J. Mol. Evol. 11, 47–56 (1978)

Journal of Molecular Evolution
© by Springer-Verlag 1978

Relationship Between Gene Function and Gene Location in *Escherichia coli*

Monica Riley, Lawrence Solomon, and David Zipkas

Department of Biochemistry, State University of New York at Stony Brook, Stony Brook, New York, 11794, USA

Summary. Genes of *Escherichia coli* were grouped according to the "biochemical relatedness" of the enzymes they specifiy, using two schemes to determine relatedness: similarity of reaction or similarity of reactants. The tendency of biochemically related genes as so defined to lie approximately 90^{o} or 180^{o} from one another on the circular genetic map was analyzed statistically. Of the classes analyzed, only the genes for the enzymes of glucose catabolism showed a significant departure from random distribution in this respect. The glucose catabolism genes showed a pronounced tendency to lie either 90^{o} or 180^{o} from one another (P = ca. 10^{-9}), and, furthermore, most of these genes were found to lie in only four gene clusters on the *E. coli* genome. The significance of this observation is discussed in relation to evolutionary mechanisms and to mechanisms of gene expression.

Key words: Genome duplication – Genome topography – Evolution – Gene expression

Introduction

Previously we analyzed the relationship between gene location and gene function in *Escherichia coli* (1), using the information that was assembled by Taylor and Trotter in 1972 (2). Since then, Bachman, Low, and Taylor have recalibrated the *E. coli* map (3), more genes have been mapped, and the biochemical activities of more gene products are now well defined. With significantly more accurate and more extensive information available on *E. coli* genes, we have undertaken to analyze the relationship between gene function and gene location using restated definitions of biochemical relatedness that seem stricter than the criteria used before. Our earlier analysis (1) indicated that many groups of seemingly functionally related genes fall 90^{o} or 180^{o} from one another on the Taylor-Trotter map. The present, more narrowly defined analysis indicates that the genes concerned with glucose catabolism are located 90^{o} or 180^{o} from one another on the new map. Other genes, however, when they are grouped by the criteria that are defined below, do not show this relationship.

0022–2844/78/0011/0047/$ 02.00

Definitions of Relatedness

We have asked whether, as previously proposed, *E. coli* genes whose gene products are biochemically related tend to lie 90^{o} or 180^{o} apart on the circular genome. To examine this question, "biochemical relatedness" of enzymes was arbitrarily defined in two alternative ways. As one alternative, two enzymes were considered to be "related" if they catalyse similar reactions as defined by the Enzyme Commission (International Union of Biochemistry) classification system. As another alternative, enzymes were considered to be "related" if there exists a common molecule which is related to the enzymes as either a substrate or a product.

Two lists of *E. coli* gene pairs were made using these criteria. To make the first list, genes were grouped according to the Enzyme Commission classification of the gene product. Within each such classification group, all possible gene pairs were listed.

To make the second list, metabolic pathways for *E. coli* were set down and were divided into major groupings such as carbohydrate metabolism, or synthesis of metabolites of the aspartate family or the glutamate family. Where the genes for the enzymes were known and mapped, the symbol for the gene was entered beside the relevant reaction. A list of gene pairs was then made such that the two enzymes produced by the genes can be seen to interact with a common molecule. The common molecule was specified to be related to the enzymes either as a substrate or as a product. Coenzymes were excluded from consideration. By this definition, enzymes catalyzing sequential reactions in a metabolic pathway are "biochemically related", since the substrate of one enzyme is a product of the other. Isozymes are also defined to be related, as are enzymes which act upon the same substrate or produce the same product. In some cases, more than two enzymes are related to a common molecule by this definition. In these cases, all possible gene pairs within such a group were included in the list.

For both lists, the map distances between each gene pair, derived from the recalibrated map (3), was recorded to the nearest 0.1 min (even though it was recognized that mapping accuracy may not be this great in all cases).

