A *Pneumocystis carinii* multi-gene family with homology to subtilisin-like serine proteases

Elena B. Lugli, Andrew G. Allen† and Ann E. Wakefield

Author for correspondence: Ann E. Wakefield. Tel: +44 1865 222344. Fax: +44 1865 222626.
e-mail: wakefi@worf.molbiol.ox.ac.uk

Copies of a multi-gene family, named *PRT1* (protease 1), encoding a subtilisin-like serine protease were cloned from the opportunistic fungal pathogen *Pneumocystis carinii*. Comparison of the nucleotide sequence of a genomic clone and a cDNA clone of *PRT1* from *P. carinii f. sp. carinii* revealed the presence of seven short introns. Several different domains were predicted from the deduced amino acid sequence: an N-terminal hydrophobic signal sequence, a pro-domain, a subtilisin-like catalytic domain, a P-domain (essential for proteolytic activity), a proline-rich domain, a serine/threonine-rich domain and a C-terminal hydrophobic domain. The catalytic domain showed high homology to other eukaryotic subtilisin-like serine proteases and possessed the three essential residues of the catalytic active site. Karyotypic analysis showed that *PRT1* was a multi-gene family, copies of which were present on all but one of the *P. carinii f. sp. carinii* chromosomes. The different copies of the *PRT1* genes showed nucleotide sequence heterogeneity, the highest level of divergence being in the proline-rich domain, which varied in both length and composition. Some copies of *PRT1* were contiguous with genes encoding the *P. carinii* major surface glycoprotein.

**Keywords:** *Pneumocystis carinii*, subtilisin-like serine protease, *PRT1*, multi-gene family, major surface glycoprotein

**INTRODUCTION**

The fungal pathogen *Pneumocystis carinii* causes potentially fatal pneumonia in the immunocompromised, including those receiving immunosuppressive therapy for organ transplantation, those with advanced malignancy and in particular those with HIV infection. The lack of an effective *in vitro* culture system still remains a major obstacle in the understanding of the biology of *P. carinii* and its interactions with its host. Molecular techniques have been employed in the study of the organism, and a number of genes have now been cloned. Among these is the multi-gene family encoding the major surface glycoprotein (MSG or gpA) of the parasite. In this paper we describe the cloning and characterization of a second *P. carinii* multi-gene family, *PRT1* (protease 1), some copies of which are contiguous with MSG.

The *P. carinii* major surface glycoprotein is highly mannosylated and is antigenically distinct in organisms isolated from different mammalian host species (Lundgren *et al.*, 1991; Gigliotti, 1992). The MSG multi-gene family has been identified in the genome of *P. carinii f. sp. carinii* (rat-derived *P. carinii*) (Kovacs *et al.*, 1993; Wada *et al.*, 1993; Sunkin *et al.*, 1994), *P. carinii f. sp. mustelae* (ferret-derived *P. carinii*) (Haidaris *et al.*, 1992; Wright *et al.*, 1995), *P. carinii f. sp. hominis* (human-derived *P. carinii*) (Stringer *et al.*, 1993; Garbe & Stringer, 1994) and *P. carinii f. sp. muris* (mouse-derived *P. carinii*) (Wright *et al.*, 1994). The different copies of the *P. carinii f. sp. carinii* MSG genes are of similar size but heterogeneous in sequence. They have been found on multiple chromosomes and often organized in tandem arrays. The majority of MSG genes are located in the subtelomeric regions of the *P. carinii f. sp. carinii* chromosomes (Underwood *et al.*, 1996; Sunkin & Stringer, 1996). The expression of MSG genes has been shown to be mediated by the upstream conserved sequence (UCS) which is found on a single chromosome situated in the subtelomeric region (Wada *et al.*, 1995). Different copies of the MSG multi-gene family have been shown to be linked to the UCS. It has been postulated...
that this differential expression of MSG may represent a strategy to evade the immune response of the host by antigenic variation (Sunkin & Stringer, 1996).

The genes comprising the novel P. carinii PRT1 multigene family described in this paper show high levels of homology with subtilisin-like serine proteases. These are a group of endoproteases which have been characterized from a wide variety of organisms including bacteria, fungi and higher eukaryotes. Some have been found to function in the specific endoproteolytic processing of pro-proteins at cleavage sites of paired basic amino acid residues, to generate regulatory proteins in a mature and biologically active form. The pro-hormone processing enzyme kexin, encoded by the KEX2 gene of Saccharomyces cerevisiae, has been characterized and found to cleave the precursors of the α-mating-factor and the killer toxin (Fuller et al., 1989). Genes encoding a similar processing endoprotease have been identified in a number of other fungi, the KEX1 gene from the yeast Kluyveromyces lactis (Tanguy-Rougeau et al., 1988), the gene encoding the KEX2-related protease (kpr) from Schizosaccharomyces pombe (Davey et al., 1994) and the XPR6 gene from Yarrowia lipolytica (Enderlin & Gryndizzly, 1994). Mammalian homologues have also been identified, including the human fur gene (fes upstream region) in the region upstream of the fes proto-oncogene, encoding the enzyme furin (van den Ouweland et al., 1990). The genes Dfur1 and Dfur2 from the insect Drosophila melanogaster encoding furin-like proteins (Roebrock et al., 1992) and the bli-4 gene from the nematode Caenorhabditis elegans have also been studied. Many other members of the subtilisin-like serine protease family have been identified and the specific endoproteolytic activity of some of them has been elucidated. However, for many others, the precise biological function has not yet been determined.

