A Comparative Study on the Genes for Three Porins of the *Escherichia coli* Outer Membrane

DNA SEQUENCE OF THE OSMOREGULATED *ompC* GENE

(Received for publication, January 24, 1983)

Takeshi Mizuno†, Mei-Yin Chou, and Masayori Inouye

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

The DNA sequence of the *ompC* gene which encodes one of the outer membrane porins has been determined. The gene appears to encode a secretory precursor of OmpC protein consisting of a total of 387 amino acid residues with a signal peptide of 21 amino acid residues at its NH₂-terminal end. The 5' end noncoding region including the promoter of the *ompC* gene is unusually short (A-T)-rich, and the codon usage in the *ompC* gene is unusual as are those in genes for other abundant outer membrane proteins. The promoter sequence of the *ompC* gene was compared with that of the *ompF* gene, both of which are controlled by the osmoregulatory operon, *ompB*. The deduced amino acid sequence of the OmpC protein showed extensive homology with that of the other porins (OmpF and PhoE proteins). The homology in the primary amino acid sequences, as well as the coding DNA sequences among the porins, indicates that the structural genes for the three porins evolved from a common ancestral gene. Comparison of the amino acid sequences among the OmpC, OmpF, and PhoE porins will be discussed with regard to structure and function.

The matrix proteins, OmpC and OmpF, are major outer membrane proteins of *Escherichia coli*, and are known as porins because they form passive diffusion pores which allow small molecular weight hydrophilic materials across the outer membrane (for review see Ref. 1). In addition to these porins, a new porin, PhoE protein, has been recently reported, which forms a passive diffusion pore across the outer membrane preferentially for organic and inorganic phosphate (2). Although the genes for these porins are completely unlinked on the *E. coli* chromosome (6, 21, and 47 min on the chromosome for the *phoE* (3), *ompF* (4), and *ompC* (4) genes, respectively), these porin proteins share a few common features: (a) they have similar amino acid compositions and molecular weights; (b) they are immunologically cross-reactive; and (c) they form passive diffusion pores of similar diameters (1.3 nm for OmpC protein (5), 1.4 nm for OmpF protein (5), and 1.2 nm for PhoE protein (6)).

These results strongly suggest that the genes for these proteins may have evolved from a common ancestral gene. The DNA sequences for both the *ompF* and the *phoE* genes have been recently determined by Inokuchi et al. (7) and Overbeeke et al. (8), respectively. In this paper, we report the DNA sequence of the *ompC* gene. Comparison of the DNA sequences of these genes with those of the major amino acid sequences will provide insight as to not only how they are related to each other and how they are evolved but also how the expressions of these genes are regulated. It is known that the *ompC* and *ompF* genes are regulated by another independent operon, *ompB*, consisting of the *ompR* and *envZ* genes, in such a way that the *ompC* gene is preferentially expressed in cells grown in a media of high osmolarity, while the *ompF* gene is expressed in cells grown in a media of low osmolarity (for review see Ref. 9). The DNA sequence encompassing the entire *ompB* operon has been determined (10, 11). In contrast to the osmoregulation of the *ompC* and *ompF* genes, the *phoE* gene is known to be controlled by phosphate concentrations in a culture medium via the *phoB* gene (12). A preliminary report concerning the cloning and the identification of the *ompC* gene has been published recently (13).

EXPERIMENTAL PROCEDURES

Materials—Restriction endonucleases were obtained from either Bethesda Research Laboratories or New England BioLabs. DNA polymerase I large fragment was from New England Nuclear. [a-³²P] dNTPs (2000–3000 Ci/mmol) were purchased from Amersham Corp. Bacterial Strains and Plasmids—*E. coli* T19 recA (F*′* tsr-354 ompB) was used for most of the experiments (14). Plasmid pMY111 carrying the entire *ompC* gene was used for DNA sequencing (15). The cloning of the *ompC* gene has been reported previously (13).

Preparation of DNA and Restriction Endonuclease Digestion—Plasmid DNA was purified as described previously (14). Electrophoresis of DNA fragments on 0.7% agarose gels or 5.0% polyacrylamide gels has been described previously (15). DNA fragments were isolated from agarose gels by freeze-squashing, or from polyacrylamide gels by electroelution. Conditions for restriction enzyme digestions were described previously (16).

