JOURNAL OF BACTERIOLOGY, Mar. 1972, p. 953-965
Copyright © 1972 American Society for Microbiology

Vol. 109, No. 3
Printed in U.S.A.

Polynucleotide Sequence Divergence Among Strains of *Escherichia coli* and Closely Related Organisms

DON J. BRENNER, G. R. FANNING, F. J. SKERMAN, AND STANLEY FALKOW

Division of Biochemistry, Walter Reed Army Institute of Research, Walter Reed Army Medical Center, Washington, D.C. 20012, and Department of Microbiology, Georgetown University Schools of Medicine and Dentistry, Washington, D.C. 20007

Received for publication 1 December 1971

Polynucleotide sequence similarity tests were carried out to determine the extent of divergence present in a number of *Escherichia coli* strains, obtained from diverse human, animal, and laboratory sources, and closely related strains of *Shigella*, *Salmonella*, and the Alkalescens-Dispar group. At 60 C, relative reassociation of deoxyribonucleic acid (DNA) from the various strains with *E. coli* K-12 DNA ranged from 100 to 36%, with the highest level of reassociation found for three strains derived from K-12, and the lowest levels for two "atypical" *E. coli* strains and *S. typhimurium*. The change in thermal elution midpoint, which indicates the stability of DNA duplexes, ranged from 0.1 to'14.5 C, with thermal stability closely following the reassociation data. Reassociation experiments performed at 75 C, at which temperature only the more closely related DNA species form stable duplexes, gave similar indications of relatedness. At both temperatures, Alkalescens-Dispar strains showed close relatedness to *E. coli*, supporting the idea that they should be included in the genus *Escherichia*. Reciprocal binding experiments with *E. coli* BB, 02A, and K-12 yielded different reassociation values, suggesting that the genomes of these strains are of different size. The BB genome was calculated to be 9% larger than that of K-12, and that of 02A 9% larger than that of BB. Calculation of genome size for a series of *E. coli* strains yielded values ranging from 2.29×10^9 to 2.97×10^9 daltons. *E. coli* strains and closely related organisms were compared by Adansonian analysis for their relatedness to a hypothetical median strain. *E. coli* 0128a was the most closely related to this median organism. In general, these data compared well with the data from reassociation experiments among *E. coli* strains. However, anomalous results were obtained in the cases of *Shigella flexneri*, *S. typhimurium*, and "atypical" *E. coli* strains.

Complementary polynucleotide sequences are present throughout the *Enterobacteriaceae*. Such diverse groups as *Proteus* and *Serratia* (3), *Erwinia* and *Edwardsiella* (Brenner and Fanning, *unpublished data*) contain significant stretches of nucleotides that specifically reassociate with deoxyribonucleic acid (DNA) from *Escherichia coli* K-12. DNA held in common between *E. coli* and these distantly related groups has diverged greatly as shown by the thermal lability of reassociated heteroduplex molecules. For example, at conditions optimal for specific reassociation, DNA from *P. mirabilis* (enterobacterial species that has diverged from *E. coli* to the greatest extent) shows 6 to 8% heteroduplex formation with *E. coli* DNA. The product of this reaction has a thermal'stability some 16 C less than that of a homologous *E. coli* reaction. Assuming that each degree of instability is caused by 1% of unpaired bases within a reassociated duplex (1, 12), those sequences still held iri common between *P. mirabilis* and *E. coli* have diverged 17% on the average (3).

Whereas reactions between *E. coli* K-12 and distantly related enterobacteria have been examined in some detail, there is little knowledge concerning divergence among members of the *Enterobacteriaceae* that are closely related to *E. coli* K-12. This category includes other strains of *E. coli*, strains formerly contained in the Alkalescens-Dispar group of organisms, and species contained in the genus *Shigella*.

It is a reasonable view that variations exist among strains of *E. coli*. But to what extent? The question that we asked in this study was in what sense do the populations of cells which we conveniently call *E. coli* on the basis of a relatively few phenotypic characteristics diverge in overall genetic organization and in nucleotide sequence.

MATERIALS AND METHODS

Organisms and media. The strains used in this study are listed in Table 1. Bacteria were maintained on meat extract or nutrient agar slants. Brain Heart Infusion broth was used to cultivate organisms. The medium employed for labeling cells with $^{32}PO_4$ has been described (3).

DNA preparation. Both unlabeled and labeled DNA were prepared by a modification of the method of Berns and Thomas (2) as described previously (3). DNA preparations were sheared mechanically in a pressure pump at 50,000 psi to an average single-strand fragment size of approximately 125,000 daltons and filtered through metrical filter discs (6). Labeled DNA fragments were then denatured by heating and further purified by passage through a hydroxyapatite (HA) column equilibrated with 0.14 M PB (phosphate buffer, an equimolar mixture of NaH_2PO_4 and Na_2HPO_4, *p*H 6.8) and held at 60 C.

TABLE 1. *Bacterial strains employed*

Organism	Origin	Source[a]
Escherichia coli K-12	Standard strain	U. of Washington
E. coli C600	Standard strain	WRAIR
E. coli 1485	Standard strain	WRAIR
E. coli W3110	Standard strain	WRAIR
E. coli B	Standard strain	WRAIR
E. coli B/r	Standard strain	WRAIR
E. coli BB	Standard strain	WRAIR
E. coli B (Albany)	Standard strain	WRAIR
E. coli W3442 K$^-$	Standard strain	WRAIR
E. coli 0128a, 128D:H12	Kitten	CDC
E. coli 418	Fish	England
E. coli 121	Goat	California
E. coli 190	Human	California
E. coli 01:K1:H7	Standard strain	CDC
E. coli 888	Horse	Pennsylvania
E. coli 674	Bovine	California
E. coli 020:H11	Quail	CDC
E. coli 0111:B4	Human	Maryland
E. coli 0125	Human	Georgetown U.
E. coli 3122-66	Mouse	CDC
E. coli 3541-67	Human	CDC
E. coli 3360-66	Human	CDC
E. coli 128	Human	Japan
E. coli 0102:H21	Rabbit	CDC
E. coli 075:H5	Human	CDC
E. coli 06:K2:H1:Bi	Standard strain	CDC
E. coli 786	Pig	Georgetown U.
E. coli 01A	Pig	CDC
E. coli 02A	Pig	CDC
E. coli 04	Pig	CDC
E. coli 07	Pig	CDC
E. coli 09	Pig	CDC
E. coli 025	Pig	CDC
E. coli E148B (serotype 08)	Bovine	AUS
E. coli 233 (serotype 08)	Bovine	AUS
E. coli 4A (serotype 08)	Human	AUS
E. coli E215B (serotype 08)	Bovine	AUS
E. coli 69C (serotype 08)	Bovine	AUS
E. coli 25A (serotype 08)	Human	AUS
E. coli 41A (serotype 08)	Human	AUS
E. coli 243B (serotype 08)	Bovine	AUS
E. coli 2B-(serotype 08)	Human	AUS
E. coli 179C (serotype 08)	Bovine	AUS
E. coli E156A (serotype 088)	Bovine	AUS
E. coli 52A (serotype 088)	Bovine	AUS
E. coli E242 (serotype 088)	Bovine	AUS
E. coli E114 (serotype 088)	Bovine	AUS
E. coli E68 (serotype 088)	Equine	AUS
E. aurescens ATCC 12814	Pig	ATCC
E. freundii	Pig	WRAIR
Alkalescens-Dispar 01 111-69	Pig	CDC
A-D 02 7040-59	Pig	CDC
A-D 03 3203-59	Pig	CDC
A-D 03 (ceylonensis)	Pig	WRAIR
A-D 04 774-56	Pig	CDC
A-D 05 3121-63	Pig	CDC
A-D 06 4878-54	Pig	CDC
A-D 07 1868-62	Pig	CDC
A-D 08 1143-51	Pig	CDC
Shigella flexneri 2a 24570	Standard strain	WRAIR
Sh. sonnei	Standard strain	WRAIR
Sh. sonnei (virulent)	Human	WRAIR
Sh. sonnei (avirulent)	Human	WRAIR
Sh. flexneri (Newcastle)	Human	WRAIR
Sh. boydii 1	Human	WRAIR
Sh. boydii 7 (etousa)	Human	WRAIR
Sh. dysenteriae 1	Human	WRAIR
Sh. dysenteriae 2	Human	WRAIR
Sh. dysenteriae 3	Human	WRAIR
Salmonella typhimurium LT2	Standard strain	NIH

