A serpin in the cellulosome of the anaerobic fungus *Piromyces* sp. strain E2

Peter J. M. Steenbakkers[a], James A. Irving[b], Harry R. Harhangi[a], Willem J. C. Swinkels[a], Anna Akhmanova[a,1], Rembrandt Dijkerman[a], Mike S.M. Jetten[a], Chris van der Drift[a], James C. Whisstock[b], Huub J. M. Op den Camp[a,*]

[a] Department of Microbiology, IWWR, Radboud University Nijmegen, Toernooiveld 1, NL-6525ED Nijmegen, The Netherlands

[b] Department of Biochemistry and Molecular Biology, Monash University, Clayton Campus, Melbourne, VIC 3800, Australia and The Victorian Bioinformatics Consortium, Monash University, Melbourne VIC 3800, Australia.

[1] Present address: MGC Department of Cell Biology and Genetics, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands.

* Corresponding author
Mailing address: Dr. H.J.M. Op den Camp
                      Department of Microbiology, IWWR
                      Radboud University Nijmegen
                      Toernooiveld 1
                      NL-6525ED Nijmegen
                      The Netherlands
                      Phone: +31 24 3652657
                      E-mail: h.opdencamp@science.ru.nl

Abstract

A gene encoding a novel component of the cellulolytic complex (cellulosome) of the anaerobic fungus *Piromyces* sp. strain E2 was identified. The encoded 538 amino acid protein, named Celpin, consists of a signal peptide, a positively charged domain of unknown function followed by two fungal dockerins, typical for components of the extracellular fungal cellulosome. The C-terminal end consists of a 380 amino acid serine proteinase inhibitor (or serpin) domain homolog, sharing 30 % identity and 50 % similarity to vertebrate and bacterial serpins. Detailed protein sequence analysis of the serpin domain revealed that it contained all features of a functional serpin. It possesses the conserved amino acids present in more than 70 % of known serpins and it contained the consensus of inhibiting serpins. Because of the confined space of the fungal cellulosome inside plant tissue and the auto-proteolysis of plant material in the rumen, the fungal serpin is presumably involved in protection of the cellulosome against plant proteinases. The Celpin protein of *Piromyces* sp. strain E2 is the first non-structural, non-hydrolytic fungal cellulosome component. Furthermore, the Celpin protein of *Piromyces* sp. strain E2 is the first representative of a serine proteinase inhibitor of the fungal kingdom.

Introduction

Anaerobic fungi are known for their ability to degrade plant cell wall material (Dijkerman *et al*. 1997a). During growth on sugar polymers, anaerobic fungi produce high molecular mass complexes specialized in the degradation of crystalline cellulose (Dijkerman *et al*. 1997b). These cellulolytic complexes resemble the cellulosomes produced by several *Clostridium* species (Bayer *et al*. 1998). Clostridial cellulosomes contain a central, organizing protein, named scaffoldin, that adheres to cellulose and is able to bind the catalytic subunits of the complex through its cohesin domains (Miras *et al*. 2002). The catalytic components in turn require a dockerin domain for incorporation into the cellulosome (Pages *et al*. 1997). Although the model of the cellulosome has been derived from studying the *Clostridium thermocellum* complex, the fungal cellulosome is considered to be very similar. The structure of a fungal dockerin was recently resolved (Raghothama *et al*. 2001) and putative scaffoldins have been identified in extracellular protein samples of *Neocallimastix*, *Piromyces* and *Orpinomyces* species (Fanutti *et al*. 1995; Fillingham *et al*. 1999; Steenbakkers *et al*. 2001).

Anaerobic fungi have only been identified in the digestive tract of herbivores where they are part of the cellulolytic microflora, although Lockhart *et al*. (2006) detected anaerobic fungi in landfill sites using molecular methods. Next to the anaerobic fungi also rumen bacteria like *Ruminococcus* species produce cellulosome-like complexes (Ding *et al*. 2001). Despite an apparent similarity in cellulolytic systems, anaerobic fungi degrade and ferment the bulk of ingested feed (Gordon & Phillips 1998; Lee *et al*. 2000). The higher fraction of degraded plant cell wall material can largely be explained by the intrinsic characteristics of fungal growth. Anaerobic fungi grow as saprophytes and, unlike rumen bacteria, develop a mycelium that penetrates the plant particle (Orpin 1977). The fungal cellulosome is released at the hyphal tips (Ljungdahl *et al*. 1999), enabling the mycelium to degrade the local surrounding plant tissue and extend their hyphae. This results in an increased surface within the plant particle that is subjected to fungal degradation. Little is known about the exact interactions between cellulolytic organisms and the plant tissue that is digested in the rumen. Plants have an extensive array of defence mechanisms against herbivores, insects and pathogenic fungi (Lindsey *et al*. 2002). Once ingested, the damaged plant tissue is exposed to an elevated temperature in a dark and chemically reduced environment. It was recently found that these conditions cause autolysis of plant cells and release of large amounts of proteinases (Kingston-Smith *et al*. 2000; Zhu *et al*. 1999). The absence of plant proteins may also explain

why *Piromyces* sp. strain E2 is not proteolytic and can use ammonium but not protein as a nitrogen source (Dijkerman *et al*. 1996).

Serine proteinase inhibitors or serpins generally act by serving as suicide substrates for their respective target proteinases. Upon cleavage, a large conformational change is triggered within the serpin which completely disrupts the proteinase, resulting in a covalently linked serpin-proteinase complex (Pike *et al*. 2002). Since the initial description of prokaryotic serpins (Irving *et al*. 2002) it became clear that serpins occur in many archaeal and bacterial genera without being widespread within these genera (Roberts et al. 2004). In a recent study it was reported that the *Clostridium thermocellum* cellulosome harbours two serpins (Kang et al. 2006). In this paper we report the isolation of *Piromyces* sp. strain E2 cDNA encoding a cellulosomal serine proteinase inhibitor, a representative of a protein superfamily not previously isolated from a member of the fungal kingdom.

Experimental procedures

Immunoscreening

The previously described *Piromyces* sp. strain E2 cDNA library (Akhmanova *et al*. 1998) in lambda ZAP II (Stratagene, La Jolla, CA U.S.A.) was screened with anti-cellulosome antibodies (Steenbakkers *et al*. 2002b). From 50 immunoreactive (rescreenend) clones, 32 were subcloned and sequenced using pBluescript SK vector primers (see below).

