while *L. major* and *T. brucei* lack AP-4 or AP-2, respectively, which suggests that the repertoire of AP complexes in kinetoplastids is variable and species-specific. Although low similarity of the *E. histolytica* components to either yeast or mammalian orthologues make unequivocal assignment of *Entamoeba* AP complexes challenging, tentative assignments have been made. It is likely that *E. histolytica* encodes four kinds of AP complex corresponding to APs 1–4.

6.5. Glycosylation and protein folding

6.5.1. Asparagine-linked glycan precursors

Mammals, plants, *Dictyostelium* and most fungi synthesise asparagine-linked glycans (N-glycans) by means of a common 14-sugar precursor dolichol-PP-$Glc_3Man_9GlcNAc_2$ (Figs. 2.7 and 2.8) (Helenius and Aebi, 2004). This lipid-linked precursor is made by at least 14 glycosyltransferases,

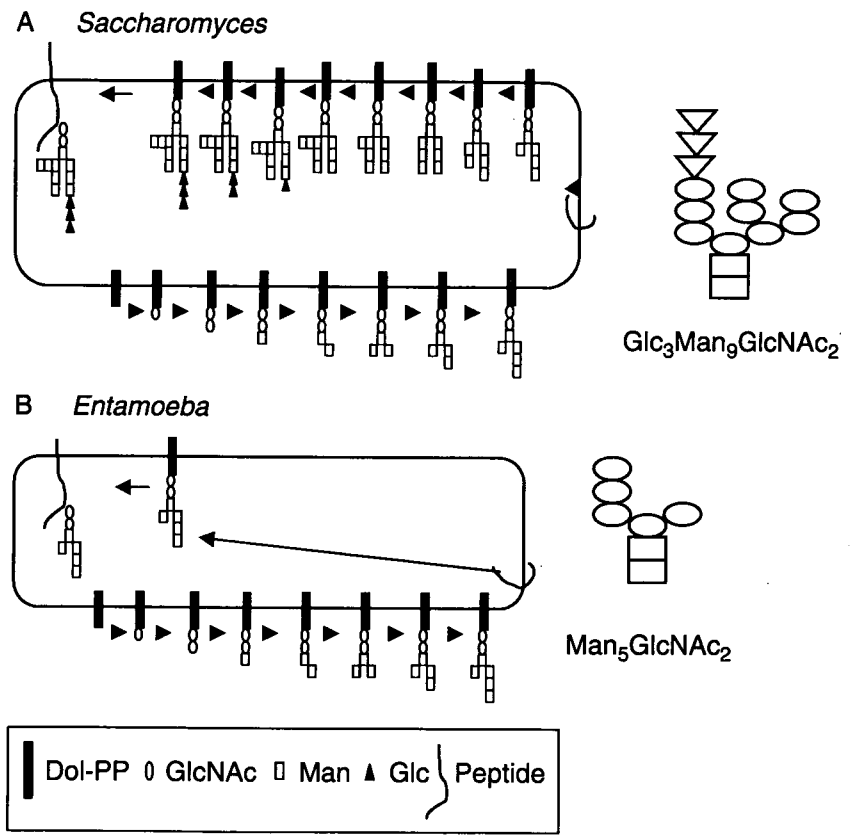

FIGURE 2.7 Synthesis of N-glycan precursors by *S. cerevisiae* (A) and *E. histolytica* (B). The N-glycan precursor of *S. cerevisiae* contains 14 sugars ($Glc_3Man_9GlcNAc_2$), each of which is added by a specific enzyme. The *E. histolytica* N-glycan precursor contains just seven sugars ($Man_5GlcNAc_2$), as the protist is missing enzymes that add mannose and glucose in the lumen of the ER. The figure is modified from Figure 1 of Samuelson *et al.* (2005). Glc = Glucose; GlcNAc = *N*-acetyl glucosamine; Man = Mannose.

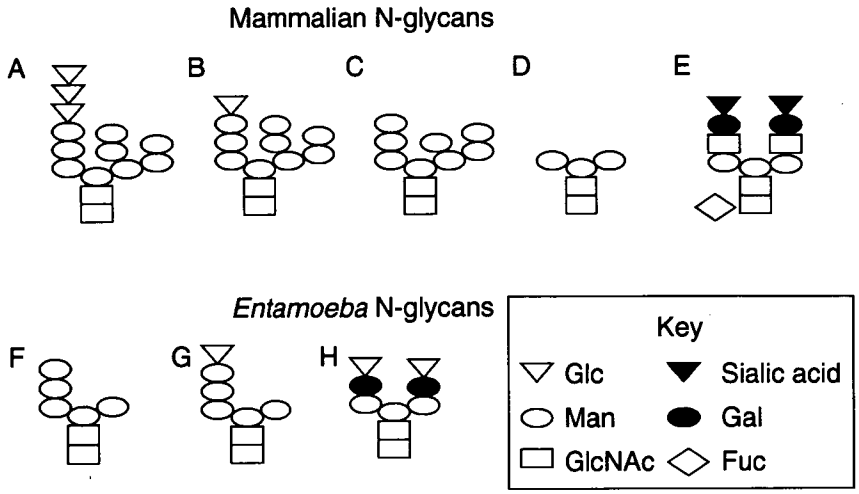

FIGURE 2.8 Selected N-glycans of mammals (A–E) and *Entamoeba* (F–H). Precursors transferred to nascent peptide (A and F). Glycosylated products involved in N-glycan-associated QC of protein folding (B and G). Mannosidase product involved in N-glycan-associated protein degradation (mammals only) (C). Trimmed product that is building block for complex N-glycans (mammals and *Entamoeba*) (D). Complex N-glycans made in the Golgi (E and H). Glc = Glucose; GlcNAc = *N*-acetyl glucosamine; Man = Mannose; Gal = Galactose; Fuc = Fucose.

which are present in the cytosolic aspect or lumen of the ER. The reducing end of the glycan contains two *N*-acetylglucosamines, while nine mannoses are present on three distinct arms. Three glucoses are added to the left arm, which is the same arm that is involved in the quality control (QC) of protein folding (see next section) (Trombetta and Parodi, 2003).

Entamoeba is missing luminal glucosylating and mannosylating enzymes and so makes the truncated, seven-sugar N-glycan precursor dolichol-PP-$Man_5GlcNAc_2$ (Figs. 2.7 and 2.8) (Samuelson *et al.*, 2005). Five mannoses on this N-glycan include the left arm, which is involved in the quality control of protein folding. In contrast, *Entamoeba* is missing the middle and the right arms, which are involved in N-glycan associated QC of protein degradation (see next section). Because *Dictyostelium*, which is phylogenetically related to *Entamoeba*, makes a complete 14-sugar N-glycan precursor, it is likely that *Entamoeba* has lost sets of glycosyltransferases in the ER lumen (Samuelson *et al.*, 2005). Similarly, secondary loss of glycosyltransferases best explains the diversity of N-glycan precursors in fungi, which contain 0–14 sugars, and apicomplexa, which contain 2–10 sugars (Samuelson *et al.*, 2005).

