Molecular characterization of PcpA: 
a novel choline-binding protein of *Streptococcus pneumoniae*

Ana R. Sánchez-Beato, Rubens López, José L. García

Department of Molecular Microbiology, Centro de Investigaciones Biológicas (Consejo Superior de Investigaciones Científicas), Velázquez 144, 28006 Madrid, Spain

Received 31 March 1998; revised 6 May 1998; accepted 10 May 1998

Abstract

The gene *pcpA* that encodes a novel pneumococcal choline-binding protein has been cloned and characterized. Northern blot analysis revealed that *pcpA* is expressed during the exponential phase of growth of pneumococci as a monocistronic transcript of about 2.3 kb. The transcription start site has been located 132 bp upstream of the start codon and the proposed −35 and −10 boxes that are highly similar to those of the typical σ70 promoters from *Escherichia coli*. This gene encodes a putative 79 kDa protein that contains a typical C-terminal choline-binding domain (ChBD). The ChBD of PcpA is built up by 11 identical motifs of 20 amino acids plus a tail of 19 amino acids, which represents the longest ChBD that has been characterized so far. Interestingly, two tandem arrays of five characteristic amphipatic leucine reach repeats (LRRs) of 22–26 amino acids in length have been found in the N-terminal region of PcpA. Since LRRs have been proposed to be involved in protein-protein and protein-lipid interactions our finding suggests a role for PcpA in pneumococcal adhesion.

Keywords: Adhesin; Choline-binding domain; Leucine-rich repeat; Pneumococcus

1. Introduction

*Streptococcus pneumoniae* is the major pathogen causing pneumonia and a leading etiologic agent of meningitis and otitis media. Although the mechanism whereby pneumococci cause death is still unclear several virulence factors, e.g. the capsular polysaccharide, pneumolysin, neuraminidase, autolysin and proteases, are thought to contribute to the final outcome of the disease [1]. In addition, it has been recently shown that adherence of pneumococci to extracellular matrix proteins, such as fibronectin, could play a role in the invasion of injured epithelia [2]. Since *S. pneumoniae* is a Gram-positive bacterium containing a thick cell wall, surface proteins in contact with the environment should interact with peptidoglycan, and teichoic or lipoteichoic acids. This class of proteins includes choline-binding proteins which interact with choline residues of the pneumococcal envelope. Pneumococcal autolysins (LytA amidase and glucosaminidase) [5], the surface-binding of PspA [6] and the recently described surface proteins CbpA [3] and SpsA [4] are members of this protein family. These proteins contain a choline-binding domain (ChBD), that is composed of...
4 to 10 repeated units each about 20 amino acid long. The ChBD appears to be an interchangeable module that can be transferred among the pneumococcal strains and its phages as well as to other bacteria as *Clostridium* [5].

Taking into account that the ChBDs are homologous we used specific DNA probes to identify a new member of the choline-binding protein family.

2. Materials and methods

2.1. Bacterial strains, phage, plasmids, and growth conditions

*Escherichia coli* TG1 (Amersham) and *E. coli* C600 [7] have been utilized as hosts for genetic transformation. The *S. pneumoniae* strains used were the strain R6, and the autolysin-deficient mutants M31 [8]. Plasmids used were: pGL80 [9], pKDE543 [10], and pUC18 (Pharmacia). *E. coli* was grown in Luria-Bertani (LB) medium at 37°C with shaking [7]. *S. pneumoniae* was grown in C medium supplemented with yeast extract (0.8 mg/ml; Difco Laboratories) (C+Y medium) at 37°C without shaking [8]. Growth of *S. pneumoniae* was monitored with a Coleman nephelometer (1 N = 7.5×10⁵ CFU/ml).

