Proc. Nat. Acad. Sci. USA
Vol. 72, No. 4, pp. 1354–1358, April 1975

Proposal Concerning Mechanism of Evolution of the Genome of *Escherichia coli*

(gene duplication/concatemer DNA/polyploidy)

DAVID ZIPKAS AND MONICA RILEY

Biochemistry Department, State University of New York at Stony Brook, Stony Brook, N.Y. 11794

Communicated by Bentley Glass, January 28, 1975

ABSTRACT **Many pairs of genes whose gene products are functionally related lie either 90° or 180° apart on the circular map of the *E. coli* chromosome. A mechanism of evolution is proposed that involves two sequential duplications of an ancestral genome, followed by mutation and divergence of function of replicate genes.**

The process of total genome duplication (as distinguished from individual gene duplication) may have played a role in evolution. Wallace and Morowitz (1) have noted a discontinuity in the frequency distribution of genome sizes of *Mycoplasma* species and have proposed that genome duplications were involved in the evolution of prokaryotes. Hopwood (2) has taken note of 180° relationships between related genes on the circular map of *Streptomyces coelicolor* and has speculated on the possibility of genome duplication as a mechanism of evolution of this organism. We have inspected the genetic map of *Escherichia coli* K12 and have found that biochemically related pairs of genes have a tendency to lie either 90° or 180° apart on the map. We propose that the genome of *E. coli* underwent two sequential genome duplications in the past.

The *E. coli* map is still incomplete: we are still far from being able to lay out the entire genetic atlas of *E. coli*. Nevertheless, using currently available information, we have prepared a list of mapped genes whose gene products could be considered to be biochemically related.

We proceeded in the following way: using the list of genes and relevant literature citations collected by Taylor and Trotter (3), and without prejudice as to map position, we extracted those genes whose gene products have been characterized to an extent that allowed us to make a judgment concerning the functional relatedness of pairs or groups of genes. Genes were considered to be related if their gene products fell into one of the following categories:

(*1*) isozymes catalyzing the same reaction (e.g., aspartokinases I, II, and III);

(*2*) functionally related enzymes that share some feature of specificity, such as (*a*) sequential enzymes in metabolic conversions (e.g., the enzymes of purine biosynthesis), (*b*) enzymes that catalyze similar, but not identical, reactions involving the same or similar small molecules (e.g., tryptophan synthetase B protein and tryptophanase);

(*3*) related cellular structures (e.g., ribosomal proteins) or functions (e.g., potassium transport).

Using these criteria, a list of pairs or groups of related genes was assembled. From this list, we eliminated those sets of functionally related genes all members of which fall into a single operon cluster (e.g., *his* and *ara* genes), although we retained individual members of an operon such as *thrC* if the gene seemed to have a biochemical relationship to one or more genes lying outside of the operon, such as, in this instance, *ilvA*.

Genes that were not included in the list include genes whose degree of relatedness in biochemical terms could not be readily assessed, such as genes for cell shape. Also excluded from consideration were regulatory genes, genes for phage attachment and phage resistance, and genes for suppressors and transfer RNAs. Certain arbitrary, and to some extent subjective, judgments were made. For instance, although the gene products of the potassium transport system genes have not been characterized biochemically, these genes were included.

Of the 460 genes listed by Taylor and Trotter (3), 154 genes were judged to be biochemically related to at least one other gene that has been mapped at a different location. A list of these 154 genes is presented in Table 1, grouped according to biochemical function. Instances in which seemingly related genes are about 180° (42–48 min) apart or about 90° (20–25 min) apart are indicated.

Since not all enzymes of intermediary metabolism have been correlated with mapped genes, only a portion of the full metabolic capacity of *E. coli* is represented on this list. Inspection of the list shows that the largest group of mapped genes that are involved in consecutive metabolic conversions is the group concerned with carbohydrate and fatty acid metabolism. These metabolic conversions are shown in Fig. 1, together with the gene designations for each reaction. It can be seen that many metabolically related genes are either about 90° or about 180° from one another.

In order to facilitate analysis, we used the numbers of gene "locations" rather than the total numbers of genes. Some genes for related functions are clustered on the map, and for purposes of examining the topographical relationships between genes of related function such grouping can be considered as one location. For example, the *argB*, *C*, *E*, and *H* genes are all at 79 min on the map, and are considered here as one gene location at 79 min. In these terms, our list of 154 genes reduces to 125 gene locations. For brevity, we will refer, in the discussion to follow, to gene locations as genes. Of these 125 genes, 92 (74%) are involved in either about 90° or about 180° relationships or both.