Subcloning procedure of rescreened immunoreactive cDNA clones

To rapidly subclone and sequence clones that were identified by immunoscreening, a subcloning procedure for pBluescript vectors was developed. The method resembles nested deletions in that it directionally deletes nucleic acid stretches but does not require the same high quality starting material. Vector with insert were first linearised with *SmaI* and a stretch of the 5' end of the insert was removed by cutting once with a second, diluted restriction enzyme *RsaI*, also rendering a blunt terminus. This restriction enzyme cuts on average once very 254 base pairs and is therefore likely to cut inserts that are too long for single sequence run. For the subcloning about 1 μg of plasmid DNA (Flexiprep kit) was digested with *SmaI* followed by heat-inactivation of the enzyme. The DNA was aliquoted into three tubes containing 1, 0.2 and 0.04 units *RsaI* and was incubated for 30 min at 37 °C. All three restriction incubations were heat-inactivated and approximately 100 ng of each partial digestion was analysed on a 0.7 % agarose gel. The partial *RsaI* digestion was diluted and

self-ligated overnight. After ligation the mixture was digested with 10 U *PstI* to select against plasmids which were not *RsaI* digested. Next, the digested ligation mixture was transformed and plated on LB plates containing X-gal, IPTG and ampicillin to select against plasmids that were cut in the *RsaI* site within the *KpnI* site of the multiple cloning site, the *bla* gene or the origin of replication. White colonies were grown overnight in 5 ml of Luria-Bertani broth and screened for insert size by the rapid screening procedure. This procedure generated approximately 45 subclones from 15 cDNA clones.

Isolation of the full-length celpin cDNA sequence

Immunoscreening of the *Piromyces* sp. strain E2 cDNA library resulted in the isolation of a partial cDNA clone encoding a cellulosomal serine-protease inhibitor (Serpin) homolog. Inverse PCR was used to isolate the remainder of the gene. Approximately 2 µg *Piromyces* genomic DNA was digested overnight with *TaqI* and self-ligated overnight at 4 °C in 12 µl final volume. Two µl of the ligation mixture served as a template in inverse PCR using the following primers: 5'-GTTATACAGTCCATTATCTATAGAATATGC-3' and 5'-CATGGGTAGGATGGTTTTAATTTTTGTGAGAAAC-3'. PCR was performed with 30 cycles of 1 min at 95°C, 1 min at 53°C and 5 min at 72°C using Platinum Taq (Invitrogen). The resulting 1 kb PCR product was cloned into pGEM T-Easy (Promega) and sequenced. Primers were designed to amplify the complete coding part of the serpin gene with the cDNA library as a template using *pfu* polymerase: Serpin Forw. 5'-AATATGAAGAACTTATCATTATATC-3' and Serpin Rev. 5'-TTTACTTAAAATCTAGAGGTTGG-3'. The resulting PCR product was tailed with dATP and cloned into the pGEM-T Easy vector (Promega) and sequenced.

Sequence analysis

Celpin sequences were assembled using vector NTI Suite software (Informax, Golden, CO U.S.A). Further analysis was performed at default settings using the on-line facilities of the Swiss ExPasy server (http://www.expasy.ch). For comparison serpin sequences selected from Roberts *et al*. (2004) were obtained from GenBank (http://www.ncbi.nlm.nih.gov) and aligned using ClustalW.

Southern blot analysis of the celpin gene

Eight µg of *Piromyces* sp. strain E2 genomic DNA, isolated as described before (Brownlee 1994), was digested overnight at 37°C with different restriction enzymes: *TaqI* (65°C),

HincII, RsaI, ScaI, XbaI and *XmnI*. The digested gDNA was separated overnight on a 0.7% w/v agarose gel (6 mA, 18V). After separation DNA was transferred to a membrane as described before (Harhangi *et al*. 2002). A radioactively labeled probe that covered the Serpin-encoding part of the isolated cDNA clone was synthesized by PCR (hexanucleotide labeling kit, Promega) according to the manufacturer´s instructions using the following primers: Probe Forw. 5'-TATGTTACAAGAAGGTGC-3' and Probe Rev. 5'-TGCGGTCATATTTTCACC-3'. Hybridization and washing of the Southern blot were performed as described before (Harhangi *et al*. 2002) except that washing conditions were less stringent (washing at 42°C with 2 × SSC).

Results and Discussion

Identification of genes encoding cellulosome components

Immunoscreening of a *Piromyces* sp. strain E2 cDNA library was used to identify genes encoding cellulosome components. A total of 60,000 plaques were screened using a polyclonal antibody raised against the cellulosome fraction of extracellular protein (Steenbakkers *et al*. 2002b). Fifty immunoreactive plaques were rescreened and a total of 32 cDNA clones were investigated in detail. The selected cDNA clones were subcloned using a newly developed subcloning procedure (see Experimental procedures) and partially sequenced. A comparison of the resulting partial sequences to the CAZy database indicated the isolation of genes encoding glycoside hydrolases belonging to families 3, 5, 9, 10, 26, 45, 48, 67 according to the classification by Henrissat & Davies (1997) (Fanutti *et al*. 1995; Steenbakkers *et al*. 2002a; 2002b; 2003; Eberhardt *et al*. 2000). These gene sequences showed a domain structure and all encoded one or more copies of a fungal dockerin domain, which indicated that all were cellulosome components (Steenbakkers *et al*. 2001). Next to the isolation of genes encoding glycoside hydrolases, also one partial cDNA clone was identified that encoded the C-terminal half of a fungal dockerin. The dockerin was followed by a 380 amino acid domain showing similarity to serine proteinase inhibitors (serpin). The gene encoded a cellulosomal serpin homolog and was named *celpin* and the encoded protein Celpin.

Isolation of full-length celpin

The 5' end of the *celpin* gene, which was missing in the cDNA clone, was isolated by inverse PCR. The PCR product obtained was of the expected size and the missing sequence was

determined. Primers were designed to amplify and isolate the full-length *celpin* sequence from the cDNA library using a proofreading polymerase. Southern blot analysis appointed the use of *TaqI* for generating a suitable template for inverse PCR. Southern blot analysis, which was performed with low stringency washing, revealed that *celpin* was present most likely as a single copy gene (Fig. 1). This is in contrast to recent results on non-cellulosomal and cellulosomal glycoside hydrolases of Piromyces sp. strain E2, which appeared to be present as multicopy gene families easily detectable even after high stringency washing (Harhangi *et al*. 2003).

The celpin *nucleotide sequence*

Analysis of the partial genomic and the full-length *celpin* cDNA sequence (Fig. 2) indicated the presence of a typical A/T-rich upstream and downstream region (88 %). No additional coding region could be recognized in the upstream genomic DNA sequence. The expression of cellulosomal cellulases of anaerobic fungi was shown to be catabolite-regulated by glucose and cellobiose (Morrison *et al*. 1990). Previously, putative regulatory sites were identified in the upstream regions of the *cel48A* genes of *Piromyces* sp. strain E2 and *P. equi* (Steenbakkers *et al*. 2002a). Here, two highly conserved palindromes were located in an otherwise unconserved upstream region at equal distances to the start codon. To identify possible regulatory transcription or translation sites, the upstream region of the *celpin* gene was analyzed for the presence of transcription elements, palindromes and repeat sequences (Fig. 2). Three putative CCAAT boxes (ending at nucleotides -359, -447 and -454) and two TATAAAA boxes (ending at positions -287 and -379) could be identified. Furthermore, a 34 nt palindrome as well as a partially overlapping 18 nt tandem repeat were located in the immediate upstream region (ending at positions -89 and -54, respectively). Further sequence analysis identified a 181 nt intron in the partial genomic *celpin* sequence, located in between codons 30 and 31 (Fig. 2). The relative position of the intron at the 5' end of the coding region, the intron size and A/Tcontent (92 %) were consistent with the characteristics of previously identified introns in genes from anaerobic fungi (Durand *et al*. 1995; Harhangi *et al*. 2003; Steenbakkers *et al*. 2002b. Also the conserved 8 nt palindrome, TATTAATA, presumably involved in lariat formation during gene splicing was found at 39 nt of the 3' end. Introns are rare in anaerobic fungi and always occur at the 5'-end of the genes. Analysis of codon usage gave results consistent with other genes from anaerobic fungi and displayed the characteristic preference for an A or T residue in the wobble position and a GC content of 31 % of the coding region.