[a] Abbreviations: WRAIR, Walter Reed Army Institute of Research, Washington, D.C.; CDC, Center for Disease Control, Atlanta, Ga.; NIH, National Institutes of Health, Bethesda, Md.; ATCC, American Type Culture Collection, Bethesda, Md.; AUS, University of Queensland, Brisbane, Australia.

The material that bound to HA was discarded since under these conditions single-stranded DNA does not bind to HA. This procedure decreased the "zero-time" binding (label bound to HA immediately after the DNA is denatured from about 2.5% to less than 1%).

DNA reassociation. The conditions employed for DNA reassociation and separation of single-stranded from reannealed DNA on HA have been presented in detail elsewhere (3). These techniques are described below in general terms.

^{32}P-labeled DNA and a series of unlabeled DNA species in 0.14 M PB plus 0.05 M ethylenediaminetetraacetic acid (EDTA) were denatured by heating at 100 C for 3 to 4 min and then immediately cooled in an ice bath. Labeled DNA (0.1 μg/ml) was added to 400 μg of each unlabeled DNA per ml, and the mixtures were incubated at 60 or 75 C for 16 hr. These conditions permit essentially complete reassociation of the unlabeled DNA fragments with each other or with labeled fragments but preclude significant reassociation of labeled fragments with one another. Reaction mixtures were either immediately added to HA or frozen until use.

Separation of reassociated DNA from single-stranded DNA was accomplished by adding samples to HA by a batch method (4). The HA was equilibrated with 0.14 M PB plus 0.005 M EDTA and 0.4% sodium dodecyl sulfate (SDS) and held at the temperature at which the samples had been incubated (60 or 75 C). EDTA was added to PB to chelate any MG^{2+} and thereby preclude deoxyribonuclease activity. SDS was used to prevent nonspecific binding of DNA to HA. The HA was washed six times with 15-ml portions of 0.14 M PB plus EDTA and 0.4% SDS to remove DNA fragments that had not reannealed and therefore had not adsorbed to HA. In 0.14 M PB, double-stranded DNA binds to HA, while single-stranded DNA is not bound.

The DNA adsorbed to HA was then washed with 15-ml portions of 0.14 M PB at temperatures increasing in increments of 2.5 C to 100 C. As the temperature exceeded the dissociation temperature of DNA duplexes bound to HA, the resultant single-stranded DNA was eluted, and a thermal elution profile was obtained. The HA was finally washed twice with 15-ml portions of 0.4 M PB to elute any material that remained bound to HA. Neither double- nor single-stranded DNA remained bound to HA in 0.4 M PB. When thermal stability of reassociated DNA duplexes was not of interest, the amount of heteroduplex formation was determined by eluting bound DNA with 0.14 M PB at 95 and 100 C, followed by two washes with 0.4 M PB. All eluates are collected in counting vials and assayed by Cerenkov counting (7).

Spectrophotometric determination of genome size. The spectrophotometric estimation of the molecular complexity of bacterial DNA (genome size) was performed essentially as described by Gillis et al. (10). Sheared DNA in 0.1 × SSC (SSC = 0.15 M sodium chloride plus 0.015 M sodium citrate) was denatured by heating in a boiling water bath for 5 min. The DNA was immediately diluted into a quantity of 10× SSC to yield a final concentration of 2× SSC at an absorbancy of 2.000 ± 0.010. The DNA was then transferred to 10-mm cuvettes in a 75 C chamber of a Gilford 2000 spectrophotometer. A solution of 0.1 M EDTA treated identically to the DNA samples was included in most runs as a control to monitor the time required for the DNA solutions to achieve thermal equilibrium. The change in absorbancy of EDTA at 260 nm as a function of temperature is a useful parameter to include in all DNA reassociation and thermal denaturation studies. In the main, the total elapsed time between the denaturation, mixing, and subsequent transfer to the cuvettes was less than 2 min. Temperature equilibration to 75 C within the cuvettes was generally achieved within 4 additional min. The absorbancy of each sample was recorded automatically at 20-sec intervals at 260 nm on a Minneapolis Honeywell recorder for a total of 40 to 60 min. The decrease in absorbancy from minute 20 to minute 40 after the initiation of renaturation was employed in the calculation of the genome size as described below. Each determination was carried out at least six times.

The thermal denaturation of DNA for the determination of the guanine plus cytosine (G + C) content was determined on native DNA as described previously (3).

Procedures used in Adansonian analysis. A total of 191 characters were studied and coded for computer analysis. These studies were undertaken in collaboration with R. R. Colwell who performed the calculations of S value (see below) and major cluster analysis. The hypothetical median organism of the *E. coli* cluster was calculated by the method described by Liston et al. (13). The strain, *E. coli* 0128a, which had the highest computed relationship with the hypothetical median strain, was employed as the source of ^{32}P-labeled nucleic acid in one series of DNA binding studies.

RESULTS

DNA-agar experiments using DNA species from a few strains of *E. coli* suggested that these strains were too closely related to be distinguished on the basis of comparative binding (15; Brenner, *unpublished data*). These results were accepted, as it was generally assumed that various common *E. coli* strains differed largely by single base changes. In occasional "control experiments," both we and D. E. Kohne found that the reaction between DNA from *E. coli* strains K-12 and BB was less extensive than either the homologous K-12 or BB reaction. These observations prompted us to determine the nucleotide sequence relatedness between a series of common laboratory and clinical strains of *E. coli*, as well as strains from uncommon and widely distributed sources. Also included in this study are strains formerly contained in the Alkalescens-Dispar group, and strains of several *Shigella* species that are closely related to *E. coli* on the basis of both biochemical and preliminary reassociation data.