In this paper we report the identification and characterization of the P. carinii f. sp. carinii PRT1 multigene family. We demonstrate high levels of homology of the PRT1 sequences to subtilisin-like serine proteases. We also show that different copies of the PRT1 genes display DNA sequence heterogeneity and some copies are contiguous with MSG genes.

**METHODS**

**P. carinii DNA extraction.** P. carinii infection was induced in Sprague–Dawley rats by steroid immunosuppression. The organisms were isolated and enriched from infected rat lung tissue by the method described by Peters et al. (1992). Total DNA was extracted from the enriched parasite preparation by digestion with proteinase K (1 mg ml⁻¹) in the presence of 0.5% SDS and 10 mM EDTA (pH 8.0) at 50 °C for 16 h, followed by phenol:chloroform extraction and ethanol precipitation. Samples of DNA for use in PFGE experiments were digested with proteinase K (1 mg ml⁻¹) in the presence of 0.5% SDS and 10 mM EDTA (pH 8.0) at 50 °C for 16 h, followed by phenol:chloroform extraction and ethanol precipitation. Samples of DNA for use in PFGE experiments were prepared in SeaPlaque GTG agarose as described by Banerji et al. (1993). The library was screened with a cDNA clone containing a region of a P. carinii f. sp. carinii MSG gene (EMBL accession number 20870, donated by C. J. Delves and F. Volpe, as part of a study examining subtelomeric sequences in P. carinii. A relatively high number of recombinant plaques gave positive hybridization signals compared to the number when the library was screened with a probe derived from the single copy aom locus (Banerji et al., 1993). Five recombinant plaques were isolated from a tertiary screen; the recombinant DNA was subcloned into the plasmid vector pBluescript II prior to sequencing.

To isolate a full cDNA clone, a P. carinii f. sp. carinii cDNA library constructed in ZAPII (donated by C. J. Delves and F. Volpe, see Dyer et al., 1992) was screened with PCR products derived from amplification of the 5' end of the gene with oligonucleotide primer pair pcprot9 and prp4r (9/4r product), and of the 3' end of the gene with pcprot13/R1 and pcprot12/R1 (13/12 product) (Fig. 1). A primary screening was carried out using both probes, and secondary and tertiary screens were carried out using only the 9/4r product. The number of positive clones when screening the cDNA library with the two probes appeared to be relatively high when compared to the number obtained using a single copy gene. Four recombinant phage isolated from the cDNA library were partially characterized. The recombinant DNA was recovered from the i phage by in vitro excision as pBluescript plasmid DNA. The size of the recombinant DNA ranged from 27 kb to 2.9 kb, and sequence analysis revealed that all four clones contained a poly(A) tail. One recombinant, 73j, was selected for further analysis and the recombinant DNA was sequenced in full from both strands.

**DNA amplification.** Oligonucleotide primers were designed to hybridize to various regions of the P. carinii PRT1 nucleotide sequences (Fig. 1, Table 1). Some oligonucleotides had an EcoRI restriction endonuclease site incorporated at the 5' end to facilitate cloning of the amplification products into EcoRI-digested plasmid vectors pBluescript SK(−) (Stratagene) or pUC18 (Pharmacia). The final concentration of the amplification reaction mix was 50 mM KCl, 10 mM Tris (pH 8.0), 0.1% Triton X-100, 3 mM MgCl₂, 400 μM (each) deoxy-nucleoside triphosphate, 1 μM oligonucleotide primer and 0.025 U Taq polymerase ml⁻¹ (Promega). With primer pair pcprot9 and pcprot10, 40 cycles of amplification was performed at 94 °C for 1.5 min, 55 °C for 1.5 min, and 72 °C for 2.0 min. With primer pair pcprot9 and pcprot10r the same conditions were used, except an annealing temperature of 50 °C was used. With all other primer pairs, ten cycles of amplification were carried out at 94 °C for 1.5 min, 55 °C for 1.5 min and 72 °C for 2.0 min. Negative controls were included in each experiment.

The entire putative gene was amplified as three overlapping fragments, Prp5e (1626 bp), M14 (1279 bp) and Prp2g (251 bp) (Fig. 1). Oligonucleotide primer pairs pcprot9 with pcprot10, followed by pcprot6/R1 with pcprot4/R1 were used in a nested PCR to amplify the 5' fragment, designated Prp5e, of length 1626 bp. The second portion, called M14, spanning 1279 bp of the central region of PRT1, was amplified using a nested PCR with primer pairs pcprot2/R1 with pcprot4/R1, followed by pcprot7/R1 with pcprot2/R1. The third fragment, Prp2g, encompassing the 3' end of the sequence (251 bp), was amplified using oligonucleotide primers pcprot13/R1 and pcprot14/R1 (Fig. 1, Table 1).