We have tested the statistical significance of the 90° and 180° relationships by performing a G Test (for the goodness of fit) on the distribution of gene location pair distances in the following manner. We defined pairs of gene locations as being of two classes, biochemically related and biochemically nonrelated. We determined the map distances separating all 125 locations one from another, expressing these in terms of map minutes, proceeding always in a clockwise direction. The distribution of these distances was then determined (that is, the number of gene pairs that lie 0, 1, 2, . . . , 88, and 89 min apart). A computer program was written to carry out these processes.

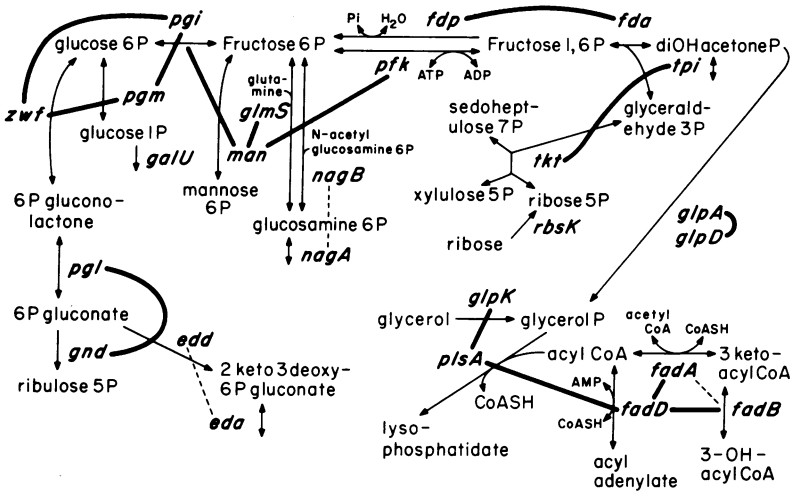

FIG. 1. Map relationships of metabolically related genes. Metabolic conversions are indicated with light arrows. The genes determining the enzyme for each reaction are shown beside the reaction arrow. Heavy lines connect related genes that lie either about 90° or about 180° apart on the *E. coli* map. Broken lines connect genes that are contiguous on the map.

We determined by hand the distances between biochemically related gene pairs. Two methods of identifying related gene pairs were used. In the case of groups of genes concerned with a single function (such as the potassium transport genes) or the genes of a simple biosynthetic pathway (such as the *arg* or *dap* genes), the distances were calculated for all possible gene pairs in the group. In the case of wide-ranging and metabolically complex groups (such as the group in Fig. 1 concerned with carbohydrate and fatty acid metabolism), distances were determined only for the gene pairs that have a close biochemical relationship. The distribution of these biochemically related gene pair distances was determined as described above.

To find the distribution of biochemically nonrelated gene pairs, we subtracted the distribution of related gene pairs from the distribution for all 125 gene locations. The significance of the departure of the observed distributions from a random distribution was then tested.

The results of the G Tests are presented in Table 2. The expected frequencies were determined by calculating that 21.35% of all distances will be equivalent to 90°, 180°, and 270° relationships as we define them (20–25 min, 42–48 min, and 65–70 min). Note that 270° clockwise is equivalent to 90° counterclockwise. As a consequence, 78.65% of all distances should fall outside of these limits. These values assume a random distribution of distances. The observed frequencies are the sums of the observed distributions within the 90°, 180°, and 270° limits as defined above (In Limits) and the sums of those not within the limits (Out of Limits).

As can be seen in Table 2A and B, the G Test values are much larger than the χ^2 value for $p_{0.05} = 3.841$, indicating statistical significance. The distributions for both the biochemically related and nonrelated gene pair separations do not fit the expected random distribution. In the case of biochemically related gene pairs, more distances fall into the 90°, 180°, and 270° limits than expected, whereas in the case of biochemically nonrelated gene pairs fewer fall within these limits than expected. Table 2C presents the results of a G Test for independence of the two distributions. Again, the G Test value is larger than the χ^2 value for $p_{0.05} = 3.841$, indicating that the two distributions are not independent.