Relatedness of E. coli strains, with E. coli K-12 used as the reference strain. Results obtained from reacting ^{32}P-labeled *E. coli* K-12 DNA fragments with DNA fragments from closely related organisms are summarized in Table 2. In these reactions, 0.1-μg portions of denatured, labeled fragments (5×10^4 to 3×10^5 counts per min per μg) from K-12 were added to 400 μg of unlabeled, denatured DNA fragments per ml from the homologous or a heterologous source. The DNA species were incubated on 0.14 M PB plus EDTA for 21 hr at either 60 or 75 C and then passed through HA to separate the reassociated DNA duplexes from unreacted DNA. This incubation criterion insures essentially complete reaction of unlabeled DNA fragments with each other and with labeled fragments of complementary nucleotide sequence. The concentration of labeled DNA fragments is sufficiently small to preclude reaction of labeled fragments with one another. In practice, we obtained 85 to 95% reassociation in homologous reactions and usually less than 1.5% reaction in labeled preparations incubated in the absence of unlabeled DNA. Table 2 lists the relative binding of *E. coli* DNA to DNA from other strains at 60 C, which is the optimal temperature for DNA reassociation in these organisms, and at 75 C, at which temperature only closely related sequences can form stable reassociation products. The data tell how much of the *E. coli* K-12 genome is capable of forming duplexes with DNA from other strains, and therefore, the extent of nucleotide sequence relatedness that exists between two given strains at this point in evolution. Thermal elution mid-points from these reactions, relative to that from a homologous *E. coli* reaction ($\Delta T_{m(e)}$), serve as an index to divergence in related nucleotide sequences. It is assumed that each decrease in $T_{m(e)}$ of 1.0 C is the result of 1.0% unpaired bases within a reassociated DNA duplex (12). The thermal binding index (TBI) is also given in Table 2. TBI is the ratio of binding at 75 C to that at 60 C. In closely related organisms the TBI approaches 1.0 (0.85 to 1.0), whereas in organisms that are only moderately or distantly related, the TBI is below 0.4.

E. coli strains C600, 1485 T$^-$, and W3110 are derivatives of strain K-12, and they exhibit virtually complete reactions with K-12 DNA at the optimal 60 C incubation temperature. The close relationship among these strains is also evident from the minimal decrease in duplex stability ($T_{m(e)}$ values are only 0.1 to 0.3 C less than that of the homologous K-12 reaction). DNA from four *E. coli* B strains behave identically, exhibiting 89 to 94% reassociation with K-12 DNA at 60 C and 88 to 90% reassociation with K-12 DNA at 75 C. TBI values for these strains are 0.94 or higher, and the reaction products are only slightly less stable than the homologous reaction product. Another laboratory strain, W3442K$^-$, and a strain of *E. aurescens* also showed reactions with K-12 similar to those exhibited by the B strains.

The remainder of *E. coli* reactions were carried out with some two dozen strains of diverse origin. DNA species from certain of these strains, such as 0102, 128, 128a, 418, and 888, react virtually completely with K-12 DNA at both incubation temperatures. DNA from the majority of strains exhibits 90% or higher reassociation with *E. coli* K-12 DNA at 60 C. These duplexes are predominantly stable at 75 C, as indicated by the high TBI values. Several strains (04, 06, 190, 3360–66) show less than 90% binding to K-12 DNA at 60 C, and a decrease of 3 C or more in stability occurs in the reaction products formed between DNA from several strains and K-12 (01A, 04, 07, 025). At the stricter 75 C incubation temperature, up to 22% of the DNA from these strains cannot form stable duplexes with DNA from K-12. It is also evident that the thermal stability of these reactions increases at the higher incubation temperature.

DNA from three cultures received as "atypical" *E. coli* strains was also tested. One of these, 3360-66, gave 88% reaction with K-12 at 60 C. The other two strains, 3122-66 and 3451-67, had reactions of 45% and 36%, respectively, with K-12. The $\Delta T_{m(e)}$ of these reaction products indicated an average nucleotide sequence divergence of approximately 14%. Only 13% relatedness was evident between these two strains and K-12 at 75 C. Based on the extent of reaction with *E. coli* and biochemical data, these strains are clearly not *E. coli*, but may be *Citrobacter* or *Enterobacter* strains. Interspecific DNA reassociation reactions with members from these genera would identify the group to which these two strains belong.

Extent of divergence in the species E. coli. In order to assess the maximum divergence occurring within a species, a series of strains from diverse animal, laboratory, and human sources was studied. Reassociation reactions were carried out at 75 C to restrict duplex formation to only highly complementary nucleotide sequences. The binding of DNA from *E. coli* strain 0128a to DNA from other strains ranged from 100 to 72% as shown in Table 3. Thus, 25% or more of DNA in many *E. coli* strains has diverged to a point

TABLE 2. *Relative reassociation of DNA from E. coli K-12 with DNA from closely related organisms*

Source of unlabeled DNA	Per cent relative binding at 60 C	$\Delta T_{m(e)}$[a] at 60 C	Per cent relative binding at 75 C	$\Delta T_{m(e)}$ at 75 C	TBI[b]
E. coli K-12	100		100		
E. coli C600	98	0.1	96	0.3	0.98
E. coli 1485T⁻	99	0.3	100	0.3	1.0
E. coli W3110 (21)	97	0.1	96	0.1	0.99
E. coli B	94	0.8	88	1.0	0.94
E. coli BB	92	0.8	89	0.6	0.97
E. coli B/r	94	0.8	90	1.1	0.96
E. coli B (Albany)	89	0.7	89	0.4	1.0
E. coli W3442K⁻	94	0.8	89	1.0	0.95
E. aurescens	93	0.3	93	0.0	1.0
E. coli 01A	94	3.2	80	2.8	0.85
E. coli 02A	91	1.9	80	1.7	0.88
E. coli 04	85	3.1	78	2.0	0.92
E. coli 06	87	2.1	88	1.7	1.0
E. coli 07	91	3.6	82	2.6	0.90
E. coli 09	93	1.8	90	1.7	0.97
E. coli 025	92	3.0	88	2.2	0.95
E. coli 075	91	2.2	89	1.7	0.98
E. coli 0102			100	0.0	
E. coli 0111B4	94	1.9	92	0.8	0.98
E. coli 0125B15	93	1.2	92	0.8	0.99
E. coli 0128a	97	1.0	94	0.5	0.97
E. coli 121			90	0.7	
E. coli 128	98	0.4	98	0.4	1.0
E. coli 190	87	2.5	82	2.0	0.94
E. coli 418	96	0.8	98	0.6	1.0
E. coli 674	96	1.5	93	1.2	0.97
E. coli 786	89	2.9	83	2.2	0.93
E. coli 888	99	0.5	98	0.1	0.99
E. coli 3360-66	88	2.0	88	0.9	1.0
E. coli 3122-66	45	13.9	13	6.4	0.29
E. coli 3541-67	36	14.5	13	5.5	0.36
S. typhimurium LT2	45	12.3	11	4.5	0.24
Alkalescens-Dispar 01	89	0.9	82	1.2	0.92
A-D 02	91		86		0.95
A-D 03	88	2.2	78	1.7	0.89
A-D 03 (ceylonensis)	90	1.8	88	1.2	0.98
A-D 04	93	0.5	94	0.6	1.0
A-D 05	95	1.5	89	1.3	0.94
A-D 06	88		84		0.95
A-D 07	90		89		0.99
A-D 08	94		92	0.9	0.98
Sh. boydii 1	80		71		0.88
Sh. boydii 7 (etousa)	89	1.3	85	1.0	0.96
Sh. dysenteriae I	82	1.3	78	0.9	0.95
Sh. dysenteriae II	89	1.7	85	0.5	0.98
Sh. dysenteriae III	80		76		0.95
Sh. flexneri	84	1.0	79	0.8	0.94
Sh. flexneri (Newcastle)	85	1.0	83	1.0	0.98
Sh. sonnei	87	0.7	85	0.8	0.98
Sh. sonnei (avirulent)	84	0.9	83	0.7	0.99
Sh. sonnei (virulent)	87	0.5	79	0.7	0.91

