High-Molecular-Weight Protein 2 of Yersinia enterocolitica Is Homologous to AngR of Vibrio anguillarum and Belongs to a Family of Proteins Involved in Nonribosomal Peptide Synthesis

INGRID GUILVOUT,1 ODILE MERCEREAU-PUIJALON,2 SERGE BONNEFOY,2 ANTHONY P. PUGSLEY,3 AND ELISABETH CARNIEL1*

Unité de Bactériologie Moléculaire et Médicale,1 Unité de Parasitologie Expérimentale,2 and Unité de Génétique Moléculaire,3 Institut Pasteur, 75724 Paris Cedex 15, France

Received 6 May 1993/Accepted 28 June 1993

The iron-regulated irp2 gene is specific for the highly pathogenic Yersinia species and encodes high-molecular-weight protein 2 (HMWP2). Despite the established correlation between the presence of HMWP2 and virulence, the role of this protein is still unknown. To gain insight into the function of HMWP2, the entire coding sequence and the promoter of irp2 were sequenced. Two putative −35 and −10 promoter sequences were identified upstream of a large open reading frame, and two potential Fur-binding sites were found overlapping the second −35 box. The large open reading frame is composed of 6,126 nucleotides and may encode a protein of 2,035 amino acids (ca. 228 kDa) with a pl of 5.81. A signal sequence was not present at the N terminus of the protein. Despite the existence of 30 cysteine residues, carboxymethylation prevented the formation of most if not all disulfide bonds that otherwise occurred when the cells were sonicated. The protein was composed of three main domains: a central region of ca. 850 residues, bordered on each side by a repeat of 550 residues. A high degree of identity (44.5%) was found between HMWP2 and the protein AngR of Vibrio anguillarum. The central part of HMWP2 (after removal of a loop of 337 residues) also displayed significant homology with proteins belonging to the superfamiliy of adenylate-forming enzymes and, like them, possessed a putative ATP-binding motif that is also present in AngR. In addition, HMWP2 shared with the group of antibiotic and enterochelin synthetases a potential amino acid-binding site. Six consensus sequences defining the superfamily and four defining the family of synthetases were derived from the multiple alignment of the 30 sequences of proteins or repeated domains. A phylogenetic tree that was constructed showed that HMWP2 and AngR are in a family composed of Lys2, EntF, and the tyrocidine, gramicidin, and β-lactam synthetases. This finding suggests that HMWP2 may participate in the nonribosomal synthesis of small biologically active peptides.

The environmental conditions encountered by pathogenic microorganisms in their hosts are different from those found in vitro. The organisms sense a number of parameters (temperature, pH, osmolarity, ion concentration, etc.) which may vary during the course of infection. Bacteria adapt to environmental changes through a series of induced responses (55, 61). Most if not all of these responses are regulated at the gene level: some genes or operons are coordinate switching on, while some others are turned off.

One of the best-studied bacterial responses induced in the host is the microbial adaptation to iron limitation (5). Laboratory growth media contain a large excess of iron, while in mammals, the low level of free iron (10−18 M) is not sufficient to sustain bacterial growth. To circumvent this iron limitation, most bacteria synthesize small excreted molecules called siderophores which chelate iron bound to specific eukaryotic proteins (e.g., transferrin and lactoferrin) and are then transported back into the microorganism. Some bacteria produce outer membrane proteins able to directly bind transferrin and/or lactoferrin and to capture the iron carried by these molecules (5). Finally, bacteria may also produce hemolysins which lyse erythrocytes and liberate heme-bound Fe molecules (5). Synthesis of the proteins involved in these responses is induced by iron starvation.

The genus Yersinia is composed of 11 species, only 3 of which (Yersinia pestis, the agent of plague, and two enteric pathogens, Y. enterocolitica and Y. pseudotuberculosis) are pathogenic for humans. The role of iron in the virulence of Y. pestis was described almost 40 years ago (33), but iron-regulated proteins were identified only recently (11, 54, 60). No precise function has yet been attributed to the iron-regulated polypeptides of Y. pestis and Y. pseudotuberculosis. In Y. enterocolitica, the species which synthesizes the largest number of iron-regulated proteins (11), two of the proteins, FoxA (75.7 kDa) and FcuA (81.7 kDa), are the respective receptors for ferrioxamine and ferrichrome, hydroxamate siderophores utilized but not produced by Y. enterocolitica (3, 35). Several proteins are encoded by the Y. enterocolitica hemin uptake operon. These include HemR (78 kDa), the hemin receptor, HemS (42 kDa), which is either a cytoplasmic membrane permease or a hemin-degrading enzyme, and HemP (6.5 kDa) and HemT (27 kDa), both of which have unknown functions, although it is known that the latter is not necessary for hemin uptake (62). Y. enterocolitica also produces an approximately 80-kDa outer membrane protein which is structurally related to the FepA protein of Escherichia coli and therefore may serve as an enterochelin receptor (48). Finally, a 65-kDa iron-regulated
protein (FyuA) has been shown recently to be the receptor for the siderophore yersiniabactin and probably also for the bacteriocin pesticin (29). All highly pathogenic Yersinia species produce two high-molecular weight iron-regulated proteins (high-molecular-weight proteins 1 and 2 [HMWP1 and HMWP2]) (11, 17). The irp2 gene encoding HMWP2 is absent from weakly virulent strains of Yersinia species. The link between the presence of these proteins and virulence was strengthened by two additional observations: (i) highly pathogenic Y. pestis strains which spontaneously lost the ability to produce HMWP2 were much less virulent when injected subcutaneously into mice (10), and (ii) a mutation in the irp2 gene of Y. pseudotuberculosis caused the disappearance of both HMWP1 and a marked decrease in virulence (9). Despite the relationship between pathogenicity and the presence of the HMWP2, the role of these proteins is still unknown.