We fully recognize the fact that evolutionarily meaningless groupings are undoubtedly present in our list of related genes. Evolutionary relationships between genes cannot be asserted simply on the basis of functional relationships. Nevertheless, we suggest that the results of the G Tests indicate that the frequency of observed 90° and 180° relationships between functionally related genes is much greater than would be expected by chance alone and that these topographical relationships may echo earlier events in the evolution of the *E. coli* genome.

PROPOSED MODEL FOR THE EVOLUTION OF THE *E. COLI* GENOME

Duplication of the Genome. We propose that duplication of a primitive genome led to a duplication of all genes in such a way that duplicate gene pairs were disposed 180° apart on the circular genome. A second such duplication would lead to a total of four copies of each gene, separated by 90° from one another.

There are several ways in which one can visualize genome duplications that would lead to this arrangement of duplicated genes. One way involves homologous crossing over between identical circular genomes (Fig. 2A). Another way involves head-to-tail union of two identical linear genomes which have complementary cohesive ends (Fig. 2B). A third way presupposes a mode of replication for the primitive bacterial genome that is similar to the mode used by some modern bacteriophages: production of concatemers followed by cleavage at specific points to produce unit genomes. We will call the point of cleavage "R." We visualize a mutation at one of these points of cleavage, to R′, rendering the DNA immune to attack by the (perhaps restriction-like) R-specific cleavage enzyme. A dimer genome would be produced which could be capable of circularization. Further replication of this dimer genome would yield concatemers which would thenceforth be reduced by cleavage to dimers of the genome rather than to monomers (Fig. 2C).

Following two duplications of the genome, a change in the mechanisms used for replication or recombination in the primitive organism could have removed the opportunity to undergo additional genome duplications.

Primitive bacteria containing duplicated genomes might have had a survival advantage over the haploid bacteria from

TABLE 1. *Map relationships among E. coli genes with functionally related gene products*

GENE	FUNCTION	MAP POSITION (MIN)	MAP DISTANCE (MIN)
pps	PHOSPHOENOLPYRUVATE SYNTHETASE	33	45
ppc	PHOSPHOENOLPYRUVATE DECARBOXYLASE	78	
mtlA	PHOSPHOTRANSFERASE SYSTEM, ENZYME I	71	25
ptsH	PHOSPHOTRANSFERASE SYSTEM, ENZYME II	46	
cmlA	50S RIBOSOMAL PROTEIN(S)	19	45
cmlB	50S RIBOSOMAL PROTEIN(S)	21	43
eryA,nek ram,spcA strA,fus	RIBOSOME STRUCTURE AND/OR FUNCTION	64	
rts	50S RIBOSOMAL PROTEIN	78	
bgl	β-GLUCOSIDASE	74	25
lac	β-GALACTOSIDASE	9	
ebg	EVOLVED β-GALACTOSIDASE	59	22
melA	α-GALACTOSIDASE	81	
proC	Δ-PYRROLINE-5-CARBOXYLATE REDUCTASE	10	
poa	PROLINE OXIDASE	24	
metF	N^5, N^{10}-METHYLTETRAHYDROFOLATE REDUCTASE	78	21
metK	S-ADENOSYL METHIONINE SYNTHETASE	57	
purF	PHOSPHORIBOSYL-PYROPHOSPHATE AMIDASE	44	
purG	PHOSPHORIBOSYLFORMYLGLYCINEAMIDE SYNTHETASE	47	22-3
purI	AMINOIMIDAZOLE RIBOTIDE SYNTHETASE	48	
purC	PHOSPHORIBOSYLAMINOIMIDAZOLE-SUCCINOCARBOXAMIDE SYNTHETASE	47	
purB	ADENYLOSUCCINASE	25	
purE	PHOSPHORIBOSYLAMINOIMIDAZOLE CARBOXYLASE	12	23
purH	PHOSPHORIBOSYLAMINOIMIDAZOLE CARBOXAMIDE FORMYLTRANSFERASE	79	
purD	PHOSPHORIBOSYLGLYCINEAMIDE SYNTHETASE	79	
purA	ADENYLOSUCCINIC ACID SYNTHETASE	84	
pup	PURINE NUCLEOSIDE PHOSPHORYLASE	90	
trpE	ANTHRANILATE SYNTHETASE	27	23
pheA	CHORISMATE MUTASE P-PREPHENATE DEHYDRATASE	50	
tyrA	CHORISMATE MUTASE T-PREPHENATE DEHYDROGENASE	50	
aroF	DHAP SYNTHETASE ISOZYME	50	
aroG	DHAP SYNTHETASE ISOZYME	17	
aroH	DHAP SYNTHETASE ISOZYME	32	
aroD	DEHYDROQUINASE	32	
aroB	DEHYDROQUINATE SYNTHETASE	65	45
aroE	DEHYDROSHIKIMATE REDUCTASE	64	44
aroA	3-ENOLPYRUVYLSHIKIMATE-5-PHOSPHATE SYNTHETASE	20	24
aroC	CHORISMIC ACID SYNTHETASE	44	
pyrD	DIHYDROOROTIC ACID DEHYDROGENASE	21	
pyrC	DIHYDROOROTASE	24	42
pyrE	OROTODYLIC ACID DECARBOXYLASE	72	45
pyrF	OROTODYLIC ACID PYROPHOSPHORYLASE	27	
aceB	MALATE SYNTHETASE A	80	23
glc	MALATE SYNTHETASE G	57	
popC	SYNTHESIS OF δ-AMINO LEVULINIC ACID	4	22
hem	SYNTHESIS OF δ-AMINO LEVULINIC ACID	26	