Celpin amino acid sequence

Analysis of the encoded protein sequence of Celpin revealed a modular architecture (Fig. 2). SignalP (Nielsen *et al.* 1997) predicted a 21 amino acid signal peptide, which is in agreement with the extracellular nature of the components of the cellulosome and the protein preparation that was used for raising the polyclonal antibodies. The mature protein with a calculated mass of 59 kDa consisted of a small 20 amino acid N-terminal domain of unknown function containing predominantly positively charged amino acids (6 out of 20 amino acids). The coding nucleotide sequence of this domain was interrupted by the intron. This domain was followed by a 14 amino acid linker rich in threonine and two fungal docking domains of 36 and 39 amino acids, respectively. Both dockerins were characteristically cysteine-rich and based on the number and relative position of these residues, the dockerins could be classified as type 2 and 3, respectively (Fig. 3) (Steenbakkers *et al.* 2001). The presence of a fungal dockerin within a protein is indicative that this protein is part of the cellulosome (Steenbakkers *et al.* 2001). The two fungal dockerins were followed by a 20 amino acid linker and a 380 amino acid domain resembling serpins. Sequence comparison of this C-terminal domain using BlastP indicated 25 to 30 % sequence identity and 40 to 50 % sequence similarity to serine proteinase inhibitors. The serpins form a large superfamily of 370-450 amino acid proteins whose members are predominantly involved in regulating serine and cysteine proteinase activity (Gettins 2002). Serpins have first been identified and characterized in higher eukaryotes, e.g. insects, plants and mammals where they play crucial roles in key physiological processes, and in viruses where they play a role in circumventing the infected host's immune system. Recently, also serpin homologs were identified in unicellular eukaryotes, bacterial and archaeal genomes (Irving *et al.* 2002; Roberts *et al.* 2004), but so far no serpin domain has been identified within the fungal kingdom. Serpins act by serving as suicide substrates for their target proteinases. Upon cleavage, a large conformational change is triggered within the serpin, which completely disrupts the proteinase, resulting in a covalently linked serpin-proteinase complex (Pike *et al.* 2002). The serpin inhibiting mechanism therefore intrinsically requires a finely tuned set of structural interactions within the serpin to trigger full conformational change only at the covalent intermediate stage. This is reflected in a conserved set of structural features that are involved in the inhibiting process. A typical serpin core structure consists of 3 ß-sheets (numbered A to C), 8 α-helices (numbered A to H) and an exposed loop (the reactive center loop or RCL), that serves as substrate bait for target proteinases. The RCL and the largest ß-sheet (ß-sheet A) are

pivotal in the proteinase inactivation mechanism and the other structural elements are considered to stabilise the over-all structure or to regulate serpin activity (Belzar *et al*. 2002; Gettins, 2002). Upon cleavage of the scissile bond, a flexible hinge region allows the N-terminal RCL amino acids to rapidly integrate into ß-sheet A, forming an additional strand. This structural shift causes a disruption of the proteinase (Pike *et al*. 2002).

Conservation of serpin residues

To further investigate the novel fungal serpin, the 380 amino acids serpin domain of Celpin was aligned to human α_1-antitrypsin and a selection of representatives sequences (Roberts *et al*. 2004) (Fig. 4). Inspection of the alignment showed that the *Piromyces* sequence had two short extensions in between structural elements, but helix D was reduced in size. Deviations from the consensus of the alignment were also observed for the *Bifidobacterium longum* sequence, which had a 15 amino acid insertion between helix G and helix H. The Celpin serpin domains were analysed for the presence of the 51 amino acids conserved in > 70 % of known serpins, which are thought to reflect the serpin core structure (Irving *et al*. 2000). The sequences contained 39 of these 51 conserved amino acids, which is within the observed range for other serpin family members (44 ± 6). When chemically similar amino acids at the conserved positions were included, the number of conserved residues was 43. These results, in addition to the overall sequence similarity, further suggest that the identified sequence resembles a functional serpin. A comparison of the respective hinge regions to the consensus for inhibiting serpins (Irving *et al*. 2000) indicated that the identified serpin presumably will be involved in proteinase inhibition (Fig. 4). The *Piromyces* sp. strain E2 P_1-P_1' amino acids of the RCL, of which the P_1 amino acid mostly reflects the recognition site of the target proteinase (Irving *et al*. 2002), consisted of a lysine and a serine, respectively. Similar residues have been found for antithrombin III from different hosts, which suggests that the fungal serpin targets serine proteinases. Whether the fungal serpin has a broad or limited specificity remains to be resolved. Heterologous expression could be a helpful tool for biochemical characterization. However, thus far all attempts were not successful.

Physiological role of Celpin

Recently, three serpin homologs were identified in the cellulosome-producing thermophilic bacterium *Clostridium thermocellum* (Kang *et al*. 2006). The two largest *C. thermocellum* serpins shared 78% sequence similarity and were located tandemly on the genome suggesting recent gene duplication. Interestingly they both had an identical modular architecture and

contained a putative signal peptide, a 100 amino acid fibronectin type III-like domain (Kataeva *et al*. 2002), a bacterial cellulosome docking domain (Lytle *et al*. 2001) and a C-terminal serpin Apparently both bacterial and fungal cellulosomes contain serpins. The identification of the *celpin* gene from a cDNA library of the anaerobic fungus *Piromyces* sp. strain E2 offers some insights into the possible physiological role of these serpins. *A*naerobic fungi (chytrids) have only been cultured from the digestive tract of herbivorous mammals where they contribute to plant cell wall degradation (Lee *et al*. 2000; Orpin 1977). Cellulosomal serpins presumably protect the cellulosome against proteolytic degradation by proteinases from the directly surrounding environment. Several potential proteinase sources can be identified for the *Piromyces* sp. strain E2 serpin: the host's digestive system, competing microflora or the plant cell substrate. Recent studies on the molecular responses of damaged plant tissue to the rumen environment indicated that the dark, relatively warm, anoxic and chemically reduced environment triggered the massive release of proteolytic activity within plant cells and eventually caused cell lysis (Kingston-Smith *et al*. 2000; Lindsay *et al*. 2002).