To gain insight into the role of HMWP2, the irp2 gene and its promoter region were sequenced. In this work, the nucleotide sequence of irp2, the deduced amino acid sequence of HMWP2, and some features of the irp2 promoter and of the structure of HMWP2 are reported. The existence of potential disulfide bridges in HMWP2 was examined, and the location of the protein within the bacterium is discussed. We show that HMWP2 displays significant homology to AngR of Vibrio anguillarum and shares similarities with a superfamily of adenylate-forming enzymes. A role for HMWP2 is proposed and discussed.

**MATERIAL AND METHODS**

**Bacterial strains and culture media.** The wild-type strain Ye8081 of Y. enterocolitica biotype 1B, serotype O:8, and two strains of E. coli, DH5α and JM101 (1), were used in this study. pIR2 is a pUC18 recombinant with an 8-kb ClaI insert containing the entire irp2 gene from strain Ye8081 (9). Phages M13mp18 and M13mp19 (1) were used for single-stranded DNA sequencing. The Yersinia strain was grown at 28°C in peptone broth, and E. coli was grown in Luria broth, in 2xYT, or on LB agar plates at 37°C (1). The chemically defined medium used for the extraction of the HMWP5 was prepared as described previously (11). Iron repletion or starvation was obtained by addition of 150 μM FeCl₃ or 50 μM α-α'-dipyridyl (Sigma Chemical Co.), respectively, to the chemically defined medium. Ampicillin (100 μg/ml) was added when needed.

**Cloning procedures and sequencing.** To sequence the irp2 gene of Ye8081 in pIR2, suitable overlapping inserts were obtained by different strategies. First, several small subclones were directly ligated into the replicative form of bacteriophage M13mp18 and M13mp19. Second, large inserts were treated with exonuclease III (Erase-a-Base system; Promega Biotec) to generate a set of smaller fragments (1) which were inserted in M13mp phage DNA. Third, oligonucleotides were synthesized for nonoverlapping regions, using a model 391 DNA synthesizer (Applied Biosystems, Inc., Foster City, Calif.). The nucleotide sequence of the two DNA strands was determined by the dideoxy-chain termination method (1) with use of a modified T7 DNA polymerase (Sequenase; United States Biochemical Corp.). Sequence ambiguities were resolved by using dITP instead of dTP. The nucleotide and amino acid sequences were analyzed with various softwares: DNA Strider (42), ClustalV (30), and the FastA, Bestfit, Motifs, and Pileup programs from the version 7 UNIX of the Genetics Computer Group package (18).

**DNA manipulations.** DNA electrophoresis, electroleution, ligation, and electroporation were carried out as previously described (13). Double-stranded plasmids were extracted according to the procedure of Birnboim and Doly (4a), and M13 single-stranded DNA was isolated by the method of Messing as described in reference 1.

**Primer extension.** Y. enterocolitica Ye8081 and E. coli (pIR2) were grown for 3 days in iron-rich or iron-depleted chemically defined medium, and total RNA was extracted by the hot phenol method of Von Gabain et al. (68), with some modifications (12). Primer extension experiments were performed with reverse transcriptase (Promega Biotec) (1). Three oligonucleotides were used as primers (Fig. 1): P1 (positions 241 to 270), P2 (445 to 469), and P3 (505 to 534). The DNA strands obtained after primer extension were electrophoresed alongside the DNA sequence reaction products obtained with the same primers.

**Cell fractionation and analysis of proteins.** To obtain total proteins, bacteria were washed three times in saline, and the pellet was heated to 100°C in a denaturing solubilization buffer (1). Sonication of the cells and membrane extraction were performed as previously described (8). When needed, the bacteria were suspended in 5 mM MgCl₂-25 mM K₂HPO₄ (pH 7) and incubated for 10 min at room temperature in the presence of either 30 mM (for membrane proteins) or 10 mM (for total proteins) iodoacetamide prior to sonication or extraction of total proteins. All samples were diluted in sodium dodecyl sulfate (SDS) solubilization buffer in the presence or absence of β-mercaptoethanol and heated for 5 min at 100°C prior to electrophoresis on an SDS-7.5% polyacrylamide gel. After migration, the gel was stained with Coomassie brilliant blue.

**Nucleotide sequence accession number.** The nucleotide sequence reported in this study has been submitted to GenBank under accession number L18881.

**RESULTS**

**Nucleotide sequence of the irp2 gene.** Sequencing of the irp2 gene revealed a large open reading frame (ORF) of 6,126 nucleotides (nt) (Fig. 1). The irp2 ORF was followed by the beginning of another ORF in the same reading frame (data not shown). There was no significant ORF in any other reading frame on the same DNA strand or on the complementary strand. The mol% G+C of the irp2 gene is 59.86%, which is higher than the overall value of 47 to 50 mol% G+C for the Y. enterocolitica chromosome (4).