The 14-sugar N-glycan precursor of mammals, plants, *Dictyostelium* and most fungi is transferred to the nascent peptide by an oligosaccharyltransferase (OST), which is composed of a catalytic peptide and six to seven non-catalytic peptides (Kelleher and Gilmore, 2006). In contrast, the *Entamoeba* OST contains a catalytic peptide and just three non-catalytic peptides, while other protists (e.g., *Giardia* and *Trypanosoma*) have an OST

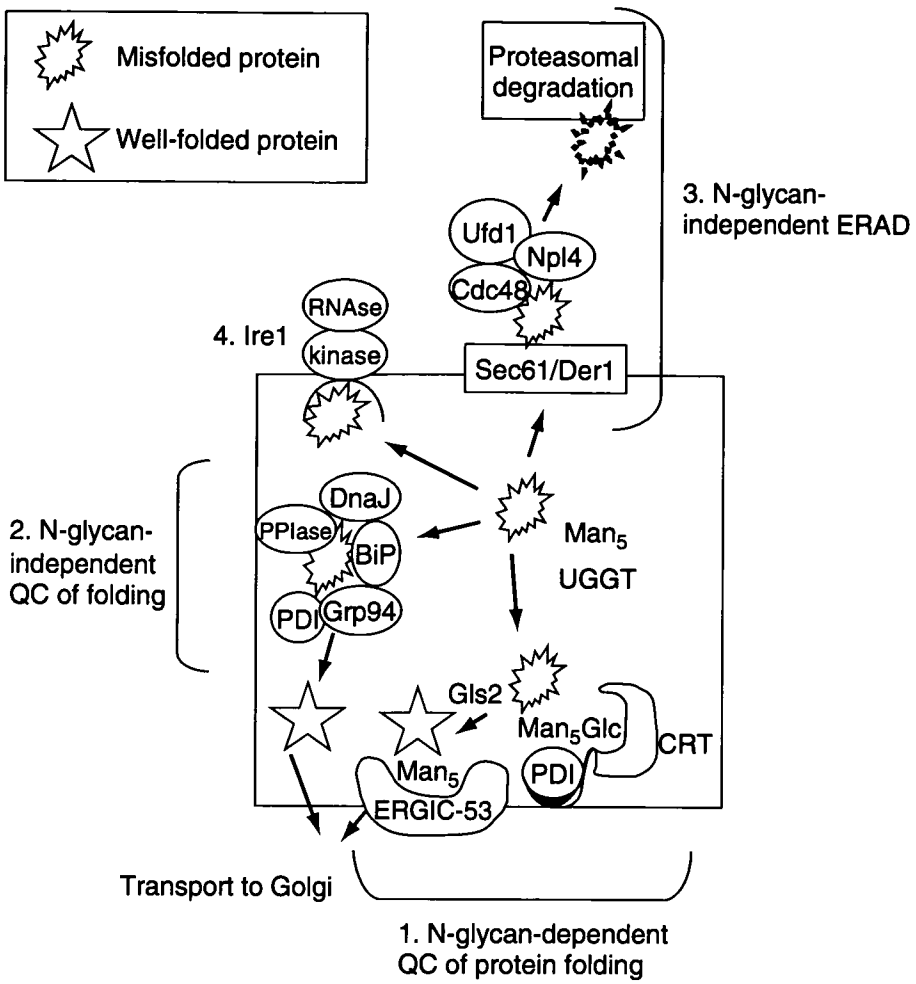

FIGURE 2.9 Model of quality control of protein folding in *Entamoeba*. 1. N-glycan-dependent QC of protein folding. 2. N-glycan-independent QC of protein folding. 3. N-glycan-independent ERAD. 4. Ire1 and unfolded protein response (see text for details).

with a single catalytic peptide. This reduced complexity does not likely affect the site of N-glycan addition to the nascent peptides, which is NxS or NxT (the so-called sequon) (Kornfeld and Kornfeld, 1985).

6.5.2. N-glycans and quality control of protein folding

Protein folding in the lumen of the ER is a complex process that involves N-glycan-dependent and N-glycan-independent QC systems (Helenius and Aebi, 2004; Trombetta and Parodi, 2003). *Entamoeba* has four of five systems present in higher eukaryotes for protein folding (Fig. 2.9).

1. *Entamoeba* has the minimum component parts for N-glycan-dependent QC of protein folding (Helenius and Aebi, 2004; Trombetta and Parodi, 2003). These include a UDP-glucose-dependent glucosyltransferase (UGGT), which adds a single glucose to the

left arm of the N-glycans of misfolded proteins and so forms $GlcMan_5GlcNAc_2$ (Fig. 2.7). The glucosylated N-glycan is then bound and refolded by the lectin calreticulin (CRT), which is a chaperone that works with a protein disulphide isomerase (PDI) to make and break disulphide bonds. A glucosidase (Gls2) removes glucose from the well-folded protein, which is transferred to the Golgi by a mannose-binding lectin (ERGIC-53). The *Entamoeba* system is similar to that of mammals and fungi, which add glucose to the $Man_9GlcNAc_2$ precursor to make $GlcMan_9GlcNAc_2$ (Fig. 2.7). Mammals have a second glucosidase to remove glucose from the $Glc_3Man_9GlcNAc_2$ precursor (Fig. 2.7).

2. *Entamoeba* has N-glycan-independent QC of protein folding within the lumen of the ER, which includes the chaperones Hsp70 and Hsp90 (also known as BiP and Grp94, respectively) (Fig. 2.9) (Helenius and Aebi, 2004; Trombetta and Parodi, 2003). Also involved in this QC system are PDIs; DnaJ proteins that increase the ATPase activity of Hsp70 and Hsp90; and peptidyl-prolyl *cis-trans* isomerases (PPIases). This N-glycan-independent QC system for protein folding is present in all eukaryotes (S. Banerjee, J. Cui, P. W. Robbins, and J. Samuelson, unpublished data).
3. *Entamoeba* and all other eukaryotes have a N-glycan-independent system for ER-associated degradation (ERAD) of misfolded proteins (Fig. 2.9) (Hirsch *et al.*, 2004). This system is composed of proteins (Sec61 and Der1) that dislocate misfolded proteins from the ER lumen to the cytosol. There a complex of proteins (Cdc48, Npl4 and Ufd1) ubiquinate misfolded proteins, which are then degraded in the proteasome. In contrast, *Entamoeba* and the vast majority of eukaryotes are missing an N-glycan-dependent system of ERAD of misfolded proteins (Helenius and Aebi, 2004; Trombetta and Parodi, 2003). In this system, the middle arm of $Man_9GlcNAc_2$ is trimmed to $Man_8GlcNAc_2$, which is recognised by a unique mannose-binding lectin (EDEM) before dislocation into the cytosol for degradation (Fig. 2.9).
4. *Entamoeba* has a transmembrane kinase (Ire1) which recognises misfolded proteins in the lumen of the ER and triggers the unfolded protein response (Fig. 2.9) (Patil and Walter, 2001). The *Entamoeba* unfolded protein response is likely to be different from those of mammals and fungi, because *Entamoeba* is missing an important downstream target, which is a transcription factor called Hac1.

6.5.3. Unique N-glycans

Mammals make complex N-glycans in the Golgi by trimming back the precursor to $Man_3GlcNAc_2$ and then adding *N*-acetyl glucosamine, galactose, sialic acid and fucose (Fig. 2.8) (Hubbard and Ivatt, 1981). In each case, the activated sugars (UDP-GlcNAc, UDP-Gal, CMP-sialic acid and

GDP-fucose) are transferred from the cytosol to the lumen of the Golgi by a specific nucleotide-sugar transporter (NST) (Hirschberg *et al.*, 1998). In turn, each activated sugar is added to the N-glycans by a specific glycosyltransferase.