GENE	FUNCTION	MAP POSITION (MIN)	MAP DISTANCE (MIN)
pyrB	ASPARTATE TRANSCARBAMYLASE	85	21
sdh	SUCCINIC ACID DEHYDROGENASE	16	(23)
aspA	ASPARTASE	(83)	
asn	ASPARAGINE SYNTHETASE	74	
lysC	ASPARTOKINASE III	80	
metL	ASPARTOKINASE II	78	
metM	HOMOSERINE DEHYDROGENASE II	78	
thrA	ASPARTOKINASE I, HOMOSERINE DEHYDROGENASE I	0	24
asd	ASPARTIC SEMIALDEHYDE DEHYDROGENASE	66	
metA	HOMOSERINE-O-TRANSSUCCINYLASE	80	
thrB	HOMOSERINE KINASE	0	
thrC	THREONINE SYNTHETASE	0	
ilvA	THREONINE DEAMINASE	75	
metB	CYSTATHIONINE SYNTHETASE	78	20
metC	CYSTATHIONASE	58	
trpB	TRYPTOPHAN SYNTHETASE B PROTEIN	27	46
tna	TRYPTOPHANASE	73	
trkA	POTASSIUM ION TRANSPORT	64	48
trkB	POTASSIUM ION TRANSPORT	64	
kdp	REQUIREMENT FOR HIGH CONCENTRATION OF K	16	
kac	DEFECT IN POTASSIUM ION UPTAKE	(16)	
trkC	POTASSIUM ION TRANSPORT	1	
trkD	POTASSIUM ION TRANSPORT	75	43
trkE	POTASSIUM ION TRANSPORT	28	
uvrA	DETERMINES A STEP IN DIMER EXCISION	81	45
uvrC	DETERMINES A STEP IN DIMER EXCISION	36	
uvrB	DETERMINES A STEP IN DIMER EXCISION	18	
gnd	6-PHOSPHOGLUCONATE DEHYDROGENASE	39	22
pgl	6-PHOSPHOGLUCONOLACTONASE	17	
edd	6-PHOSPHOGLUCONATE DEHYDRASE	36	
eda	2-KETO-3-DEOXYGLUCONATE-6-P ALDOLASE	36	
zwf	GLUCOSE-6-PHOSPHATE DEHYDROGENASE	36	21
pgm	PHOSPHOGLUCOMUTASE	15	25
pgi	PHOSPHOGLUCOISOMERASE	80	41
man	PHOSPHOMANNOSE ISOMERASE	31	47
pfk	PHOSPHOFRUCTOKINASE	78	43
glmS	GLUTAMINE-FRUCTOSE-6-PHOSPHATE TRANSAMINASE	74	
nagA	GLUCOSEAMINE-6-PHOSPHATE DEACETYLASE	15	
nagB	GLUCOSEAMINE-6-PHOSPHATE DEAMINASE	15	
fdp	FRUCTOSE DIPHOSPHATASE	85	(24)
fda	FRUCTOSE-1,6-DIPHOSPHATE ALDOLASE	(61)	
tpi	TRIOSE PHOSPHATE ISOMERASE	78	23
tkt	TRANSKETOLASE	(55)	
rbsK	RIBOKINASE	74	
glpA	L-α-GLYCEROPHOSPHATE DEHYDROGENASE (ANAEROBIC)	43	23
glpD	D-α-GLYCEROPHOSPHATE DEHYDROGENASE (AEROBIC)	66	
glpK	GLYCEROLKINASE	78	24
plsA	GLYCEROL PHOSPHATE ACYLTRANSFERASE	12	23
fadD	ACYL-COENZYME A SYNTHETASE	35	42
fadB	HYDROXYACYL-COENZYME A DEHYDROGENASE	77	
fadA	THIOLASE I	77	