Next, the tendency for the gene pairs on both of these lists to lie either 90^{o} or 180^{o} apart on the *E. coli* map was tested statistically.

Method of Analysis

On a 100-min map, separations between two genes of 25, 50, and 75 minutes correspond to 90^{o}, 180^{o}, and 270^{o} separations respectively. Since 90^{o} clockwise is equal to 270^{o} counterclockwise, we define two genes of a gene pair to have a $90^{o}/180^{o}$ relationship if their map distances are 25, 50, or 75 min apart. If the two genes have such a topographical relationship, the pair is considered to be a Hit; if not, the separation is considered a Miss, except for one case, a null case, which is mentioned below.

Because some movement of genes along the chromosome is expected over evolutionary time (cf. *Salmonella typhimurium* and *E. coli* maps) (10), we did not look for separations of exactly 90^{o} or 180^{o}, but instead added an arbitrary range of $\pm$ 3 min within which a separation could fall and be considered a Hit. This latitude, when applied to map distances that are recorded to the nearest 0.1 min, entails an effective

range of 3.5 min. Thus, separation of 21.5–28.5 and 46.5–53.5 min in either a clockwise or a counterclockwise direction are considered to be Hits.[1] All other separations were considered Misses, except that separations of 0 ± 3.5 min were considered neither Hits nor misses, and were removed from consideration as a null class.

Using these limits, the probability of a given separation being a Hit (assuming a random distribution) was computed. This figure was determined by assuming one of the pair of points to be fixed. Three seven-minute zones exist in which the second point could lie and make the separation a Hit. Since only 93 min are available for the separation to be considered a Hit or Miss (separations from a fixed point of ± 3.5 min fall into the null class), the *a priori* probability of a Hit is 3 x 7/93 = 0.23.

For both of the lists of gene pairs described above, frequencies of gene separations that were Hits or Misses were determined for each list as a whole and for sub-groupings of genes within each list. The chance that a random group of the same number of genes would give as great or greater a percentage of Hits was then computed. The exact probability, P, that such an occurrence could occur by chance was found by expanding the binomial formula, *viz.*, where k = sample size (excluding null cases),

$$P = \sum_{n}^{k} \frac{k!}{n!\,(k-n)!} p^n\, q^{(k-n)}$$

n = observed number of Hits, p = expected frequency of Hits and $q = 1 - p$ = expected frequency of Misses. Values of $P \leqslant 0.05$ were considered significant.

Results

When the list of gene pairs which were grouped by reaction type was tested for 90° or 180° relationships, none of the major subdivisions contained more gene pair separations that were Hits than is expected by chance (Table 1). Genes whose enzymes are biochemically related by gross similarity of the reaction catalysed show no tendency to lie 90° or 180° from one another. The analysis was also performed on Enzyme Commission classification subgroups that have more narrowly defined biochemical similarities (Table 1). For only two of these subgroups, C-C lyases and glycosyl transferases, were more gene pairs separated by 90° or 180° than would be expected by chance. This observation will be discussed further below.

When the second list of gene pairs, as related by a common substrate or product, was tested in this way, more gene pair separations were found that were Hits than would be expected by chance (Table 2). The overall value of P for all genes in this list was 0.024. Analysis was also performed on metabolic subgroups such as the genes of carbohydrate metabolism, or synthesis of metabolities of the aspartate or glutamate families. Of the subgroups, when the distance between genes having common reactants were used in the analysis, only the genes of carbohydrate metabolism showed a significant deviation towards 90° or 180° relationships ($P = 2.4 \times 10^{-3}$) (Table 2). No other metabolic subgroup yielded P values smaller than 0.22.

---

[1] Separations that were 25.0, 20.0, or 75.0 ± exactly 3.5 min were considered half Hits and half Misses.