[a] $T_{m(e)}$, Thermal elution midpoint; that temperature at which 50% of the DNA bound to HA (at the 60 or 75 C incubation temperature) is eluted. $\Delta T_{m(e)}$ is the decrease in $T_{m(e)}$ between heterologous reactions and the homologous K-12 reaction. The $T_{m(e)}$ for *E. coli* DNA in our system is between 90 and 91 C.

[b] TBI, Thermal binding index; relative binding at 75 C divided by relative binding at 60 C.

TABLE 3. *Relative reassociation at 75 C for DNA from E. coli strains*[a]

Strain	Source	Per cent relative binding at 75 C	$\Delta T_{m(e)}$[b]
E. coli 0128a	Animal	100	
E. coli 418	Fish	100	0.0
E. coli 0102	Animal	95	0.1
E. coli 128	Animal	93	0.1
E. coli K-12	Lab strain	91	0.1
E. coli 075	Human	90	0.1
E. coli 121	Human	89	0.1
E. coli 888	Animal	88	1.0
E. coli 020	Human	88	0.8
E. coli 06	Lab strain	86	0.5
E. coli 0125	Human	85	0.0
E. coli 0111B4	Human	83	0.2
E. coli 674	Animal	83	0.7
E. coli 190	Human	80	2.2
E. coli 786	Animal	76	1.2
E. coli 3360-66	Human	75	0.7
E. coli 01	Lab strain	72	1.0
Sh. flexneri	Lab strain	80	0.8

[a] *E. coli* strain 0128a was the source of labeled reference DNA in all reactions.

[b] See Table 2 for a definition of $\Delta T_{m(e)}$.

where it no longer reassociates at stringent criteria.

There is no apparent correlation between the source of isolation of these strains and their relatedness. Strain 0128a, which is of animal origin, reacts completely with a strain isolated from a fish, 76 to 95% with other strains of animal origin, 72 to 91% with laboratory strains, and 75 to 90% with strains isolated from humans. The high thermal stability of these reaction products is expected in heteroduplexes formed between closely related strains, especially at stringent criteria for reassociation ($\Delta T_{m(e)}$ values from 0.0 to 2.2 C).

Polynucleotide sequence relatedness among *E. coli* strains may be summarized as follows. (i) At optimal reassociation criteria, less than one-fifth of DNA can no longer react with DNA from K-12. (ii) Reaction products from 60 C incubations contain as much as 4% unpaired bases (1.0% unpaired bases per 1.0 C drop in thermal stability). (iii) At incubation conditions designed to preclude formation of all but highly complementary polynucleotide sequences (75 C), as much as one-third of the DNA in many *E. coli* strains is unable to form duplexes with DNA from strains K-12 and 0128a. Those duplexes that do form still contain up to 3% unpaired bases. (iv) The TBI (last column, Table 2) is useful in detecting the presence or absence of highly related genetic material in heterologous reassociation reactions. A low TBI indicates that most of the duplexes formed at optimal conditions are not stable, and therefore not highly complementary at stringent reassociation conditions. The high TBI values obtained in reactions between strains of *E. coli* indicate that most of the polynucleotide sequences contain only minimal amounts of unpaired bases. (v) No strict correlation exists between source of isolation and relatedness among strains.

E. coli reactions with Shigella and Alkalescens-Dispar strains. Data obtained from K-12–Alkalescens-Dispar group and K-12–*Shigella* sp. reassociation reactions are presented in Table 2. At optimal reassociation conditions, DNA from members of the Alkalescens-Dispar group is 88 to 95% related to *E. coli* K-12 DNA. The relative stability of these reaction products indicates divergence of 0.5 to 2.5% in the *E. coli*-Alkalescens-Dispar heteroduplexes. With the possible exception of AD-03, there is little decrease in binding at more stringent incubation conditions. On the basis of these data, the Alkalescens-Dispar strains are as closely related to *E. coli* K-12 as are most *E. coli* strains. This group seems to be indistinguishable from *E. coli* strains and our data support the suggestion of Ewing (9) that, on the basis of biochemical and serological studies, the Alkalescens-Dispar strains be considered as strains of *E. coli.*

The shigellae tested form a close group with respect to relatedness to *E. coli* K-12. Under optimal conditions, 80 to 89% interspecies duplex formation is obtained with $\Delta T_{m(e)}$ values indicating 0.5 to 2.0% divergence in the duplexes. Seventy-one to 85% of *E. coli* DNA forms stable heteroduplexes with DNA from *Shigella* strains at stringent conditions. The range of binding percentages obtained is slightly lower than the average obtained in reactions involving *E. coli* strains of members of the Alkalescens-Dispar group. The values obtained with shigellae are, however, within the limits of divergence evident between *E. coli* strains tested against *E. coli* strains K-12 and 0128a. These data support the bulk of taxonomic observations that indicate close relationship between shigellae and *E. coli.*

Frequency distribution of relatedness of E. coli, Shigella, and Alkalescens-Dispar strains. The present study assayed duplex formation of DNA from *E. coli* K-12 with DNA species from 47 *E. coli*, *Shigella*, and Alkalescens-Dispar strains. In Fig. 1 relative relatedness data from 60 C reactions are plotted as a frequency distribution. The mode for *E. coli* K-12 reaction with *E. coli* strains is

90 to 94%, and the range of relatedness values around this mode approximates a Gaussian distribution. Relative reactions with all Alkalescens-Dispar strains fall well within the distribution pattern seen with *E. coli* strains, again emphasizing the validity of including these strains within the species *E. coli*. Relatedness with *Shigella* strains is distributed mainly within the range obtained for *E. coli* strains, although the mode is significantly shifted towards lower relatedness. If the frequency distribution pattern for *E. coli* strains is equated with the amount of divergence tolerable within a species, then, we suppose, one can argue that the shigellae be included as one or more species within the genus *Escherichia*.

