Systematic sequencing of the *Escherichia coli* genome: analysis of the 0–2.4 min region

Takashi Yura, Hirotada Mori, Hiroki Nagai, Toshio Nagata, Akira Ishihama¹, Nobuyuki Fujita¹, Katsumi Isono², Kiyoshi Mizobuchi³ and Atsuo Nakata⁴

Institute for Virus Research, Kyoto University, Kyoto 606-01, ¹National Institute of Genetics, Mishima, Shizuoka 411, ²Department of Biology, Kobe University, Kobe 657, ³Department of Biophysics and Biochemistry, University of Tokyo, Tokyo 113 and ⁴Institute of Microbial Diseases, Osaka University, Osaka 565, Japan

Received April 30, 1992; Accepted June 5, 1992

DDBJ accession no. D10483

ABSTRACT

A contiguous 111,402-nucleotide sequence corresponding to the 0 to 2.4 min region of the *E. coli* chromosome was determined as a first step to complete structural analysis of the genome. The resulting sequence was used to predict open reading frames and to search for sequence similarity against the PIR protein database. A number of novel genes were found whose predicted protein sequences showed significant homology with known proteins from various organisms, including several clusters of genes similar to those involved in fatty acid metabolism in bacteria (e. g., *betT*, *baIF*) and higher organisms, iron transport (*sfuA*, *B*, *C*) in *Serratia marcescens*, and symbiotic nitrogen fixation or electron transport (*fixA*, *B*, *C*, *X*) in *Azorhizobium caulinodans*. In addition, several genes and IS elements that had been mapped but not sequenced (e. g., *leuA*, *B*, *C*, *D*) were identified. We estimate that about 90 genes are represented in this region of the chromosome with little spacer.

INTRODUCTION

The knowledge of complete nucleotide sequence of the entire genome of a given organism is undoubtedly of an enormous value contributing to, and indispensable for, deeper understanding of the biological phenomena in numerous spheres. The ultimate goal may be deciphering the complete sequence of the human genome, but as a stepping stone and a model, the *Escherichia coli* genome is no doubt one of the ideal subjects for such an endeavor.

The genome of *E. coli* K-12 contains more than 1,400 identified genes (1) and consists of a DNA molecule of 4,700 kb (2). Whereas great many sequence data have accumulated in the past which amounted to at least 32% of the genome (3, 4), much efforts (experiments and data compilation) are required for achievement of the complete analysis. Such efforts are important to gain further insight into the structural and functional organization and evolution of the genome. Bearing these in mind, we have initiated systematic sequencing of the *E. coli* genome, starting from 0 min, using the ordered λ phage clones constructed by Kohara *et al.* (2). Here we report the complete sequence of

111,402 nucleotides spanning from 0 to 2.4 min of the *E. coli* chromosome (accession number: D10483).

EXPERIMENTAL

Phage clones

Lambda phage clones used were from the mini-set Kohara clones (2) and include: # 101 (9E4), # 102 (6H3), # 103 (22B12), # 104 (2F7), # 106 (8D2) and # 108 (8H11).

Determination of the nucleotide sequence

An initial survey of the literature and DNA databases permitted us to localize the chromosomal segments that had not previously been sequenced. Specific λ clones were chosen to obtain six separate DNA segments. After subcloning their fragments into an appropriate plasmid or M13 phage vectors and generating appropriate deletion derivatives, their nucleotide sequences were determined by the dideoxy chain termination method (5) using automated DNA sequencers. The sequencing was performed for both strands to ensure high accuracy.

Computer analysis of the sequence data

Open reading frames (orf's) were predicted by using the program ODEN developed by Ina (6). Search for sequence homology was carried out using the algorithm of Lipman and Pearson (7, 8) and the PIR (Protein Identification Resources) database (Release 30).

RESULTS AND DISCUSSION

Construction of the contiguous nucleotide sequence

After sequencing the six λ clones, new sequence data which amounted to 45 kb were assembled with preexisting sequences in the DNA databases into a contiguous sequence of 111,402 bp (Fig. 1). Relevant sequence data from the GenBank (Release 69) and EMBL databases (Release 28) were first assembled with one another on the basis of overlapping sequences. When two or more data entries existed for a given region, longer sequences were generally adopted; if there were discrepancies in the overlapping

Fig. 1. Assembly of contiguous nucleotide sequence data (ca. 111.4 kb divided into 20 kb portions) for the 0 to 2.4 min region of the *E. coli* genetic map, compiled by one of us (H. M.) and deposited to the DNA database. Thin bars shown below indicate the sequence data taken from the GenBank (Release 69) or EMBL (Release 28) database with their EMBL entry names. Filled portions of thin bars represent the data actually used (with accession numbers) to construct the contiguous sequence. Hatched thick bars indicate the regions sequenced in this study. Arrows indicated approximate positions of ends of chromosomal segments carried by the λ clones.