Further biochemical studies of the novel serpin-like domains from prokaryotic and eukaryotic microorganisms and the putative cellulosomal serpin domains from the chytrid *Piromyces* sp. strain E2 and *C. thermocellum*, which might play a role in cellulosome defence, will shed new insights into the evolution of serpins, cellulosome functioning and the molecular interactions taking place within a very complex ecosystem like the rumen.

REFERENCES

Akhmanova A, Voncken FG, Harhangi HR, Hosea KM, Vogels GD, Hackstein JH, 1998. Cytosolic enzymes with a mitochondrial ancestry from the anaerobic chytrid *Piromyces* sp. E2. *Molecular Microbiology* **30**: 1017-1027.

Bayer EA., Chanzy H, Lamed R, Shoham Y, 1998. Cellulose, cellulases and cellulosomes. *Current Opinion in Structural Biology* **8**: 548-557.

Dijkerman R, Ledeboer J, Verhappen ABM, Op den Camp HJM, van der Drift C, Vogels GD, 1996. The anaerobic fungus *Piromyces* sp strain E2: Nitrogen requirement and enzymes involved in primary nitrogen metabolism. *Archives of.Microbiology* **166**: 399-404.

Dijkerman R, Bhansing DCP, Op den Camp HJM, van der Drift C, Vogels GD, 1997a. Degradation of structural polysaccharides by the plant cell-wall degrading enzyme system from anaerobic fungi: An application study. *Enzyme Microbial Technology* **21**: 130-136.

Dijkerman R, Op den Camp HJM, van der Drift C, Vogels GD, 1997b. The role of the cellulolytic high molecular mass (HMM) complex of the anaerobic fungus *Piromyces* sp. strain E2 in the hydrolysis of microcrystalline cellulose. *Archives of Microbiology* **167**: 137-142.

Ding SY, Rincon MT, Lamed R, Martin JC, McCrae SI, Aurilia V, Shoham Y, Bayer EA, Flint HJ, 2001. Cellulosomal scaffoldin-like proteins from *Ruminococcus flavefaciens*. *Journal of Bacteriology* **183**: 1945-1953.

Durand R, Fischer M, Rascle C, Fevre M, 1995. *Neocallimastix frontalis* enolase gene, enol: First report of an intron in an anaerobic fungus. *Microbiology* **141**: 1301-1308.

Eberhardt RY, Gilbert HJ, Hazlewood GP, 2000. Primary sequence and enzymic properties of two modular endoglucanases, Cel5A and Cel45A, from the anaerobic fungus *Piromyces equi*. *Microbiology* **146**: 1999-2008.

Fanutti C, Ponyi T, Black GW, Hazlewood GP, Gilbert HJ, 1995. The conserved noncatalytic 40-residue sequence in cellulases and hemicellulases from anaerobic fungi functions as a protein docking domain. *Journal of Biological Chemistry* **270**: 29314-29322.

Fillingham IJ, Kroon PA, Williamson G, Gilbert HJ, Hazlewood GP, 1999. A modular cinnamoyl ester hydrolase from the anaerobic fungus *Piromyces equi* acts synergistically with xylanase and is part of a multiprotein cellulose-binding cellulase-hemicellulase complex. *Biochemical Journal* **343**: 215-224.

Gettins PGW, 2002a. Serpin structure, mechanism, and function. *Chemical Reviews* **102**: 4751-4804.

Gettins PGW, 2002b. The F-helix of serpins plays an essential, active role in the proteinase inhibition mechanism. *FEBS Letters* **523**: 2-6.

Gordon GLR, Phillips MW, 1998. The role of anaerobic gut fungi in ruminants. *Nutrition Research Reviews* **11**, 133-168.

Harhangi HR, Steenbakkers PJM, Akhmanova A, Jetten MSM, van der Drift C, Op den Camp HJM, 2002. A highly expressed family 1 β-glucosidase with transglycosylation capacity from the anaerobic fungus *Piromyces* sp. strain E2. *Biochimica et Biophysica Acta* **1574**: 293-303.

Harhangi HR, Akhmanova A, Steenbakkers PJM, Jetten MSM, van der Drift C, Op den Camp HJM, 2003. Genomic DNA analysis of genes encoding (hemi-)cellulolytic enzymes of the anaerobic fungus *Piromyces* sp. E2. *Gene* **314**: 73-80.

Henrissat B, Davies G, 1997. Structural and sequence-based classification of glycoside hydrolases. *Current Opinion in Structural Biology* **7**: 637-644.

Irving JA, Pike RN, Lesk AM, Whisstock JC, 2000. Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function. *Genome Research* **10**: 1845-1864.

Irving JA, Steenbakkers PJM, Lesk AM, Op den Camp HJM, Pike RN, Whisstock JC, 2002. Serpins in prokaryotes. *Molecular Biology and Evolution* **19**: 1881-1890.

Kang S, Barak Y, Lamed R, Bayer EA, Morrison M, 2006. The functional repertoire of prokaryote cellulosomes includes the serpin superfamily of serine proteinase inhibitors. *Molecular Microbiology* **60**: 1344-1354.

Kataeva IA, Seidel RD, Shah A, West LT, Li XL, Ljungdahl LG, 2002. The fibronectin type 3-like repeat from the *Clostridium thermocellum* cellobiohydrolase CbhA promotes hydrolysis of cellulose by modifying its surface. *Applied and Environmental Microbiology* **68**: 4292-4300.

Kingston-Smith AH, Theodorou MK, 2000. Post-ingestion metabolism of fresh forage. *New Phytologist* **148**: 37-55.

Lee SS, Ha JK, Cheng K, 2000. Relative contributions of bacteria, protozoa, and fungi to in vitro degradation of orchard grass cell walls and their interactions. *Applied and Environmental Microbiology* **66**: 3807-3813.

Lindsey K, Casson S, Chilley P, 2002. Peptides: new signalling molecules in plants. *Trends in Plant Sciences* **7**: 78-83.

Ljungdahl LG, Blum DL, Chen H, He Y, Kataeva I, Li XL, Ximenes E, 1999. The cellulase/hemicellulase system of the anaerobic fungus *Orpinomyces* and aspects of further cellulase research, in: Ohmiya K, Hayashi K, Sakka K, Kobayashi Y, Karita S, Kimura T (Eds), *Cellulose degradation*. Uni Publishers Co., Tokyo, pp. 495-506.

Lockhart RJ, Van Dyke MI, Beadle IR, Humphreys P, McCarthy AJ, 2006. Molecular biological detection of anaerobic gut fungi (Neocallimastigales) from landfill sites. *Applied Environmental Microbiology* **72**: 5659-5661.

Lytle BL, Volkman BF, Westler WM, Heckman MP, Wu JH, 2001. Solution structure of a type I dockerin domain, a novel prokaryotic, extracellular calcium-binding domain. *Journal of Molecular Biology* **307**: 745-753.