Five different overlapping regions of the PRT1 gene were also amplified, cloned and the DNA sequences determined. The
Fig. 1. Schematic representation of a P. carinii PRT1 gene. (a) PRT1 domains: HR, hydrophobic region; PRO-, pro-domain; CATALYTIC, catalytic domain; P-, P-domain; PROLINE-RICH, proline-rich region (the box indicates length and sequence variation in different copies of PRT1); STR, serine/threonine-rich region; 0, catalytic active site residues D, H, and S; *, potential glycosylation sites; I, conserved cysteine residues. (b) A genomic copy of PRT1(Paga) showing the positions of the seven introns (I–VII), a cDNA copy of PRT1(73j) and the products of amplification of different regions of PRT1 genes. (c) Position of oligonucleotide primers used in the amplification of different regions of PRT1 genes.

first region amplified with primer pair pcprot1/RI and pcprot3/RI spanned approximately half of the subtilisin-like catalytic domain, the second region amplified with primer pair pcprot2/RI and pcprot4/RI spanned the end of the subtilisin-like catalytic domain and the start of the P-domain, the third region amplified with primer pair pcprot7/RI and pcprot8/RI spanned the proline-rich domain and the fifth region amplified with primer pair pcprot13/RI and pcprot14/RI spanned the C-terminal hydrophobic domain (Fig. 1, Table 1). The sequences Prp1a, Prp3a, Prp7a, Prp2c, Prp3c, Prp4c, Prpta2, Prp4, Prp5f, Prpg3 and Prp5g were amplified from the P. carinii cDNA library, and sequences Par-19, Par-14, Par-5, Par-3, Par-1, and Par-1 were amplified from the P. carinii genomic DNA (Fig. 2).

DNA sequence analysis. DNA sequence analysis was performed using the dideoxy chain-termination method (Sanger et al., 1977). Sequence data were obtained in full from both strands for all sequences. Analysis of the sequence data was carried out using the University of Wisconsin Genetics Computing Group (UWGCG) Sequence Analysis Software Package, version 8 (Genetics Computer Group, Madison, WI, USA).

PFGE. P. carinii f. sp. carinii was isolated from an infected rat lung and the chromosomes were separated by PFGE using a Contour Clamped Homogeneous Electric Field (CHEF) DRII apparatus (Bio-Rad) operated at 4 °C. Electrophoretic separation was achieved using 0.9% Seakem agarose gel with initial switching time of 10 s, increasing to a final switching time of 60 s at 180 V for 48 h. A karyotype corresponding to P. carinii f. sp. carinii form I was observed (Cushion et al., 1993).

Southern hybridization. Southern blotting and hybridization were carried out using standard techniques (Sambrook et al., 1989). PFGE blots were hybridized with three probes derived from different domains of the PRT1 gene. The product 9/4r was derived from amplification of the 5' end of the PRT1 gene with primer pair pcprot9 and pcprot4/RI, product 2/4 from amplification of the central catalytic region with primer pair
Table 1. Oligonucleotide primer sequences

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>prp4t</td>
<td>5‘-GCTTGTCACTATAAAAAC-3’</td>
</tr>
<tr>
<td>p3/R1</td>
<td>5‘-GGGAATTCTGAAGCTTTTCGAGTGGTTG-3’</td>
</tr>
<tr>
<td>36ex/R1</td>
<td>5‘-GGGAATTCTAACCTCAGGACAGTCCA-3’</td>
</tr>
<tr>
<td>pctl2</td>
<td>5‘-AAGTCAGTGCTCTCTTGTTGCTA-3’</td>
</tr>
<tr>
<td>msgterm1</td>
<td>5‘-AATGGTTGTTGGAGGTATGG-3’</td>
</tr>
<tr>
<td>pcprot1/R1</td>
<td>5‘-GGGAATTCTTATTCTTGAGCAGGACGAC-3’</td>
</tr>
<tr>
<td>pcprot2/R1</td>
<td>5‘-GGGAATTCTTCTACACCTCTTGCTGC-3’</td>
</tr>
<tr>
<td>pcprot3/R1</td>
<td>5‘-GGGAATTCAAGCCCATGTAAGATTAGA-3’</td>
</tr>
<tr>
<td>pcprot4/R1</td>
<td>5‘-GGGAATTCTAATGTTAGGATATCCGCG-3’</td>
</tr>
<tr>
<td>pcprot5/R1</td>
<td>5‘-GGGAATTCTATGTGAAATGGTGCCATGAG-3’</td>
</tr>
<tr>
<td>pcprot6/R1</td>
<td>5‘-GGGAATTCTTTTTTITTTAACATTTACATG-3’</td>
</tr>
<tr>
<td>pcprot7/R1</td>
<td>5‘-GGGAATTCTCTGTATTAGGAAACTAGAGTG-3’</td>
</tr>
<tr>
<td>pcprot8/R1</td>
<td>5‘-GGGAATTCAAGGTTAGCATCCAGATCG-3’</td>
</tr>
<tr>
<td>pcprot9</td>
<td>5‘-AGAATTCTAATTTAAAAAGTTAAG-3’</td>
</tr>
<tr>
<td>pcprot10</td>
<td>5‘-AACACCAACATACCTGAAAC-3’</td>
</tr>
<tr>
<td>pcprot12/R1</td>
<td>5‘-GGGAATTCTTATAGTACATGAAAGCTTTTCG-3’</td>
</tr>
<tr>
<td>pcprot13/R1</td>
<td>5‘-GGGAATTCTTCATCTCATCTCGACTTT-3’</td>
</tr>
<tr>
<td>pcprot14/R1</td>
<td>5‘-GGGAATTCTATAGGTTAAAAAGTACCC-3’</td>
</tr>
</tbody>
</table>