A 10-bp inverted repeat sequence (IR1) is present 23 bp downstream of the TAG termination codon (Fig. 1). This sequence may form a stem-loop structure with a calculated ΔG (25°C) of -20.8 kcal (1 kcal = 4.184 kj) (64). The absence of a poly(T) stretch immediately downstream of the stem-loop structure and its relatively low energy suggest that it is not a typical rho-independent transcription terminator (46). In contrast, two perfect inverted repeats of 14 bp (IR2) and 12 bp (IR3), with calculated energies of −36.2 and −29.4 kcal, respectively, are located upstream of the irp2 gene and could form tight hairpin structures (Fig. 1). IR2 was followed by a stretch of three thyminide residues and may represent a rho-independent transcription termination signal for the ORF upstream of irp2.

**irp2 promoter region.** A potential promoter region composed of a −35 (TTGTTA) and a −10 (TATTAT) sequence separated by 19 bp and quite close to the canonical sequence of E. coli (27) was identified upstream of the first ATG codon, between nt 141 and 171 (Fig. 1). This region con-
FIG. 1. Nucleotide and amino acid sequences of the ipr2 gene. −10 and −35, promoter sequences; SD, Shine-Dalgarno ribosome-binding site; IR1 to IR3, nucleotide inverted repeats; Fur, Fur-binding sites; a, e and h, locations of termination of the primer extension products; DR1 to DR5, highly conserved stretches of amino acids repeated twice in HMWP2; H1 and H2, hydrophobic domains which may be membrane associated; LZ, possible leucine zipper; SBS, putative SBS sharing homology to the consensus sequence F(Y)XYLG(H/D)IS(L) which has been shown to bind the activated amino acyladenylate either directly or via the 4'—phosphopentehayne cofactor; P-loop, region containing the consensus sequence SGQTPXPKG.

tained two overlapping potential Fur boxes; Fur1 (nt 137 to 155) and Fur2 (nt 143 to 161) exhibit the dyad symmetry classically found in the Fur-binding sequences (45). The sequence AGGAGG 4 bp upstream of the first ATG codon (Fig. 1) might serve as a Shine-Dalgarno ribosome-binding site.

To map the transcriptional start point of the ipr2 gene, we performed a primer extension analysis using three different oligonucleotides (P1, P2, and P3) as primers (Fig. 1) and total RNA of E. coli(pIR2) grown under iron-rich or iron-poor conditions as the template. No extended cDNA product was found with primers P2 and P3, perhaps because they were too far from the transcriptional start point. With primer P1, extended products were obtained with RNA extracted from iron-starved E. coli only (Fig. 2A). This result is in accordance with previous data showing that the ipr2 transcript is synthesized under iron limitation but not iron repletion conditions in Yersinia spp. (12) and that the ipr2 gene is also iron regulated in E. coli (9). At least five different extended products (a to e) ending with an adenine residue were obtained (Fig. 2A). A potential promoter sequence was identified in the region from nt 141 to 171, located 17 bp upstream of extended products d and e, suggesting that this sequence may indeed function as an RNA polymerase-binding site in E. coli (Fig. 1). The results also suggested the presence of another RNA polymerase-binding site upstream of the a to e transcriptional start points. A potential promoter region composed of a −35 (TTAGCCA) and a −10 (TAAACC) hexamer, separated by 18 bp, was identified 19 bp upstream of sites a to e (nt 92 to 121) (Fig. 1). Although the degree of identity between this region and the canonical consensus sequence is low, the sequence seemed to be sufficiently conserved to be recognized by the E. coli RNA polymerase. Therefore, two potential promoter sequences were identified.
upstream of the irp2 gene; one is close to the consensus sequence of *E. coli* promoters, while the second is somewhat different.

Since transcriptional signals may differ in *E. coli* and *Y. enterocolitica*, and since transcription of the cloned irp2 gene may differ from that of its chromosomal counterpart, primer extension was performed with primer P1 and with total RNA from iron-replete or iron-starved *Y. enterocolitica* Ye8081. As expected, cDNA bands were present only when the mRNA of iron-starved bacteria was used as the template (Fig. 2B). As in *E. coli*, several transcriptional start points were observed, but only extended products a to c were common to *E. coli* and *Y. enterocolitica*. This could mean that the nonconsensus promoter sequence from nt 92 to 121 acts as an RNA polymerase-binding site in both *Yersinia* spp. and *E. coli* but the region from nt 141 to 171 is not functional in *Yersinia* spp. However, given the potential for hairpin formation in this region, it also is possible that products d and e represent strong pause sites for the RNA polymerase. Additional longer extended products (f to h) were also observed in *Y. enterocolitica* (Fig. 2B). Product h would correspond to a transcriptional start point at position 41, i.e., immediately upstream of the tight IR3 hairpin structure (Fig. 1), and product g would correspond to a start point approximately 200 bp upstream of nt 1 in Fig. 1. The end of extended product f was too distant to be correctly located. Whether the f to h transcriptional start points far upstream the irp2 coding sequence are used to generate a long messenger overlapping irp2 is not clear. Alternatively but less likely, these extended products may result from the hybridization of P1 to other iron-dependent *Yersinia* mRNAs sharing sequence homologies to P1.