Fig. 2. Prediction of possible orf's based on the contiguous nucleotide sequence data. The coordinates are the same as in Fig. 1. Orf's consisting of 75 or more codons were predicted by the program ODEN (6) and subjected to homology analysis against the PIR datanase. Putative genes ($\Rightarrow$) predicted by homology analysis and other uncharacterized orf's ($\rightarrow$) are shown along with the previously sequenced genes ($\Rightarrow$). The sizes of orf's are based on the maximum number of codons (see Table 1). Broken arrows ($\Rightarrow$) indicate a pair of neighboring orf's on the basis of sequence data alone but may represent single genes on the basis of homology analysis (see text).

Table 1. Putative genes whose products showed sequence homology with known proteins

Orf	First codon ^a (nt.)	Number of codons max. ^b	plaus. ^c	Gene ^d	PIR entry ^e	Description	Homology score	% match
8	7677	495	476	(<i>tet</i>)	YTECT0	Tetracycline resistance protein— <i>Escherichia coli</i>	130	20.0
9	7827	207	183		*YSCTAL1	Saccharomyces cerevisiae TAL1 gene for transaldolase	497	54.4
10	8365	162	95		*YSCTAL1	Saccharomyces cerevisiae TAL1 gene for transaldolase	393	51.9
28	15083	374	370		QQEC47	IS186 47K hypothetical protein— <i>Escherichia coli</i>	1766	96.2
29	15830	103	0		S04881	Hypothetical protein II (insertion sequence IS421) — <i>Escherichia coli</i>	537	99.0
54	35586	183	104		S06477	Enoyl-CoA hydratase precursor, mitochondrial —Rat	220	32.1
55	35927	194	194		S06477	Enoyl-CoA hydratase precursor, mitochondrial —Rat	120	37.3
56	37523	531	522		S01667	4-Coumarate-CoA ligase— <i>Parsley</i>	531	29.3
57	38517	320	245	(<i>baiF</i>)	F37844	baiF 47.5K protein— <i>Eubacterium</i> sp.	514	36.5
58	38791	114	108	(<i>baiF</i>)	F37844	baiF 47.5K protein— <i>Eubacterium</i> sp.	188	44.2
59	40047	381	380		C34252	Isovaleryl-CoA dehydrogenase precursor, hepatic—Rat	461	29.2
60	41593	505	504	(<i>betT</i>)	S10898	betT protein— <i>Escherichia coli</i>	592	25.7
61	41975	285	268	(<i>fixA</i>)	S14070	fixA protein— <i>Azorhizobium caulinodans</i>	209	26.2
62	43207	193	76	(<i>fixB</i>)	S14071	fixB protein— <i>Azorhizobium caulinodans</i>	299	46.5
63	43827	432	428	(<i>fixC</i>)	S14072	fixC protein— <i>Azorhizobium caulinodans</i>	768	37.0
64	45020	129	95	(<i>fixX</i>)	S14073	fixX protein— <i>Azorhizobium caulinodans</i>	221	38.3
65	45460	168	166	(<i>tet</i>)	YTBSU6	Tetracycline resistance protein— <i>Bacillus</i> <i>cereus</i> plasmid pBC16	147	20.2
66	46108	230	220		A25949	Glucose-transporter protein—Rat	172	22.4
67	46893	181	176		A30879	NAD(P)H dehydrogenase (quinone)—Human	195	28.6
81	56946	297	271		A33618	Heat shock protein npl1/translocation protein sec63— <i>Saccharomyces cerevisiae</i>	125	36.8
99	70958	271	237	(<i>dedA</i>)	XMECAD	dedA protein— <i>Escherichia coli</i>	184	28.0
100	72607	239	232	(<i>sfuC</i>)	QRSEUC	Nucleotide-binding protein sfuC— <i>Serratia</i> <i>marcescens</i>	334	36.0
101	74183	537	536	(<i>sfuB</i>)	QRSEUB	sfuB protein— <i>Serratia marcescens</i>	241	23.8
102	75181	341	323	(<i>sfuA</i>)	QRSEUA	sfuA protein precursor— <i>Serratia marcescens</i>	134	22.6
104	77198	419	392		YOECNQ	Sodium-glutamate symport carrier protein — <i>Escherichia coli</i>	119	20.5
105	79156	216	201	(<i>leuD</i>)	S07306	leuD protein— <i>Salmonella typhimurium</i>	946	92.0
106	80567	481	466	(<i>leuC</i>)	S10171	Isopropylmalate isomerase chain leuC— <i>Salmonella typhimurium</i>	1908	86.3
107	81781	418	363	(<i>leuB</i>)	DEBSIC	3-Isopropylmalate dehydrogenase— <i>Bacillus</i> <i>coagulans</i>	909	53.0
108	83304	562	523	(<i>leuA</i>)	S08431	2-Isopropylmalate synthase— <i>Salmonella</i> <i>typhimurium</i>	1813	78.1

