JOURNAL OF BACTERIOLOGY, Mar. 1995, p. 1585–1588
0021-9193/95/$04.00+0

Vol. 177, No. 6

# Widespread Protein Sequence Similarities: Origins of *Escherichia coli* Genes

BERNARD LABEDAN[1] AND MONICA RILEY[2]*

*Institut de Génétique et Microbiologie, Université de Paris-Sud, 91405 Orsay Cedex, France,[1] and Marine Biological Laboratory, Woods Hole, Massachusetts 02543[2]*

Received 24 August 1994/Accepted 8 January 1995

**To learn more about the evolutionary origins of *Escherichia coli* genes, we surveyed systematically for extended sequence similarities among the 1,264 amino acid sequences encoded by chromosomal genes of *E. coli* K-12 in SwissProt release 26 by using the FASTA program and imposing the following criteria: (i) alignment of segments at least 100 amino acids long and (ii) at least 20% amino acid identity. Altogether, 624 extended alignments meeting the two criteria were identified, corresponding to 577 protein sequences (45.6% of the 1,264 *E. coli* protein sequences) that had an extended alignment with at least one other *E. coli* protein sequence. To exclude alignments of questionable biological significance, we imposed a high threshold on the number of gaps allowed in each of the 624 extended alignments, giving us a subset of 464 proteins. The population of 464 alignments has the following characteristics expressed as median values of the group: 254 amino acids in the alignment, representing 86% of the length of the protein, 33% of the amino acids in the alignment being identical, and 1.1 gaps introduced per 100 amino acids of alignment. Where functions are known, nearly all pairs consist of functionally related proteins. This implies that the sequence similarity we detected has biological meaning and did not arise by chance. That a major fraction of *E. coli* proteins form extended alignments strongly suggests the predominance of duplication and divergence of ancestral genes in the evolution of *E. coli* genes. The range of degrees of similarity shows that some genes originated more recently than others. There is no evidence of genome doubling in the past, since map distances between genes of sequence-related proteins show no coherent pattern of favored separations.**

It was proposed long ago (6) and is generally accepted today that ancestral genes duplicated and diverged to give rise to families of proteins with related functions in modern organisms. Ancestral proteins are believed to have been fewer in number early in evolution, with broader specificity than the descendant, diverged proteins of today, which have narrower specificity and more-focused functions (11).

A higher percentage of all of its protein sequences are available for *Escherichia coli* than for any other organism, providing the opportunity to test for compatibility of the sequence relationships among *E. coli* proteins with accepted ideas of mechanisms of molecular evolution. *E. coli* is also the organism for which the greatest proportion of chromosomal genes which have been sequenced to date corresponds to genes previously well characterized in terms of gene product function and regulation. In this study, translated protein sequences of *E. coli* were systematically compared to determine how many ancestral relationships could be detected and how important the mechanism of duplication and divergence has been during the evolution of the *E. coli* genome and its encoded proteins. No other organism provides the opportunity to make such a study that the massively sequenced *E. coli* genome does.

To quantify all instances of extended sequence similarities among *E. coli* proteins, we undertook a systematic search for similarity of each *E. coli* protein to all others, applying uniform criteria to define extended sequence similarity and to eliminate short sequence similarities arising from small motifs. Our approach put to one side the examination and analysis of short patterns and motifs in protein sequences as, for instance, in the work of Kister et al. (5) and McCaldon and Argos (7); rather, it focused on extended alignments that seem more likely to identify descendants of past whole-gene duplications.

Sequences were obtained from the SwissProt database (1), and the FASTA algorithm, designed to detect homologous proteins, to create an alignment for the largest possible segment (8), was used for rapid protein sequence comparisons.