**HMWP2 amino acid sequence.** Assuming that the first ATG codon following the potential Shine-Dalgarno sequence is the translational initiation codon, HMWP2 is a protein of 2,035 amino acids with a calculated molecular mass of 228,566 Da, somewhat higher than the molecular mass of 190,000 Da previously estimated by SDS-polyacrylamide gel electrophoresis (PAGE). Identification of the translation initiation codon could not be performed because the NH₂-terminal amino acid of the protein was blocked. The deduced isoelectric point of HMWP2 is 5.81, but the carboxy terminal of HMWP2 is very rich in arginine residues (7 of the last 12 amino acids [aa]).

Upon bacterial sonication and ultracentrifugation, the HMWPs are found predominantly in the pellet (membrane)
fraction (8). Analysis of the predicted NH₂-terminal amino acid sequence of HMWP2 did not reveal a potential signal sequence, suggesting either that the protein is not exported or that another export signal is used. The computer-derived hydropathicity profile of HMWP2 (38) displayed a succession of short hydrophobic and hydrophilic fragments (data not shown). Two relatively long hydrophobic regions, H1 (24 aa, located at positions 630 to 653) and H2 (22 aa; positions 1725 to 1746) were found and might correspond to membrane-anchored domains (Fig. 1).

One striking feature of the HMWP2 sequence was the presence of two approximately 550-aa direct repeats (DR) at each extremity of the protein (positions 3 to 547 and 1395 to 1919). These repeats are aligned in Fig. 3. They share 35.6% identical residues and 54.6% similar amino acids. Within these regions, five highly conserved sequences are found; DR1 (23 of 33 aa identical), DR2 (9 of 12), DR3 (11 of 13), DR4 (8 of 9), and DR5 (13 of 15) (Fig. 1 and 3).

**Disulfide bridges.** The deduced amino acid sequence of HMWP2 includes 30 cysteine residues. When the HMWP2s from sonicated cells are heated to 100°C in SDS buffer in the absence of a reducing agent, they do not enter the polyacrylamide gel, suggesting the existence of intermolecular disulfide bridges (8). However, the possibility that the disulfide bonds observed were artificially formed upon sonication, when the proteins were not in their natural environment, could not be excluded. To clarify this point, iron-starved cells of strain Ye801 were washed several times in saline, then resuspended in SDS dissociation buffer with or without β-mercaptoethanol, and heated to 100°C. The two samples were electrophoresed on a polyacrylamide gel alongside the corresponding membrane fraction obtained after sonication, centrifugation, and dissociation in SDS buffer with or without β-mercaptoethanol. When the membrane fraction was subjected to SDS-PAGE, the HMWP2s in the membrane fraction entered the gel only when heated with β-mercaptoethanol (Fig. 4, lanes 3), confirming the results of previous experiments (8). In our case, the absence of a reducing agent did not alter the migration of the two polypeptides present in the total bacterial extract (Fig. 4, lanes 1). These results could mean that the disulfide bridges formed only after sonication. To test this hypothesis, the cells were incubated with iodoacetamide prior to sonication or extraction of total proteins. Iodoacetamide penetrates the bacteria and alkyl-

FIG. 1—Continued.
lates free cysteine sulfhydryl groups, thus preventing the subsequent formation of artificial disulfide bridges between the free SH radicals upon sonication. Incubation of the bacteria with iodoacetamide prior to sonication permitted the HMWP s to migrate into the SDS-polyacrylamide gel, even in the absence of a reducing agent in the solubilization buffer (Fig. 4, lanes 4). The fact that migration of the HMWP s in membrane extracts of iodoacetamide-treated bacteria or in total protein extracts was not affected when β-mercaptoethanol was omitted from the dissociation buffer suggests that despite their high number, the cysteine sulfhydryl groups are not disulfide bonded in situ.

Homology of HMWP 2 to AngR from V. anguillarum. To determine whether HMWP 2 had significant homology with known protein sequences, a search in the Swiss-Prot data base was performed with the FastA program (44). The highest percentage of similarity or identity was found with the protein AngR of V. anguillarum (24). AngR is a 118-kDa protein which has been reported to activate the iron uptake system of V. anguillarum (50). When the overall amino acid sequences of HMWP 2 and AngR were compared, a stretch of 337 aa (1011 to 1347) in HMWP 2 was found to be completely absent from AngR. In addition, the 100 amino-proximal and the 550 carboxy-proximal amino acids of HMWP 2 are also absent from AngR (see Fig. 6). Otherwise, the complete sequence of AngR could be aligned with a contiguous segment of HMWP 2, and the two proteins exhibited a high percentage of similarity or identity (44.5% identical residues and 61.7% identical plus similar amino acids on a length of 1,071 aa; Fig. 5). Some segments of HMWP 2 and AngR are more similar than others: the central regions of the two proteins (residues 594 to 1010 in HMWP 2) exhibit over 66% identity, while the amino and carboxy flanking regions exhibit only 28 and 34% identity, respectively.

FIG. 1—Continued.

FIG. 2. Primer extension analysis performed with oligonucleotide P1 as the primer and total RNAs isolated from iron-replete (+Fe) and iron-starved (−Fe) E. coli(pIR2) (A) and Y. enterocolitica Ye8081 (B) as templates. Lanes A, C, G, and T correspond to the labeled nucleotides. a to h are the products of the primer extension experiments.
Amino acids 873 to 984 of AngR are similar to the helix-turn-helix domain of the DNA-binding domain of the Cro protein of *Salmonella* phage P22 (24). The corresponding region in HMWP2 (984 to 1005) is immediately upstream of the loop and shares 12 of 22 identical amino acids with AngR, including a central glycine residue (Fig. 5). The helix-turn-helix motif of HMWP2 has a basic calculated pI (9.37), as does the corresponding region in AngR and DNA-binding domains in prokaryotic regulatory proteins. However, the predicted score of 31 determined by the method of Dodd and Egan (21) for this segment of HMWP2 is much lower than that for helix-turn-helix motifs known to be involved in DNA binding.