Table 1. Probability Values for Distribution of Gene Pair Separations when Grouped by Reaction Type

Group	*E.C.*[a] *Number*	*Number of Gene Pairs*[b]			
		Hits	*Misses*	*Total*	*P*
A. Major Groups					
Oxidoreductases	1.X.X.X.[c]	111	499	610	0.71
Transferases	2.X.X.X.	471	2128	2599	0.91
Hydrolases	3.X.X.X.	33	149	182	0.59
Lyases	4.X.X.X.	176	724	900	0.39
Isomerases	5.X.X.X.	8	43	51	0.66
Ligases	6.x.X.X.	74	348	422	0.78
B. Sub Groups					
Oxidoreductases acting on hydroxyl group	1.1.X.X.	27	138	165	0.79
aldo-keto groups	1.2.X.X.	4	11	15	0.32
NAD or NADP	1.X.1.X.	58	283	341	0.82
Transferases					
methyl-	2.1.X.X.	6	47	53	0.90
acyl-	2.3.X.X.	3	7	10	0.30
glycosyl-	2.4.X.X.	22	53	75	0.022
amino-	2.6.X.X.	5	38	43	0.86
phospho-	2.7.X.X.	81	396	477	0.88
Hydrolases					
ester bonds	3.1.X.X.	10	35	45	0.36
glycosyl compounds	3.2.X.X.	2	4	6	0.32
non-peptide C-N	3.5.X.X.	1	14	15	0.81
Lyases					
C-C	4.1.X.X.	57	161	218	0.007
C-O	4.2.X.X.	19	98	117	0.75
C-N	4.3.X.X.	0	6	6	0.71
Isomerases					
aldo-keto	5.3.1.X.	6	13	19	0.14
Ligases					
tRNA synthetases	6.1.X.X.	13	76	89	0.83
C-N	6.3.X.X.	14	61	75	0.47

[a] Enzyme Commission
[b] Null class not included
[c] X represents any integer

The genes for carbohydrate metabolism were further subdivided to isolate the genes for the enzymes of the major pathways of glucose degradation, as these were viewed as the basic, central enzymes necessary for the production of energy and metabolic intermediates in *E. coli.* Genes for synthetic reactions were excluded from this group.

Table 2. Probability Values for Distribution of Gene Pair Separations when Grouped by Metabolic Pathway

Metabolic Groups	*Numbers of Gene Pairs*[a]			
	Hits	*Misses*	*Total*	*P*
Glucose catabolism[b]	23.5	14.5	37.0	1.0×10^{-6}
Other Carbohydrate[c]	22.0	73.0	95.0	0.48
Total Carbohydrate	44.5	87.5	132.0	2.4×10^{-3}
Aspartate Family[d]	15.5	43.5	59.0	0.30
Glutamate Family[e]	8.0	21.0	29.0	0.32
Aromatic Compounds[f]	6.0	30.0	36.0	0.85
Other[g]	3.0	18.0	21.0	0.88
Total Non-Carbohydrate	32.5	112.5	145.0	0.55
Grand Total	77.0	200.0	277.0	0.024

[a] Null class not included. The numbers of gene pairs designated as null in each group are: glucose, 7; other carbohydrate 17; aspartate, 21; glutamate, 5; aromatic, 9; other, 3.
[b] Emden-Meyerhof, pentose phosphate, and Entner-Douderhoff pathways; tricarboxylic acid and glyoxylate cycles.
[c] Degradation of carbohydrates other than glucose; synthetic reactions.
[d] Biosynthesis of lysine, asparagine, methionine, isoleucine, valine, leucine, pyrimidines.
[e] Biosynthesis of glutamine, proline, arginine, putrescine, protoporphyrinogen.
[f] Biosynthesis of phenylalanine, tyrosine, tryptophan, ubiquinone.
[g] Biosynthesis of serine, glycine, quinolate ribonucleotide, murein, fatty acids, histidine, purines.