Entamoeba N-glycans are remarkable for two properties. First, the most abundant N-glycan is unprocessed $Man_5GlcNAc_2$ (Fig. 2.7) (P. E. Magnelli, D. M. Ratner, P. W. Robbins and J. Samuelson, unpublished data). This N-glycan is recognised by the mannose-binding lectin Concanavalin A, which caps glycoproteins on the *Entamoeba* surface (Silva *et al.*, 1975). Unprocessed $Man_5GlcNAc_2$ is also recognised by the anti-retroviral lectin cyanovirin, which binds $Man_9GlcNAc_2$ on the surface of gp120 (Adams *et al.*, 2004). This result suggests the possibility that the anti-retroviral lectin may be active against numerous protists.

Second, complex N-glycans of *Entamoeba*, which are built upon the same $Man_3GlcNAc_2$ core as higher eukaryotes, contain just two additional sugars (galactose and glucose) (Fig. 2.9, D and H) (P. E. Magnelli, D. M. Ratner, P. W. Robbins and J. Samuelson, unpublished data). Galactose is added first to both arms of $Man_3GlcNAc_2$ and then glucose is added to galactose. To make these complex N-glycans, *Entamoeba* has NSTs for glucose (UDP-Glc) and galactose (UDP-Gal) (Bredeston *et al.*, 2005). Glucose is also transferred to N-glycans during the QC of protein folding in the ER, while both galactose and glucose are transferred to proteophosphoglycans (PPGs) (see next section) (Moody-Haupt *et al.*, 2000). Because the complex N-glycans of *Entamoeba* are unique, it is possible that they may be targets of anti-amoebic antibodies.

6.5.4. O-glycans and GPI anchors

The surface of *E. histolytica* trophozoites is rich in glycoconjugates as shown by the ability of many lectins and carbohydrate-specific antibodies to recognise the cell surface (Srivastava *et al.*, 1995; Zhang *et al.*, 2002). Proteophosphoglycans constitute the major glycoconjugate of the *E. histolytica* cell surface. PPG is anchored to the cell surface through a GPI moiety (Bhattacharya *et al.*, 1992). The structure of the PPG GPI has been tentatively determined (Moody-Haupt *et al.*, 2000). In most eukaryotes, PI is glycosidically linked to the reducing end of de-acetylated glucosamine followed by three mannoses that are in turn attached to the ethanolamine that links the protein to the GPI. However, the GPI anchor of *E. histolytica* PPG was found to have a unique backbone that is not observed in other eukaryotes, namely Gal-Man-Man-GlcN-*myo*-inositol. The intermediate and light subunits of the *E. histolytica* Gal/GalNAc lectin, among other cell surface molecules, are anchored to the cell surface through GPI anchors. Though the structure of the GPI anchors is not known, they are thought to be functionally important (Ramakrishnan *et al.*, 2000).

In humans, 23 genes are known to participate in the biosynthesis of GPI anchors. However, only 15 of these were identified in *E. histolytica* (Vats *et al.*, 2005). Interestingly, all the catalytic subunits were identified in *E. histolytica*, the missing genes encoding the accessory subunits suggesting that the biosynthetic pathway may not be significantly different from that in other eukaryotes. The presence of the pathway was also confirmed by detecting the biochemical activities of the first two enzymes—*N*-acetyl glucosamine transferase and deacetylase. In addition, antisense inhibition of the deacetylase blocked GPI anchor biosynthesis and reduced virulence of the parasite (Vats *et al.*, 2005). A novel GIPL (glycosylated inositol phospholipid) was also identified in *E. histolytica* (Vishwakarma *et al.*, 2006). Structural studies indicate that a galactose residue is attached to glucosamine as the terminal sugar instead of mannose. This suggests that *E. histolytica* is capable of synthesising unusual GPI-containing glycoconjugates not observed in other organisms.

In PPG, glycans are attached to a peptide backbone by an O-phosphodiester-linkage (O-P glycans). The *E. histolytica* O-P-glycans have galactose at the reducing end followed by a chain of glucoses. *E. invadens* also has O-P-glycans on its cyst wall proteins but the reducing sugar is a deoxysugar rather than galactose (Van Dellen *et al.*, 2006b). While *Dictyostelium* has also O-P-glycans on glycoproteins in its spore wall, glycoproteins with O-P-glycans are absent from the vast majority of animals and plants (West, 2003).

6.5.5. Significance

The unique glycans of *Entamoeba* lead to three important evolutionary inferences. First, much of the diversity of eukaryotic N-glycans is due to secondary loss of enzymes that make the 14-sugar lipid-linked precursor, which was present in the common ancestor to extant eukaryotes. Despite the truncated N-glycan precursor, *Entamoeba* has conserved the relatively complex N-glycan-dependent QC system for protein folding. Third, the unique N-glycans and O-P-linked glycans are based on a novel set of glycosyltransferases, which are present in *Entamoeba* and remain to be characterised biochemically.

7. PROTEINS INVOLVED IN SIGNALLING

7.1. Phosphatases

The combined actions of protein kinases and phosphatases regulate many cellular activities through reversible phosphorylation of proteins. These activities include such basic functions as growth, motility and metabolism. Although it was once assumed that kinases played the major regulatory

role, it is now clear that phosphatases can also be critical participants in some cellular events (Li and Dixon, 2000). There are few publications on the role of phosphatases in *E. histolytica*; however, several investigators have established a role for phosphatases in proliferation and growth. Chaudhuri *et al.* (1999) observed that there was an increase in phosphotyrosine levels in serum starved, growth inhibited, *E. histolytica* cultures. Upon the addition of serum and subsequent growth simulation, an increase in tyrosine phosphatase activity occurred. These investigators also demonstrated that genistein, a tyrosine kinase inhibitor, had no effect on growth, while the addition of sodium orthovanadate, a phosphatase inhibitor, produced a major decrease in cell proliferation. Membrane-bound and secreted acid phosphatase activities have been detected in *E. histolytica* (Aguirre-Garcia *et al.*, 1997; Anaya-Ruiz *et al.*, 1997). The secreted acid phosphatase activity is absent from *E. dispar* (Talamas-Rohana *et al.*, 1999). This secreted acid phosphatase was found to have phosphotyrosine hydrolase activity, and caused cell rounding and detachment of HeLa cells (Anaya-Ruiz *et al.*, 2003), suggesting that phosphatase activity contributes to the virulence of the organism.

There are four families of phosphatases (Stark, 1996). Members of the PPP (protein phosphatase P) family are serine/threonine phosphatases and include PP1, PP2A and PP2B (calcineurin-like) classes. The PPM (protein phosphatase M) family phosphatases also dephosphorylate serine/threonine residues but are unrelated to the PPP family proteins. A third family consists of protein tyrosine phosphatases (PTP) and dual phosphatases. Low molecular weight phosphatases make up the fourth family. In eukaryotic cells, greater than 99% of protein phosphorylation is on serine or threonine residues (Chinkers, 2001). Human cells have about 500 serine/threonine phosphatases and 100 tyrosine phosphatases (Hooft van Huijsduijnen, 1998; Hunter, 1995). *S. cerevisiae* has 31 identified or putative protein phosphatases (Stark, 1996). *E. histolytica* has over 100 putative protein phosphatases. Only a few of these phosphatases have potential transmembrane domains. Some *E. histolytica* phosphatases have varying numbers of LRRs. The LRR domain is thought to be a site for protein–protein interactions (Hsiung *et al.*, 2001; Kobe and Deisenhofer, 1994). LRR domains have been found in a few kinases, but had not been identified in any phosphatases until recently (Gao *et al.*, 2005).