Novel orf's that showed homology scores of more than 100 (optimized) with one or more proteins are presented. The nucleotide number of the first codon (a) was counted from the A of the initiation codon (ATG) of *thrA*, clockwise along the *E. coli* genetic map. The maximum number (b) of contiguous sense codons, namely, the number of sense codons spanned by two adjacent nonsense codons, and the plausible number (c) of sense codons counted from the first ATG codon are indicated. Gene symbols (d) of *E. coli* or other organisms together with PIR entries (e) of the highest homology scores are shown. Asterisks (*) before the entry names indicate the GenBank entries of DNA sequence that were used to detect homology in deduced amino acid sequence.

data, the 5' and/or 3' ends of the sequence were avoided as much as possible to minimize errors due to inclusion of vector sequences. Figure 1 indicates the individual entry names and accession numbers used as well as the regions whose sequence was determined in this study.

Prediction of possible open reading frames (orf's)

The 111.4 kb contiguous sequence data were then used to predict possible orf's. Orf's containing 75 or more consecutive sense codons were searched in 3 phases for both orientations by the computer program ODEN (6), with which detection of possible sequence errors was facilitated: orf's that were interrupted by an apparently erroneous nonsense codon could be picked up and corrected during subsequent analysis for sequence homology (see below). The orf's thus predicted in the newly sequenced regions together with known genes are illustrated in Fig. 2. All possible orf's except for those overlapping with known genes and putative

genes were numbered, starting from the *thrA* gene and extending clockwise along the *E. coli* genetic map. The size of orf's was indicated as the 'maximum' numbers of codons in order not to exclude possible orf's with initiation codons other than ATG. The 'plausible' number of codons counted from the first ATG codon was also employed to indicate the plausible size of proteins that may be encoded (see Table 1).

Search for sequence similarity to known protein sequences

The whole set of 461 orf's was subjected to search for sequence similarity against the PIR database (Release 30) using the FASTA program (7, 8). (A list of 461 orf's in a computer readable form is available upon request.) A number of putative genes and orf's were found to exhibit significant degrees of homology with one or more of the known proteins from various organisms (Table 1). Among the genes that had been mapped but not sequenced a cluster of four genes (*leuA*, *B*, *C*, *D*), required for leucine

biosynthesis, was identified at 78.5–83.3 kb (orf's 105–108), on the basis of a very high homology (53–92% identity) with the corresponding *leu* genes of *Salmonella typhimurium* (9–11) and *Bacillus coagulans* (12). Two insertion elements one homologous to IS186 (13) and a copy of IS1 previously mapped in this region (14) were located at 15.0 kb and 19.5 kb, respectively. In addition, a number of putative genes were found whose products exhibited significant sequence similarity with known proteins as described below.

Putative genes predicted by homology analysis

An analogue of *betT*, mapped at min 7.5 on the *E. coli* genetic map and involved in choline transport (1), was found at 41.6 kb (orf60). Downstream of the *betT* analogue are present at least 4 putative genes (orf's 54–59) whose products show significant homology with bacterial and mammalian enzymes involved in the metabolism of fatty acids and other lipophilic compounds. Two overlapping orf's 54 and 55 also exhibited modest homology with a mitochondrial enzyme, enoyl-CoA hydratase (15). Similarly, a pair of orf's 57 and 58 showed homology with a bile acid inducible protein (BaiF) of *Eubacterium sp* (16). Although the two overlapping orf's could in each case encode functionally related but separate proteins, an alternative explanation is that a fortuitous sequence error separated a single orf into two. Another cluster of orf's 61–64 exhibited fairly high homology (and the same order) with a group of genes (*fixA*, *B*, *C*, *X*) required for symbiotic nitrogen fixation, possibly involved in electron transport, in *Azorhizobium caulinodans* (17).

Significant homology was also found between three adjacent orf's 100–102 and the gene cluster (*sfuA*, *B*, *C*) of *Serratia marcescens* involved in periplasmic-binding-protein-dependent iron transport (18), suggesting that they might form a single operon. Finally, a pair of orf's 9 and 10 were found to exhibit very high homology scores with the transaldolase of *Saccharomyces cerevisiae* (19), suggesting that they might represent a single gene with similar function in *E. coli*. The physiological significance of all these genes, however, remains unknown.