Genes for degradation of carbohydrates other than glucose were excluded. Genes that connect carbohydrate catabolic intermediates with other pathways, such as *ser A, asp A* or *met A*, were also excluded. The central pathways of glucose degradation, as so defined, and the relevant genes are shown in Figure 1. The position of these genes on the *E. coli* map, taken from references (3) and (4), are shown in Figure 2.

The probability that the number of Hits which are observed for the genes of glucose catabolism would arise by chance was calculated to be 1.0×10^{-6} (Table 2). It appears that within this metabolic group, pairs of genes for enzymes that share a common reactant have a pronounced tendency to lie either 90° or 180° apart on the circular map.

To pursue this coincidence further, we next considered the question of whether *all* possible, pairwise combinations of glucose catabolism genes lie 90° or 180° apart. Of the 378 gene separations tested, 44 fell into the null class and were removed from consideration. Of the rest, 126 were Hits, 208 were Misses. This would occur randomly only at the rate of one in 300 million trials ($P = 3 \times 10^{-9}$). Therefore, all members of this group of genes, not simply those pairs that are related by common reactant, tend to lie 90° or 180° from one another.

A pictorial representation is given in Figure 3. A plot of the distribution of genes within each 5-minute interval of the map is presented, commencing arbitrarily at 0 min. Four major peaks can be seen, approximately equidistant from one another. The genes for glucokinase (*glk*), the glucose phosphotransferase system (*pts G*), pyruvate oxidative decarboxylation (*ace E, ace F*), and lactate dehydrogenase (*lct*), all seem to lie

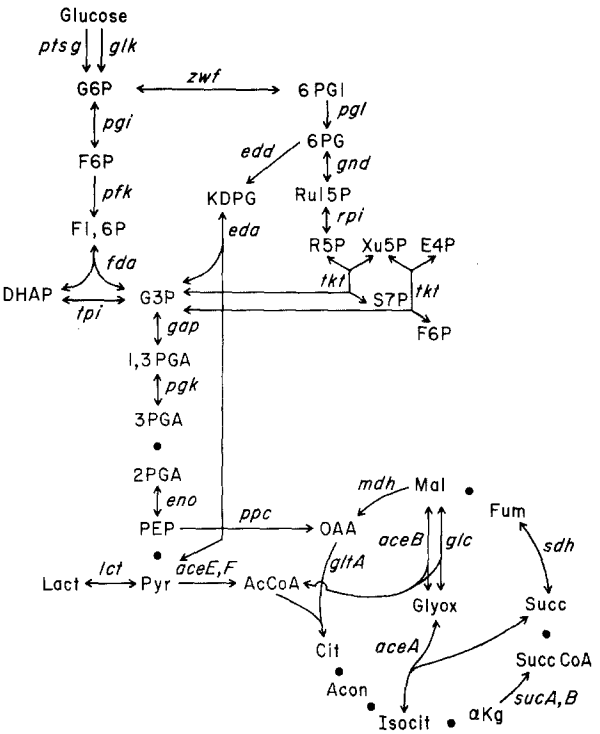

Fig. 1. *Genes of the pathways of glucose catabolism in C. coli.* Includes the genes of the Embden-Meyerhoff, pentose phosphate, and Entner-Douderoff pathways, and the genes for the tricarboxylic acid cycle and the glyoxylate cycle. Mapped genes are indicated beside reaction arrows. Dots represent reactions for which no mapped gene is known