7.1.1. Serine/threonine protein phosphatases

Members of the PPP family of protein phosphatases are closely related metalloenzymes, and complex with regulatory subunits. In contrast, PPM family members are generally monomeric, ranging 42–61 kDa in size. By BLAST analysis, the serine/threonine protein phosphatases of *E. histolytica* are most closely related to PPP phosphatases PP2A, PP2B and PPM phosphatase PP2C.

7.1.1.1. PP2A and PP2B (Calcineurin-like) serine/threonine phosphatases PP2A phosphatases are trimeric enzymes consisting of catalytic, regulatory and variable subunits (Wera and Hemmings, 1995). Calcineurin is a calcium-dependent protein serine/threonine phosphatase (Rusnak and Mertz, 2000). Orthologues of calcineurin are widespread from yeast to mammalian cells. Calcineurin is a heterodimeric complex with catalytic (CaNA) and regulatory (CaNB) subunits. CaNA ranges in size from 58 to 64 kDa. Its conserved domain structure includes a catalytic domain, a CaNB-binding domain, a calmodulin binding domain and an auto-inhibitory (AI) domain. The binding of CaNB and calmodulin activates CaNA. CaNB subunit is 19 kDa, contains 4 EF hand calcium binding motifs, and has similarity to calmodulin. The binding of calmodulin releases the auto-inhibitory domain and results in activation of the phosphatase. Deletion of the AI domain results in a constitutively active protein. Calcineurin is specifically inhibited by cyclosporin A and FK506. Cyclosporin A and FK506 first bind to specific proteins, cyclophilin A and FK506BP, respectively, then bind to CaNA at the CaNB binding site. Cyclophilin A has been identified in *E. histolytica* and treatment with cyclosporin A decreases growth and viability (Carrero *et al.*, 2000, 2004; Ostoa-Saloma *et al.*, 2000).

The *E. histolytica* genome has 51 PP2A and calcineurin-like protein phosphatases. The Pfam motif that classifies proteins as PPP phosphatases is Metallophos (PF00149, calcineurin-like phosphoesterase). This motif is also found in a large number of proteins involved in phosphorylation, including DNA polymerase, exonucleases and other phosphatases. The genome annotation identifies three loci as CaNA orthologues. However, due to the similarity among this family of phosphatases, it is difficult to tell by sequence analyses alone those that are calcium-dependent. Identification of CaNA will have to be confirmed experimentally.

Two of the PPM phosphatases contain a tetratricopeptide repeat (TPR) domain (PF00515). TPR is thought to be involved in protein–protein interactions (Das *et al.*, 1998). Activities that have been ascribed to TPR include regulatory roles, lipid binding and auto-inhibition.

7.1.1.2. PP2C phosphatases PP2C phosphatases are also widespread and are often involved in terminating/attenuating phosphorylation during the cell cycle or in response to environmental stresses such as osmotic and heat shock (Kennelly, 2001). Thirty-five genes were identified as PP2C phosphatases. These proteins can be divided into three broad categories: (1) PP2C domain only small (235–381 amino acids), (2) PP2C domain only large (608–959 amino acids) and (3) PP2C with LRR domains.

7.1.2. Tyrosine phosphatases (PTP)

Tyrosine phosphorylation-dephosphorylation is a key regulatory mechanism for many aspects of cell biology and development (Li and Dixon, 2000). PTPs are a large class of enzymes that have catalytic domains of ~300 amino acids. Forty of these residues are highly conserved (Hooft van Huijsduijnen, 1998). PTPs can be divided into membrane (receptor) and non-membrane (soluble) PTPs (Li and Dixon, 2000). The soluble PTP group includes those that contain conserved SH2, PEST, Ezrin, PDZ or CH2 domains. Two other classes of PTPs are the low molecular weight and dual phosphatases. *S. cerevisiae* lacks classic PTPs but does contain dual phosphatases such as the MAP kinase kinases.

E. histolytica has only four potential PTPs, none of which are receptor PTPs (i.e., PTPs with recognisable transmembrane spanning regions). Two of the PTPs (XM_650778, XM_645883) are 350 and 342 amino acids in length and share 48% identity. Neither of these phosphatases has any other recognisable conserved domain. Non-receptor type 1 PTPs are the closest match to these proteins (Li and Dixon, 2000). Membrane and secreted forms of a PTP that cross-react with anti-human PTP1B have been reported in *E. histolytica* (Aguirre-García *et al.*, 2003; Talamas-Rohana *et al.*, 1999). Both forms have an apparent molecular weight of 55 kDa and disrupt host actin stress fibres. However, since none of the putative PTPs identified by the genome project appear to encode secreted or membrane forms, it is unlikely that these loci represent these previously reported PTP1B cross-reacting proteins.

A third PTP contains a protein tyrosine phosphatase like protein (PTPLA) domain (PF04387). The PTPLA domain is related to the catalytic domains of tyrosine kinases, but it has an arginine for proline substitution at the active site (Uwanogho *et al.*, 1999). It is not yet clear whether this family of proteins actually has phosphatase activity or serves some other regulatory role.

An orthologue of a low molecular weight PTP has also been identified. Low molecular weight protein tyrosine phosphatases have been found in bacteria, yeasts and mammalian cells (Ramponi and Stefani, 1997). They are not similar to other PTPs except in the conserved catalytic domain.

7.1.3. Dual-specificity protein phosphatases

Dual-specificity PTPs (DSP) can hydrolyse both tyrosine and serine/threonine residues, though they hydrolyse phosphorylated tyrosine substrates 40–500-fold faster (Zhang and VanEtten, 1991). In other organisms, DSPs are mostly found in the nucleus and have roles in cell cycle control, nuclear dephosphorylation and inactivation of MAP kinase.

The *E. histolytica* genome has 23 sequences related to DSPs. They fall into three main subclasses: those with the DSP domain only, those with

DSP plus a variable number (one to five) of LRRs and those with the Rhodanese homology domain (RHOD; IPR001763). Rhodanese is a sulphurtransferase involved in cyanide detoxification. Its active site, RHOD, is also found in the catalytic site of the dual-specificity phosphatase CDC25 (Bordo and Bork, 2002).

7.1.4. Leucine rich repeats

LRRs are tandem arrays of 20–29 amino acid, leucine-rich motifs. LRRs have been found in a number of proteins with varied functions including enzyme inhibition, regulation of gene expression, morphology and cytoskeleton formation (Kobe and Deisenhofer, 1994). LRRs are thought to provide versatile sites for protein–protein interactions and have been found linked to a variety of secondary domains. Most LRRs form curved horseshoe-shaped structures with "a parallel β sheet on the concave side and mostly helical elements on the convex side" (IPR001611).