pcprot2/R1 and pcprot4/R1, and product 13/12 from amplification of the 3' end of the gene with primer pair pcprot13/R1 and pcprot12/R1 (Fig. 1). The amplification products were gel-purified (GeneClean II, Bio101) and labelled with [α-32P]dCTP by random priming (Megaprime, Amersham). Hybridization was carried out at 45 °C and stringency washing at 60 °C in 0.2× SSC, 0.1% SDS.

Southern blots of genomic P. carinii DNA digested with restriction endonuclease PstI or BamHI were probed with oligonucleotide probes pcprot3/R1, pcprot5/R1, pctl2 and msgterm (Table 1) labelled with [α-32P]dATP using polynucleotide kinase. Hybridization was carried out at 46 °C and stringency washing at 52 °C in 5× SSC, 0.5% SDS.

**RESULTS**

Analysis of DNA and derived amino acid sequence of copies of the PRT1 gene

We have identified a family of genes in the P. carinii f. sp. carinii genome which shows homology to the subtilisin-like serine proteases. We have named this gene family PRT1 (protease 1). A copy of the PRT1 gene (Paga) was isolated from a P. carinii genomic library, the ORF (3069 bp) containing seven short putative intervening sequences. A copy of the PRT1 gene (73j) of length 2370 bp was also isolated from a cDNA library. Portions of the gene were amplified by PCR from the cDNA library as three overlapping fragments at the 5' end (Prp53e), the central region (M14) and the 3' end (Prp2g). Five other regions of the gene were also amplified, from either the P. carinii cDNA or genomic libraries, to determine the extent of diversity among different copies of the multi-gene family at different domains within the gene sequence.

Analysis of the DNA sequence of the copy of the PRT1 gene from the genomic library, PRT1(Paga), and of the copy from the cDNA library, PRT1(73j), confirmed the presence of seven short introns in the genomic DNA sequence. The introns ranged in length from 38 bp to 45 bp, with a base composition ranging from 71 to 84 mol% A + T. In all seven introns, the dinucleotide GT was present at the 5' splice donor site and AG at the 3' splice acceptor site. The sequence YTRAT, which has been identified as the putative lariat-forming motif in other P. carinii f. sp. carinii introns (Zhang & Stringer, 1993), was present in the first, second, fourth, fifth and seventh intron. The eukaryotic lariat consensus sequence, YYRAY, was identified in the third and sixth intron.

The sequence of the cDNA clone, PRT1(73j), contained an ORF of 2370 bp, which on translation resulted in a peptide of 790 amino acids (Fig. 2). The deduced amino acid sequence was compared to sequences in the GenBank and EMBL databases and showed homology to fungal and other eukaryotic subtilisin-like serine proteases. The A + T content of the ORF was 64 mol%, with a high A + T content at the third base position of the codons. The base composition of the 5' upstream sequence was 74 mol% A + T and of the 3' downstream sequence 75 mol% A + T. A consensus polyadenylation signal, AATAAA, was observed 68 bp downstream of the stop codon.

The deduced amino acid sequence of the genomic clone PRT1(Paga), the cDNA clone PRT1(73j), the three fragments obtained by PCR amplification of the cDNA library and the other recombinant clones generated by DNA amplification were compared (Fig. 2). Several regions of homology were found and also a number of regions in which significant divergence was observed. These data suggested that the sequences were derived from different copies of the PRT1 multi-gene family.
Comparison of \textit{P. carinii} PRT1 with other subtilisin-like serine proteases

The deduced amino acid sequence of the cDNA clone PRT1(73j) was aligned with nine other subtilisin-like serine proteases including fungal, mammalian, insect and nematode sequences (Fig. 3). The PRT1 sequences showed homology with all the other sequences, with a high level of homology in the subtilisin-like catalytic domain. The three essential residues of the catalytic active site, aspartic acid (Asp_{114}), histidine (His_{227}) and serine (Ser_{422}) [residues are numbered with reference to PRT1(73j)] were conserved in all the PRT1 sequences. The highest levels of homology between all the sequences were around these residues.