## MATERIALS AND METHODS

The FASTA algorithm (8), implemented on a VAX minicomputer as part of the University of Wisconsin Genetics Computer Group program package, was used to compare each sequence of a customized *E. coli* data bank to the entire set of sequences of this *E. coli* data bank. In a first step, we used the Dataset program of the Genetics Computer Group package to assemble into a data library the entire set of *E. coli* protein sequences present in SwissProt release 23 and then successively deleted all of the nonchromosomal (i.e., plasmid-encoded proteins) or non-K-12 (i.e., prophage proteins or proteins from non-K-12 *E. coli* strains) sequences, repeated elements such as Rhs sequences, or fragments or hypothetical proteins (i.e., all of the sequences determined by genes whose designations begin with the letter *y*). This customized *E. coli* data library was then updated to the equivalent of SwissProt release 26 by adding the new sequences appearing in the updates at the NCBI and EMBL database servers. To find potential similarities within this data set, the sequences were divided into groups of around 100 and each group was compared with the complete set and successively processed in the following way. First, an overnight automatic procedure allowed a FASTA comparison without alignment. Only the sequences displaying a significant optimized FASTA score against at least one other *E. coli* sequence were submitted to a second round of overnight automatic FASTA comparisons with alignment. The FASTA comparisons were made by using the following standard parameters: a word size (ktup) of 2 to identify identities in the initial step, use of the PAM250 matrix to determine the similarities in the second step, and a gap penalty of 20. Finally, we retained any pair showing 20% or better amino acid identity in an aligned region of at least 100 residues. Subsequently, we retained only alignments with a normalized alignment score (NAS) (3) of at least 180. The NAS expresses the percent identity of amino acids moderated by a credit for matched cysteine pairs and a debit for gaps as follows: NAS = [(number of identical residues except Cys in alignment × 10) + (number of cysteine residues × 20) − (number of gaps × 25)]/ number of residues aligned × 100.

* Corresponding author. Phone: (508) 548-3705. Fax: (508) 540-6902. Electronic mail address: mriley@hoh.mbl.edu.

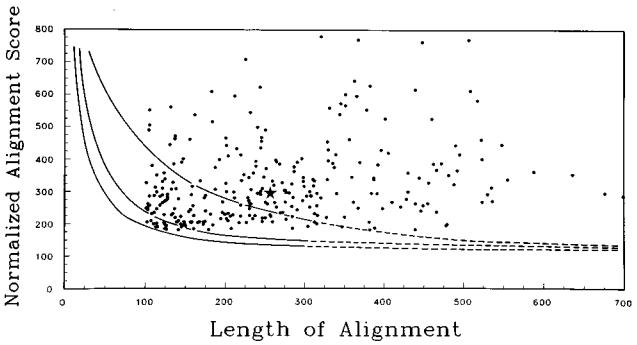


FIG. 1. Significance of NAS as a function of alignment region length. The formula for calculating the NAS is given in Materials and Methods. The empirical overlaid curves were determined by Doolittle (4) to be used as an easy framework allowing investigators to check directly the statistical significance of an experimental alignment through its determined NAS and the length of the alignment regions. Statistical significance was calculated as follows. Experimental pairs of various lengths were used to randomize the possible NASs, i.e., each sequence of each pair was repeatedly jumbled by the computer, and each jumbled sequence was aligned with each of the jumbled versions of the other sequence of the same pair, allowing us to average the scores and calculate the standard deviation for each experimental pair. The score of a given sequence pair was then compared with the mean of each corresponding randomized set and how much greater (or less) is expressed in standard deviation units. The original curves proposed by Doolittle were limited to 300 residues aligned. For clarity, these curves have been extended with dashed lines. Areas defined by the curves increase in relative significance of sequence similarity as they move up and outward from the origin and are labelled by Doolittle (4), in turn, as improbable (below 3 standard deviation units), marginal (between 3 and 5 standard deviation units), probable (between 5 and 10 standard deviation units), and certain (above 10 standard deviation units). Data for six sequence pairs are off the scale and are not shown; they all have length-of-alignment values greater than 700 and NAS greater than 270. The median alignment length and NAS is represented by the star.

## RESULTS

Of the 1,264 *E. coli* K-12 chromosomally determined protein sequences extracted from the SwissProt database through release 26 and compared by using the FASTA program under standard conditions (see Materials and Methods), 624 pairs of sequences were aligned over at least 100 amino acids with at least 20% amino acid identity. Since some proteins had multiple pairwise relationships, forming families of interrelated proteins, some proteins appeared more than once in the list of 624. A subset containing only one instance of each protein was assembled by choosing in each case to retain the pair with the highest degree of similarity. The unique set contained 577 proteins. That is, of the 1,264 protein sequences surveyed, 577 (45.6%) could be aligned with at least one other *E. coli* protein sequence over an extended region and some were aligned with more than one partner.