**Homology between HMWP2 and a group of antibiotic synthetases.** A group of proteins involved in the nonribosomal biosynthesis of antibiotics also displayed significant homology to HMWP2. This group includes tetrocidine synthetase I (TY1) (69), gramicidin S synthetase I (GS1) (31, 37) and GS2 (32, 67) from *Bacillus brevis*, and the 5-(l-O-a-minoadipyl)-l-cysteinyl-d-valine synthetases (ACVSs) from *Penicillium chrysogenum* (20, 58), *Aspergillus nidulans* (41), *Cephalosporium acremonium* (26), and *Nocardia lactamurans* (15). All of these proteins belong to the same superfamily of adenylyl-forming enzymes (67).

The entire sequences of TY1 (69) and GS1 (37) were similar to the segment of HMWP2 from aa 500 through the 5' carboxy terminus, except for the 337-aa loop (Fig. 6). The entire sequence of GS2 (67) is composed of four repeated domains of about 600 aa, separated by nonhomologous sequences of approximately 500 aa. The homology between HMWP2 (without the 337-aa loop) and GS2 extends throughout the HMWP2 sequence (Fig. 6). However, when each of the four repeated domains of GS2 (da to dD) was aligned separately with HMWP2, the highest matches were found in each case with the central part of HMWP2 (approximately between residues 540 and 1470), and the percentage of amino acid identity was of the same order of magnitude (30.1, 33, 33.9, and 34.7% for da to dD, respectively; Fig. 6). Strikingly, this region corresponds mainly to the sequence between the two repeats in HMWP2, the region with the highest degree of identity to AngR (Fig. 6).

The ACVSs are very large enzymes (>400 kDa) involved in the biosynthesis of β-lactam antibiotics. All of the sequenced ACVSs contain three functional repeats of approximately 570 aa exhibiting significant homology to TY1, GS1, and GS2. The homology with the different ACVSs was found throughout the entire length of HMWP2 (Fig. 6). However, when the three repeats (da to dC) from the ACVS sequences were separately aligned with the HMWP2 sequence, they, like the four GS2 repeated domains, matched the central part of HMWP2 (Fig. 6).

The consensus sequence SGTGTGXPKG, which is highly conserved in this group of antibiotic synthetases, is present in each repeat of GS2 and the ACVSs (32). It has been suggested that this conserved stretch of amino acids may be a new class of phosphate-binding loop (P loop) (32). This sequence is also present in different enzymes involved in the catalysis of ATP-P, exchange. A highly similar sequence (SGSTGTPKG) was also found in HMWP2 (positions 713 to 721) and in AngR (positions 602 to 610), in the same position relative to the aligned sequences (Fig. 1, 5, and 6).

Recently, it was shown that the sites for covalent substrate amino acid binding (SBS) in each repeat of GS2 (52) share the same consensus sequence: IF(Y)XLLG(H,D) S(L,I). The serine residue present in this motif has been shown to bind the substrate, which is either the activated...
FIG. 5. Homology between HMWP2 and AngR performed with the Bestfit program (see the legend to Fig. 3). A loop of 337 aa (1011 to 1347) was withdrawn from the HMWP2 sequence in order to obtain the optimal alignment.

... amino acid or more probably the cofactor 4'-phosphopantetheine, which bears a sulfhydryl group that would be involved in thioester formation with the amino acid substrates and in peptide transfer (52). This consensus sequence, which is present in GS1, in TY1, and in the three repeated domains of the different ACVSs, was also found at positions 1432 to 1440 in the second direct repeat (DR) of HMWP2 and was well conserved (FQOOGGDSL) (Fig. 1 and 6). A related sequence, LIQAGLDIS, was identified in the first direct repeat at positions 45 to 53 (Fig. 3). While less similar to the consensus sequence, it includes the critical serine residue at the appropriate position. Curiously, the related sequence FFOLSSGDAY is present at positions 993 to 1001 in AngR but does not contain the primordial serine residue at position 8 in the consensus sequence (Fig. 5).

Homology of HMWP2 with EntF from E. coli. In addition to the group of antibiotic synthetases, HMWP2 shares significant homology with another member of the superfamily of adenylate-forming enzymes: the enterobactin synthetase component (F EntF) of E. coli (47). EntF is a 142-kDa protein involved in biosynthesis of the siderophore enterobactin. Like the antibiotic synthetases, EntF is involved in the synthesis of peptides through protein template mechanisms and shares sequence homologies with TY1 and GS1 (47). The entire EntF sequence can be aligned with aa 112 to 1803 of HMWP2 (except for the 337-aa loop section) (Fig. 6). The putative binding site for 4'-phosphopantetheine (FALGGLHS) in EntF (positions 999 to 1007) (47) matches with the corresponding sites in HMWP2 and the different synthetases (Fig. 6). Similarly, the potential P-loop sequence of EntF (STGSRGPKG) located at positions 604 to 612 aligns with the other P-loop sequences (Fig. 6).