The structural organization of the fungal sequences showed domains characteristic of this class of endopeptidases, a hydrophobic signal sequence, a pro-domain that may be cleaved by autoproteolysis, a subtilisin-like catalytic domain, a P-domain (known as such because it is essential for proteolytic activity), a serine/threonine-rich domain which may potentially be modified by O-linked glycosylation, a C-terminal hydrophobic transmembrane domain and a C-terminal tail with acidic residues (Van de Ven & Roebroek, 1993) The \textit{P. carinii} PRT1 sequences showed a similar putative structural organization but unlike the nine other subtilisin-like serine proteases, they also had a proline-rich domain preceding the serine/threonine-rich domain and the C-terminal hydrophobic domain (Fig. 1). The \textit{P. carinii} PRT1(73j) sequence had a hydrophobic signal sequence at the N-terminus, followed by a putative pro-domain, a subtilisin-like catalytic domain from Ser_{171} to His_{474}, a P-domain from residue Tyr_{472} to Ser_{631}, a proline-rich domain from residue Pro_{641} to Pro_{707}, a serine/threonine-rich domain from residues Thr_{708} to Ser_{865}, and a C-terminal hydrophobic domain from residues His_{711} to Phe_{790}.

Analysis of subtilisin-like catalytic domain

The three-dimensional structures of four subtilisin-like serine proteases have been determined, subtilisin BPN’/Novo from \textit{Bacillus amylo liquefaciens} (Hirn et al., 1984; Bott et al., 1988), subtilisin Carlsberg from \textit{Bacillus licheniformis} (McPhalen & James, 1988), thermotase from \textit{Thermoa ctoinomyces vulgaris} (Gros et al., 1989; Teplyakov et al., 1990) and proteinase K from \textit{Tritirachium album} (Betzel et al., 1988). The amino acid sequence of these four proteases has been compared to that of 31 other subtilisin-like serine proteases isolated from bacteria, fungi and higher eukaryotes and the essential core structure of the catalytic domain of this group of molecules has been identified (Siezen et al., 1991).

We have compared the deduced amino acid sequence of the \textit{P. carinii} PRT1(73j) gene with the multiple sequence alignment of the other subtilisin-like serine proteases and have identified, by homology, the three essential residues of the catalytic active site, aspartic acid, histidine and serine, in the PRT1 sequence (Asp_{114}, His_{227} and Ser_{422}). On the basis of the sequence alignment, the \textit{P. carinii} PRT1 sequence could be assigned to the class I subtilases, within the subgroup I-E which contains the pro-hormone processing proteases from yeasts and higher eukaryotes (Siezen et al., 1991).

Eight \(\alpha\)-helical domains and nine \(\beta\)-sheet regions of the sequences have been defined as the structurally conserved regions within the essential core structure. The variable regions which connect the core segments have been found to differ both in length and in amino acid sequence (Siezen et al., 1991). High levels of homology were observed between the PRT1 sequences and the other sequences in the regions of the two conserved internal helices, helix C (residues 252–262) and helix F (residues 422–438) [Fig. 3; residues are numbered with reference to the PRT1(73j) sequence]. Eleven amino acid residues have previously been found to be totally conserved in all the characterized subtilisin-like serine proteases, and most but not all are conserved in the PRT1 sequences. These amino acid residues are at the active site Asp_{114}, His_{227} and Ser_{422} [found in all the PRT1 sequences except PRT1(Prp7a)] and in the internal helices at residues Gly_{253}, Gly_{258} and Pro_{437}. The residues Ser_{510}, Gly_{512}, Gly_{431}, Gly_{432} and Thr_{422} involved in substrate binding, were conserved in all the PRT1 sequences except Thr_{422} which was found only in two sequences generated by PCR, PRT1(Prp3a) and PRT1(Prp7a).

In addition to the totally conserved residues, seven other amino acid residues have been identified which are highly conserved. Of these, six were conserved in the \textit{P. carinii} PRT1 sequences and included the oxyanion hole residue (Asn_{339}), residues Gly_{216} and Thr_{254} near the active site, and also residues Gly_{208}, Gly_{217} and Gly_{243}. Seven conserved cysteine residues were found in all the \textit{P. carinii} PRT1 sequences: Cys_{256}, Cys_{258}, Cys_{260}, Cys_{269}, Cys_{269}, Cys_{269} and Cys_{269}. Nineteen variable regions, generally located in loops on the surface of the molecule, have been identified in the subtilase family, of which 14 were found in the \textit{P. carinii} PRT1 sequences. Three positions have been identified at which charge is totally conserved in all the subtilisin-like proteases examined, and these were also conserved in the \textit{P. carinii} PRT1 sequences: the positive charge on Arg_{227} and the negative charges on residue Asp_{214} (active site) and Asp_{227}.

\textbf{Fig. 2.} Alignment of the \textit{P. carinii} PRT1 deduced amino acid sequences from the genomic clone Papa, the cDNA clone 73j and the three overlapping PCR products amplified from a cDNA library corresponding to the 5′ region (Prp5e), the central region (M14) and the 3′ region (Prp2g). The deduced amino acid sequences of PCR products amplified from five different regions of the PRT1 gene family were also aligned: the catalytic domain (Prp1a, Prp3a, Prp7a); the boundary of the catalytic domain and the P-domain (Prp2c, Prp3c, Prp4c); the P-domain (Prptaf2, Prp4f, Prp5f); the proline-rich region (Prp-19, Prp-14, Prp-5, Prp-3, Prp-1, Lam-1); and the C-terminal region (Prpg4, Prpg3, Prpg5). Gaps were introduced to maximize homology; identical amino acids are boxed.
Fig. 2. For legend see p. 2227.
Fig. 2 (cont.) For legend see p. 2227.
It has been proposed that the high specificity of the class I–E subtilisin-like serine proteases for paired basic residues Lys-Arg or Arg-Arg may be facilitated by a high density of negative charge at the substrate-binding face, provided by nine highly conserved Asp residues and one Glu residue (Siezen et al., 1991). Two of the Asp residues, Asp94 and Asp95, were found in all the P. carinii PRT1 sequences and also the Glu93 residue. In addition, four other Asp residues were found in some but not all of the copies of PRT1.