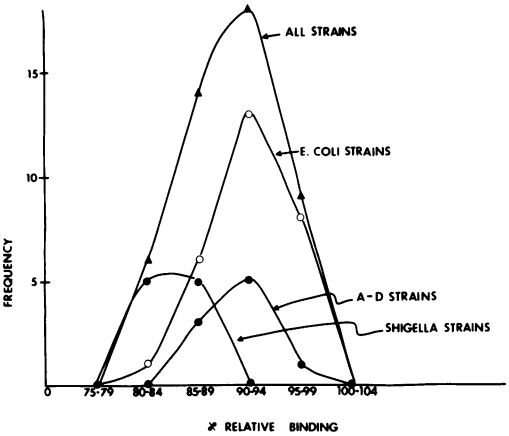

FIG. 1. *Frequency distribution of relatedness between E. coli, Shigella and Alkalescens-Dispar (A-D) strains. Values shown were obtained from 60 C reactions in which E. coli K-12 was the reference strain.*

Relatedness values obtained by using E. coli BB and 02A as reference strains. A single strain of *E. coli*, K-12 or B, has been utilized as the reference organism in almost all relationship studies of the *Enterobacteriaceae* (3, 5, 15). To depart from complete dependence upon one strain as the sole parameter of relatedness among closely related enteric bacteria, reassociation studies were carried out using ^{32}P-labeled *E. coli* BB and 02A as reference strains. In general, heteroduplex formation with BB DNA (Table 4) was similar to that seen with K-12 DNA. *E. coli*, Alkalescens-Dispar, and *Shigella* strains form extensive and stable reassociation products with BB at 60 C. The relative binding percentages at 75 C are only slightly less than those obtained at 60 C, and the thermal stability of heteroduplexes formed at 75 C is only slightly higher than that of the heteroduplexes formed at 60 C. A closer look at these data indicates that, while the $\Delta T_{m(e)}$ values for duplexes formed between BB or K-12 and any given strain are comparable, the amount of duplex formation is lower with BB DNA than with K-12 DNA as reference. Unexpectedly, *E. coli* strains B/r and B Albany showed only 82 and 83% reaction with BB.

An examination of the data obtained with 02A used as reference DNA (Table 5) reveals the same pattern of stability evident in BB and K-12 reactions. In this case heteroduplex formation with any given strain is significantly less than that obtained with K-12 or BB DNA.

For purposes of comparison, the relative binding percentages and $\Delta T_{m(e)}$ values from 60 C reactions found in Tables 2, 4, and 5 are summarized in Table 6. $\Delta T_{m(e)}$ values from reactions of *E. coli*, Alkalescens-Dispar, and *Shigella* strains with *E. coli* strains K-12, BB, and 02A are not markedly different, although interspecies duplexes formed with 02A may contain a slightly higher amount of unpaired bases. A marked disparity is evident in relative binding of each group of strains to K-12, compared to BB and 02A. *E. coli* strains average 9% less binding to BB and 15% less binding to 02A than to K-12. Similar decreases are seen in the binding of Alkalescens-Dispar and *Shigella* strains to BB and 02A, as compared to K-12.

Reciprocal reactions between K-12 and BB, K-12 and 02A, and BB and 02A are identical in thermal stability but clearly show nonreciprocal binding. *E. coli* BB shows 8% less binding to K-12 than the reciprocal reaction, 02A binds to K-12 17% less than its reciprocal reaction, and 02A binds 6% less to BB than in the reciprocal case.

Conceivably, the apparent nonreciprocal binding between DNA species from K-12, BB, and 02A reflects either inherent technical differences in the purity of one or more labeled or unlabeled DNA preparations or genuine differences in the genome sizes of these organisms. Additional reciprocal binding experiments, using two or three different labeled and unlabeled DNA preparations from K-12, BB, and 02A, again exhibited the same extent of nonreciprocity as seen in the initial reactions. Furthermore, DNA species from all three strains (Table 7) form stable homologous duplexes to approximately the same extent. It is, therefore, quite unlikely that the differences in reciprocal reassociation are explainable on the basis of differences in the DNA preparations or differences in the extent to which these DNA species form stable reassociation products.

Reciprocal binding percentages are only obtained from reactions between organisms

TABLE 4. *Reactions with DNA from E. coli strain BB*

Source of unlabeled DNA	Per cent relative binding at 60 C	$\Delta T_{m(e)}$[a] at 60 C	Per cent relative binding at 75 C	$\Delta T_{m(e)}$ at 75 C	TBI[a]
E. coli K-12	84	0.5	82	0.2	0.98
E. coli C600	84	0.4	79	0.5	0.94
E. coli B/r	83	0.7	78	0.6	0.94
E. coli B (Albany)	82	0.7	78	0.6	0.95
E. coli W3442K⁻	91		87	0.4	0.96
E. coli 01A	84	3.5	71	2.7	0.85
E. coli 02A	84	2.1	76	1.7	0.90
E. coli 04	77		69	1.7	0.90
E. coli 07	80		70	2.6	0.88
E. coli 09	84		81	1.0	0.97
E. coli 025	87	2.4	84	2.1	0.97
AD 01	81		76	1.7	0.94
AD 02	81		79	0.7	0.98
AD 03	76	1.7	69	2.0	0.91
AD 03 (ceylonensis)	82	1.3	78	1.1	0.95
AD 04	79	0.1	84	0.2	1.0
AD 05	83		76	0.9	0.92
AD 06	83		74	0.6	0.89
AD 07	83	0.6	79	0.8	0.96
AD 08	80		77	0.6	0.96
Sh. boydii (etousa)	81	1.1	79	1.3	0.98
Sh. dysenteriae 2	78		74	1.2	0.95
Sh. dysenteriae 3	76		70	0.9	0.92
Sh. flexneri	76		70	1.0	0.92

[a] See Table 2 for a definition of $\Delta T_{m(e)}$ and TBI.

TABLE 5. *Reactions with DNA from E. coli strain 02A*

Source of unlabeled DNA	Per cent relative binding at 60 C	$\Delta T_{m(e)}$[a] at 60 C	Per cent relative binding at 75 C	$\Delta T_{m(e)}$ at 75 C	TBI[a]
E. coli K-12	74	1.8	71	1.4	0.96
E. coli C600	77		68		0.88
E. coli BB	78	2.1	70	1.3	0.90
E. coli B/r	75		68	2.0	0.91
E. coli B (Albany)	75		70	1.5	0.93
E. coli W3442K⁻	81		72	1.7	0.89
E. coli 01A	83	3.2	63	2.8	0.96
E. coli 04	70	2.0	64	1.4	0.91
E. coli 07	87	1.9	72	1.8	0.83
E. coli 09	82		73	1.4	0.89
E. coli 025	80		70		0.88
AD 01	84	1.6	77	0.9	0.92
AD 02	79		72		0.91
AD 03	81		75	1.3	0.93
AD 03 (ceylonensis)	82	1.5	79	0.5	0.96
AD 04	80		81		1.0
AD 05	84		78	1.6	0.93
AD 06	70	1.9	63	1.0	0.90
AD 07	81		75	1.7	0.93
Sh. boydii (etousa)	70	1.8	65	1.5	0.93
Sh. dysenteriae 2	77		70	1.8	0.91
Sh. dysenteriae 3	73		64	1.7	0.88
Sh. flexneri	78		66	2.3	0.85

[a] See Table 2 for a definition of $\Delta T_{m(e)}$ and TBI.