Homologies and multiple alignment. As expected, HMWP2 also displayed homology with other members of the superfamily of adenylate-forming enzymes, including the amino adiopate-semialdehyde dehydrogenase (Lys2) of Saccharomyces cerevisiae (14), the acetyl coenzyme A (acetyl-CoA) ligases of A. nidulans (14), Neurospora crassa (14), and Methanothrix soehngenii (31), the 4-coumarate-CoA ligases of Petroserinia crispum (40) and Oryza sativa (71), the O-succinylbenzoic acid-CoA ligase (MenE) of Bacillus subtilis (22), the luciferase of Photinus pyralis (19), and EntE of E. coli (59). Multiple alignment of the amino acid sequences of these enzymes, of HMWP2, AngR, TY1, and EntF of each repeated domain of GS2, and of ACVS was performed with the Pileup program, using a gap weight of 3. Two main groups of proteins (A and B) were identified. Family A included all of the antibiotic synthetases, EntF, Lys2, AngR, and HMWP2, while family B was composed of luciferase, EntE, and the various ligases (Fig. 7). Some stretches of amino acids were highly conserved among the 30 sequences studied, while others were specific for family A (Fig. 7). Among the most conserved regions, we identified at least six consensus sequences (CS I to VI) defining the whole superfamily (Fig. 8). CS I corresponded to the new class of P loop which, by our criteria (see the legend to Fig. 7), differed slightly from that previously described. CS II to VI were absent from MenE because the protein sequence is short and ends before CS II. The TGD motif found within CS III is present in a wide variety of ATPases, in which it may function as a nucleotide-binding fold (63). Although the exact function of most of these amino acid stretches is still unknown, their high degree of conservation among the entire superfamily suggests that their presence is of prime importance. Within the superfamily, four other CSs (CS VII to X) are clearly specific for family A (Fig. 8). CS X, the known...
SBS, is absent from Lys2, which ended upstream of it. Interestingly, CS VII is highly conserved in all group A proteins except AngR and HMWP2 (Fig. 7). Although CS IX overlapped CS IV, it contained a significant number of residues which were not found in group B. Analysis of the multiple alignment thus suggested that HMWP2 and AngR belong to family A and share some common features which are not present among the other members of this family.

**Phylogenetic tree.** To determine the distance between the different proteins within the superfamily, a phylogenetic tree was constructed by using the neighbor-joining (distance) method of Saitou and Nei (49) from the ClustalV program. The octapeptide repeat antigen sequence (ORA) of *Plasmodium falciparum* (25) was used as an outgroup to root the tree. This sequence was chosen because it possessed the characteristic new class of P loop but was the most distantly related to the other members of the superfamily. As shown on Fig. 9, the existence of two families suggested by the multiple alignment was clearly seen on the phylogenetic tree. Although the size, number, and nature of the sequences used for the multiple alignment, and the program used for constructing our tree, were different from those reported by Turgay et al. (67), the two trees are very similar, suggesting that they are not fortuitous. With the exception of Lys2, the proteins or protein domains constituting family A are much larger than those in family B. This accounts for the absence

![Diagram of phylogenetic tree](https://example.com/diagram.png)
of the important SBS (CS X) in the latter group. However, length is not the only trait which distinguishes the two groups of proteins. Indeed, CS VII to IX, which are a hallmark of family A, are located in a region that is present in all of the proteins (except Hmne for CS VIII and IX) but differs drastically between the two families. Analysis of the phylogenetic tree confirmed that Lys2 is the most distantly related sequence in family A. It also confirmed that HMWP2 and AngR are very close to each other and belong to family A but, interestingly, constitute a specific branch which sets them apart from the antibiotic synthetases and from EntF (Fig. 9).

Other features of HMWP2. The 337-aa loop segment is almost completely absent from all of the proteins homolo-
CONSENSUS SEQUENCES DEFINING:

the superfamily of adenylate-forming enzymes:

\[
\begin{align*}
&I \quad (Y,T,R,S,Q,T,S,T,S)X PKG(V,I) \\
&II \quad Y(G,W)X TE \\
&III \quad (L,M,W)Y(T,B)QD(A,V,G) \\
&V \quad (V,C,A,G,A,Y,F) \\
\end{align*}
\]

the family A:

\[
\begin{align*}
\end{align*}
\]

**FIG. 8.** Consensus sequences derived from the alignment of HMWP2 with the 29 sequences that are most similar. Positions of these sequences are shown in Fig. 9. Letters in parentheses represent positions where conserved amino acids were alternatively found. Boldface letters indicate amino acids or association of amino acids highly conserved among the members of the superfamily (28 of 30) or of family A (21 of 22). CS I corresponds to the position of the new type of P loop, and CS X corresponds to the position of the activated SBS.

gous to HMWP2. A search in the protein data base did not reveal any important identity. The last 32 aa of the loop share some homology with the ACVSs, indicating that the loop may be of only 300 aa when aligned with this family of synthetases. The characteristic feature of this loop is the presence of a long leucine zipper-like sequence (LXGx6Lx6Lx6L) at positions 1251 to 1279 (Fig. 1). This motif has been shown to facilitate dimerization of many gene-regulatory proteins (6, 39). However, since proline residues, which are usually absent from leucine zippers, are present within the leucine-rich segment, and since leucine zipper-like sequences have been identified in a wide variety of proteins in which their role in dimerization has not been demonstrated, the presence of a leucine zipper-like sequence in HMWP2 should be interpreted with caution.