DNA Sequence Determination—Restriction DNA fragments were labeled at their 5' end by the method of Sakano et al. (17), using [³²P] dNTPs and DNA polymerase I large fragment (Klenow fragment). Singly end-labeled DNA fragments were obtained either by digestion with a second restriction enzyme or by electrophoretic strand separation (18). DNA sequences were determined by the method of Maxam and Gilbert (18, 19), using 20, 10, and 6% polyacrylamide gels in 7 M urea (0.04 × 18 × 40 and 0.04 × 18 × 100 cm).

RESULTS

Cloning of the *ompC* Gene—As reported previously (13), we have cloned a 2.7-kilobase pair HindIII fragment of *E. coli* K-12 chromosomal DNA into a plasmid vector pBR322 (plasmid pMY111, Fig. 1). The promoter region and the coding region for the NH₂-terminal part as well as the signal peptide of OmpC protein was found in the 500-base pair *MspI-BglII* fragment of the cloned DNA (13). Amino acid analysis and molecular weight analysis of OmpC protein (20) indicate that

6932
OmpC protein consists of 340–350 amino acid residues which correspond to a DNA sequence of 1020–1050 base pairs. For this reason, the 2.7-kilobase pair HindIII fragment is considered to contain the entire ompC gene as shown in Fig. 1.

Restriction Enzyme Mapping and DNA Sequencing—In order to determine the entire DNA sequence of the ompC gene, DNA sequencing was carried out within the 1.6-kilobase pair MspI–HindIII fragment. The restriction map and the sequencing strategy are shown in Fig. 2. DNA sequences at the junction regions of two DNA fragments were determined by sequencing another DNA fragment covering the junctions. The complete nucleotide sequence of the 1626 base pairs MspI–HindIII fragment is shown in Fig. 3. The DNA sequence shows that an open translational reading frame can be extended to 1101 nucleotides from the translation initiation codon ATG (nucleotides 404–406) to the termination codon TAA (nucleotides 1505–1507). This is the only possible open reading frame which could sufficiently encode a protein the size of OmpC protein. There are as many as 42 out of phase termination codons (14 TAA, 6 TAG, 22 TGA) scattered through the entire coding region. From these results, together with the previous finding (13), we conclude that the present DNA sequence encompasses the entire ompC gene. The amino acid sequence of pro-OmpC protein was deduced from the DNA sequence and is shown in Fig. 3.

DISCUSSION

Structure of the ompC Gene—The DNA sequence of the ompC gene shown in Fig. 3 indicates that the gene encodes a secretory precursor of OmpC protein (pro-OmpC) with a signal peptide (21 amino acid residues) at its NH2-terminal end, and that pro-OmpC protein consists of total 367 amino acid residues. Since the amino acid sequence (Ala-Glu-Val-Tyr-Asn-Lys-Asp-Gly-Asn; Fig. 3) is exactly the same as the NH$_2$-terminal sequence of the mature OmpC protein which has been determined by Ichihara and Mizushima (20), pro-OmpC protein appears to be processed at the specific cleavage site between Ala-Ala residues, resulting in the formation of the mature OmpC protein consisting of 346 amino acid residues.

The translation initiation codon ATG (nucleotides 404–406) is preceded by the Shine-Dalgarno sequence GAGG (nucleotides 391–394), a possible Pribnow box, GAGAAT (nucleotides 313–318), and a possible RNA polymerase recognition site, TGTTG (nucleotides 288–292) as described previously (13). The termination codon, TAA (nucleotides 1505–1507), is followed by a [T]$_{-1}$-rich sequence (nucleotides 1542–1556). As shown in Fig. 4, an extremely stable stem-loop structure with the oligo-[T] at the end can be constructed in this region, which appears to have the well known features of p-factor independent transcription termination sites in prokaryote (21). It is most likely that this sequence is the transcription termination signal for the ompC gene.

The distribution of [A-T] content in the ompC gene was found to be nonrandom. The DNA sequence of 120 base pairs encompassing the RNA polymerase recognition site and the Pribnow box mentioned above (nucleotides 200–319) has a significantly high [A-T] content (71%), and the preceding region of 200 base pairs (nucleotides 1–199) is also [A-T]-rich (66%) in contrast to the average [A-T] content of the E. coli genome.
chromosomal DNA which is 49% (22). However, the [A-T] content decreases to 49% in the ompC coding region (nucleotides 404-1504). [A-T]-rich promoters have been reported for several other E. coli genes (7, 15). Particularly, the [A-T] content of the promoter for the lpp gene (gene for the major outer membrane lipoprotein) has been shown to be as high as 80% (15). The high [A-T] content of the ompC promoter may be important for efficient transcription of the gene, since an
abundance of [A-T] base pairs is considered to facilitate unwinding of the duplex DNA helical structure, resulting in efficient initiation of mRNA transcription.