2.2. Recombinant DNA techniques

Standard techniques were used for the preparation and analysis of plasmid DNA [7]. DNA fragments were purified by using the Geneclean II kit (Bio 101). The 1.2 kb *Hind*III fragment of pGL80 containing the *lytA* gene was used as the *lytA* probe. The C-CSPA probe, which encodes the ChBD of the CspA protein of *Clostridium beijerinckii*, corresponds to the *Pvu*II-*Ssp*I fragment of plasmid pKDE543 [10]. Probes were labelled with [α-³²P]dTTP and hybridizations were performed as described [7]. All primers for PCR and nucleotide sequencing were synthesized on a Beckman Oligo 1000M synthesiser. RNA was extracted following a procedure previously described [11]. Northern blot analyses were performed using a vacuum blotter (Bio-Rad Laboratories) and nylon membranes (Schleicher and Schuell) [7]. RNA standards for estimating the size of mRNAs were from Promega. Pulsed field gel electrophoresis was carried out as previously described [11]. Primer extension was performed as previously described [11] using the oligonucleotide OEK5 (5‘-ATAACCGCAGCTGTA3’) as primer. DNA sequencing was carried out on supercoiled plasmid DNA by the dideoxy chain-termination method using the Pharmacia T7 sequencing kit. Computer analyses were performed using the programs provided by the servers of the National Institute for Biotechnology Information and the Baylor College of Medicine-Human Genome Center. The nucleotide sequence of *pcpA* has been deposited in the GenBank/EMBL data banks under the accession number Z82001.

3. Results

3.1. Identification and cloning of DNA fragments encoding choline-binding proteins

Although the pneumococcus contains several choline-binding proteins, only two genes of this family, *lytA* [9] and *pspA* [6] had been identified when this work was started. Based on the homology found between the ChBDs previously sequenced a Southern blot hybridization was performed using the *lytA* gene as a probe to identify novel pneumococcal genes encoding ChBDs. Surprisingly, using this approach we were able to detect only the *lytA* gene when tested either at low or at high stringency (data not shown). Therefore, a different experimental approach was attempted using the DNA fragment encoding the ChBD (C-terminal region) of the *cspA* gene of *C. beijerinckii* as a probe (C-CSPA probe) [10]. This probe hybridized with two *Hinc*II fragments of about 4.4 and 2.2 kb, respectively (data not shown). Interestingly, none of these bands had the expected size of the *lytA* gene as deduced from the reported sequence [5]. Moreover, a Southern blot analysis carried out using M31 DNA, that lacks the *lytA* gene, also showed the same two bands (data not shown) discarding the possibility that they could correspond to this gene. Although the largest band might be ascribed to the *pspA* gene [6], these results demonstrated that at least one of the hybridization bands should correspond to a new gene probably encoding a ChBD.

To analyze the open reading frames (orfs) present
Fig. 1. Genetic map and nucleotide sequence of the chromosomal region of \textit{S. pneumoniae} containing the \textit{pcpA} gene. Panel A: Open reading frames (\textit{orf}s) are represented by arrows indicating the direction of transcription, and their locations are shown relative to those of some relevant restriction endonuclease sites, i.e. \textit{Ev}, \textit{EcoRV}, \textit{Hc}, \textit{Hin}cII, and \textit{X}, \textit{XbaI}. Panel B: Nucleotide and derived amino acid sequence of the noncoding sequences and the sequence of the 5'- and 3'-end coding regions of the cloned fragment. The complete sequence has been submitted to the database (accession number Z82001). Short arrows indicate the direction of transcription. Asterisks indicate the stop codons. The putative ribosome binding site (RBS) is underlined. The —35 and —10 boxes of the \textit{pcpA} promoter are doubly underlined. The transcription start site is indicated by a boldface letter and the symbol (+1). The potential rho-independent terminator of the \textit{pcpA} gene is underlined with a discontinuous line. Direct (D1, D2, and D3) and inverted (I1, and I2) repeats are marked by arrows. The IR sequences of IS1381 are boxed. Abbreviations: aa, amino acids; n, nucleotides.
in these hybridizing fragments a pneumococcal Hind-cII library constructed in E. coli TG1 using SmaI-digested pUC18 vector was screened with the radioactively labelled C-CSPA probe. This analysis rendered one positive plasmid, named pEK1 (7.1 kb), which was further characterized. The nucleotide sequence of the cloned fragment revealed the presence of four orfs (Fig. 1). Comparison of the deduced amino acid sequences encoded by these orfs with the proteins available at the data banks showed that orf3a and orf3b shared homology with a truncated transposase of IS381 (accession number Z77725), whereas those encoded by orf1 and orf2 did not display any significant similarity with other known proteins. The largest orf, i.e. orf4 (hereafter named pcpA), encoded a protein containing a typical ChBD consisting of 11 identical motifs of 20 amino acids plus a tail of 19 amino acids. The motifs