Estimated numbers of genes and orf's

In spite of the limited information available, an approximate number of genes and orf's present in this region of the chromosome could be estimated. The number of known genes and orf's in the previously sequenced regions is 57. By using the present sequence data, we have predicted the nature of at least 26 orf's on the basis of similarity of protein sequence (Table 1). In addition, 10 other orf's could be considered as likely candidates for functional genes. The relationship between these newly found orf's and the genes which have already been mapped in this region but not assigned to specific orf's (1) remain to be investigated. We thus estimate that approximately 90 genes including those encoded by IS elements may be represented. In line with the previous expectation (1,2), the agreement between the genetic map and the physical order and distance among the known genes is in general excellent. It is also noteworthy that the region is fully occupied by the genes and orf's with little spacer.

Future prospects

Having the present contiguous sequence data at hand, we anticipate that functional analysis of orf's by various techniques including the use of mini-transposon (*Th5supF*) (20) for isolation

of null mutants, will provide further insights into the structure and function of the genes located in this region of the chromosome.

As to prospects of systematic sequencing of the entire genome, the results reported here are encouraging and warrant continuing efforts. Further sequencing of the adjacent region (2.4–7 min) is currently under way. Besides, sequencing from the *terC* region (at ca. 34 min counter-clock-wise toward 0 min) is being carried out by K. Isono and his associates at Kobe University. With further coordination of efforts world-wide, proper supports, and strategic and technological improvements, the complete sequence of the *E. coli* genome should be forth-coming in the near future.

ACKNOWLEDGEMENTS

We are grateful to B.J. Bachmann for unpublished information. This work was supported in part by Grants for the Priority Research Field 'Comprehensive Analysis of the *E. coli* Genome' (T.Y.) and Grants for Human Genome Analysis (K. Matsubara, Osaka University) from the Ministry of Education, Science and Culture, Japan. Nucleotide sequencing was carried out by a group (headed by Y. Ishino) at the Biocenter, Takara Shuzo Co., Kyoto, and a group (headed by S. Kanai) at Toray Research Center, Inc., Kamakura, Japan. We thank generous supports and expert assistance of many individuals during the course of this study.

REFERENCES

1. Bachmann, B. J. (1990) Microbiol. Rev., 54, 130–197.
2. Kohara, Y., Akiyama, K., and Isono, K. (1987) Cell, 50, 495–508.
3. Kröger, M., Wahl, R., and Rice, P. (1991) Nucleic Acids Res., 19, (supplement) 2023–2043.
4. Medigue, C., Viari, A., Hénaut, A., and Danchin, A. (1991) Mol. Microbiol., 5, 2629–2640.
5. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) Proc. Natl. Acad. Sci. USA, 74, 5463–5467.
6. Ina, Y. (1991) User's Manual distributed by the DNA database of Japan (DDBJ).
7. Lipman, D. J., and Pearson, W. R. (1985) Science, 227, 1435–1441.
8. Pearson, W. R., and Lipman, D. J. (1988) Proc. Natl. Acad. Sci. USA, 85, 2444–2448.
9. Friedberg, D., Rosenthal, E. R., Jones, J. W., and Calvo, J. M. (1985) Mol. Gen. Genet., 199, 486–494.
10. Ricca, E., and Calvo, J. M. (1990) Nucleic Acids Res., 18, 1290.
11. Rosenthal, E. R., and Calvo, J. M. (1990) Nucleic Acids Res., 18, 3072.
12. Sekiguchi T., Ortega-Cesena, J., Nosoh, Y., Ohashi, S., Tsuda, K., and Kanaya, S. (1986) Biochim. Biophys. Acta, 867, 36–44.
13. Chong, P., Hui, I., Loo, T., and Gillam, S. (1985) FEBS Lett., 192, 47–52.
14. Umeda, M., and Ohtsubo, E. (1991) Gene 98, 1–5.
15. Minami-Ishii, N., Taketani, S., Osumi, T., and Hashimoto, T. (1989) Eur. J. Biochem., 185, 73–78.
16. Mallonee, D. H., White, W. B., and Hylemon, P. B. (1990) J. Bacteriol., 172, 7011–7019.
17. Arigoni, F., Kaminski, P. A., Hennecke, H., and Elmerich, C. (1991) Mol. Gen. Genet., 225, 514–520.
18. Angerer, A., Gaisser, S., and Braun, V. (1990) J. Bacteriol., 172, 572–578.
19. Schaaff, I., Hohmann, S., and Zimmermann, F. K. (1990) Eur. J. Biochem., 188, 597–603.
20. Phadnis, S. H., Kulakauskas, S., Krishnan, B. R., Hiemstra, J., and Berg, D. E. (1991) J. Bacteriol., 173, 896–899.