Comparison of DNA Sequences of the Promoter and Noncoding Regions among the ompC, ompF, and phoE Genes—In contrast to the extensive DNA sequence homologies in the coding regions of the ompC, ompF, and phoE genes as will be discussed later, there are few homologies in the promoters and 5' end noncoding regions as well as in the 3' end noncoding regions among these three porin genes. Since the regulation of phoE gene expression is controlled by the phoB gene (3) and it is totally different from the regulation of the ompC and ompF promoters, it is not surprising that the phoE promoter bears little resemblance to the ompC and ompF promoters. On the other hand, the expressions of both ompC and ompF genes are controlled by the ompR operon consisting of the ompR and envZ genes (9). The ompR gene product is thought to be modified by the envZ gene product in an osmolarity-dependent manner such that the ompR gene product would work as a positive regulatory factor for either the ompF gene in a low osmolarity culture medium, or for the ompC gene in a high osmolarity culture medium. Therefore, it is possible that both of these genes may have identical or homologous sequences to facilitate the binding of the ompR gene product in their 5' end noncoding regions.

Fig. 5A shows the DNA sequence of the 191 nucleotides upstream of the translation initiation codon for both the ompC and ompF genes. From the analysis of these sequences, one can find three identical or homologous sequences in this region. The first one is an identical sequence of 9 nucleotides except for mismatch in the center: CATCTATAG (nucleotides -165 to -157) in theompC gene and CATCAATAG (nucleotides -86 to -78) in theompF gene (a small letter represents the mismatched base). The sequences are located 40 and 13 base pairs upstream of the translation initiation codon of the ompC and ompF genes, respectively.
tides (TTATTGAC) found at the putative -35 region (23) is the upstream of the initiation codon for the binding site for the are three directly repeated sequences (a, ATCTTAAAA pairs downstream of the sequence c. In the case of the DNA sequence corresponding to the NH₂ terminus of the ompF gene (13) and TATTGAC (nucleotides -58 to -64) for the ompC gene (23). Each of these identical sequences is part of inverted repeats as indicated in Fig. 5A, and each may be involved as the recognition site for the ompR gene product. Since the ompR gene product has been purified to homogeneity in our laboratory,¹ it remains to be seen how and where the ompR gene product interacts with the ompC and ompF genes. The second set of homologous sequences are shown in Fig. 5B. These sequences are found 34 and 21 base pairs upstream of the initiation codon for the ompC and the ompF genes, respectively. Besides an extra 6 additional nucleotides and 1 missing nucleotide in the ompC sequence consisting of 37 nucleotides, there are only four mismatches within the ompF sequence consisting of 32 nucleotides. These homologous sequences may also serve as a binding site for the ompR gene product, or may be the result of evolution from the same ancestral gene. A third homologous sequence (GAGGGT) is found in the region of the Shine-Dalgarno sequence (GAGG) in both genes (Fig. 5A).

Besides the homologous sequences described above, there are three directly repeated sequences (a, ATCTTTAAAA; b, TGAACATCAT; and c, GAGGTTAAAT) in the 5' end noncoding region of the ompC gene (Fig. 5A). The repeating sequence, b, overlaps the sequence CATCTATAG, a possible binding site for the ompR gene product as described above. The repeating sequence, c, contains the Shine-Dalgarno sequence, and the initiation codons (ATG) are found 4 or 3 base pairs downstream of the sequence c. In the case of the ompF gene, it should be pointed out that a sequence of eight nucleotides (TTATTGAC) found at the putative -35 region (23) is repeated 5 base pairs downstream of the -35 region (Fig. 5A).

After the 5' end noncoding regions, one can observe a dramatic increase in homology among all three of the genes. Fig. 6A shows an example of such homology in a region of each DNA sequence corresponding to the NH₂ terminus of the mature protein through the 24th amino acid residue. In these regions, there is 81% homology between ompC and ompF and 65% homology between ompC and phoE. Fig. 6B shows the DNA sequences encompassing the regions containing the translation termination codon for the three genes. It is again clear that homologies are completely lost immediately after the termination codon, not only between ompC and phoE, but also between ompC and ompF. In all cases, inverted repeat sequences are found after the coding sequences, and at least for ompC and ompF, they appear to constitute ρ-factor independent transcription termination sites (Fig. 6B and Refs. 7 and 8).