TABLE 6. *Summary of reassociation data*

Source of unlabeled DNA	Per cent relative binding at 60 C			$\Delta T_{m(e)}$[a] at 60 C		
	K-12	BB	02A	K-12	BB	02A
E. coli strains	93	84	78	1.5	1.5	2.2
Alkalescens-Dispar strains	91	81	80	1.4	0.9	1.7
Shigella strains	85	78	75	1.1	1.1	1.8
E. coli K-12	100	92	91		0.8	1.9
E. coli BB	84	100	84	0.5		2.1
E. coli 02A	74	78	100	1.8	2.1	

[a] See Table 2 for a definition of $\Delta T_{m(e)}$.

TABLE 7. *Reassociation of DNA from E. coli strains K-12, BB, and 02A*[a]

Strain	Per cent reassociation at 60 C	$T_{m(e)}$[b] at 60 C	Per cent reassociation at 75 C	$T_{m(e)}$ at 75 C
K-12	90	89.8	83	89.5
BB	91	89.7	83	90.6
02A	87	90.0	81	89.6

[a] Values given are an average of at least 10 reactions with each strain and were obtained by using at least two different labeled and unlabeled DNA preparations.

[b] See Table 2 for a definition of $T_{m(e)}$.

whose genome sizes are essentially equal. The nonreciprocal binding observed in K-12, BB, and 02A reactions implies that these organisms have different genome sizes with BB some 9% larger than K-12, and 02A some 9% larger than BB and 19% larger than K-12.

Differences in genome size. Comparatively small differences in genome size are difficult to measure chemically or radiologically. The only practical and sensitive techniques were those of reciprocal reassociation, especially reciprocal competition experiments as employed by Hoyer and McCullough (11), which are sensitive enough to determine differences in genome size at the 1 to 2% level. Gillis, et al. (10) recently measured comparative genome sizes in a variety of organisms. Their method consisted of spectrophotometrically measuring differences in the initial rates of DNA reassociation. Since it is well known that DNA reassociation mimics a collision-dependent, second-order reaction (6), differences in initial rates of reassociation should reflect corresponding differences in genome size. This method is especially useful because it is rapid and simple to carry out in the laboratory. DeLey and his colleagues (10) demonstrated that, in organisms of known genome size, the initial rate of reassociation multiplied by the genome size $\times 10^{-7}$ bears a straight-line relationship to the G + C content of the DNA. In our hands the least-square fit for the data corresponds to:

$$k'(M \times 10^{-7}) = 97.31 - 0.91(G + C) \quad (1)$$

where k' = the decrease in optical density at 260 nm per minute per millimole of DNA $\times 10^2$, and M = molecular weight of DNA. Since the degree of hyperchromicity is dependent on G + C content and affects k', a second equation including a correction of hyperchromicity was derived. The least-square fit for the data corrected for hyperchromicity yields:

$$[k'(M \times 10^{-7})]/H = 2.669 - 0.0197(G + C) \quad (2)$$

where H = hyperchromicity = 41.1 − 0.21(G + C) (from reference 10).

The second equation was used in the determination of genome sizes reported here, although since the *E. coli* strains are reasonably similar in G + C content, the values of genome size are essentially identical using either equation. For example, *E. coli* K-12 DNA is $2.58 \times 10^9 \pm 0.11$ daltons using equation 1 and $2.56 \times 10^9 \pm 0.11$ daltons when equation 2 is employed.

DeLey's group found that a strain of *E. coli* B contained an 8% larger genome than that of a K-12 strain. Using their method, the genome size of the BB strain used in our study is 2.78×10^9 daltons, 9% larger than that of K-12. Strain 02A has a genome size of 2.97×10^9 daltons, 16% larger than that of K-12. These optical reassociation rate data are in excellent agreement with the HA reassociation data which indicate that the BB genome is 9% larger and the 02A genome is 18 to 19% larger than that of K-12.

Molecular weight determinations were carried out on a number of *E. coli* strains representing serotypes 08 and 088 isolated from human and animal sources in Australia (see Table 1). The values obtained (Table 8) indicate an average molecular weight of 2.43×10^9 daltons for the 08 strains and 2.58×10^9 dal-

TABLE 8. *Calculated genome sizes of E. coli strains*

Organism	Per cent G + C[a]	K[b]	Calculated genome size[c] × 10^{-9} daltons
E. coli E148B (08)	50.7	20.92 ± 1.4	2.43
E. coli 233 (08)	51.6	20.73 ± 0.6	2.42
E. coli 4A (08)	52.1	20.65 ± 0.5	2.40
E. coli E215B (08)	49.6	21.51 ± 1.1	2.41
E. coli 64C (08)	49.9	21.84 ± 1.1	2.36
E. coli 25A (08)	49.7	22.54 ± 1.0	2.30
E. coli 41A (08)	50.9	22.13 ± 1.3	2.29
E. coli 243B (08)	51.6	20.25 ± 1.0	2.47
E. coli 2B (08)	51.3	10.95 ± 0.5	2.52
E. coli 179C (08)	50.0	19.26 ± 0.6	2.67
Average	50.7		2.43
E. coli E156 (088)	49.7	21.88 ± 1.0	2.43
E. coli 52A (088)	49.6	20.32 ± 0.1	2.56
E. coli E242 (088)	48.5	20.33 ± 1.0	2.60
E. coli E114 (088)	50.3	19.22 ± 0.5	2.66
E. coli E683 (088)	50.2	18.85 ± 1.5	2.72
Average	49.7		2.58
E. coli K-12	50.5	19.9 ± 0.9	2.56
E. coli BB	50.3	18.4 ± 0.7	2.78
E. coli 02A	50.5	17.2 ± 0.6	2.97
Average	50.4		

[a] Tm calculated from thermal transition data obtained spectrophotometrically (14).

[b] k' = optical density at 260 nm per minute per millimole of DNA at 75 C.

[c] $[k' (M \times 10^{-7})]/H = 2.069 - 0.0197 (G + C)$; H = hyperchromicity = $41.1 - 0.21 (G + C)$. This equation was obtained by thermally denaturing DNA species with known G + C values in a Gilford spectrophotometer. The thermal denaturation was carried out with 100 μg of DNA per ml contained in 0.1 M SSC.

tons for the 088 strains. The strains thus far studied (including K-12, BB, and 02A) have genome sizes between 2.29 × 10^9 and 2.97 × 10^9 daltons, and G + C contents of 48.5 to 52.1%. It is apparent, even from this small sampling, that *E. coli* strains vary at least 23% in genome size and 4% in G + C content.

Frequency distribution of genome sizes in 08 and in all strains thus far tested (ten 08 strains, five 088 strains, K-12, BB, and 02A) are shown in Fig. 2. Genome sizes in the 08 strains approximate a normal distribution. The 08 strains, on the average, contain significantly less DNA than either the 088 strains or K-12, BB, and 02A. When enough additional strains are examined one might expect one normal frequency distribution between 2 × 10^9 and 3 × 10^9 daltons. Furthermore, it is apparent that representative *E. coli* strains of the same serotype, possessing (insofar as tested) identical phenotypic characteristics and isolated from the same geographical area, may, nonetheless, display significant heterogeneity in genetic organization. On the basis of DNA relatedness, G + C content, and genome size data, it may be possible to include physical parameters in the definition of a species.