A

Chaoptin 409 5GDNNDPTNLQTLRMTFMVEWMMRSLISRLGSVGEFDKDF-GVELEDLQTR-ASLSG 467

PcpA 130 FSSITIPSSIKQKGKESGSKARTTIDFDGASKXLEKIEDRAFD-PSELEBEILIF--ASLEY 187

Fshr 25 CSNGVFLQCSKVTEIPDPSANVLEFVLTKLAAPKGASGPELDEKIEISQNDVLE 85

Chaoptin 460 IQSHAFKHVRGLKRLPFSENQ---ISIIENDAFHEMRIVISLSKMSHGYSGERSSLAPFRLHER 527

PcpA 188 IGTASAFSGQKLKFTSFSSSLKESAFANLHLE-KLTLTPK---SVKTLGSLFRUT 245

Fshr 86 IEANVSNLPKLHEIRIEKANNLLYIDPOAFQNLPHNRLLYLISNT-VGKHLPA-VHK-E 142

Chaoptin 528 SLQELDFSNHHISMSDTFPHLKRLHELDHNR----------IEQVSLGTQFGDIHSLK 581

PcpA 246 SLNMLMLGNGIVASVQGVSQKTQLYVYPSQKQNWESYKTPE-ETKELASYSFNKSLGKLK 305

Fshr 143 SLQKVLNLQGINITHYVRSNFML-SFSEMILWLSK-----GIREIHCAFNL-GTFQ 197

Chaoptin 562 ISLRFNH-LTSISQHTFPDFDLAERKTHLDDNOKIERAFMNQNLDELEYLSLRGKINNL 641

PcpA 306 LELNEG--LQKGTDFADATKLEELISLESLF--NSLETI 338

Fshr 198 LNLSDNNDLLEELNDVFGAS--GPVIL-DISRTRHLSFYSGLNKDLRAST--YNLK 253

B

I 143 Q K K G F H G S - K A K T I I F D K G SQ L E K I 166

II 167 E D R A F D F S E - L E E I E L P A S - L E Y I 188

III 189 G T S A F S F S Q K L K K L T F S S S S K L E L I 213

IV 217 S H E A F A N L S N L E K L T L P F K S - V K T L 236


VI 293 A S Y F S F N K N S Y L K K L E L - - N E G L Q K I 315

VII 316 G T F A F A D A T K L E E I S L P - N S L E T I 338

VIII 339 E R L A F Y G N L E L K E L L L F D N - V K N F 360

IX 361 G K H V M N G L P K F - - L T L S G N N I N S L 383

X 384 P S F F L S G V L D S L K E I H I K N K ST E F S V 409

CONSENSUS G S F A F S G S T K L K E L T F G N S L L E S

Fig. 2. Sequence comparison of PcpA, chaoptin, and follicle-stimulating human receptor Fshr and proposed amphipatic leucine reach repeats (LRRs) of PcpA. Panel A: Protein alignment of PcpA, chaoptin and Fshr. Vertical lines and colons indicate identical and conserved residues, respectively. Dashes indicate gaps introduced to optimize the alignment. Panel B: The ten LRR repeats (I-X) of PcpA are aligned. A putative LRR consensus repeat is shown in the bottom. Numbers indicate amino acid positions. The amino acids conserved in at least five repeats are underlined and indicated by boldface letters.
showed a nearly perfect identity at the nucleotide level with the sole exception of the codons corresponding to the amino acids Arg and Gly that varied in few occasions between AAT or AAC, and GGT or GGC, respectively. The \( \text{pcpA} \) gene appeared preceded by a consensus ribosome-binding site (AAG-GA) located 7 bp upstream of the ATG codon. PepA was found to contain 708 amino acid residues resulting in a predicted \( M_r \) of 79 000 and a calculated \( pI \) of 9.3. The deduced polypeptide showed a high content in threonine and serine (17%) almost evenly distributed within the 469 amino acids of the N-terminal region preceding the ChBD. The peculiar hydrophobic profile of the C-terminal region of PepA could be ascribed to properties of the 11 repeated motifs of the ChBD that extend from Gly-470 to Ile-708, whereas that of the N-terminal region was more complex. Just following the N-terminal peptide (Met-1-Val-14) resembling the structure of a typical signal peptide, we observed two identical motifs of 10 amino acids (ILAAYVPNEP). The hydrophobic profile of the N-terminal region also revealed the presence of 10 repeated motifs each about 22–26 amino acid long (Fig. 2B). The motifs are organized in two blocks of five repeats, block 1 (Gln-143-Val-260) and block 2 (Ala-293-Val-409) connected by a