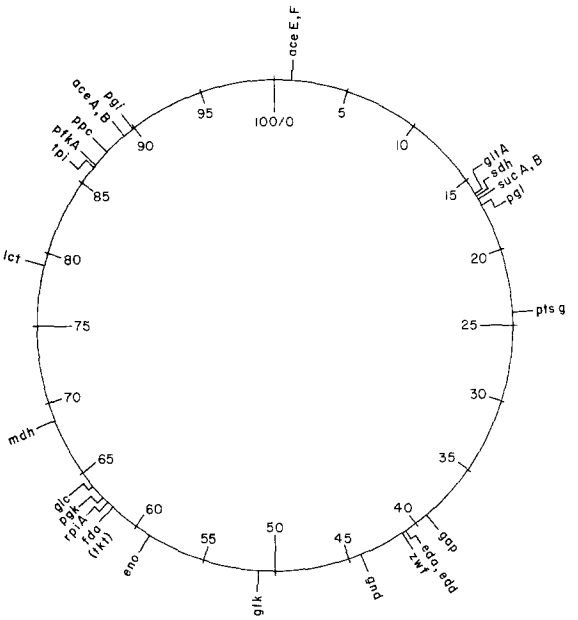

Fig. 2. *Map positions of the genes of glucose catabolism in E. coli*

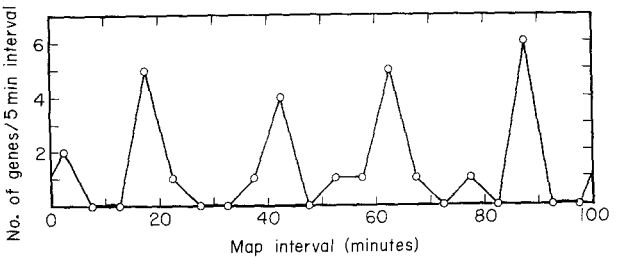

Fig. 3. *Distribution frequency of glucose degradation genes on the E. coli genome.* The number of genes in each five-minute segment was summed and plotted at the mid-point of each map segment

well outside of the clusters. The remaining 23 out of 28 genes (or if one excepts *eno, mdh,* and *gnd,* 20 out of 28) seem to lie in clusters at approximately 90° intervals on the map.

We were interested in establishing in a quantitative way whether the spacing of these clusters is compatible with chance, or whether there is a pronounced tendency for the clusters themselves to lie 90° or 180° from one another. There is no standard statistical approach for testing the existence of, boundaries of, or interrelationships among clusters. However, by inspection one can identify four clusters spanning less than 4 min, each of which contains closely spaced genes (intergene distances of 2 min or less). The genes in these four clusters are: (1) *gltA, sdh, sucA, sucB, pgl,* (2) *gap, eda, edd, zwf,* (3) *tkt, fda, rpiA, pgk, glc,* (4) *tpi, pfkA, ppc, aceA, aceB, pgi.* One finds that this group of 20 genes accounts for 82% of all of the Hits in the "all pairwise combinations" test carried out above.

To test whether these four clusters of closely grouped genes lie 90° or 180° from one another, the midpoint of the map span of each cluster was determined to be as follows: (1) 16.5, (2) 40.0, (3) 63, and (4) 88.2 min. The six separations between these midpoints were calculated, and all were found to fall within ± 3.5 min of 90° or 180°. Therefore, in this favorable case which examines the relative position of the most closely spaced 20 genes out of the total of 28, the genes are found to be disposed in four clusters which lie 90° or 180° from one another. This arrangement would not be expected to occur by chance ($P = 0.0019$).

The absence of 90° clustering among the several subsets of biosynthetic genes, as we have defined them in the second list (Table 2) is puzzling. We considered the possibility that these genes might be topographically related in another way, such as perhaps at 45° from one another. Mizobuchi and Saito (5) have suggested that the *E. coli* genome has undergone three sequential duplications in the past, and that ancestrally related genes are found 45° apart on the map. Using the map positions of Bachman, Low, and Taylor, we tested whether the gene pairs in each of the metabolic subsets had a tendency to lie 45° apart (12.5 ± 2 min). In no case was significance found at the 0.05 probability level.