The LRR_1 Pfam is the second most abundant Pfam domain found in the *E. histolytica* genome (Table 2.3). The LRR motifs in *E. histolytica* most closely resemble the LRR found in BspA (Section 2.7; Davis *et al.*, 2006). Several *E. histolytica* proteins that contain LRRs are associated with other recognised domains. These include the protein phosphatases PP2C and DSP, as well as protein kinase (PK), F-box (PF00646), gelsolin/villin headpiece (IPR007122), DNA J (IPR001623), Band 41 (B41;IPR000299), WD-40 (IPR001680) and zinc binding (IPR000967) domains. The association of LRRs with phosphatases is unusual. One published example is the phosphatase that dephosphorylates the kinase Akt (Gao *et al.*, 2005). Fungal adenylate cyclases have both LRR and PP2C-like domains, but this is not a widespread feature of adenylate cyclases in other species (Mallet *et al.*, 2000; Yamawaki-Kataoka *et al.*, 1989). The LRR may be a site for interaction with phosphorylated residues in *E. histolytica*. This speculation is supported by the example of the Grr1 protein of yeast, which contains an F-box and an LRR (Hsiung *et al.*, 2001). Grr1 is involved in ubiquitin-dependent proteolysis. The LRR domain of Grr1 binds to phosphorylated targets in the proteasome complex. Another example is the fission yeast phosphatase regulatory subunit, Sds22, which also has LRRs (MacKelvie *et al.*, 1995). The LRR containing phosphatases of *E. histolytica* may represent fusions of regulatory and catalytic subunits.

7.2. Kinases

7.2.1. Cytosolic kinases

Eukaryotic protein kinases are a superfamily of enzymes which are important for signal transduction and cell-cycle regulation. Six families of serine/threonine kinases (STKs), which include AGC, Ste, CK1, CaMK, CMGC and TKL (tyrosine kinase-like), have conserved aspartic acid

and lysine amino acids in their active sites and phosphorylate serine or threonine on target proteins (Hanks and Hunter, 1995). Tyrosine kinases (TKs), which lack active site lysine, phosphorylate tyrosine on target proteins. Phosphorylated tyrosine is in turn recognised by Src-homology 2 (SH2) domains that are present on some kinases and other proteins. All seven families of protein kinases are present in metazoa and in *D. discoideum*, while plants lack TK, and *S. cerevisiae* lacks both TK and TKL.

Over 150 predicted *E. histolytica* cytosolic kinases, those that lack signal peptides and *trans*-membrane helices, can be identified, including representatives of each of the 7 groups of kinases (AGC, CAMK, CK1, CMGC, STE, TKL and TK) (Loftus *et al.*, 2005). Two predicted *E. histolytica* TKs, which group with human TKs in phylogenetic trees, contain an AAR peptide in the active site and a Kelch domain at the C-terminus (Gu and Gu, 2003). Four cytosolic protein kinases contain C-terminal SH2 domains, which bind phosphorylated tyrosine residues. Phosphotyrosine has been identified in *E. histolytica* using specific antibodies (Hernandez-Ramirez *et al.*, 2000). The 35 predicted cytosolic *E. histolytica* TKLs include some that contain LRRs and ankyrin repeats at their N-termini. In contrast, the vast majority of *Entamoeba* cytosolic kinases lack accessory domains.

7.2.2. Receptor kinases

Five distinct families of eukaryotic proteins have an N-terminal ectoplasmic domain, a single TMH and a C-terminal cytoplasmic kinase domain (Blume-Jensen and Hunter, 2001). Ire-1 transmembrane kinases, which are present in *S. cerevisiae*, plants and metazoa, detect unfolded proteins in the lumen of the ER and help splice a transcription factor mRNA by means of a unique C-terminal ribonuclease (Patil and Walter, 2001). Receptor tyrosine kinases (RTKs), which include growth hormone and epidermal growth factor (EGF) receptors, are restricted to metazoa and have a diverse set of N-terminal ectoplasmic domains and a conserved C-terminal cytosolic TK (Schlessinger, 2000). Receptor serine/threonine kinases (RSK) of metazoa and receptor-like kinases (RLKs) of plants each contain a C-terminal TKL domain (Massague *et al.*, 2000; McCarty and Chory, 2000; Shiu and Bleecker, 2001). Phylogenetic analyses suggest that plant RLKs, animal RSKs and animal RTKs each form monophyletic groups and that plant RLKs closely resemble cytosolic TKLs of animals called Pelle or IRAK (Shiu and Bleecker, 2001).

E. histolytica contains >80 novel receptor RSKs, each of which has a N-terminal signal sequence, a conserved ectoplasmic domain, a single TMH and a cytosolic kinase domain (Beck *et al.*, 2005). The largest group of *E. histolytica* RSKs has a CXXC-rich ectoplasmic domain with 6–31 internal repeats that each contains 4–6 cysteine residues (Fig. 2.10). Very similar CXXC-rich domains are present in the ectoplasmic domain

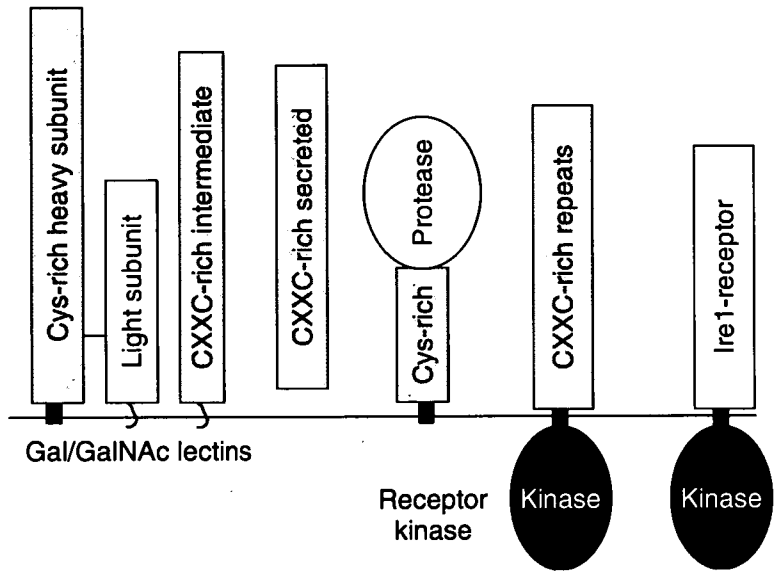

FIGURE 2.10 Structure of cysteine-rich plasma membrane proteins of *E. histolytica*. These proteins include the various subunits of the Gal/GalNAc lectin, a cysteine protease and numerous receptor kinases. Ire1, which is involved in the unfolded protein response, is also a receptor kinase but has no Cys-rich domain.

intermediate subunit of the Gal/GalNAc lectin (see Section 3.1.3). CXXC-rich domains are also present in hypothetical secreted proteins of *E. histolytica*, while cysteine-rich domains are also present in the heavy subunit of the Gal/GalNAc lectin and at the cytosolic aspect of some cysteine proteases (Fig. 2.10).

Ectoplasmic domains of other large families of *Entamoeba* RSKs have one or two 6-Cys domains at the N-terminus and four 6-Cys domains proximal to the plasma membrane. There are no plasma membrane proteins or secreted proteins with similar domains. A minority of RSKs do not contain Cys-rich ectoplasmic domains. Numerous *Entamoeba* RSKs are expressed at the same time, but the specific ligands for the *Entamoeba* RSKs have not been identified (Beck *et al.*, 2005).

As discussed in the section on protein folding (Section 6.5.2), *Entamoeba* has an Ire1 transmembrane kinase, which recognises misfolded proteins in the lumen of the ER and triggers the unfolded protein response (Fig. 2.8).

7.2.3. Significance

While most protists lack TK, TKL, receptor-kinases and Ire1 *E. histolytica* has all four. It is very likely that the *E. histolytica* receptor kinases, which are extensively duplicated, will have important roles in pathogenesis (Beck *et al.*, 2005; Okada *et al.*, 2005). Similarly, trimeric G-proteins and

the associated adenyl-cyclases likely have important roles in cyst formation and virulence (Coppi *et al.*, 2002; Frederick and Eichinger, 2004).