**Analysis of the domains flanking the subtilisin-like catalytic domain**

The putative domains of the PRT1(73j) polypeptide are summarized in Fig. 1. A hydrophobicity plot of the PRT1(73j) sequence revealed a hydrophobic region at the N-terminus, suggesting that this may be a signal sequence. Residues 1–23 of the N-terminus of the sequence showed a high level of homology to the N-terminus of the P. carinii f. sp. carinii multifunctional folic acid synthesis fas gene which encodes dihydropteroate synthase and dihydrofolate reductase (Volpe et al., 1992, 1993). This region was followed by the presumptive pro-domain, which may be cleaved by autocatalysis. Potential autocatalytic sites of paired basic residues were identified in the PRT1(Paga) and PRT1(Prp5e) sequences at Lys115–Arg116 and Arg126–Arg127, but were absent in the PRT1(73j) sequence. Five other semi-conserved autocatalytic sites were found in some, but not all, copies of the P. carinii PRT1 sequences, two in the catalytic domain (Lys490–Arg491, Arg492–Arg493) and three in the P-domain (Arg532–Arg533, Arg554 or Lys556–Arg557, Arg576–Arg577). One potential autocatalytic site at the start of the C-terminal hydrophobic region (Lys769–Arg770) was found in all the sequences. The PRT1(73j) sequence contained two of the potential autocatalytic sites, Arg556–Arg557 and Lys769–Arg770.

The PRT1 sequences showed homology with the other subtilisin-like serine proteases in the region of the P-domain, the highest homology being with the derived amino acid sequence of the *Schiz. pombe krp* gene. Four potential sites for N-linked glycosylation were observed in all the PRT1 sequences, three in the subtilisin-like catalytic domain (Asn194, Asn277, Asn442), and one in the P-domain (Asn603).

A serine/threonine-rich region was also identified in the PRT1(73j) sequence from residue Thr706 to Ser765, and the hydrophobicity plot of the PRT1(73j) sequence revealed a hydrophobic region at the C-terminal end, residues His771–Phe789, suggesting a membrane-associated domain. Unlike most other serine protease sequences, however, all the copies of the PRT1 polypeptide contained a proline-rich region downstream of the P-domain.

**Genetic organization of the PRT1 multi-gene family**

Analysis of the alignments of the DNA and the deduced amino acid sequences of copies of the PRT1 gene from genomic DNA, the cDNA sequence and the three fragments obtained by PCR of the cDNA library revealed domains in the PRT1 gene which were highly conserved and also regions where significant divergence was observed, again suggesting that PRT1 comprises a multi-gene family (Fig. 2). The subtilisin-like catalytic domain and the P-domain appeared to be conserved whereas high levels of heterogeneity were observed in the proline-rich domain and the C-terminal domain. The variation in this region was both in length and in sequence. A number of repeated DNA sequence motifs were found in the proline-rich region. Nucleotide sequences encoding polyproline were found in all the sequences, and also the dipeptides Pro-Glu and Pro-Gln and the tetrapeptides Pro-Glu-Pro-Gln and Pro-Glu-Thr-Gln. The order and number of tandem repeats varied in each sequence. The overall length of this region varied from 67 amino acid residues in the shortest sequence, PRT1(73j), to 233 residues in the longest sequence, PRT1(M14).

To further substantiate the presence within the *P. carinii* genome of multiple copies of the PRT1 gene, *P. carinii f. sp. carinii* chromosomes were separated by PFGE and a karyotype corresponding to *P. carinii f. sp. carinii* form 1 was observed (Cushion et al., 1993). Under the conditions used, the rat chromosomes were too large to be resolved and remained at the band of limiting mobility. The separated chromosomes were analysed by hybridization with three probes derived from different domains of PRT1. All three probes showed similar patterns of hybridization, annealing at high stringency to all the chromosome bands except for one, the third smallest in size, approximately 350 kb (Fig. 4). This provided further evidence that the *P. carinii f. sp. carinii* genome contained many copies of the PRT1 gene, which were present on most of the *P. carinii f. sp. carinii* chromosomes.