A search in the Prosite dictionary of A. Bairoch (University of Geneva), using the Motifs program from the Genetics Computer Group package, was performed to determine whether additional motifs of interest were present in HMWP2. Two motifs deserve attention: a sugar transport protein signature and an adipokine hormone family signature. The first, (L,I,L,V,M,S,T,D,E)X(L,L,V,M,F,A)GR(R,K)X(4,6)G, is a signature sequence found in members of a family of integral membrane proteins involved in sugar transport (50) which possesses the consensus pattern. A perfectly conserved pattern (IEFLGQRKDVKVQKV) was found in HMWP2 at positions 953 to 966 and also in AngR. Interestingly, with the exception of the expected R or K residue at position 7 of the sugar transport protein consensus sequence, this pattern is well conserved in family A, being part of CS IX described above (Fig. 7 and 8). The significance of this similarity is still unclear.

The second interesting motif is the consensus sequence Q(L,V)(N,T,Y)(F,Y)(S,T)XXW of the adipokine hormone family (51). These small hormones (8 to 10 aa) are produced by arthropods and cause the release of diglycerides from the fat body. A typical sequence (QLFTTEDW) was present at positions 422 to 429 in HMWP2. The relationship between these very small hormones produced by insects and the large iron-regulated protein of *Yersinia* spp. is unknown, but the fact that no other protein sequence bearing this motif has been detected hitherto in the Swiss-Prot library should be noted.

**DISCUSSION**

The *irp2* gene of *Y. enterocolitica* coding for the iron-regulated protein HMWP2 is an ORF of 6,126 nt encoding a protein of approximately 228,000 Da. A typical ribosome-binding site and two potential -35 and -10 promoter regions were identified upstream of this ORF. Two Fur-binding sites overlap the second -35 region, in line with our previous observation that *irp2* is under the control of the Fur repressor in *E. coli* (9). The Fur2 box was the most highly conserved, with 16 of 19 nt (84.2%) identical to the *E. coli* consensus sequence (70). Fur boxes have also been identified in the promoter regions of the *foxA* gene coding for the ferrioxamine receptor (3), the *foxC* gene encoding the ferri-chrome receptor (35), and the hemin uptake operon (*hem*) of *Y. enterocolitica* (62) and of the *fur* gene of *Y. pestis* (60).

The level of nucleotide conservation between Fur2 and the three 19-bp-long Fur boxes of *foxA*, *fcuA*, and *hemP* is high (15 of 19, 14 of 19, and 12 of 19 conserved nucleotides, respectively), indicating that the Fur recognition signal is well conserved among different iron-regulated genes of *Yersinia* spp.

The beginning of a second ORF was found immediately downstream of and in the same reading frame as the HMWP2 coding sequence. Since previous results suggested that the genes encoding HMWP1 (*irpl*) and HMWP2 form an operon with *irp2* located upstream of *irpl*, the ORF downstream of *irp2* may correspond to the *irpl* gene.

The HMWP2s were previously found in the particulate (membrane) fraction of sonicated cell extracts (8). We show here, however, that the HMWP2s may form large aggregates with intermolecular disulfide bonds when cells are sonicated. This observation, together with the absence of a typical signal peptide from the N-terminal region of HMWP2, raises doubts about previous conclusions that the HMWP2s are located in the outer membrane (8). However, HMWP2 does have two regions of relatively high hydrophobicity that might span the cytoplasmic membrane. The subcellular location of HMWP2 is currently under investigation.

A characteristic feature of HMWP2 is the presence of two 550-aa-long repeated domains at each extremity of the protein, suggesting a duplication of a fragment of *irp2* during evolution. The existence of five highly conserved stretches of amino acids (DR1 to DR5) in each repeat may indicate that they are important for the function, conformation, or location of HMWP2. Curiously, the region between these two large repeats is the segment that is the most highly conserved among related proteins and corresponds to the domains that are repeated three times in the ACVSs and four times in GS2. The presence of two repeats flanking the central conserved region and of a loop of more than 300 aa located within the conserved region but absent from the other family members are characteristic features unique to HMWP2. This special organization might confer some specific properties upon HMWP2.

A very high degree of similarity or identity was found...
between the entire lengths of AngR and HMWP2. In addition to their sequence homology, HMWP2 and AngR share other properties: (i) their synthesis is stimulated under iron deprivation, and Fur boxes are present upstream of their genes (9, 24); (ii) AngR and HMWP2 are both acidic (pI 6.34 and 5.81, respectively); and (iii) mutations in angR or irp2 reduce virulence (9, 57). The high degree of similarity or identity found between HMWP2 and AngR might be considered surprising since *Y. enterocolitica* and *V. anguillarum* are distantly related. However, it was shown recently that the iron transport system mediated by plasmid pJM1 of *V. anguillarum* shares significant homology with those found in members of the family *Enterobacteriaceae* (36). Furthermore, Koebnik et al. (35) recently reported that the ferri-chrome receptor (FcuA) of *Y. enterocolitica* shares sequence similarity with several other siderophore receptors and especially with the anguibactin receptor (FatA). Therefore, it appears that the iron-regulated genes from these two organisms are much closer than expected from their phylogenetic distance.