7.3. Calcium binding proteins

Ca^{2+}signalling plays a crucial role in the pathogenesis of many protozoan parasites, including *E. histolytica* (Ravdin *et al.*, 1985). Many of the calcium-mediated processes are carried out with the help of calcium binding proteins (CaBPs). CaBPs have been identified and characterised in almost all eukaryotic systems. Some of these, such as calmodulin (CaM) and troponin C, have been studied extensively. A number of CaBPs have also been identified in *E. histolytica*. Among these are two related EF-hand containing proteins, grainin 1 and granin 2, which are likely to be localised in intracellular granules (Nickel *et al.*, 2000). Another protein, URE3-BP, was shown to have a transcription regulatory function (Gilchrist *et al.*, 2001). The CaM-dependent secretion of collagenases from electron dense granules has been demonstrated using *E. histolytica* lysate. However, there is as yet no direct molecular evidence for the presence of CaM in *E. histolytica* (de Muñoz *et al.*, 1991). The CaM-like protein EhCaBP1 has four canonical EF-hand Ca^{2+} binding domains but no functional similarity to CaM (Yadava *et al.*, 1997). Inducible expression of EhCaBP1 antisense RNA demonstrated this protein's role in actin-mediated processes (Sahoo *et al.*, 2004).

Analysis of the whole genome revealed presence of 27 CaBPs with multiple EF-hand calcium binding domains (Bhattacharya *et al.*, 2006). Many of these proteins are architecturally very similar but functionally distinct from CaM. Moreover, functional diversity was also observed among closely related CaBPs such as EhCaBP1 and EhCaBP2 (79% identical at the amino acid level; Chakrabarty *et al.*, 2004). Analysis of partial EST and proteomic databases combined with Northern blots and RT-PCR shows that at least one-third of these genes are expressed in trophozoites, suggesting that many if not all of the 27 are functional genes (Bhattacharya *et al.*, 2006).

What are the roles of these proteins in the context of *E. histolytica* biology? At present the function of only two EhCaBPs are known, EhCaBP1 and URE3-BP. The rest of the proteins are likely to be Ca^{2+} sensors involved in a number of different signal transduction pathways. After binding Ca^{2+} these may undergo conformational changes and the bound form then activates downstream target proteins. It is not clear why *E. histolytica* would need so many Ca^{2+} sensors when many other organisms do not. It is likely that with Ca^{2+} being involved in many functions, some of which are localised in different cellular locations, the

various CaBPs may participate in different functions that are spatially and temporally separated.

8. THE MITOSOME

One of the expectations for the *E. histolytica* genome project was that it would identify the function of the mitochondrial remnant known as the mitosome (Tovar *et al.*, 1999) or crypton (Mai *et al.*, 1999). Under the microscope mitosomes are ovoid structures smaller than 0.5 μm in diameter (Leon-Avila and Tovar, 2004). While it is now clear that no mitochondrial genome still persists, from both genome sequencing and cellular localisation data (Leon-Avila and Tovar, 2004), the protein complement of the organelle is still somewhat obscure. The number of identifiable mitosomal proteins remains very small and does not provide great insight into the organelle's function. Genes encoding mitochondrial-type chaperonins (cpn60, hsp10 and mt-hsp70) have been identified and appear to be synthesised with amino-terminal signal sequences. The importation machinery has been shown to be conserved with that in true mitochondria (Mai *et al.*, 1999; Tovar *et al.*, 1999), but none of the proteins involved in mitosomal protein import have been identified with certainty.

Other genes encoding putative mitosomal proteins include pyridine nucleotide transhydrogenase (which moves reducing equivalents between NAD and NADP, and acts as a proton pump (Clark and Roger, 1995); only an incomplete gene is present in the assembly), an ADP/ATP transporter (Chan *et al.*, 2005), a P-glycoprotein-like protein (Pgp6), and a mitochondrial-type thioredoxin, although the latter two are identified based largely on their amino terminal extensions. The only enzymatic pathway that is normally mitochondrial in location is iron–sulphur cluster synthesis. Genes encoding homologues of both IscS/NifS and IscU/NifU proteins are present, but uniquely among eukaryotes the *E. histolytica* homologues are not of mitochondrial origin, having been acquired by distinct LGT from an ε-proteobacterium (Ali *et al.*, 2004b; van der Giezen *et al.*, 2004). The location of these proteins appears to be cytoplasmic as determined by immunofluorescence, using antibodies against both the native proteins as well as detection of epitope-tagged proteins in transformed *E. histolytica* (V. Ali and T. Nozaki, unpublished data). The same pathway has been localised to mitosomes in *Giardia* and is also retained in all other organisms with remnant mitochondria. Given the apparently unique non-compartmentalised nature of iron–sulphur cluster synthesis in *E. histolytica* the location of the proteins needs to be confirmed by immuno-electron-microscopy; such experiments are currently under way (V. Ali and T. Nozaki, unpublished data). The function of the *E. histolytica* mitosome therefore remains an enigma.

9. ENCYSTATION

The infectious stage of *E. histolytica*, and also that most often used for diagnosis, is the quadrinucleate cyst. Because it is not possible to encyst *E. histolytica* in axenic culture, *E. invadens*, which is a reptilian parasite, has been used as a model organism for encystation (Eichinger, 2001; Wang *et al.*, 2003). The *E. invadens* cyst wall is composed of three parts: deacetylated chitin (also known as chitosan), lectins that bind chitin (e.g., Jacob and Jessie) or cyst wall glycoproteins (e.g., plasma membrane Gal/GalNAc lectin), and enzymes that modify chitin or cyst wall proteins (e.g., chitin deacetylase, chitinase and cysteine proteases) (Fig. 2.11).

9.1. Chitin synthases

Chitin fibrils, which are homopolymers of β-1,4-linked *N*-acetyl glucosamine (GlcNAc), are synthesised by chitin synthases. Chitin synthases share common ancestry with cellulose synthases and hyaluronan synthase. They are transmembrane proteins with a catalytic domain in the cytosol (Bulawa, 1993), where UDP-GlcNAc is made into a homopolymer and is threaded through the transmembrane domains into the extracellular space. In *S. cerevisiae*, four accessory peptides, encoded by the Chs4–7 genes, are necessary for the function of its chitin synthases (Trilla *et al.*, 1999). Remarkably, the *E. histolytica* chitin synthase 2 (EhChs2) complements a *S. cerevisiae* chs1/chs3 mutant and the function of EhChs2 is independent of the four accessory peptides (Van Dellen *et al.*, 2006a). This result suggests the possibility that chimaeras of *E. histolytica* and *S. cerevisiae* chitin synthases may be used to map domains in the *S. cerevisiae* chitin synthase that interact with the accessory peptides.

9.2. Chitin deacetylases

Chitin fibrils in the cyst wall are modified by deacetylases and chitinases (see Section 9.3). There are two *E. invadens* chitin deacetylases, which convert chitin into chitosan (Das *et al.*, 2006). Chitosan is a mixture of *N*-acetyl glucosamine and glucosamine and so has a positive charge. It is also present in spore walls of *S. cerevisiae* and in lateral walls of *Mucor* (Kafetzopoulos *et al.*, 1993; Mishra *et al.*, 1997). It is likely that the positive charge of chitosan fibrils contributes to the binding of cyst wall proteins, all of which are acidic (de la Vega *et al.*, 1997; Frisardi *et al.*, 2000; Van Dellen *et al.*, 2002b). Monosaccharide analyses of the *E. invadens* cyst walls following treatment with SDS to remove proteins strongly suggest that chitosan is the only sugar homopolymer present (Das *et al.*, 2006).