**Codon Usage**—The codon usage for OmpC protein is shown in Table I. It is unusual as were those found for other major membrane proteins such as the lipoprotein (24), OmpA protein (25), and OmpF and PhoE proteins (Table I). Nineteen out of 59 sense codons are not used at all, and 8 codons are used only once. In particular, the following codons are preferentially used: CUG for leucine (10/10); AAC for asparagine (32/32); AAA for lysine (17/17); GAA for glutamic acid (11/11); CAG for glutamine (20/21); ACU or ACC for threonine (24/24), CGU or CCC for arginine (13/13). From the data in Table II, it is thus apparent that E. coli genes encoding abundant proteins selectively use major isoaccepting species of transfer RNAs (26). The codon usage in the ompC gene appears to be more uneven than those in the ompF and phoE genes.

**Comparison of Primary Protein Structures among OmpC, OmpF, and PhoE Proteins**—From the DNA sequence, OmpC protein was found to be produced from a secretory precursor, pro-OmpC protein, with a peptide extension (signal peptide) of 21 amino acid residues. The amino acid composition of OmpC protein is shown in Table II, which is in good agreement with that determined from purified OmpC protein (20). The calculated molecular weight is also in good agreement with that previously determined (20). Table II also lists the amino acid compositions of OmpF and PhoE proteins. One can see that they are very similar. One significant difference is the net charge of the mature proteins: -14 for OmpC, -11 for OmpF, and -9 for PhoE protein. This result indicates that all these porins are negatively charged at physiological pH, with PhoE protein having the least negative charge. The entire amino acid sequences of the three porin proteins are

---

¹ T. Mizuno and M. Inouye, manuscript in preparation.
The amino acid compositions are deduced from the nucleotide sequences of the ompC gene (this study), the ompF gene (7), and the phoE gene (8). Parentheses indicate the amino acid compositions of the signal peptides. At the level of the DNA sequences, 663 nucleotides (69%) are conserved between the ompC and ompF genes, and 622 nucleotides (60%) are conserved between the ompC and PhoE genes are functionally identical amino acid residues, respectively. These differences are the net results of deletions and insertions of amino acid residues throughout these sequences. It is interesting to point out that the number of amino acids of OmpF protein which forms a pore with a 1.4-nm diameter is smaller than that of OmpC protein which forms a pore with a 1.3-nm diameter (5).

It is known that these proteins are transmembranous intrinsic proteins (for review see Ref. 9). In all three proteins, however, there is no sequence of longer than 10 hydrophobic amino acid residues. The longest hydrophobic sequence of each protein is located at the COOH-terminal end, e.g. the sequence of +337 +346 of Ile-Val-Ala-Leu-Gly-Leu-Val-Tyr-Gln-Phe of OmpC protein.

Two hundred ten amino acid residues (61%) are conserved between OmpC and OmpF proteins as well as between OmpC and PhoE proteins. Furthermore, 42 residues (12%) between OmpC and OmpF proteins and 52 residues (15%) between OmpC and PhoE proteins are functionally identical amino acid residues. At the level of the DNA sequences, 663 nucleotides (69%) are conserved between the ompC and ompF genes, and 622 nucleotides (60%) are conserved between the ompC and phoE genes. The primary structures can be classified into three region types. The first region type (C region for conserved regions) is assigned to regions which are composed of the same or functionally identical amino acid residues in all three proteins (Fig. 7, solid lines). The second region type (SC region for semiconserved regions) is assigned to regions functionally conserved only in any two of the three proteins (dotted lines). The third region type (V region for variable regions) is assigned to regions which are composed of different kinds of amino acids in all three proteins (open numbered spaces). C regions are found throughout the sequences and 238 amino acid residues are located away from these sequences.
FIG. 7. Comparison of the primary amino acid sequences among OmpC, OmpF, and PhoE proteins. Primary amino acid sequences of PhoE (upper line), OmpC (middle line), and OmpF (lower line) are shown using the following abbreviations: A, alanine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tyrosine. For PhoE and OmpF proteins, only the residues which differ from those of OmpC protein are indicated. Signal peptide sequences of pro-PhoE, pro-OmpC, and pro-OmpF proteins are also shown in the top three lines, and their cleavage sites are indicated by a large arrow. Insertions of amino acid residues are indicated by small arrows with letters. The sequences which are substantially different from the OmpC sequence are boxed. The amino acid residues are numbered starting from the NH₂ group.
protein). Perhaps these regions have been conserved because they are essential for basic functions required for all three porins. Amino acid substitutions in C regions are found at 36 residues between OmpC and OmpF proteins, while 48 residues between OmpC and PhoE proteins and 50 residues between OmpF and PhoE proteins are substituted. This suggests that porins. Amino acid substitutions in C regions are found at 36 residues between OmpC and OmpF proteins, while 48 residues between OmpC and PhoE proteins and 50 residues between OmpF and PhoE proteins are substituted. This suggests that the phoE gene may have evolved first from a common ancestral gene. SC regions account for a total of 77 residues in the

