

A computer program has been developed for sorting mutants into clusters of mutually noncomplementing mutants on the basis of their qualitative complementation behaviour. This program can analyse batches of up to 270 groups of mutants at a time on Atlas, although this upper limit will vary from computer to computer.

Examination of data at several different loci has shown that the clusters produced by this method are similar to the clusters that have been previously produced by inspection of complementation maps. After the program has indicated which mutants lie in which clusters, further study of the data is possible in two main ways. (a) The classification can be used to facilitate the construction of a

A.

B.

C.

D.

FIGURE 1.—The frequency distribution of cluster interaction probabilities for clusters at the *ad-2* locus (Distribution A) and for clusters from three comparable random trials. (Distributions B, C and D). Each rectangle in a histogram represents the probability that complementation will occur between any two mutants selected at random from any two different clusters. The random trials were made using matrices with random entries which had the same frequency of positive and negative tests as the *ad-2* matrix.

complementation map. (b) Tables may be drawn up showing the estimated probability of a mutant chosen at random from one cluster complementing positively with one chosen at random from another cluster, and these estimated probabilities can be studied (see Tables 5, 7, 9 and 11). These tables are called the cluster interaction matrices.

The diagonal elements in the cluster interaction matrix are necessarily zero, since cluster members do not complement positively with each other. If we assume that the clusters we produce correspond to damage at particular sites, then zeroes or low probabilities not on the diagonal indicate that the sites of damage defining the two clusters concerned are fairly closely related in the protein monomer; whereas high probabilities close to 1.0 indicate that the sites of damage are distant from one another; intermediate percentages indicate intermediate degrees of relatedness. It was particularly interesting that these probabilities were found to be skewed for the *ad-2* data when compared with the probabilities obtained from analysis of comparable random matrices.

In map-making the question of which mutants are to be called exceptions is more subjective than in the computer classification into clusters, and the outcome of the choice affects the final results more strongly. The exceptions that we find to the clustering structure may have their own individual entries in the interaction probability table, representing the probabilities of positive complementation between them and the various clusters. If the data are represented as a map, exceptions may be omitted from the map entirely, or may be included in the map with certain of their reactions ignored, or they may be included properly in the map, in which case they alter the large-scale features of the map substantially.

For very complex data, for instance *ad-2*, the interaction probability table is more comprehensible and more accurate than the map, but it cannot tell us if two clusters interact as at the *leu-2* locus to form a circular rather than a linear map. However, the value of the two different methods will be assessable only later when physico-chemical methods of analysis have been applied to complementing proteins. It may then be possible to answer the following two questions: 1. Do the clusters suggested both by the computer and by the maps consist of mutants at particular intersections of the protein monomer with the axes of symmetry of the multimer? 2. Do the interactions between different clusters, as represented on the complementation map, best describe the interconnections of the polypeptide chain in the protein monomer, or are these interconnections better described by the figures in the cluster interaction matrix?

Although it may be some time before these questions can be answered, we believe that our method provides a cheap and objective means of analysing complementation data and so of comparing mutational defects at what we might call the functional level.

We should like to acknowledge the interest and encouragement of Dr. DAVID MANNION.

SUMMARY

A computer method has been developed for detecting 'clusters' of similarly

reacting mutants from complementation data. This has been applied to five sets of data from four loci; the *leu-2* locus of *Neurospora crassa*, the *ad-5/7* and *ad-2* loci of *Saccharomyces cerevisiae* and the *His-B* locus of *Salmonella typhimurium*. It was found possible in each case to divide the mutants up into a small number of mutually noncomplementing clusters of mutants (or complementation groups) and to study the interaction of these clusters. In the case of the *ad-2* data a comparison was made with analyses of randomised matrices showing that the *ad-2* data were nonrandom. The assignment of complementation groups to clusters by the computer method gave similar but not identical results to those obtained previously by simple inspection. The reasons for differences between the different methods are discussed.

LITERATURE CITED

- CODDINGTON, A. and J. R. S. FINCHAM, 1965 Proof of hybrid enzyme formation in a case of inter-allelic complementation in *Neurospora crassa*. J. Mol. Biol. **12**: 152-161.
- CODDINGTON, A., J. R. S. FINCHAM and T. K. SUNDARAM, 1966 Multiple active varieties of *Neurospora* glutamate dehydrogenase formed by hybridisation between two inactive mutant proteins *in vivo* and *in vitro*. J. Mol. Biol. **17**: 503-512.
- COSTELLO, W. P. and E. A. BEVAN, 1964 Complementation between *ad 5/7* alleles in yeast. Genetics **50**: 1219-1230.
- CRICK, F. H. C. and L. E. ORGEL, 1964 The theory of interallelic complementation. J. Mol. Biol. **8**: 161-165.
- DORFMAN, B., 1964 Allelic complementation at the *ad 5/7* locus in yeast. Genetics **50**: 1231-1243.
- FINCHAM, J. R. S., 1966 *Genetic Complementation*. Benjamin, New York.
- GILLIE, O. J., 1966 The interpretation of complementation data. Genet. Res. **8**: 9-31. ——— 1968 Interpretations of some large non-linear complementation maps. Genetics **58**: 543-555.
- GROSS, S. R., 1962 On the mechanism of complementation at the *leu-2* locus of *Neurospora*. Proc. Natl. Acad. Sci. U.S. **48**: 922-930.
- LOPER, J. C., M. GRABNER, R. C. STAHL, Z. HARTMAN and P. E. HARTMAN, 1964 Genes and proteins involved in histidine synthesis in *Salmonella*. Brookhaven Symp. Biol. **17**: 15-52.
- SCHLESINGER, M. J. and C. LEVINTHAL, 1963 Hybrid protein formation of *E. coli* alkaline phosphatase leading to *in vitro* complementation. J. Mol. Biol. **7**: 1-12.
- WOODS, R. A. and E. A. BEVAN, 1966 Interallelic complementation of the *ad-2* locus of *Saccharomyces cerevisiae*. Heredity **21**: 121-130.