Miras I, Schaeffer F, Beguin P, Alzari PM, 2002. Mapping by site-directed mutagenesis of the region responsible for cohesin-dockerin interaction on the surface of the seventh cohesin domain of *Clostridium thermocellum* CipA. *Biochemistry* **41**: 2115-2119.

Morrison M, Mackie RI, Kistner A, 1990. Evidence that cellulolysis by an anaerobic ruminal fungus is catabolite regulated by glucose, cellobiose and soluble starch. *Applied and Environmental Microbiology* **56**: 3227-3229.

Nielsen H, Engelbrecht J, Brunak S, von Heijne G, 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering* **10**: 1-6.

Orpin CG, 1977. The rumen flagellate *Piromonas communis*: its life-history and invasion of plant material in the rumen. *Journal of General Microbiology* **99**: 107-117.

Pages S, Belaich A, Belaich JP, Morag E, Lamed R, Shoham Y, Bayer EA, 1997. Species-specificity of the cohesin-dockerin interaction between *Clostridium thermocellum* and *Clostridium cellulolyticum*: Prediction of specificity determinants of the dockerin domain. *Proteins* **29**: 517-527.

Pike RN, Bottomley SP, Irving JA, Bird PI, Whisstock JC, 2002. Serpins: finely balanced conformational traps. *IUBMB Life* **54**: 1-7.

Raghothama S, Eberhardt RY, Simpson P, Wigelsworth D, White P, Hazlewood GP, Nagy T, Gilbert HJ, Williamson MP, 2001. Characterization of a cellulosome dockerin domain from the anaerobic fungus *Piromyces equi*. *Nature Structural Biology* **8**: 775-778.

Roberts TH, Hejgaard J, Saunders NFW, Cavicchioli R, Curmi PMG, 2004. Serpins in unicellular *Eukarya*, *Archaea*, and *Bacteria*: Sequence analysis and evolution. *Journal of Molecular Evolution* **59**: 437-447.

Steenbakkers PJM, Li XL, Ximenes EA, Arts JG, Chen H, Ljungdahl LG, Op den Camp HJM, 2001. Noncatalytic docking domains of cellulosomes of anaerobic fungi. *Journal of Bacteriology* **183**: 5325-5333.

Steenbakkers PJM, Freelove A, van Cranenbroek B, Sweegers BMC, Harhangi HR, Vogels GD, Hazlewood GP, Gilbert HJ, Op den Camp HJM, 2002a. The major component of the cellulosomes of anaerobic fungi from the genus *Piromyces* is a family 48 glycoside hydrolase. *DNA sequence* **13**: 313-320.

Steenbakkers PJM, Ubhayasekera W, Goossen HJ, van Lierop EM, van der Drift C, Vogels GD, Mowbray SL, Op den Camp HJM, 2002b. An intron-containing glycoside hydrolase family 9 cellulase gene encodes the dominant 90 kDa component of the cellulosome of the anaerobic fungus *Piromyces* sp. strain E2. *Biochemical Journal* **365**: 193-204.

Steenbakkers PJ, Harhangi HR, Bosscher MW, van der Hooft MMC, Keltjens JT, van der Drift C, Vogels GD, Op den Camp HJM, 2003. The β-glucosidase in the cellulosome of

the anaerobic fungus *Piromyces* sp. strain E2 is a family 3 glycoside hydrolase. *Biochemical Journal* **370**: 963-970.

Zhu WY, Kingston-Smith AH, Troncoso D, Merry RJ, Davies DR, Pichard G, Thomas H, Theodorou MK, 1999. Evidence of a role for plant proteases in the degradation of herbage proteins in the rumen of grazing cattle. *Journal of Dairy Science* **82**: 2651-2658.

Legends to the figures

Fig 1 - Southern blot of the *celpin* gene from *Piromyces* sp. strain E2. Restriction enzymes used are given at the top of the lanes. The probe used covered the serpin domain. The fragments observed correspond with the sequence of the *celpin* gene. No additional fragments are observed despite low stringency washing, which indicates that the *celpin* gene is a single copy gene.

Fig 2 - Composite genomic (AY437869) and cDNA sequence (AY437868) of *celpin* and the encoded amino acid sequence of the Celpin protein. The nucleotide numbering is given on the left with the first nucleotide of the startcodon numbered as +1. The putative transcription and translation regulatory sites are underlined and depicted in bold. The start codon and intron sequence are given in bold and the 8 nt palindrome in the intron is underlined. The *TaqI* site used in inverse PCR is highlighted in grey. The amino acid sequence is given underneath the DNA sequence with the signal peptide depicted in bold. Linker sequences are underlined. The stop codon is indicated by an asterisk (*).

Fig 3 - Alignment of the dockerin domains of Celpin to other dockerins from anaerobic fungi. Positions of the dockerins in the proteins are given by the numbers. Residues conserved in 3 out of 4 dockerins are highlighted. The presence of the 4 or 6 cysteine residues (arrowheads) and their respective sequence context is typical for type 2 and type 3 dockerins, respectively (Steenbakkers *et al*. 2001).

Fig 4 - ClustalW alignment of the 380 amino acids serpin domain of Celpin to human α_1-antitrypsin and a selection of representatives sequences (Roberts *et al*. 2004). Assignment of secondary structures of the cleaved form of human α_1-antitrypsin is included above this sequence. Levels of shading indicate conservation of residues. For clarity, a gap was introduced in all sequences after the P_1 residue. Accession numbers are: *Homo sapiens* (AAH11991); *Entamoeba histolytica* (XP650262); *Methanococcoides burtonii* (YP565417); *Bifidobacterium longum* (NP695337); *Clostridium thermocellum* (YP001037695); *Oceanobacillus iheyensis* (NP694094); *Thermobifida fusca* (YP289989).