The sequences of the PRT1 gene family showed high levels of homology with ORF3, an ORF which was reported to encode a *P. carinii* protein of unknown function and was demonstrated to be contiguous with a copy of the gene encoding the major surface glycoprotein *MSG100* (Wada & Nakamura, 1994). This gene arrangement was reported in 15 other *λ* clones, in which a gene showing high homology to ORF3 was located downstream of a copy of *MSG* (Wada & Nakamura, 1994). Most copies of the *MSG* genes have been demonstrated to be located in the *P. carinii f. sp. carinii* subtelomeric regions (Underwood et al., 1996; Sunkin & Stringer, 1996). The copy of the PRT1 gene encoded by the PRT1(Paga) sequence was cloned from a λEML3 genomic library as a single 14 kb fragment and was approximately 1150 bp downstream of a copy of *MSG*. Four other *λ* clones isolated from the same library contained a copy of PRT1 contiguous with a copy of *MSG*. 

2230
**DISCUSSION**

We describe the cloning and characterization of copies of the PRT1 multi-gene family from *P. carinii* f. sp. *carinii*. A copy of the PRT1 gene was isolated from a *P. carinii* f. sp. *carinii* genomic library. A different copy was isolated from a cDNA library, indicating that this copy of the gene was transcribed, and also identifying the presence of seven short introns in the genomic sequence. Consistent with many other *P. carinii* genes, the coding region and the flanking sequences of the PRT1 sequences showed a strong bias for adenine or thymine, and in particular at the third base position of the codons. Similarly, the presence of short A + T-rich introns has been reported in other *P. carinii* genes. In the PRT1 sequences, the introns were not distributed throughout the gene, but six of the seven introns were found in the subtilisin-like catalytic domain and the seventh in the P-domain. It is possible that the introns may play a role in restricting the variation in this region of the gene, whereas no introns were observed in the highly heterogeneous proline-rich region.

The high level of homology of the *P. carinii* PRT1 sequences to subtilisin-like serine proteases, and in particular in the region of the catalytic domain, strongly suggested that this gene encoded a protease of this type. The predicted *P. carinii* PRT1 polypeptide sequences possessed the three essential residues of the catalytic active site as well as many other highly conserved motifs. The domain organization of the PRT1 gene strongly resembled that of the fungal prohormone-processing proteases, with the exception of the proline-rich domain. This proline-rich region is very uncommon in the subtilisin-like serine protease superfamily, although the KRP6 gene from *Y. lipolytica* is reported to contain a short region of a tetrapeptide repeat, the consensus sequence of the four amino acids being Glu-(Asp/Glu)-Lys-Pro (Enderlin & Ogrzydiak, 1994). A proline-rich region has also been found in the C-terminal tail domain of the mammalian serine protease acrosin, a proteolytic enzyme of sperm cells, located in the acrosome at the apical end of the spermatozoan (Klemm et al., 1991).

In the African trypanosome, *Trypanosoma brucei*, a proline-rich domain has been identified in the procyelic acidic repetitive proteins (PARPs). These proteins are found on the cell surface of the insect form of the parasite and are encoded by a family of polymorphic genes which contain a variable region with heterogeneity in both length and sequence. The variable region contains the proline-rich domain and is primarily composed of the dipeptide Glu-Pro (Roditi et al., 1989).

Unlike any of the other fungal prohormone-processing proteases, which appear to be single-copy genes, the data reported in this study suggested that the PRT1 sequence is present in many copies, which are similar but not identical, in the genome of *P. carinii* f. sp. *carinii*. The relatively large number of recombinants present in both the genomic and the cDNA libraries suggested a multi-copy gene and this was substantiated by PFGE data revealing that at least one copy of a PRT1 gene was present on all but one of the *P. carinii* chromosomes. Southern hybridization of restriction endonucleolytic digests of *P. carinii* f. sp. *carinii* DNA probed with PRT1 sequences also confirmed the presence of many copies of the gene. Analysis of sequence data generated by the amplification of the locus showed heterogeneity, suggesting that a variety of different copies of the gene were present in the *P. carinii* genome. Some domains, including the subtilisin-like catalytic domain and the P-domain, were highly conserved between gene copies, whereas the highest levels of divergence were observed in the proline-rich domain, which varied both in length and in sequence.

Of five genomic clones analysed in this study, all possessed a copy of PRT1 contiguous with an MSG gene. It has been reported that 15 independent genomic clones which encoded MSG were contiguous with the ORF3 sequence, which from our analysis appears to encode the proline-rich domain of PRT1 (Wada & Nakamura, 1994). It has been demonstrated that most copies of MSG are subtelomeric (Underwood et al., 1996; Sunkin & Stringer, 1996). It is therefore highly likely that many copies of the PRT1 multi-gene family are located in the subtelomeric regions of the *P. carinii* f. sp. *carinii* genome. However, PFGE analysis has shown that not every *P. carinii* f. sp. *carinii* chromosome contains a copy of PRT1, and the preliminary characterization of a
Fig. 3. For legend see p. 2231.
Fig. 3 (cont.) For legend see p. 2231.
E. B. LUGLI, A. G. ALLEN and A. E. WAKEFIELD

Fig. 4. Southern hybridization analysis of P. carinii chromosomes separated by PFGE and probed with three different regions of the PRT1(Paga) gene. Lanes: 1, 3, 5 and 7, Sacch. cerevisiae DNA; 2, 4, 6 and 8, P. carinii DNA; 1 and 2: ethidium-bromide-stained gel; 3 and 4, probed with the 5' end of the PRT1(Paga) gene (9/4r product); 5 and 6, probed with part of the catalytic domain (24/24 product); 7 and 8, probed with the 3' end of the gene (13/12 product). The arrow indicates the P. carinii chromosome which did not hybridize to the PRT1 probes.

clone of one of the subtelomeric regions of P. carinii f. sp. carinii has not revealed a copy of PRT1 (Underwood & Wakefield, unpublished results). Hybridization of MSG and subtelomeric probes to endonuclease-digested P. carinii f. sp. carinii DNA resulted in positive hybridization to fragments greater than approximately 7 kb in size. Probes derived from the PRT1 sequence hybridized to these bands but also to low-molecular-mass fragments, again suggesting that not all copies of PRT1 are subtelomeric.