Finally, we can return for a moment to consider those subsets of the first ("type of reaction") list of genes that seemed to show significant departures from randomness

(Table 1). In retrospect, it seems clear that the apparent tendency of the C-C lyase genes to lie 90^{o} or 180^{o} apart can be attributed solely to the glucose catabolism genes within that group. After extraction of the glucose degradation genes from the C-C lyase group, no significance in 90^{o} or 180^{o} gene location remains ($P = 0.21$). The non-random distribution of the genes for glycosyl transferases remains unexplained, since no central glucose degradation genes, as they are defined here, reside in this group.

Discussion

We conclude that there is a strong tendency for the genes governing the reactions of glucose catabolism to lie 90^{o} or 180^{o} from one another, and that most of these genes fall into four clusters which themselves have a marked tendency to lie 90^{o} or 180^{o} from one another on the *E. coli* genome. In the other metabolic groupings that were tested, gene pairs which are related by a common substrate or product, or gene pairs which are related by similar reaction mechanisms, do not show this tendency.

Although we do not yet know the nature of the underlying mechanisms or the specific factors that have led to the establishment and maintenance of these non-random gene locations, the groupings of the glucose catabolism genes that we see here might reflect, as we suggested previously (1), earlier events in the evolution of the present-day genome of *E. coli*, such as perhaps two sequential duplications of a primitive genome that was about 1/4 the size of the present genome. After genome duplication, mutation and divergence of function of the duplicated genes could lead to an expansion of genetic capacity and, in the absence of gene movement, could leave evolutionarily closely related genes positioned either 180^{o} (one genome duplication) or 90^{o} (two genome duplications) apart. Genome duplication was suggested earlier by Hopwood to explain the location of biochemically related genes in *Streptomyces coelicolor* (6, 7). Recently Sparrow and Nauman (8) have proposed that the process of genome doubling has played a major role in the evolution of eukaryotes as well as procaryotes.

If some of the genes of glucose catabolism have evolved by divergence from others of this group, we are not able, by this analysis, to distinguish whether such ancestrally related gene pairs are functionally related by similarity of reaction mechanism or by similarity or specificity toward metabolites or by some other relationship, since the *P* value for *all* possible gene pairs within the group of glucose catabolism genes is highly significant. In terms of the genome duplication model, ancestor-descendant relationships for individual gene pairs cannot be deduced by the approach that has been applied here.

The absence of 90^{o} or 180^{o} (or 45^{o}) placement of the genes that govern sequential reactions in biosynthetic pathways seems to imply that if genome duplication did take place in an ancestral enteric organism, there was no tendency for any set of four ancestrally related gene copies to evolve into a set of genes that at the present day govern sequential reactions within a single biosynthetic pathway. Descendants of any one ancestral biosynthetic gene may not have evolved into genes that now function within one single synthetic pathway; instead they may appear in up to four separate pathways in the modern cell. As a corollary, the genes of each synthetic pathway may well be derived from more than one ancestral gene. It also seems possible that further analysis

of the *E. coli* genome might reveal, among sets of biosynthetic genes that lie 90° apart, other sorts of evolutionary relationships which are neither a common reaction type as specified in our first list nor a sharing of specificity toward a common small molecule as specified in our second list.

Returning to consideration of the glucose catabolism genes, we recognize that it is possible that the observed 90° grouping of these genes may not reflect evolutionary events at all, but may reflect instead some present-day functional requirement. There may be a constraint on gene location that is not at present understood, whether imposed by cellular control mechanisms or by modes of gene expression. Transcription or translation of certain related genes may entail or require a specific spatial clustering of the genes. Although it is known that translocated genes can be expressed in their new locations (e.g. ref. 9), genome topography might play a role in gene expression that we do not yet appreciate.