Steenbakkers et al. Fig. 1
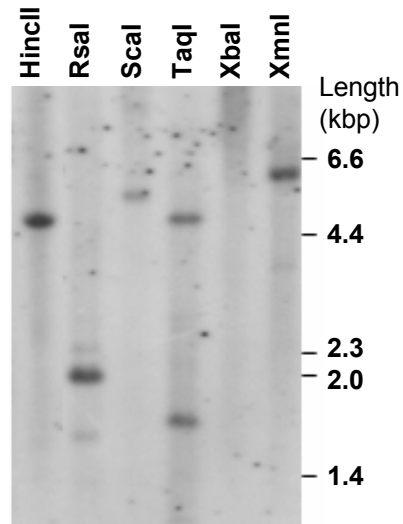

Fig 1 - Southern blot of the *celpin* gene from *Piromyces* sp. strain E2. Restriction enzymes used are given at the top of the lanes. The probe used covered the serpin domain. The fragments observed correspond with the sequence of the *celpin* gene. No additional fragments are observed despite low stringency washing, which indicates that the *celpin* gene is a single copy gene.

```
-749  TTTATTGTATTCTAATTATTTATTTTTTTTTTGCTAATTAAAATATAACTACTCAAGTTAAGGTCAAAGTCTTAAACTTTTTTTTCCCTT
-659  TTACTCTATAAAAGGTTTATCTATAAAAAAATATAGAAATATTAAAAAAAAAAACATATTTAAAAATTTAACATTTTTTTCTTTTATATAT
-569  TTATATATATTTATATATTTTTAATATTATTTAATTAATGACACATTTTATTTATAGATTAAAAATACTATTGTTATAATCATTTACAAT
-479  AACAATAATATAACACCATAACCAATAACCAATAACCATAAAAAAGATATTAAATCAACAGAGATGTAATCGTATAACATTTCTCATTTT
-389  TTAATATAAAAACATTAAATTTTATCCAATAATAATAAAAAAATTTTATAAATTCTCTATAAGATAAAGTTTAATATTAAAAAAAAAA
-299  TATAAATATAAAAAGAAAATATTTTGCTCATTTTTTTTATTTTTTTCTTTTTTTTTTTATTTAATTACTTAACTATTTAAAATATAA
-209  AAATTGTTCTTTTTTTTTTTTTTTTTTTTTAATCAATAGGAATTAAAATTAAATTTTTTATTGATATAATAAAACATTTAAAGGAAAGAAAAA
-119  AAAAAAAAAACTTTAAAGTTTTGTTTTTTTTATATATTAATTAGAATATAATAATTAGAATATAATAATAAAAAGAATTATTATAACAAT
-29   TATTTGTAATTAATAAAAAATTTTAAAAATATGAAGAACTTATCATTATATCTTTTAGGATTGGTATTTATTTATATACCAGCATGTTTAT
                                        M  K  N  L  S  L  Y  L  L  G  L  V  F  I  Y  I  P  A  C  L

62    CCCAAAAGAATATCAGATTTGACCCAAACGTAAGAAGAAGAATAATTAAAAAAAAAAAAAATAAAATAAAATAAAAAATAAATATAAATAT
      S  Q  K  N  I  R  F  D  P  N

152   AAATATATATATATATTATCTAATAAAAAATTATATCTAATATATACATTATATCTAATTTGAATATTTTTATTATTAATATAAACTTAT

242   TAATATAAACTTGGTAAAAAATAAAAAAAGAATGGTAAAATTAACCCATGGAATAAAATAAAGAAGACAACTACTACTACTACTACTACT
                                                                     N  G  K  I  N  P  W  N  K  I  K  K  T  T  T  T  T  T  T

332   CAAACTAGTAGTCCTACTTGTTTCTCACAAAAATTAAAACCATCCTACCCATGTTGTAAAGGTAATAAAGTTATTTATACCGATAAAGAT
      Q  T  S  S  P  T  C  F  S  Q  K  L  K  P  S  Y  P  C  C  K  G  N  K  V  I  Y  T  D  K  D

422   GGATATTGGGGTATAGAAAATGGAAAATGGTGTGGTATTGATCCTTCTCTCAATTCCTGTTTCTCTGTCGCTCTTGGCTATCCATGTTGT
      G  Y  W  G  I  E  N  G  K  W  C  G  I  D  P  S  L  N  S  C  F  S  V  A  L  G  Y  P  C  C

512   AAATCTTGTAAGGTATATGTCACTGATAAATATGGAGATTGGGGTGTAGAAAATAATAAATGGTGTGGTATTAAGAAAAGTTGTCCCACC
      K  S  C  K  V  Y  V  T  D  K  Y  G  D  W  G  V  E  N  N  K  W  C  G  I  K  K  S  C  P  T

602   TCCAATAACAATGGAGGTCAAGGTGATATAATAGAAGAACCTGCCCAAAATAACACCGATGTTTCCACCACCGAATTTGAATTTTCATTC
      S  N  N  N  G  G  Q  G  D  I  I  E  E  P  A  Q  N  N  T  D  V  S  T  T  E  F  E  F  S  F

692   TTAAAAATGGAAAACAAAAAGAAAAATATGTTATACAGTCCATTATCTATAGAATATGCTTTAAATATGTTACAAGAAGGTGCTAATAAG
      L  K  M  E  N  K  K  K  N  M  L  Y  S  P  L  S  I  E  Y  A  L  N  M  L  Q  E  G  A  N  K

782   AATACTTACACGGAAATAAATAAAGTGATAGGAAATGCAGAACTTCCAAAATACGCTAGCATCGAAAAAGTTTTATCCTTAGCCAATGGA
      N  T  Y  T  E  I  N  K  V  I  G  N  A  E  L  P  K  Y  A  S  I  E  K  V  L  S  L  A  N  G

872   TTATTTGTAAGAGACACTTACTACGAAAATGTCAGAACTGAATATATCAAAACATTAAAAGAAAAATATGATGCAGAAATAATAAAGGAT
      L  F  V  R  D  T  Y  Y  E  N  V  R  T  E  Y  I  K  T  L  K  E  K  Y  D  A  E  I  I  K  D

962   GAATTTAAAGATGCTAACAATGCTAATCAATGGATTGAAGATAAAACTTTAGGAATTATTAAAAATATGTTAGAAGATGATTTAGTACAA
      E  F  K  D  A  N  N  A  N  Q  W  I  E  D  K  T  L  G  I  I  K  N  M  L  E  D  D  L  V  Q

1052  AATCCAAACCTTGTAATGCTTATTATTAATGCCCTCGCCATTGATATGGAATGGGCTACCCAATTTAGTACTGAAAAAACTTACGGGGAA
      N  P  N  L  V  M  L  I  I  N  A  L  A  I  D  M  E  W  A  T  Q  F  S  T  E  K  T  Y  G  E

1142  ACATTTTATAAGGATGATGGTGAAAATATGACCGCAACAATGATGTTTAAAAAAGAAGAATATAGCGATGATATTGCCTATTACATGGAT
      T  F  Y  K  D  D  G  E  N  M  T  A  T  M  M  F  K  K  E  E  Y  S  D  D  I  A  Y  Y  M  D

1232  GATGACGTAACAGTTCTCTCTATGAATTTAGAGAAATATAGTGGAACCCAATTTGAATTTTTAGCCATCATGCCAAAAGAAAACTTAAGT
      D  D  V  T  V  L  S  M  N  L  E  K  Y  S  G  T  Q  F  E  F  L  A  I  M  P  K  E  N  L  S

1322  GGTTATGTTGAAAATGTAACCAAAGAACAAATAAATGAATTAGATACAAAACTTAAATTATCTTCAAATGAACCAGACGGTGTTAACATA
      G  Y  V  E  N  V  T  K  E  Q  I  N  E  L  D  T  K  L  K  L  S  S  N  E  P  D  G  V  N  I

1412  AAAATTCCAAAATTCAAATTTAGCTATGATTTAGAATTAAAAAAAGATTTAATAAATTTAGGTATAAAAGATGCTTTTGATGAAAAGGCA
      K  I  P  K  F  K  F  S  Y  D  L  E  L  K  K  D  L  I  N  L  G  I  K  D  A  F  D  E  K  A

1502  GCTGACTTTACAAAGATGGCTTACAGAAATGAACTTTATGTGTCCGAAGCACTTCATAAAGCTGATATTGAATTTACCGAAAAGGGAGTA
      A  D  F  T  K  M  A  Y  R  N  E  L  Y  V  S  E  A  L  H  K  A  D  I  E  F  T  E  K  G  V

1592  AAAGCAGCCGCAGTAACCGTCTTCGGAATGATGACTAAATCTGCTATGCCTATTCAAAGAAAACCAGTTAATGTTGTCATTAACAAACCA
      K  A  A  A  V  T  V  F  G  M  M  T  K  S  A  M  P  I  Q  R  K  P  V  N  V  V  I  N  K  P

1682  TTTATGTTTGTTATTAGAGATAAAGTACAAAGGGTATATGGTTCACTGGTACAGTATATGAACCAAATTCATGGGAAAAGGATAAGGAT
      F  M  F  V  I  R  D  K  S  T  K  G  I  W  F  T  G  T  V  Y  E  P  N  S  W  E  K  D  K  D

1772  AGTTACATTCCAACCTCTAGATTTTAAGTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
      S  Y  I  P  T  S  R  F  *
```