How significant is the sequence relatedness of all members in this set of 577 instances? Significance of percent amino acid identity increases with the length of the alignment region (2, 4). The longer the alignment region and the fewer the introduced gaps, the more significant is the sequence relatedness, as graphically shown by Doolittle (3). The NAS (3) was calculated for each pair (see Materials and Methods). Next, the data were progressively restricted for NAS. Setting limits on the NAS progressively reduced the numbers of gaps allowed at the lower percent identities of amino acids. With no restriction, there were 577 in the set; setting a lower NAS limit of 150 reduced the number in the set to 536, and a lower NAS limit of 180 reduced the number to 464. The NAS for each pair of this set of 464 was plotted against alignment region length (Fig. 1). Curves representing calculated data obtained by Doolittle for alignments of randomized sequences (3) were overlaid to indicate four significance ranges (Fig. 1, legend). The values for all of the *E. coli* sequence-related pairs in our data set of 464 fell in the three more significant ranges of the plot, the largest number in the most significant range, and the fewest in the marginally significant range. For further analysis, we adopted the most stringent criterion of an NAS of at least 180, yielding 464 of the 1,264 proteins that are related by sequence to at least one other protein by these criteria.

Median values for the parameters of the 464 extended alignments were determined. The NASs for the 464 alignments ranged from 181 to 1,009 and had an arithmetic mean of 352 and a median of 299. By stipulation, the percent identity of amino acids in the alignments was 20% or greater and the values ranged from 20 to 100%. The arithmetic mean percent identity of amino acids was 37.7%, and the median was 33% identity. The lengths of all alignments were 100 amino acid residues or greater by stipulation. Most were two to three times this length, ranging from 101 to over 1,000 for the very big proteins. The mean length for the 464 alignments was 290 amino acids, and the median was 254 amino acids. Most of the extended alignments involved nearly the whole length of the protein, although they ranged from 10% for very large proteins to 100%. Almost half of all of the alignments involved over 90% of the length of the protein. The median percentage aligned was 86%. The median number of gaps per 100 amino acids of alignment was 1.1.

To summarize, the median values for the 464 extended alignments between pairs of *E. coli* protein sequences were 254 amino acids aligned, representing 86% of the length of the protein, with 33% identity of amino acids, 1.1 gaps introduced per 100 amino acids, and an NAS of 299. This median entity is symbolized by a star in Fig. 1.

Statistical significance of sequence alignments should be modified by information on the biological significance of the data. To this end, referring to a previously prepared compilation of *E. coli* gene product function (9), we assessed the degree of functional relatedness of the 464 paired proteins. Each pair was placed in one of the following three categories of functional relatedness: (i) functionally related, some very closely related, such as a pair of isozymes that catalyze the same reaction, and some not as closely related, such as two transport functions specific for different molecules; (ii) different, that is, functions having no apparent connection; or (iii) not determined, that is, inadequate information for assessment of functional relatedness. Examples of functionally related pairs are isozymes, such as the three aspartokinases whose identities in the SwissProt database are AK1H, AK2H, and AK3, all assigned Enzyme Commission number EC 1.1.1.3. Pairs with related functions that are less closely similar are aspartate aminotransferase and aromatic aminotransferase (EC 2.6.1.1 and EC 2.6.1.57, respectively). A pair considered to differ in function are adenylate kinase and fumarase (EC 2.7.4.3 and EC 4.2.1.2, respectively). For pairs whose relationships are deemed to be not determined, the function of at least one of the two proteins is not well known or is only predicted through sequence similarity.

When these categories were used to broadly classify the 464 sequence-related pairs, 398 were deemed functionally related, 8 seemed to have quite different functions, and for 58 pairs there was insufficient information for classification. Thus, of the protein pairs whose functional relationships could be assessed, 98% were functionally related and only 2% seemed to have unrelated functions.