---

**Fig. 3.** Transcriptional analysis of the \( \text{pcpA} \) gene. Panel A: Northern blot analysis was carried out using total RNAs from \( S. \) pneumoniae R6 grown in C+Y medium that were extracted at 168 N (lane 1), 344 N (lane 2), 800 N (lane 3) and 904 N (lane 4). The filters were probed with the labelled \( \text{XbaI-EcoRV} \) fragment of \( \text{pcpA} \) (see Fig. 2). Sizes of the molecular mass markers are indicated at the left. Panel B: Mapping of the transcription start site of \( \text{pcpA} \) by primer extension analysis. RNA was isolated from \( S. \) pneumoniae R6 grown in C+Y medium at 344 N (lane 1) and 800 N (lane 2). The \( \text{pcpA} \) start site was determined by comparison of the extended cDNA with the sequence of plasmid pEK1. The elongation and the sequence were carried out using the same oligonucleotide OEK5, therefore the position of the transcriptional start site can be read directly from the gel.
strong hydrophilic peptide (Asp-261-Leu-292) (Fig. 2B). These repeats were very rich in Leu and other hydrophobic amino acids (Ile, Val and Phe) which alternated with hydrophilic residues conferring these motifs an amphipatic character. This structure resembled that of superfamily of proteins with leucine-rich amphipatic motifs (LRRs) [12]. A comparative analysis in the protein data banks showed that this region was highly similar to chaoptin (68% similarity, 24% identity), a surface glycoprotein found in the photoreceptor cells of the flies [13] and to the gonadotrophic receptors such as the follicle-stimulating human receptor Fshr (67% similarity, 23% identity) (Fig. 2A) [14]. This comparison can be also extended to other proteins containing LRRs (data not shown).

3.2. Transcriptional analysis of pcpA

To determine whether the pcpA gene is expressed in pneumococcus, total RNA was isolated from S. pneumoniae R6 at various times of the growth curve, and Northern blots of the gels were hybridized with the XbaI-EcoRV fragment of the pcpA gene (Fig. 3A). The results suggested that pcpA was transcribed during the exponential phase, but not during the stationary phase of growth. The size of the transcript was about 2.3 kb which agreed with the length calculated according to the proposed promoter and terminator elements of pcpA (Fig. 1). This result also suggested that pcpA was expressed as a monocistronic gene.

The transcription start site of pcpA was determined by primer extension using total RNA extracted from pneumococcus at two different times during the exponential phase of growth. This analysis revealed that the transcription start site was located 132 bp upstream of the start codon (Fig. 3B). Upstream of this site we found the putative −35 (TTGACt) and −10 (TATAcT) boxes of the pcpA gene that appeared to be highly similar to those of the consensus E. coli σ70 promoter. Just upstream of the −35 box we found a rare structure built up by three direct and two inverted repeats that might form a hairpin-loop with a ΔG of −41.2 kcal/mol (Fig. 1). Interestingly, a putative rho-independent terminator (ΔG = −14.9 kcal/mol) was found 87 bp downstream of the TAG codon which could define the limit of the transcriptional unit, in agreement with the data obtained by Northern analysis.