(*Continued at the top of next page*)

TABLE 1. *Map relationships among E. coli genes with functionally related gene products (continued)*

GENE	FUNCTION	MAP POSITION (MIN)	MAP DISTANCE (MIN)
pdxA	TWO FUNCTIONS IN PYRIDOXINE SYNTHESIS	1	43
pdxB	THREE FUNCTIONS IN PYRIDOXINE SYNTHESIS	44	
pdxC	TWO FUNCTIONS IN PYRIDOXINE SYNTHESIS	20	24
polA	DNA POLYMERASE I	76	
polB	DNA POLYMERASE II	2	
polC	DNA POLYMERASE III	4	
rnsA	RIBONUCLEASE I	14	43
rnsB	RIBONUCLEASE II	57	
endA	ENDONUCLEASE I	57	(41)
endB	ENDONUCLEASE I	(16)	
cysH	ADENOSINE-3'-PHOSPHATE-5'-SULPHATOPHOSPHATE REDUCTASE	(53)	
cysC	ADENOSINE-5'-SULPHATOPHOSPHATE KINASE	52	
cysQ	SULPHITE REDUCTASE	(53)	
cysP	SULPHITE REDUCTASE AND PERMEASE	(53)	
cysG	SULPHITE REDUCTASE	65	
dapB	DIHYDROPICOLINIC ACID REDUCTASE	0	47
dapA	DIHYDROPICOLINIC ACID SYNTHETASE	47	
dapE	N-SUCCINYLDIAMINOPIMELIC ACID DEACYLASE	47	44
dapC,D	TETRAHYDRODIPICOLINIC ACID ⟶ N-SUCCINYLDIAMINOPIMELATE	3	
serA	3-PHOSPHOGLYCERIC ACID DEHYDROGENASE	56	
serB	3-PHOSPHOSERINE PHOSPHATASE	0	20
serC	3-PHOSPHOSERINE-2-OXOGLUTARATE AMINOTRANSFERASE	20	
tpp	THYMIDINE PHOSPHORYLASE	89	
udp	URIDINE PHOSPHORYLASE	76	
tdk	THYMIDINE KINASE	27	
thyA	THYMIDYLATE KINASE	64	

GENE	FUNCTION	MAP POSITION (MIN)	MAP DISTANCE (MIN)
gltE	GLUTAMYL TRNA SYNTHETASE	72	
gltM	GLUTAMYL TRNA SYNTHETASE	(38)	
guaB	INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE	47	42
guaA	XANTHOSINE-5'-MONOPHOSPHATE DEAMINASE	47	
guaC	GUANOSINE-5'-MONOPHOSPHATE REDUCTASE	89	
gurB	METHYL-β-D-GLUCURONIDE UTILIZATION	64	(48)
gurC	METHYL-β-D-GLUCURONIDE UTILIZATION	(16)	
argF	ORNITHINE TRANSCARBAMYLASE	7	47
argA	N-ACETYLGLUTAMATE SYNTHETASE	54	
argB	N-ACETYLGLUTAMATE-5-PHOSPHOTRANSFERASE	79	25
argC	N-ACETYLGLUTAMIC-γ-SEMIALDEHYDE DEHYDROGENASE	79	
argE	L-ORNITHINE-N-ACETYLORNITHINE LYASE	79	
argH	ARGININOSUCCINIC ARGININE LYASE	79	
argD	N-ACETYLORNITHINE-δ-TRANSAMINASE	64	21
argI	ORNITHINE TRANSCARBAMYLASE	85	24
argG	ARGININOSUCCINIC ACID SYNTHETASE	61	
ubiG	2-OCTAPRENYL-3-METHYL-5-HYDROXY-6-METHOXY-1, 4-BENZOQUINONE ⟶ UBIQUINONE	42	
ubiA	4-HYDROXYBENZOATE ⟶ 3-OCTAPRENYL-4-HYDROXYBENZOATE	81	24
ubiF	UBIQUINONE BIOSYNTHETIC ENZYME	15	
ubiB,D,E	UBIQUINONE BIOSYNTHETIC ENZYMES	76	
xthA	EXONUCLEASE III	31	23
recB,C	EXONUCLEASE V	54	
sbcB	EXONUCLEASE I	38	