Addressing a different point, we used the list of functionally related and sequence-related proteins to determine whether

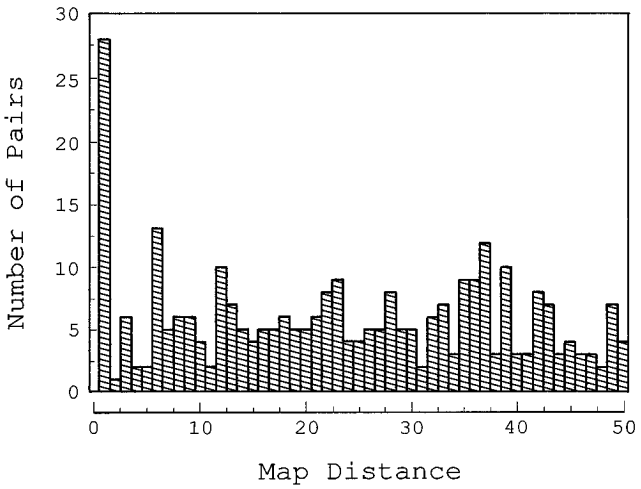


FIG. 2. Distribution of map distances between pairs of sequence-related genes in *E. coli*. Map positions were kindly provided by K. E. Rudd from the database EcoSeq6 (10). The difference in map location for each pair, expressed as the shorter distance on the circular map, was determined. The numbers of aligned pairs of sequences that are separated by each map distance shown at an interval of 1 map unit of distance between them are shown.

their corresponding genes have preferential locations on the chromosome map. The map distances between the genes for 296 sequence-related pairs were determined by using map values kindly provided by K. E. Rudd (10, 10a). Any two genes on the circular map of 100 map units have two distance arcs between them, the long and short distances, that add to 100. The distribution of the 296 gene pair distances represented by the shorter of the two arcs is shown in Fig. 2. There is an apparent excess of instances in the segment between 0 and 1 map unit that might reflect events of tandem duplication, but other than that no clear or regular pattern in the distribution is evident. Duplications do not seem to have occurred with any chromosomewide regularity.

## DISCUSSION

We have found that extended sequence similarities coupled with functional similarities are widespread among the proteins of *E. coli* whose sequences are known. Indeed, the 577 proteins found to have extended alignments with other proteins constitute 45.6% of the 1,264 sequenced proteins of *E. coli*. The somewhat more rigorously defined set of 464 proteins (Fig. 1) still represents 36.7% of the 1,264. Even though the significance of the similarity of the sequences of alignments that fall in the range of 20 to 25% amino acid identity might not be high on purely statistical grounds, the extremely high frequency of functional similarity between the two proteins argues that their sequence similarity is not accidental, but rather that the genes for these proteins are paralogous, descendants of duplicate copies of ancestral genes residing in the same genome. Although such a massive (98% of the assessable proteins) proportion of similar sequences displaying functional relatedness strongly suggests that the corresponding genes are paralogous, we could not dismiss the possibility that an undefined proportion of the present-day *E. coli* genes have gained both their sequence similarity and functional relatedness by convergent evolution. However, the simplest hypothesis explaining our results would be that a large proportion of *E. coli* proteins and their genes are related by duplication and divergence of ancestral genes and that many of these ancient relationships are still detectable today in the amino acid sequences of the proteins even when the nucleotide sequences of the genes are not seen to be related. Remarkably, these results are found to agree with prescient speculations (e.g., 6, 11) proposed well before such an analysis of sequence relationships could be done.

The large proportion of sequences in alignments that we report here is, without doubt, a minimum figure since our arbitrary cutoff criteria exclude some well-known examples of proteins believed to share evolutionary ancestry. Indeed, there seem to be biologically significant relationships at even lower levels of similarity. For instance, excluded from consideration by these criteria are pairs of proteins commonly believed to be evolutionarily related, such as the closely related chorismate mutase enzymes TYRA and PHEA. These proteins are aligned over 202 amino acids but were excluded because they are related by only 19% amino acid identity. In another example, the pair of shikimate kinase enzymes AROK and AROL align over 97 residues with 34% identity but were excluded because the alignment was less than 100 residues. Also, open reading frames were not included in this survey and a significant proportion of the 275 *E. coli* open reading frames with lengths greater than 100 amino acid residues present in SwissProt release 26 have sequence similarity to either sequences of known genes or other open reading frames (data not shown).