The P. carinii PRT1 gene family shows some striking similarities to that of MSG. Both are composed of many genes, copies of which are found on most P. carinii chromosomes and show sequence heterogeneity. Some copies of PRT1 are contiguous with MSG and are located in the subtelomeric regions of the P. carinii chromosomes.

It is interesting to note that one of the major components of the cell surface of Leishmania has proteolytic activity. The Leishmania major surface protease (msp or gp63), a zinc endoprotease, is found in all species of Leishmania and is encoded by a family of genes, some of which are tandemly arrayed (Bouvier et al., 1989; Webb et al., 1991). Expression of different copies of the gene is regulated during the development of the parasite and different isoforms of the protein are found in the promastigote stage in the gut of the sand fly and in the amastigote stage in the phagolysosomes of the macrophages (Frommel et al., 1990; Roberts et al., 1995; Ramamoorthy et al., 1995). The major surface protease is thought to play an important role in the virulence of Leishmania by involvement in the degradation of components of the extracellular matrix and by facilitating promastigote attachment to host macrophages (McMaster et al., 1994). Immunization with MSP protein confers partial protection of mice against Leishmania infection (Abdelhak et al., 1995).

The proteins encoded by the P. carinii PRT1 gene family show highest homology to subtilisine-like serine proteases. A wide diversity of different types of precursor proteins are processed by this family of proteases to mature and active regulatory proteins, but the precise function of many of these proteases has not yet been determined. Some of the fungal homologues have been shown to function in the processing of several proteins, such as the Sacch. cerevisiae KEX2 gene product which processes both the pheromone α-factor and the killer toxin (Fuller et al., 1989). The krp gene product from Schiz. pombe, which cleaves the pheromone precursor pro-P-factor to its active form, is thought to also
function in the processing of other regulatory proteins since its activity is essential for cell viability (Davey et al., 1994). The XPR6 gene product from Y. lipolytica, although not essential for cell viability, when disrupted was found to cause aberrant growth and morphology (Enderlin & Ogrydziak, 1994). The function of the products of the P. carinii PRT1 gene family is not yet understood but they are likely to play an important role in the life cycle and possibly also the pathogenicity of the organism.

ACKNOWLEDGEMENTS

This research was supported by the Medical Research Council (E. B. L.) and the Royal Society (A. E. W.), and formed a part of the European Concerted Action, Biomed I, PL941188. We should like to thank Dr C. J. Delves and Dr F. Volpe, Glaxo Wellcome Research, for the gift of the P. carinii cDNA library and the MSG clone, and the Oxford Pneumocystis Research Group for helpful discussion.

REFERENCES

Recombinant BCG expressing the leishmania surface antigen GP63 induces protective immunity against Leishmania major infection in BALB/c mice. Microbiology 141, 1585–1592.


Three-dimensional structure of proteinase K at 0·15 nm resolution. Eur J Biochem 175, 155–171.

The three-dimensional structure of Bacillus amyloliquefaciens subtilisin at 1·8 Å and an analysis of the structural consequences of peroxide inactivation. J Biol Chem 263, 7895–7906.


Isolation and characterization of krp, a dibasic endopeptidase required for cell viability in the fission yeast Schizosaccharomyces pombe. EMBO J 13, 5910–5921.


Cloning, nucleotide sequence and functions of XPR6, which codes for a dibasic processing endopeptidase from the yeast Yarrowia lipolytica. Yeast 10, 67–79.

The major surface glycoprotein (GP63) is present in both life stages of Leishmania. Mol Biochem Parasitol 38, 25–32.

Fuller, R. S., Brake, A. & Thorner, J. (1989).
Yeast prohormone processing enzyme (KEX2 gene product) is a Ca2+-dependent serine protease. Proc Natl Acad Sci USA 86, 1434–1438.


Molecular dynamics refinement of a thermolysin-eglin-c-complex at 1·98 Å resolution and comparison of two crystal forms that differ in calcium content. J Mol Biol 210, 347–367.


Acrosin, the peculiar sperm-specific serine protease. Hum Genet 87, 635–641.

Multiple genes encode the major surface glycoprotein of Pneumocystis carinii. J Biol Chem 268, 6034–6040.


Quantification of the detection of Pneumocystis carinii by DNA amplification. Mol Cell Probes 6, 115–117.


Received 21 January 1997; revised 7 March 1997; accepted 20 March 1997.