Fig 2 - Composite genomic (AY437869) and cDNA sequence (AY437868) of *celpin* and the encoded amino acid sequence of the Celpin protein. The nucleotide numbering is given on the left with the first nucleotide of the startcodon numbered as +1. The putative transcription and translation regulatory sites are underlined and depicted in bold. The start codon and intron sequence are given in bold and the 8 nt palindrome in the intron is underlined. The *TaqI* site used in inverse PCR is highlighted in grey. The amino acid sequence is given underneath the DNA sequence with the signal peptide depicted in bold. Linker sequences are underlined. The stop codon is indicated by an asterisk (*).

```
Celpin dockerin #1           57   CFSQKLKPSYPCCK-GNKVIYTDKDGYWGIENGKWCGI    93
Pir.E2 (AMM81967) Cel9A #1   655  CFSVA--QGYPCCGAGIPVSYEDDSGQWGIENGNWCGI   690
Pir.E2 (AY172977) Cel3A #2   786  CFSI--KLGYPCCK-GNEVAYTDNDGQWGFENGQWCGI   820
Pir.eq. (P55298) ManC        533  CWSEA--LGYPCCVSSSDVYYTDNDGEWGVENGDWCGI   568
                                  ▲         ▲▲                      ▲
```

```
Serpin dockerin #2           100  CFSVALGYPCCKSCKVYVTDKYGDWGVENNKWCGIKKS-C  138
Pir.E2 (AMM81967) Cel9A #3   830  CTGQNLGYPCCDTCEAIYTDESGKWGIKNGDWCGLKSS-C  866
Pir.eq. (CAB92326) Cel5A     1674 CWSIPYGYKCCDHCRVLDKDETGKWGEMNGEWCGIDTNKC  1708
Orp.pc. (AAB92679) CelC      61   CKFEALGYSCCKGCEVVYSDEDGNEGVENQQWCGIRDN-C   99
                                  ▲        ▲▲                    ▲       ▲
```