The large proportion of paralogous genes allowed us to test the idea that evolutionarily related genes lie at regular intervals on the chromosome as a result of genome doubling events in the past (12). With almost 20 times as much data available today as when this hypothesis was originally proposed, the data on map separation of pairs of genes for proteins with extended sequence similarity were examined (Fig. 2). There is a region of apparent excess of map separations of 0 to 1 map unit, possibly representing ancient tandem duplications. No clear pattern emerged from the variability in frequency of occurrence of other map distances. The data did not yield a coherent pattern of regularity expected of genomewide events. We concluded that the idea of full genome doubling did not survive the test.

In conclusion, we found that when systematically examined, a large proportion of *E. coli* K-12 proteins are related by extended sequence alignments and almost all of the pairs are also related by function. The demonstration of a large proportion of paralogous relationships among *E. coli* genes strongly suggests that duplication and divergence of genes in an ancester of *E. coli* could be the most frequently used mechanism by which a primitive genome was enlarged and biochemical complexity was introduced. We have found that many contemporary *E. coli* proteins still bear marks of their ancient relationships, allowing us to identify *E. coli* genes related not only by recent but also by distant common ancestry. When the remainder of the *E. coli* genome is sequenced, we will be able to determine sequence similarities for all *E. coli* proteins, including many previously undetected genes. This may help us to determine the relative proportions in the present-day *E. coli* chromosome of genes descending from ancestral sequences by either (i) duplication and divergence or (ii) convergent evolution or (iii) horizontal transfer. Some sequences are members of families of related proteins. Identification of all familial relationships among all *E. coli* proteins will allow us to estimate the number of ancestral sequences required to generate by duplication and divergence many of the genetic elements needed to manufacture and maintain this free-living single-cell organism.

(The data referred to in this report are part of a larger

compilation called GenProtEc [genes and proteins of *E. coli*], which is available to interested scientists either by FTP from the Marine Biological Laboratory via anonymous FTP or World Wide Web or by post from M. Riley as a simple MS-DOS application on diskette. Data on the sequence-related pairs are a subset called Simone, which is also available from B. Labedan on diskette in Macintosh format as output of the Claris FileMaker program. If this is used, kindly cite the source of data.)

## REFERENCES


1. **Bairoch, A., and B. Boeckman.** 1993. The SWISS-PROT protein sequence data bank, recent developments. Nucleic Acids Res. **21:**3093–3096.
2. **Doolittle, R. F.** 1981. Similar amino acid sequences: chance or common ancestry? Science **214:**149–159.
3. **Doolittle, R. F.** 1992. Stein and Moore Award address: reconstructing history with amino acid sequences. Protein Sci. **1:**191–200.
4. **Doolittle, R. F., D. F. Feng, M. S. Johnson, and M. A. McClure.** 1986. Relationships of human protein sequences to those of other organisms. Cold Spring Harbor Symp. Quant. Biol. **51:**447–455.
5. **Kister, A., I. Muchnik, D. Buzida, E. L. Reinherz, and T. Smith.** 1993. Efficient pattern comparative method for selecting functionally important motifs in protein sequences: application to zinc enzymes. BioSystems **30:** 233–240.
6. **Lewis, E. B.** 1951. Pseudoallelism and gene evolution. Cold Spring Harbor Symp. Quant. Biol. **16:**159–174.
7. **McCaldon, P., and P. Argos.** 1988. Oligopeptides of 2–11 residues in the PIR database: improving methods for detecting protein coding regions within nucleotide sequences. Proteins **4:**99–122.
8. **Pearson, W. R., and D. J. Lipman.** 1988. Improved tools for biological sequence comparisons. Proc. Natl. Acad. Sci. USA **85:**2444–2448.
9. **Riley, M.** 1993. Functions of the gene products of *Escherichia coli*. Microbiol. Rev. **57:**862–952.
10. **Rudd, K. E.** 1993. Maps, genes, sequences, and computers: an *Escherichia coli* case study. ASM News **59:**335–341.

10a. **Rudd, K. E.** Personal communication.

11. **Ycas, M.** 1974. On earlier states of the biochemical system. J. Theor. Biol. **44:**145–160.
12. **Zipkas, D., and M. Riley.** 1975. Proposal concerning mechanism of evolution of the genome of Escherichia coli. Proc. Natl. Acad. Sci. USA **72:**1354–1358.