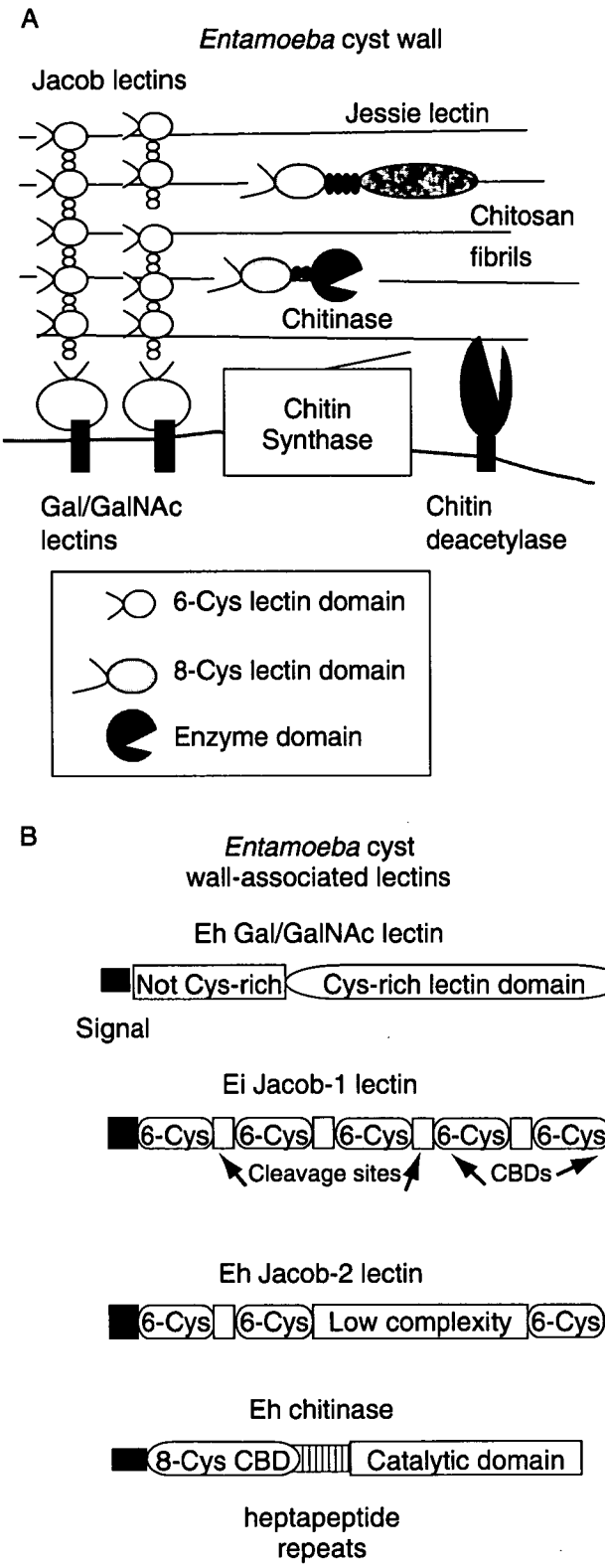

FIGURE 2.11 Model for the *Entamoeba* cyst wall derived primarily from experiments with *E. invadens*. (A) The cyst wall consists of chitosan fibrils, which are made by chitin synthase and chitin deacetylase. Wall proteins include Jacob lectins with tandem arrays of 6-Cys chitin-binding domains (CBDs), as well as chitinase and Jessie lectins that have a single 8-Cys CBD. The Gal/GalNAc lectin in the plasma membrane binds sugars on the Jacob and Jessie lectins. (B) Structures of representative lectins are illustrated in (A).

9.3. Chitinases

Entamoeba species encode numerous chitinases with a conserved type 18 glycohydrolase domain (de la Vega *et al.*, 1997). Recombinant *Entamoeba* chitinases have both endo- and exo-chitinase activities. Two other domains are important in *Entamoeba* chitinases: (1) At the N-terminus is a unique 8-Cys chitin-binding domain (CBD), which is also present as a single domain in *E. histolytica* Jessie lectins (Fig. 2.11) (Van Dellen *et al.*, 2002b). Chitinase and Jessie-3 lectin bind to the *E. invadens* cyst wall by means of this 8-Cys CBD (Van Dellen *et al.*, 2006b). This *E. histolytica* chitinase CBD has the same function as CBDs in chitinases of fungi, nematodes, insects and bacteria, but has no sequence similarity (i.e., it has arisen by convergent evolution) (Shen and Jacobs-Lorena, 1999). (2) Between the CBD and chitinase domains of *Entamoeba* species are low complexity sequences that contain heptapeptide repeats (Ghosh *et al.*, 2000). These polymorphic repeats may be used to distinguish isolates of *E. histolytica* within the same population and may be able to discriminate among isolates from New to Old World (Haghighi *et al.*, 2003). These polymorphic repeats, which are rich in serine and resemble mucin-like domains in other glycoproteins, may also be the sites for addition of O-phosphodiester linked sugars (see Section 6.5.4).

9.4. Jacob lectins

Chitin fibrils in the cyst wall of *E. invadens* are cross-linked by Jacob lectins, which contain three to five unique 6-Cys CBDs (Frisardi *et al.*, 2000). *E. invadens* has at least nine genes encoding Jacob lectins, and the mRNA levels from each gene increase during encystation (Van Dellen *et al.*, 2006b). In addition, at least six Jacob lectin proteins are present in *E. invadens* cyst walls (Van Dellen *et al.*, 2006b). Between the CBDs, Jacob lectins have low complexity sequences that are rich in serine as in the case of chitinase (Van Dellen *et al.*, 2006a). Jacob lectins are post-translationally modified in two ways. First, they are cleaved by cysteine proteases at conserved sites in the serine- and threonine-rich spacers between CBDs. Second, they have O-phosphodiester-linked sugars added to serine and threonine residues. O-phosphodiester-linked glycans are also present in PPGs on the surface of *E. histolytica* trophozoites (Moody-Haupt *et al.*, 2000).

9.5. Gal/GalNAc lectins

The Gal/GalNAc lectins present on the surface of *E. histolytica* trophozoites have been described earlier (see Section 3.1) and in the literature (Mann *et al.*, 1991; Petri *et al.*, 2002). Their possible role in encystation is suggested by two independent experiments. First, the signal for

encystation likely depends in part on aggregation of *E. invadens*, which is inhibited by exogenous galactose (Coppi and Eichinger, 1999). Aggregated *E. invadens* secrete catecholamines, which in an autocrine manner stimulate amoebae to encyst (Coppi *et al.*, 2002). Second, in the presence of excess galactose, *E. invadens* forms wall-less cysts that contain four nuclei and makes Jacob lectins and chitinase (Frisardi *et al.*, 2000). Because *E. invadens* trophozoites have a Gal/GalNAc lectin on their surface that is capable of binding sugars on Jacob lectin, and because Jacob lectins have no carboxy-terminal TMH or GPI-anchor, it is likely that the cyst wall is bound to the plasma membrane by the Gal/GalNAc lectin.