Fig 3 - Alignment of the dockerin domains of Celpin to other dockerins from anaerobic fungi. Positions of the dockerins in the proteins are given by the numbers. Residues conserved in 3 out of 4 dockerins are highlighted. The presence of the 4 or 6 cysteine residues (arrowheads) and their respective sequence context is typical for type 2 and type 3 dockerins, respectively (Steenbakkers *et al*. 2001).

```
                                                                ←——— hA ———→   <s6B>←—— hB —

H.sapiens      : ------------------MPSSVSWGILLLAGLCCLVPVSLAEDPQGDAAQKTDTSHHDQDHPTFNKITPNLAEFAFSLYRQLAHQSNSTNIFFSPVSIATAFAML : 88
E.histolytica  : ----------------------------------------------------MSMDYQDIENMQIALYKLCVDWYNSSPIKEDIVFSTHSMFIAFSLL : 46
Piromyces sp. E2: ----------------------------------------------------------EEPAQNNTDVSTTEFEFSFLKMENKKKNMLYSPLSIEYALNML : 43
M.burtonii     : ---------------------------MKKIKTLLVVTLCIVSSLFLGCIEDSTVNTKNTINADSVEEYDIATANNAFAFDMYSMIESEDENIFFSPYSIFTAMAIC : 80
B.longum       : MSEQLMEQYRLRGQRKCRNACIAAIVTVVLVLAVAGGVWWTAGDGSALVRNMFKPKATPATQPVVNSTATFAYRTAPEFLAMEAGDRGTGNVNYSPASMWMALAIA : 106
C.thermocellum : -MQIKMSCPMELYFITYVKKPVRINLTGVEKMKRTACAVLCIFVLTFLFSGCSREKKVFDELKLDTELEKKNTEFCFDIFSKLNEEDYNKNIFISPLSISTVLSMT : 105
O.iheyensis    : ------------------MKKCFLIIPILGLFLLFSACGTDDNSVTGGNSTSDNPPTTKIDTSVLVNDANQASNHLGFQALQRLEGNDEDNIFVSPLSIWLAVSMA : 88
T.fusca        : -------MLHPRAREYATCPFHNRKRSAAVSVGRVAAAPRGNPAGAGTRARSPQAPSRGVSMSGGFLRDDHLEFALHLHRRLAEAVPDGEVIWSPYSVACALGVL : 98


                   →   ←——— hC —→     ←——— hD ———→       ←—— s2A —→     ←— hE —→ <s1A>     ←——— hF ———→

H.sapiens      : SLGTKADTHDEILEGLNFNLTEIPEAQIHEGFQELLRTLNQPDSQLQLTTGNGLFLSEGL--KLVDKFLEDVKKLYHSEAFTVNFGD--TEEAKKQIND-YVEKGT : 189
E.histolytica  : YIGAAAETKTQLEKVFGFAS--IPERNFIKFLQSIIKQQDPT-MSVTVDIVNGIWASQKL--EFTEEYKKAITTL-DCQLKNVNFGN-DSENIRQEINK-FVEEAT : 144
Piromyces sp. E2: QEGANKNTYTEINKVIGNA--ELP-------KYASIE--------KVLSLANGLFVRDTYYENVRTEYIKTLKEKYDAEIIKDEFKD--ANNANQWIEDKTLGIIK : 130
M.burtonii     : YDGAEGSTKKQISNVFYYP---LSKPVLEESSKEMIDTINSANDEYDLKTANALWIRKNY--PLNEQFAHNLKIYYDGNVTNVDFRN-EPEKSKDIINE-WVSTNT : 179
B.longum       : AQGANGTTRSQLNELLGSGS--LTDSDYQSLLSSINGQYS--GAKSEMSAANSLWIDDDY--SLASDYQSTVKKMFEAEVTTLPFDDQA----AAKMSD-WIAKHT : 201
C.thermocellum : VQGAGTTTKDGMLKALKYDG--MDLDKINESYRYIIDYLSKTDKTIELEINNSIWIREGK--QIKKDFIDINKDVYNAYVTELDFSSP---NAADRINK-WISDST : 203
O.iheyensis    : YNGANGDTKDEMTNALQPEN--IDLQTLNEHNGTMMDQLNQH-DDVELNILNSIWLDPEF--TFLEDFEHQVTESYNPELGPLTT-------KEPMNE-WVSEKT : 180
T.fusca        : AAGARATTRTELTTLLG---------TDPAPLLAALDRAVA--DSPDLASRTVLWVSADV--PVRSSFRATVHDRPDSDVRTADFRTN-PEGVRATVNA-DIADAT : 189


                            ←——— s3A ———→       <s4C>        ←— s3C —→<s1B>      < s2B >       < s3B >

H.sapiens      : QGKIVDLVKELDR---DTVFALVNYIFFKGKWERPFEVKDTEEEDFH--VDQVTTVKVPMMKRLGMFNIQHCKK-----LSSWVLLMKYL--GNATAIFFLP-DEG : 282
E.histolytica  : RKVIVDFLQPGTISG-DTIAVIVNAIYFKGEWETPFKIVQK-KMKFE------GDEEVVVMKERIECSAVFTEK-------YTSVSIPYV-GNQYSMVIIMPNNMK : 234
Piromyces sp. E2: NMLEDDLVQNP-----NLVMLIINALAIDMEWATQFSTEKIYGETFY--KDDGENMTATMMFKKEEYSDDIAYYMDDDVTVLSMNLEKYS-GTQFEFLAIMPKENL : 228
M.burtonii     : NGKIKDIISDEMIDPYNTAIIITNAIYFKGKWINEFDIENTQKELFYNSSSNDEGTLIDMMYTRQYFNYGESED-------AKIVELPYK-GNDLCMYIVLP-EEN : 276
B.longum       : NGSLKPKITLRD----REVLSIINTVYADGRWKDPFEEQSTGNGTFHGEAGDAQVPMHQTFSQMAYGHDEYNT-------WQRVEIPFD--NGGNLAIVLP-AEG : 293
C.thermocellum : KKKITDIIDSPIPE--NTAMFLINAIYFKGDWAEKFKKQDTFTAKFQ--SGNGQTKEVMMMERKDTIEYGAKED-------FKVVRLPYG-KGTTSMYCVLPAKDV : 297
O.iheyensis    : NGKITDLIEKVDDD---HVAFLLNATYFNGNWKYPFDENNIQDGLFH--LDESSTTEVPFMSLDEELPYWETDE-------FQSVALPYGEEGNIQMEIFLPNENI : 274
T.fusca        : RGMIRELLPQGAVTP-DLRAILTNALWARARWTTPFEAHLIREGTFRT---PRGPKRVPFMHRTGTMPYATARG-------WGMVTLHAH--DELAVDVLLP-PGT : 281


                   ← hG →              ← hH →          <s2C>←— s6A —→← hI →                        ←——

H.sapiens      : KLQHLENE----------------LTHDIITKFLENE------DRRSASLHLPKLSITG---TYDLKSVLGQLGITKVFSNG-ADLSGVTE----EAPLKLSKAVH : 358
E.histolytica  : EFEKENMG----------------ELKEYVRRTIQE-------FSEKRNVTIPKFKIET---SFSMNQQLKQLGLINAFDER-ADFSKMAK-----GHFCVSEAIH : 308
Piromyces sp. E2: SGYVENVT----------------KEQINELDTKLKL---SSNEPDGVNIKIPKFKFSY---DLELKKDLINLGIKDAFDEKAADFTKMAY----RNELYVSEALH : 308
M.burtonii     : NIDDFENR----------------FTLSYYNKLKSSME------SEKDVRIWLPKFKFNT---KNELSNTLRNMGIVEAFTSN-ANFSSIST----RGDLSISEVIH : 353
B.longum       : HFDELAGDAEKLSWAFGTCSTASLGEGAMGCAADSMPGWGVSVNSVMVNVTLPRFTIDSMFDSEATIKAFEKLGVTDAFSAGDADFTKMIDTGSHGENLYIGSILQ : 399
C.thermocellum : SINDFIKT----------------LDVNKWEEIKNSIS------KAENVTLNIPRFKIAYG--TKELRDCLIAMGMEEAFTER-ADFSGIS-----EGLLFIDSVIH : 374
O.iheyensis    : ELQEFVNK----------------LTLEDWQNWRESMQ-------ETVGSLKMPKFSLEY---ESELQDFFQVLGMEKAFEEKEADFTNMIDPAQLRGNLYISKIMH : 355
T.fusca        : NAAAVP------------------TAPLLTALHRRS-------ASTSVELALPRFELTQ---PHQLVEVLAEAGVRTLFTAS-ADLSGIST-----VPLYVDTVIH : 353


                   s5A →      ←——— s4A ———→ P₁  P₁'   <s1C>    ←—s4B→    ←—s5B→

H.sapiens      : KAVLTIDEKGTEAAGAMFLEAIPM--SI------PPEVKFNKPFVFLMIE-QNTKSPLFMGKVVNPTQK-------------- : 418
E.histolytica  : KAVVEVDEKGTIAAAATGLALMRC--CLPL--EPPRDVIINKPYFFVIIG-EEQ-YPLFFGKVSHPRFK-------------- : 371
Piromyces sp. E2: KADIEFTEKGVKAAAVTVFGMMTK--SAMPIQRKPVNVVINKPFMFVIRDKSTKG-IWFTGTVYEPNSWEKDKDSYIPTSRF : 387
M.burtonii     : QAYIDVNEEGTEAAAATALEATDS--APMP--GQIMEFKADHPFMFFIEDK-RTGSILFMGKVGSPESEEMS---------- : 420
B.longum       : GTRIEVNEAGAKAMSFTKVGADSV--SAPV---DNVEFTVDRPFLYSYVT--PDGIPLFIGAVRNLGGVGGEN--------- : 465
C.thermocellum : KAIIEVNEDGSTAAGSTVVRMIDG--AAIG---EPLSFIADRPFLFFITE-DVTGTILFMGKLYDCEKY------------- : 437
O.iheyensis    : KTFLEVDVRGTEAAGATSVEMKEE--SAVV-SPAPFDMEINRPFLLTISDK-ETDSILFMGMIHEPTST------------- : 420
T.fusca        : QARLRVDERGAEGAAATAAMMLLA--GAMPP-RRTIRFSVDRPFHIVVRRR---GAILFLGSIADPHDPGPAQ--------- : 420
```