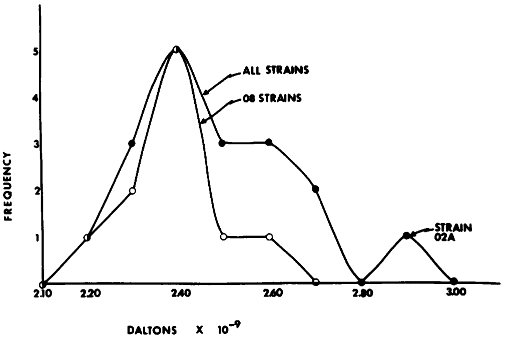

FIG. 2. *Frequency distribution of genome size in selected E. coli strains.*

Adansonian analysis of strain similarity. Sokal and Sneath (17) defined the principles of Adansonian classification as the inclusion of all information about the strains to be classified, equal weighting of each character, and establishment of taxa based on the correlation of these characters. This technique, in which taxonomic data from usually 100 or more tests are correlated by means of a computer, is referred to as Adansonian, numerical, or computer analysis (or taxonomy) and has been used in comparing tens of thousands of bacterial strains. In these procedures the per cent similarity (% S) of strains is normally compared either to each other, or to a median strain. The strains form one or more clusters depending upon the level of similarity. It has been popular to assign arbitrarily a level of similarity for relatedness at the species or genus level. The number, type, and weighting of tests varies in different laboratories, and there is some controversy about the level of relatedness that can be accurately assessed. Nonetheless, the only study, to our knowledge, in which numerical analysis and nucleic acid base sequence relatedness were carried out on the same strains (*Vibrio* strains) revealed a high correlation between the two methods (8).

Adansonian analysis of 21 strains from diverse sources was carried out to determine % S to the median strain, 0128a, and to compare these data to nucleotide sequence similarity data obtained in 75 C reassociation reactions using labeled 0128a DNA. Seventy-five per cent S or higher was considered to comprise a species cluster. Relationship at the level of genus was considered to be between 70 and 75% S.

Table 9 contains the numerical taxonomy data and 75 C relative DNA binding data obtained by using labeled 0128a DNA. In addition, there is a column containing "approximate" relative DNA binding at 60 C. It was assumed that the genomes of strains K-12 and 0128a are almost completely similar at a 60 C criterion (97% relatedness at 60 C). It was also assumed that the increase in relative binding of either K-12 or 0128a DNA to each strain at 60 C compared to that at 75 C would be essentially identical. Therefore, the increase, if any, in per cent relative binding of K-12 DNA at 60 C to each strain compared to that at 75 C to each strain was added to the observed relative binding of 0128a DNA to each strain at 75 C to get an "approximate" per cent relative DNA binding of 0128a to each strain at 60 C.

Fifteen of the 16 "typical" *E. coli* strains form a tight species cluster at 81% S. Within this group, % S and per cent relative DNA binding are reasonably correlated in all but two strains, 190 and 01, where the % S is 10 and 15% higher than the per cent relatedness values obtained at 75 C. In these two strains, however, the % S is in good agreement with the approximated per cent relatedness at 60 C. One strain, 128, is not included in the 75% S species or the 70% S genus cluster despite the fact that this organism shows 93% relatedness to 0128a at 75 C with a $\Delta T_{m(e)}$ of only 0.1 C.

Three atypical *E. coli* strains isolated from humans were included in the study. One of these, 3360-66, was below the genus level of similarity (67% S) although it showed 75% relatedness to 0128a with a $\Delta T_{m(e)}$ of 0.7 C at 75 C. A second atypical strain clustered at the *E. coli* species level; however, it reacted only 14% with 0128a DNA at 75 C and approximately 45% at 60 C. The third atypical strain (3541-67) is included at the generic level (72% S), although it is only 17% related to 0128a DNA at 75 C and approximately 38% related at 60 C.

The two most disturbing results obtained from Adansonian analysis are those for *Shigella flexneri* and *Salmonella typhimurium*. Relatedness between *E. coli* strains and both of these organisms has been extensively studied in DNA reassociation experiments (3, 5). The data obtained support the relatedness values obtained between 0128a and these organisms in the present study. *S. flexneri* is 80% related to 0128a at 75 C and approximately 85% related to 0128a at 60 C, whereas relatedness between 0128a and *S. typhimurium* DNA is 12% at 75 C and approximately 45% at 60 C. From numerical analysis, *S. typhimurium* is included in the generic cluster with 0128a, while *S. flexneri*, with an extremely low S value at 59%, is not included at the species or genus level. This study should be extended to determine whether these are isolated exceptions or the rule for *Salmonella* and *Shigella* strains and other members of the *Enterobacteriaceae*. To this end, a larger sample of enterobacterial strains with known DNA relatedness values is currently being analyzed numerically in the laboratory of R. R. Colwell.

TABLE 9. *Median organism calculation and relative DNA binding for set of 21 strains*

Strain (*E. coli* unless designated)	Per cent S to median	Per cent relative DNA binding at 75 C	Approximate[a] per cent relative DNA binding at 60 C
0128a	94.6	100	100
418	93.5	100	100
121	92.6	89	
0102	90.2	95	
190	90.2	80	85
06	88.8	86	86
01	88.3	73	87
K-12	87.3	91	94
888	87.1	88	89
075	85.0	90	92
674	84.9	83	86
786	83.9	76	82
020	82.6	88	
0111B4	82.2	83	85
0125	81.2	85	86
3122-66 (atypical)	78.5	14	45
3541-67 (atypical)	72.0	17	38
Salmonella typhimurium LT2	71.3	12	45
3360-66 (atypical)	67.0	75	75
128	67.0	93	93
Shigella flexneri 24570	59.0	80	85

[a] Only 75 C reactions were carried out with labeled 0128a DNA. The 60 C data were obtained by determining the increase in binding of K-12 DNA to each of these strains at 60 C, compared to that at 75 C, and adding that difference to the observed relative binding observed with labeled 0128a DNA at 75 C. Since 0128a and K-12 are 97% related at 60 C, these results should approximate the level of reaction expected between these strains and 0128a.

The product-moment correlation between the two methods was approximately 0.45 ($P = < 0.05$) from the data for all strains for both 75 and 60 C DNA binding. If only typical *E. coli* strains were compared, the product-moment correlation was 0.53 ($P = <0.05$) between % S and per cent DNA binding at 75 C, and 0.71 ($P = <0.01$) between % S and per cent 60 C binding. The relationship between % S and 75 C DNA binding between all strains tested may be expressed by the equation: % S = 1.22 (per cent DNA binding) − 24.

DISCUSSION

Extensive polynucleotide sequence similarity tests, both at optimal and stringent reassociation criteria, were carried out to determine the extent of divergence present in *E. coli* strains and closely related strains. Common laboratory strains, strains isolated from diverse human, mammalian, and vertebrate sources, as well as strains isolated from varied geographical areas, were included to obtain a representative sample. Based on the ability to form stable interspecies duplexes at 75 C, as much as 25% divergence occurred between strains of *E. coli*. Thus, while there is an astronomical number of distinct cultures which can be independently isolated from animals which would be almost universally accepted by microbiologists as representing the species *Escherichia coli*, it is apparent that many display substantial differences in their genetic fine structure.