The generality of the phenomenon of the non-random location of glucose catabolism genes in other bacterial species cannot be properly assessed at present. The genetic maps of other bacterial species are not as advanced as the *E. coli* map. The *Salmonella typhimurium* map is very much like the *E. coli* map (10). The *Bacillus subtilis* map is quite different (11). The few genes of central carbohydrate metabolism that have been mapped in *B. subtilis*, such as *cit* genes, do not fall 90° from one another. Some of the genes for the enzymes of glucose degradation have been located in *Pseudomonas putida* (12). These genes are all clustered in one location and are not linked to other known clusters of degradation genes, such as the cluster of genes for the degradation of aromatic acids (13). More relevant to this work is the observation that in the enteric organism, *Klebsiella pneumoniae, vpr-1*, the gene for acetoin formation from pyruvate (a gene not present in *E. coli*), has been mapped at a position that appears to lie in a region that corresponds to the upper left cluster of carbohydrate degradation genes in *E. coli* (14).

To summarize, we have observed that 20 of the 28 genes for glucose catabolism in *E. coli* appear to lie in four clusters that are approximately 90° apart on the genome. We note that these gene locations may reflect earlier events in evolution, such as genome duplications, or they may reflect mechanisms of gene expression or other cellular processes that are not fully understood at present.

Acknowledgements. This work was supported by a grant from the National Institutes of Health (GM-21316).

References

1. Zipkas, D., Riley, M.: (1975) Proposal Concerning Mechanism of Evolution of the Genome of *Escherichia coli.* Proc. Natl. Acad. Sci., Wash., **72**, 1354–1358.
2. Taylor, A.L., Trotter, C.D.: (1972) Linkage map of *Escherichia coli* strain K-12. Bacteriol. Rev. **36**, 504–524
3. Bachman, B.J., Low, K.B., Taylor, A.: (1976) Recalibrated Linkage Map of *Escherichia coli* K-12. Bacteriol. Rev. **40**, 116–167
4. Irani, M.H., Maitra, P.K.: (1976) Glyceraldehyde-3-Phosphate Dehydrogenase, Glycerate-3-Phosphate Kinase and Enolase Mutants of *Escherichia coli:* Genetic Studies. Mol. Gen. Genetics **145**, 65–71

5. Mizobuchi, K., Saito, H.: (1975) Proceedings, Molecular Biology Meeting of Japan (Tokyo) pp. 60–62
6. Hopwood, D.A.: (1967a) In discussion to F.W. Stahl's paper: Circular Genetic Maps. Cell Physiol. **70**, Suppl. 1, 1–12
7. Hopwood, D.A.: (1967b) Genetic Analysis and Genome Structure in *Streptomyces coelicolor.* Bacteriol. Rev. **31**, 373–403
8. Sparrow, A.H., Nauman, A.F.: (1976) Evolution of Genome Size by DNA Doublings. Science **192**, 524–529
9. Beckwith, J.R., Singer, E.R., Epstein, W.: (1966) Transposition of the *Lac* Region of *E. coli.* Cold Spring Harbor Symp. Quant. Biol. **31**, 393–401
10. Sanderson, K.E.: (1976) Genetic Relatedness in the Family *Enterobacteriaceae.* Ann. Rev. Microbiol. **30**, 327–349
11. Kejzlarova-Lepesant, J., Harford, N., Lepesant, J.A., Delonder, R.: (1975) In: *Spores VI,* P. Gerhardt et al., ed., pp. 592–614, Washington: Amer. Soc. Microbiology
12. de Torrontegui, G., Diaz, R., Wheelis, M.L., Canovas, J.L.: (1976) Supraoperonic Clustering of Genes Specifying Glucose Dissimilation in *Pseudomonas putida.* Mol. Gen. Genetics **144**, 307–311
13. Leidigh, B.J., Wheelis, M.L.: (1973) The Clustering on the *Pseudomonas putida* Chromosome of Genes Specifying Dissimilatory Functions. J. Mol. Evol. **2**, 235–242
14. Matsumo, H., Tazaki, T.: (1971) Genetic Mapping of *aro, pyr* and *pur* Markers in *Klebsiella pneumoniae.* Japan J. Microbiol. **15** (1), 11–20

Received October 20, 1977