which they were derived because of a gene dosage effect, or because of a reduced frequency of deleterious mutations in the essentially diploid or tetraploid bacterium, or, over a long period of time, because of the expanded genetic capacity which would provide the raw material for a variety of evolutionary changes that would ultimately lead to a greatly increased genetic capability.

Evolutionary Changes in Duplicated Genes. We picture that duplicated genes changed by mutation and divergence in function. As an example, duplicate copies of an ancestral gene concerned with tryptophan metabolism could have evolved in somewhat different ways to yield the gene for the tryptophan synthetase B protein at one locus (27 min) and the gene for tryptophanase at the other locus (73 min). The reaction catalyzed by the two enzymes are similar: the tryptophan synthetase B protein catalyzes the reversible synthesis of tryptophan from indole and serine, tryptophanase catalyzes the breakdown of tryptophan into indole, pyruvate, and ammonia and is also capable of synthesizing tryptophan from indole and serine.

CHROMOSOMAL ORGANIZATION OF FUNCTIONALLY RELATED GENES

Genes for Enzymes of Sequential Metabolic Pathways. Some functionally related genes in *E. coli* are clustered (such as the genes of the *his* operon), other functionally related genes are dispersed throughout the genome (such as the *arg* genes). Clustering of functionally related genes may have occurred either before or after the proposed genome duplications, perhaps by the process of tandem gene duplication followed by stepwise evolution of a series of related genes, as proposed by Lewis (4) and Horowitz (5). We propose that the dispersion of functionally related genes was not a random process, but rather that separations of about 90° or 180° took place as a consequence of the process of genome duplication. Complex spatial separations of related genes could have resulted from the existence of relevant genetic information at *two* loci in the primitive genome. For example, the scattered *arg* genes of *E. coli* K12 can be thought of as falling into two sets of genes. Each set is composed of genes which, within a set, are located approximately either 90° or 180° from one another. One set includes *argF*, *argA*, *argP*, and the *B-C-E-H* cluster at 7, 54, 56, and 79 min, respectively. The other set includes *argI*, *argD*, *argG*, and *argR*, at 85, 64, 61, and 63 min, respectively. Each of these sets includes a gene for ornithine transcarbamylase, *argF* in one set, *argI* in the other. These two ornithine transcarbamylase genes may have had independent origins. The *argI* ornithine transcarbamylase gene may be evolutionarily related to the immediately adjacent *pyrB* gene for aspartate transcarbamylase (6); the *argF* gene may have an entirely separate evolutionary history. We propose, then, that the

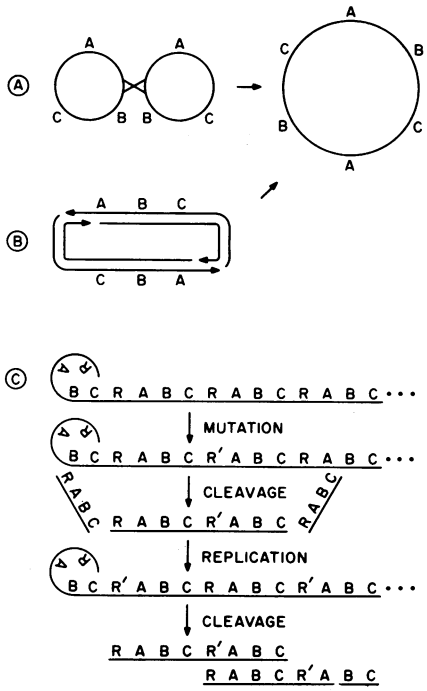

FIG. 2. Possible mechanisms for genome duplication.

multiplicity of modern *arg* genes could have arisen by two sequential duplications of a primitive genome which contained two different ancestral *arg* genes or gene clusters. Divergent evolution of the resulting duplicate genes, including perhaps tandem duplications at some loci and loss of arginine-related functions at other loci, could have resulted in the distribution pattern of the *arg* genes as they are found today.