AngR has been reported to activate expression of genes coding for the iron uptake system of *V. anguillarum* (50). The genes thought to be under the control of AngR code for the siderophore anguibactin, for the anguibactin receptor, and for a protein (p40) involved in the transport of iron into the cell (16). A second activator (Taf) is also reported to activate transcription of these genes (65). The loci coding for AngR (*angR*), for Taf (*taf*), and for the iron uptake system are located on the 65-kb plasmid pJM1 (50). However, the large size of AngR, the existence of the new class of P loop, and the presence of a gene coding for a putative S-acyl fatty acid thioesterase immediately downstream of *angR* (24) are inconsistent with a gene-regulatory function and are more consistent with a role in anguibactin biosynthesis (2). This hypothesis was strengthened by a recent report indicating that the cloned *angR* gene is able to complement an *E. coli* strain mutated in *entE* (66). Interestingly, our results and those of Turgay et al. (67) indicate that AngR and EntE belong to the same superfamly.

Moreover, the results of the multiple alignment and the phylogenetic tree show that HMWP2 and AngR may be included in family A of the superfamly of adenylate-forming enzymes (67). Except for Lys2, which is the most distantly related protein and which is too short to contain the SBS, family A is composed of different synthetases which can be divided into three groups. The first group includes the synthetases TY1, GS1, and GS2 from *B. brevis*. TY1 (115 kDa) and GS1 (126 kDa) both activate the first phenylalanine amino acid, racemize it, and form an amino acid-adenylate-enzyme complex. The aminoacyl moiety is transferred and covalently bound to another site (the SBS) on the same enzyme (34) and subsequently transferred to the second thioester-linked amino acid covalently bound to another SBS on GS2 (530 kDa). Binding of the last three amino acids also takes place on GS2. The mechanism of tyrocidine elongation is similar to that of gramicidin S except that two enzymes (TY2 and TY3) are involved. The second group of synthetases is composed of different ACVSEs found in eukaryotic
filamentous fungi and bacteria. They are involved in the first step in the biosynthesis of β-lactam antibiotics, formation of the tripeptide β-(L-α-aminoadipyl)-L-cysteinyl-D-valine. The third group contains only EntF, an E. coli protein involved in biosynthesis of the siderophore enterochelin, a macrocyclic trimer of 2,3-dihydroxybenzoic acid. EntF is associated with three other proteins (EntE, EntG, and EntH) to form a large complex which carries out a series of reactions with enzyme-bound intermediates (55). EntF itself activates L-serine via an L-serine-AMP intermediate (47) and contains a covalently bound 4'-phosphopantetheine cofactor (47). All of these enzymes, and even individual repeated domains within each enzyme, possess the new class of P loop, the SBS, and the additional eight consensus sequences.

The existence in HMWP2 and AngR of the typical new type of P loop and of the four additional consensus sequences defining the superfamily of adenylate-forming enzymes strongly suggests that they activate their substrate via an ATP-dependent process, leading to the formation of an acyladenylate-enzyme complex. Although AngR belongs to family A, the typical sequence of the potential SBS defining this family suggests that it might not be functional and may explain why AngR is functionally close to EntE, which belongs to family B. In contrast, the presence of a perfectly conserved SBS in HMWP2 suggests that the protein may catalyze the transfer of the aminoacyl moiety from the adenylyl site to the SBS, yielding an activated amino acid covalently bound either directly to the protein itself or more likely to the thiol group of 4'-phosphopantetheine cofactor.

Although the idea is entirely speculative for the moment, we propose that HMWP2 may belong to a multienzyme complex involved in the nonribosomal synthesis of a peptide. HMWP1, whose gene (irp1) probably forms an operon with irp2 (9), would combine with HMWP2 to form an enzymatic complex in a manner similar to that for GS2 with GS1, TY2 and TY3 with TY1, or EntE, EntG, and EntH with EntF. The resulting product could be either an antibiotic, a siderophore, or a new class of peptide. The fact that production of HMWP1 and HMWP2 is iron regulated would be consistent with their involvement in siderophore synthesis. A new class of siderophore (yersiniabactin) which is probably a catechol-like iron chelator has been reported in Yersinia spp. (28, 29). Like HMWP1 and HMWP2, this molecule is chromosomally encoded and is synthesized only by highly pathogenic Yersinia spp. An attractive hypothesis would be that the HMWP2s are enzymes involved in yersiniabactin biosynthesis. Preliminary attempts to distinguish between the parental Y. pseudotuberculosis strain IP2790 and its derivative mutated in irp2 (9) on siderophore indicator medium (53) were inconclusive (7). Moreover, the strain of Y. pseudotuberculosis serotype II studied by Heesemann et al. was shown to produce yersiniabactin (29), while we found that none of five strains of this serotype harbored the irp2 gene (17). These results do not eliminate the possibility that the HMWP2s are involved in yersiniabactin synthesis, but the relatively distant relationship of HMWP2 to EntE, EntF, and the antibiotic synthetases in the phylogenetic tree and its unique molecular organization may also indicate that it has a novel activity that is different from those of the well-characterized enzymes in the superfamily.

ACKNOWLEDGMENTS

We thank members of the computer service of Institut Pasteur (M. Keller, B. Caudron, F. Chauveau, and F. Tekaija) for advice during the computer analysis and J. D’Alayer (microsequencing laboratory) for his attempt to perform gas-phase sequencing of HMWP2. We also thank Annie Guiyoute for helping in reading the sequencing gels.

This work was supported in part by grant CRE 920604 from the Institut National de la Santé et de la Recherche Médicale.

REFERENCES