**protein**. Perhaps these regions have been conserved because they are essential for basic functions required for all three porins. Amino acid substitutions in C regions are found at 36 residues between OmpC and OmpF proteins, while 48 residues between OmpC and PhoE proteins and 50 residues between OmpF and PhoE proteins are substituted. This suggests that the phoE gene may have evolved first from a common ancestral gene. SC regions account for a total of 77 residues in the
OmpC sequence, which corresponds to 22% of OmpC protein. V regions thus cover only 9% of OmpC protein and are scattered throughout the entire sequence. Nine of 16 V regions are single amino acid substitutions which are not functionally related. All other V regions appear to be caused by either the deletion or insertion of amino acid residues (except for V region 8). These regions may be poorly conserved to provide specificity for the individual porins, or they may not be important for porin formation at all. In this regard, it is interesting to note that out of the 26 amino acid residues in the V regions of PhoE protein, there are 7 basic and 1 acidic amino acid residues, while out of the 31 amino acid residues in the V regions of OmpC protein there are only 2 basic and 3 acidic amino acid residues. The more basic nature of PhoE V regions may be directly associated with the function of PhoE porin, which is known to be a more specific porin towards negatively charged phosphate compounds. The regions encompassing V regions 2, 3, 4, 5, 11, 13, 15, and 16, as well as the region from amino acid residues 155–169 in OmpC protein, seem to have been the result of drastic mutations at the level of the DNA sequence. These regions are indicated by A–F in Fig. 7, and their corresponding DNA sequences are shown in Fig. 8. Drastic alterations in the DNA sequences coding for these regions appear to occur by either deletions or insertions of DNA fragments in all cases. It is also interesting to note that in many cases there are considerable direct repeats in the DNA sequence in these regions as shown by the arrows in Fig. 8. In the case of regions A, C, and D in the phoE, ompF, and ompC sequences, respectively, one can observe an inverted repeat sequence. At present it is not possible to elucidate how mutations have occurred to produce the DNA sequences in these regions. However, it is possible that the repetitive sequences found in these regions might have been involved in the DNA rearrangements.

In OmpC protein, there is an extra sequence of 15 amino acid residues not present in OmpF protein in region C. Out of these 15 amino acid residues, 5 are glycine and proline residues indicating that this region, located at the center of the OmpC protein molecule, plays an important role in the OmpC protein conformation, possibly forming a β-turn at this region (28). In contrast to region C, there are 5 more amino acid residues in OmpF protein than there are in OmpC protein in region A, and this region is extremely glycine-rich in OmpF protein (5 out of 11 residues are glycine from positions 26–36 of OmpF protein), perhaps forming a β-turn structure (28). The specific compositions of such regions in each protein may also be important to enable the porins to serve as specific phage and colicin receptors (4) in the outer membrane.

In any event, from these comparisons, it is most likely that the ompC, ompF, and phoE genes have evolved from a common ancestral gene, and somehow came to reside at totally different sites on the E. coli chromosome. Subsequently, base substitutions, insertions, and deletions have occurred to some extent during evolution, resulting in the generation of similarly, but distinct structural genes for OmpC, OmpF, and PhoE proteins. Moreover, their DNA sequences in 5' and 3' ends noncoding regions have changed extensively, and now the expression of each gene is controlled in a different manner.

Acknowledgments—The authors are grateful to P. Green for her critical reading of this manuscript, and also to Dr. M. Riley for her helpful suggestions on analyses of repetitive DNA sequences.

REFERENCES

A comparative study on the genes for three porins of the Escherichia coli outer membrane. DNA sequence of the osmoregulated ompC gene.

T Mizuno, M Y Chou and M Inouye


Access the most updated version of this article at http://www.jbc.org/content/258/11/6932

Alerts:

- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 0 references, 0 of which can be accessed free at http://www.jbc.org/content/258/11/6932.full.html#ref-list-1