Multiple Genes. We have considered the question of the origin of functionally related genes which determine separate enzymes that carry out the same or closely similar reactions. In the case of isozymes, some gene pairs seem to be spatially related in such a way that, in terms of the hypothesis presented here, they could be considered to have arisen from a common ancestral gene. For example, the two member genes comprising the gene pairs for the malate synthetases, the glycerol phosphate dehydrogenases, and the phosphotransferase system enzymes are located about 90° from each other.

On the other hand, the genes for other sets of isozymes or catalytically similar enzymes do not seem to be related to a common ancestral gene in terms of 90° or 180° separations on the map. For instance the genes for the aspartokinase isozymes, *thrA* (90 min), *lysC* (80 min), and *metL* (78 min), have no apparent relationship in these terms. Instead, each isozyme gene seems to be spatially related to a separate set of genes concerned with separate metabolic pathways. The *thrA* aspartokinase is metabolically related to *thrB* and *thrC* at the same locus (90 min) and the *asd* at 66 min (about 90°). The *metL* aspartokinase gene is metabolically related to *metM*, *B*, and *F* at the same locus (78 min) and to *metK* and *C* at 57–58 min (about 90°). The *lysC* aspartokinase gene (80 min) is metabolically related to *lysA* at 55 min (about 90°).

It seems likely that the multiple genes for some isozymes or closely related enzymes arose from a common ancestral gene and were dispersed on the genome by genome duplication, while other multiple genes arose independently by convergent evolution.

TABLE 2. *G Test data for distributions of gene pair separations*

A. *G Test for biochemically related gene pair separations*

	Frequency	
	In Limits	Out of Limits
Expected	42	157
Observed	68	131
	G = 17.347	

B. *G Test for biochemically nonrelated gene pair separations*

	Frequency	
	In Limits	Out of Limits
Expected	1612	5939
Observed	1500	6051
	G = 10.063	

C. *G Test for independence of the two distributions (2 × 2 Test)*

	Observed Frequency		
	In Limits	Out of Limits	Total
Related	68	131	199
Nonrelated	1500	6051	7551
Total	1568	6182	7750
	G = 21.592		

For A and B, $G = 2[\Sigma f_o \ln f_o - \Sigma f_o \ln f_e]$, where f_o and f_e are the observed and expected frequencies.

For C, $G = 2[\Sigma f_o \ln f_o - \Sigma f \ln f$ (row and column totals) $+ f \ln f$ (grand total)].

All tests assume one degree of freedom. Derivations of expected and observed frequencies are explained in the *text*.

DISCUSSION

If genome duplications did indeed take place in the past, subsequent changes have led to some movement of related genes in relation to one another, since many of the 180° and 90° relationships as found today are not precise. Nevertheless, such events must have occurred either infrequently or with nearly equal frequency in all parts of the genome in order to account for the persistence of the nonrandom distribution to the present time.

We wish to thank Dr. Robert Sokal and Dr. Richard Koehn for their help in the statistical analysis and Dr. Norman Arnheim for discussion and constructive criticisms. This work was supported by a grant from the National Institutes of Health, GM-21316.

1. Wallace, D. C. & Morowitz, H. J. (1973) *Chromosoma* **40**, 121–126.
2. Hopwood, D. A. (1967) *Bacteriol. Rev.* **31**, 373–403.
3. Taylor, A. L. & Trotter, C. D. (1972) *Bacteriol. Rev.* **36**, 504–524.
4. Lewis, E. B. (1951) *Cold Spring Harbor Symp. Quant. Biol.* **16**, 159–174.
5. Horowitz, N. H. (1965) in *Evolving Genes and Proteins*, eds. Bryson, V. & Vogel, H. (Academic Press, New York), pp. 15–23.
6. Legrain, C., Halleux, P., Stalon, V. & Glansdorff, N. (1972) *Eur. J. Biochem.* **27**, 93–102.