Reactions between K-12 DNA and DNA from Alkalescens-Dispar strains fall well within the limits of divergence seen among *E. coli* strains, and the A-D strains should definitely be included as part of the genus *E. coli*. The frequency distribution of relatedness between *E. coli* K-12 and *Shigella* species overlaps the distribution seen among *E. coli* strains. No strains from any other genus of enteric bacteria exhibit greater than 50% relatedness to *E. coli* at 60 C, or greater than 25% relatedness to *E. coli* at 75 C (3). Furthermore, duplexes formed at optimal reassociation criteria between *E. coli* and members of other genera of enteric bacteria exhibit $\Delta T_{m(e)}$ values of from 10 to 18 C (3), whereas $\Delta T_{m(e)}$ values of *E. coli-Shigella* reactions are less than 2 C.

Virtually complete reassociation was observed between the three K-12 derivatives (C600, 1485T$^-$, and W3110(21)) tested against the reference K-12 strain. In contrast, strains B/r and B Albany exhibited only 82 to 83% relatedness to the BB reference strain. It is possible that these B strains have diverged to a larger extent than K-12 strains, or that B/r and B Albany have significantly smaller genome sizes than that of strain BB. The second alternative is being tested by genome size determinations now in progress.

It has been implicit in all bacterial relationship studies that the organisms under test contain the same or nearly the same size genome. In fact, one obtains reciprocal relatedness values between a pair of organisms only when their genomes have essentially the same molecular weight. Nonreciprocal reactions between strains K-12, BB, and 02A suggested significant genome size differences in these strains. The extent of these differences was substantiated spectrophotometrically, by measuring initial rates of DNA reassociation (10). With the possibility of significantly different genome sizes, even in strains of the same species, it seems prudent that investigators establish genome size in any careful quantitative study of polynucleotide sequence relatedness. This can be done by reciprocal binding studies, reciprocal competition studies (11), or by a spectrophotometric determination of comparative reassociation rates (10, 16). There is a controversy about the effect of G + C content on reassociation rate and the necessity of employing a correction factor (10, 16, 18). It appears, however, that a correction factor is not critical in determining genome size (especially *comparative* genome size) in a group of organisms with comparable G + C contents.

The relatively small sample of *E. coli* strains assayed showed a 23% difference in genome size which would correspond to roughly 600 average-sized genes. We assume that DNA species from a representative number of diverse *E. coli* strains will form an essentially Gaussian frequency distribution with a range of approximately 2×10^9 to 3×10^9 daltons. This surprising variation in genome size may be important in thinking about divergence and evolutionary pathways in bacteria. One should like to know the function of the "extra" DNA in strains containing larger genomes and whether the "extra" DNA diverges at the same or at a different rate compared to the rest of the bacterial DNA.

As data accumulate with respect to DNA relatedness, per cent G + C, and genome size, these parameters should take an increasing importance in defining genus and species. These determinations would most usefully be expressed as values within a given range and are not likely subject to change by a single mutation as is the case with parameters such as sugar fermentation, enzyme activity, motility, etc.

It is clear from this study and that of *Vibrio* strains (8) that strains with a high degree of

DNA similarity almost always fall into the % S range specifying a species cluster. However, all three "atypical" *E. coli* strains, *S. typhimurium*, and *S. flexneri* exhibited falsely high or low % S values compared to the level of relatedness obtained from DNA reassociation. It appears that, while average correlations between numerical analysis and polynucleotide sequence relatedness are reasonable, both for closely and distantly related strains (8), false clustering in numerical analysis may be prevalent at the genus cluster level, and that DNA binding is the more sensitive indicator of genetic divergence. This conclusion will be tested in depth in a comparison of the two methods on several dozen strains from representative genera of enteric bacteria.

LITERATURE CITED

1. Bautz, E. K. F., and F. A. Bautz. 1964. The influence of non-complementary bases on the stability of ordered polynucleotides. Proc. Nat. Acad. Sci. U.S.A. **52:**1476-1481.
2. Berns, K. I., and C. A. Thomas. 1965. Isolation of high molecular weight DNA from *Hemophilus influenzae*. J. Mol. Biol. **11:**476-490.
3. Brenner, D. J., G. R. Fanning, K. E. Johnson, R. V. Citarella, and S. Falkow. 1969. Polynucleotide sequence relationships among members of the *Enterobacteriaceae*. J. Bacteriol. **98:**637-650.
4. Brenner, D. J., G. R. Fanning, A. Rake, and K. E. Johnson. 1969. A batch procedure for thermal elution of DNA from hydroxyapatite. Anal. Biochem. **28:**447-459.
5. Brenner, D. J., M. A. Martin, and B. H. Hoyer. 1967. Deoxyribonucleic acid homologies among some bacteria. J. Bacteriol. **94:**486-487.
6. Britten, R. J., and D. E. Kohne. 1966. Nucleotide sequence repetition in DNA. Carnegie Inst. Wash. Year B. **65:**78-106.
7. Clausen, T. 1968. Measurement of ^{32}P activity in a liquid scintaillation counter without the use of scintillator. Anal. Biochem. **22:**70-73.
8. Colwell, R. R. 1970. Polyphasic taxonomy of the genus *Vibrio*: numerical taxonomy of *Vibrio cholerae*, *Vibrio parahaemolyticus*, and related *Vibrio* species. J. Bacteriol. **104:**410-433.
9. Edwards, P. R., and W. H. Ewing. 1966. Identification of the *Enterobacteriaceae*, p. 83-89. Burgess Publishing Co., Minneapolis.
10. Gillis, M., J. DeLey, and M. DeCleene. 1970. The determination of molecular weight of bacterial genome DNA from renaturation rates. Eur. J. Biochem. **12:** 143-153.
11. Hoyer, B. H., and N. B. McCullough. 1968. Homologies of ribonucleic acids from *Brucella ovis*, canine abortion organisms, and other *Brucella* species. J. Bacteriol. **96:**1783-1790.
12. Laird, C. D., B. L. McConaughy, and B. J. McCarthy. 1969. On the rate of fixation of nucleotide substitutions in evolution. Nature (London) **224:**149-154.
13. Liston, J., W. Wiebe, and R. R. Colwell. 1963. Quantitative approach to the study of bacterial species. J. Bacteriol. **85:**1061-1070.
14. Marmur, J., and P. Doty. 1962. Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. J. Mol. Biol. **5:**109-118.
15. McCarthy, B. J., and E. T. Bolton. 1963. An approach to the measurement of genetic relatedness among organisms. Proc. Nat. Acad. Sci. U.S.A. **50:**156-164.
16. Seidler, R. J., and M. Mandel. 1971. Quantitative aspects of DNA renaturation: base composition, state of chromosome replication, and polynucleotide homologies. J. Bacteriol. **106:**608-614.
17. Sokal, R. R., and P. H. A. Sneath. 1963. Principles of numerical taxonomy. W. H. Freeman and Co., San Francisco.
18. Wetmur, J. G., and N. Davidson. 1968. Kinetics of renaturation of DNA. J. Mol. Biol. **31:**349-370.