9.6. Summary and comparisons

Similar to the cyst wall of *Giardia*, the cyst wall of *E. invadens* is a single homogeneous layer and contains a single homopolymer, chitosan (Fig. 2.11) (Frisardi *et al.*, 2000; Gerwig *et al.*, 2002; Shen and Jacobs-Lorena, 1999). In contrast, *S. cerevisiae* spore walls have multiple layers and contain β-1,3-glucans in addition to chitin, while *Dictyostelium* walls have multiple layers and contain *N*-acetyl galactosamine polymers in addition to cellulose (West, 2003; Yin *et al.*, 2005).

Similar to *Dictyostelium* and in contrast to fungi, the vast majority of *Entamoeba* cyst wall glycoproteins are released by SDS (Frisardi *et al.*, 2000; Van Dellen *et al.*, 2006b; West, 2003; Yin *et al.*, 2005). While some *Dictyostelium* cyst wall proteins have been shown to be cellulose-binding lectins, all of the proteins bound to the cyst wall of *E. invadens* have 6-Cys CBDs (Jacob lectins) or 8-Cys CBDs (Jessie 3 lectin and chitinase) (Frisardi *et al.*, 2000; Van Dellen *et al.*, 2002b). In the same way that *Giardia* cyst wall protein 2 is cleaved by a cysteine protease, Jacob lectins are cleaved by an endogenous cysteine protease at sites between chitin-binding domains (Touz *et al.*, 2002).

Like *Dictyostelium* spore coat proteins and insect peritrophins, cysteine-rich lectin domains of *E. invadens* cyst wall proteins are separated by serine- and threonine-rich domains that are heavily glycosylated (Frisardi *et al.*, 2000; West, 2003; Yin *et al.*, 2005). *S. cerevisiae* cyst wall proteins have also extensive serine- and threonine-rich domains that are heavily glycosylated (Yin *et al.*, 2005). These glycans likely protect proteins in cyst walls or fungal walls from exogenous proteases. While glycoproteins of the *E. invadens* cyst wall and *Dictyostelium* spore coat contain O-phosphodiester-linked glycans, *S. cerevisiae* wall glycoproteins contain O-glycans (Gemmill and Trimble, 1999; West *et al.*, 2005).

Like *S. cerevisiae*, *E. invadens* has enzymes in its wall that modify chitin (Yin *et al.*, 2005). Similar to chitinases of *S. cerevisiae* and bacteria, *E. invadens* chitinase has a CBD in addition to the catalytic domain (Kuranda and Robbins, 1991). It is likely that the CBD is present to localise

chitinase to the cyst wall (*E. invadens*) or cell wall (*S. cerevisiae*). Finally, while *E. invadens* uses catecholamines as autocrines for encystation, *Dictyostelium* uses cAMP as an autocrine for sporulation (Coppi *et al.*, 2002; Kriebel and Parent, 2004). An important goal of future research will be to translate what is known about the *E. invadens* cyst wall to that of *E. histolytica*.

10. EVIDENCE OF LATERAL GENE TRANSFER IN THE *E. HISTOLYTICA* GENOME

Lateral (or horizontal) gene transfer (LGT) plays a significant role in prokaryotic genome evolution, contributing up to ~20% of the content of a given genome (Doolittle *et al.*, 2003). LGT has therefore been an important means of acquiring new phenotypes, such as resistance to antibiotics and new physiological and metabolic capabilities, that may permit or facilitate adaptation to new ecological niches (Koonin *et al.*, 2001; Lawrence, 2005a; Ochman *et al.*, 2000). More recently, data from microbial eukaryote genomes suggest that LGT has also played a role in eukaryotic genome evolution, particularly among protists that eat bacteria (Andersson, 2005; Doolittle, 1998; Doolittle *et al.*, 2003; Lawrence, 2005b; Richards *et al.*, 2003). *E. histolytica* lives in the human gut, an environment that is rich in micro-organisms and where LGT is thought to be common between bacteria (Shoemaker *et al.*, 2001). The *E. histolytica* genome thus provides a nice model for investigating prokaryote to eukaryote LGT. In the original genome description (Loftus *et al.*, 2005), 96 putative cases of LGT were identified using phylogenetic analyses of the *E. histolytica* proteome. These have now been reanalysed in the light of more recently published (August, 2005) eukaryotic and prokaryotic genomes. This has allowed evaluation of how previous inferences were influenced by the sparse sampling of eukaryotic and prokaryotic genes and species available at the time of the original analysis. Sparse gene and species sampling is, and is likely to remain, a very serious problem for reconstructing global trees and inferring LGT (Andersson *et al.*, 2001; Richards *et al.*, 2003; Salzberg *et al.*, 2001). Thus, although ecologists differ in their claims for the extent of the unsampled microbial world, they all agree that those species in culture, and the even smaller subset for which genome data exist, represent the smallest tip of a very large iceberg.

10.1. How do the 96 LGT cases stand up?

As before (Loftus *et al.*, 2005), Bayesian and maximum likelihood distance bootstrap phylogenetic analyses were used to identify putative LGT using the following ad hoc conservative criteria: Putative LGT was inferred

where either no other eukaryote possessed the gene or where the *E. histolytica* sequence was grouped with bacteria and separated from other eukaryotes by at least two strongly supported nodes (bootstrap support >70%, posterior probabilities >0.95). In cases where tree topologies were more weakly supported but still suggested a possible LGT, bootstrap partition tables were examined for partitions where the *E. histolytica* sequence clustered with another eukaryote. If no such partitions were found that gene was considered to be a putative LGT. Table 2.8 lists the results of the new analyses and also gives BlastP statistics for each sequence.

A total of 41 LGTs remain as strongly supported as before based on the original criteria. For the remaining 55 tree topologies, support for recent LGT into the *Entamoeba* lineage is not as strong as before. For 27 of these 55 trees, 2 strongly supported nodes separating *E. histolytica* from other eukaryotes have been reduced to only 1 well-supported node. However, close scrutiny of the bootstrap partition tables for these trees revealed that, as before, there are no trees in which *E. histolytica* is found together with another eukaryote. Thus, LGT still remains the strongest hypothesis to explain 68 (70%) of the original 96 tree topologies. In a further 14 cases, the position of *E. histolytica* among prokaryotes and eukaryotes was not well supported. The taxonomic sampling of eukaryotes in these trees is very patchy and the trees do not depict consensus eukaryotic relationships. Thus, although the trees do not fulfil the conservative criteria for LGT, they also do not provide strong support for the alternative hypothesis that the *E. histolytica* genes were vertically inherited from a common ancestor shared with all other eukaryotes.

In nine trees *E. histolytica* either clustered with a single newly published eukaryotic sequence, or such a relationship could not be ruled out. In six of these nine trees *E. histolytica* and *T. vaginalis* grouped together, and two trees grouped *E. histolytica* with the diatom *Thalassiosira* (e.g., see Fig. 2.12). Such trees are also not easy to explain within the current consensus for eukaryotic relationships (Baldauf, 2003). Similar topologies have previously been reported for other eukaryotes (Andersson, 2005). The explanations advanced to explain the absence of the gene in other eukaryotes include massive gene loss from multiple eukaryotic lineages, or LGT between the eukaryotic lineages concerned. *Entamoeba* species can ingest both eukaryotes and prokaryotes, and it has been suggested that LGT between eukaryotes, subsequent to one lineage acquiring the gene from a prokaryote, could explain such peculiar tree topologies and sparse distribution (Andersson, 2005). The fact that six of the nine cases recover a relationship between *Entamoeba* and *Trichomonas*, whose relatives often share the same niche, is consistent with this idea. In prokaryotes, recent large-scale analyses support the hypothesis that species from the same