To determine the location of the pcpA gene within the chromosome of pneumococcus, the DNA prepared from strain R6 was digested with the restriction enzymes Smal, SacII or PspA and the resulting fragments were resolved by pulsed field gel electrophoresis and hybridized with the XbaI-EcoRV fragment of the pcpA gene (Fig. 1). The hybridization profile revealed that the pcpA gene is located in the same Apai, SmalI and SacII fragments that contain the genes hexA and malM [20] (data not shown). This finding shows that the pcpA gene is located in the chromosome far from the lytA and pspA genes [15].

4. Discussion

The use of an heterologous probe prepared from the cspA gene of C. beijerinckii allowed us to isolate the pneumococcal gene pcpA that represents a novel member of the ChBD family. PcpA is a large protein that, as all the other choline-binding proteins described so far, appears to be composed by a two-domain structure. The C-terminal domain is built by 11 identical repeats of 20 amino acids and a tail of 19 residues derived from a degenerated motif. The perfect identity of the repeated motifs is extremely surprising and poses many questions on the evolution of this particular ChBD. This identity allows us to discard the possibility that the ChBD encoded by pcpA was acquired by a direct interchange with that of the lytA, pspA or the other clostridial genes encoding ChBDs, since all the known ChBDs show some variability within their repeats. Moreover, the finding that the identity of the motifs at the nucleotide level was also nearly perfect suggests that the acquisition of this ChBD should be very recent. Therefore, it can be concluded that the ChBD of PcpA has been the result of successive duplications of a sequence of 60 nucleotides.

To establish whether the pcpA gene was essential for the growth of pneumococcus we have disrupted the pcpA gene by homologous recombination as well as increased its copy number by cloning the gene into a multicopy plasmid. However, these modifications did not produce any noticeable change in the
pneumococcal phenotype (data not shown), suggesting that this gene is not essential for its growth at least under laboratory conditions.

Although the function of PcpA remains still unknown the presence of the LRRs in its N-terminal domain might provide some clues on its role in pneumococcus. The LRRs were discovered in the glycoprotein α2, a protein of the human serum of unknown function [16]. The proteins that contain LRRs form a superfamily with different functions and subcellular localizations and are present in a great variety of organisms [12]. In spite of this functional diversity, all these proteins are involved in protein-protein interactions and many of them appear to act as signal transducers. The high similarity found between the PcpA and the chaoptin of Drosophila, a protein involved in cellular adhesion [13], suggests that the pneumococcal protein could also act as an adhesin. The existence of several LRR proteins capable to bind different components of the extracellular matrix [12] makes feasible the hypothesis that the LRRs of PcpA could facilitate similar interactions. On the other hand, it is worth noting that the LRR motifs have been only described in a few number of bacterial proteins [17–19] and some of them exhibit a function related to recognition of surface proteins of the extracellular matrix of eukaryotic cells.

Our knowledge on the surface proteins of pneumococcus is still very limited. Apart from the two pneumococcal autolysins [5], only four proteins, PspA [6], PsaA [20], CbpA [3] and SpsA [4], have been characterized as pneumococcal surface proteins. Taking all our data together, we propose that PcpA could be a surface protein of pneumococcus, that is bound to the choline-containing teichoic or lipoteichoic acid of the cell envelope through the C-terminal choline-binding domain, whereas the N-terminal repeats could be involved in cell adhesion through the interaction of its LRRs with specific proteins of the human extracellular matrix.

Acknowledgments

We thank P. García and E. García for helpful comments. We also thank to E. Cano and M. Carrasco for their technical assistance and A. Hurtado for the art work. This work was supported by Grants PB93-0015-C02-01 from the DGICYT.

References


