Available online at www.sciencedirect.com

SCIENCE DIRECT

Biochimica et Biophysica Acta 1673 (2004) 29–44

BIOCHIMICA ET BIOPHYSICA ACTA
BBA

www.bba-direct.com

Review

Cytoplasmic glycosylation of protein-hydroxyproline and its relationship to other glycosylation pathways

Christopher M. West[a,*], Hanke van der Wel[a], Slim Sassi[b], Eric A. Gaucher[c]

[a] *Department of Biochemistry and Molecular Biology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA*
[b] *Department of Anatomy and Cell Biology, University of Florida, Gainesville, FL 32610, USA*
[c] *Foundation for Applied Molecular Evolution, Gainesville, FL, USA*

Received 19 November 2003; received in revised form 14 April 2004; accepted 14 April 2004
Available online 8 May 2004

Abstract

The Skp1 protein, best known as a subunit of $E3^{SCF}$-ubiquitin ligases, is subject to complex glycosylation in the cytoplasm of the cellular slime mold *Dictyostelium*. Pro143 of this protein is sequentially modified by a prolyl hydroxylase and five soluble glycosyltransferases (GT), to yield the structure Galα1,Galα1,3Fucα1,2Galβ1,3GlcNAcα1-HyPro143. These enzymes are unusual in that they are expressed in the cytoplasmic compartment of the cell, rather than the secretory pathway where complex glycosylation of proteins usually occurs. The first enzyme in the pathway appears to be related to the soluble animal prolyl 4-hydroxylases (P4H), which modify the transcriptional factor subunit HIF-1α in the cytoplasm, and more distantly to the P4Hs that modify collagen and other proteins in the rER, based on biochemical and informatics analyses. The soluble αGlcNAc-transferase acting on Skp1 has been cloned and is distantly related to the mucin-type polypeptide *N*-acetyl-α-galactosaminyltransferase in the Golgi of animals. Its characterization has led to the discovery of a family of related polypeptide *N*-acetyl-α-glucosaminyltransferases in the Golgi of selected lower eukaryotes. The Skp1 GlcNAc is extended by a bifunctional diglycosyltransferase that sequentially and apparently processively adds β1,3Gal and α1,2Fuc. Though this structure is also formed in the animal secretory pathway, the GTs involved are dissimilar. Conceptual translation of available genomes suggests the existence of this kind of complex cytoplasmic glycosylation in other eukaryotic microorganisms, including diatoms, oomycetes, and possibly *Chlamydomonas* and *Toxoplasma*, and an evolutionary precursor of this pathway may also occur in prokaryotes.

Keywords: Glycosyltransferase evolution; Multifunctional protein; Mucin-type *O*-glycosylation; *Thalassiosira*; *Phytophthora sojae*; *Yersinia*

1. Introduction

O-glycosylation of the GalNAcα-*O*-Ser/Thr type is a prevalent but heterogeneous modification of proteins that enter the rER and pass through the secretory pathway of animals and some microbes [1,2]. Consisting of one to many sugars, these *O*-glycans are referred to as mucin-type and contribute many general and protein- and cell-specific functions [3,4]. In the cytoplasm and nucleus, many proteins are also glycosylated by a single sugar, GlcNAc, in *O*-β-linkage to Thr- and Ser-residues [5]. *O*-βGlcNAc can influence protein phosphorylation and protein–protein interactions. At present, complex glycosylation of cytoplasmic or nuclear proteins is less well known (Ref. [6], other articles in this issue). This may be because complex cytoplasmic glycans are targeted to only one or a few proteins, thereby reducing the overall abundance of the modification. Such glycans may be easily overlooked unless high-sensitivity metabolic radiotracer methods are employed [7], or high-sensitivity/high-resolution mass spectrometry methods are applied to the purified protein [8]. One such complex glycan defined by these approaches is present on a small protein, Skp1, which is expressed universally in the cytoplasmic and nuclear com-

Abbreviations: pp αGlcNAcT, polypeptide *N*-acetyl-α-D-glucosaminyltransferase; pp αGalNAcT, polypeptide *N*-acetyl-α-D-galactosaminyltransferase; α2FucT, α-2-L-fucosyltransferase; β3GalT, β-3-D-galactosyltransferase; GT, glycosyltransferase; P4H, prolyl 4-hydroxylase; HyPro, hydroxyproline; Gal, D-galactose; Fuc, L-fucose

* Corresponding author. Tel.: +1-405-271-2227; fax: +1-405-271-3139.
E-mail address: Christopher-west@ouhsc.edu (C.M. West).

0304-4165/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbagen.2004.04.007

partments of eukaryotes. Skp1 produced by the mycetezoan *Dictyostelium discoideum* is modified at a specific hydroxyproline (HyPro) residue by a pentasaccharide [9] with the structure shown in Fig. 1. The majority (>90%) of Skp1 in the cell appears to be fully modified [10], although purified Skp1 shows heterogeneity with respect to the two peripheral α-linked D-galactose (Gal) residues [11]. Skp1 is the only known *Dictyostelium* protein to be modified by this structure based on radiolabeling methods that probe for its internal L-fucose (Fuc) residue [12]. The modification is formed by the sequential action of a soluble prolyl hydroxylase and five soluble glycosyltransferases (GT), which will be reviewed below.

A function for Skp1 in cell cycle regulation and centromere organization was originally defined genetically in mammalian cell lines and yeast, and subsequent proteomics and biochemical studies have shown that it is a member of several multi-subunit protein complexes in the cell [13]. It is best known as an adaptor-like protein in the SCF-class of E3-ubiquitin ligases, a group of enzymes that contains a variable substrate recognition protein, the F-box protein [14]. Skp1 has also been implicated in complexes associated with the yeast centromere [15], V-ATPase assembly, and vesicle-docking complexes [16]. At present, Skp1 functions have been defined experimentally primarily in yeast, *C. elegans*, plants, and mammalian cells, whereas Skp1 glycosylation is defined in *Dictyostelium*. However, as will be delineated below, the genomes of certain other organisms harbor sequences predicted to encode Skp1 modification enzymes, raising the possibility that this pathway is more general. In *Dictyostelium*, pharmacological or mutational blockade of Skp1 glycosylation inhibits its nuclear accumulation [10], and cells which lack the Skp1 β-3-D-galactosyltransferase (β3GalT) enzyme as a result of gene disruption have altered cell size [17].

Prolyl hydroxylation has recently assumed a new dimension of importance with the discovery that the accumulation of the transcriptional factor subunit HIF-1α, and probably other cytoplasmic/nuclear target proteins, is tightly regulated by the oxygen-dependent hydroxylation of critical Pro-residues by soluble P4Hs [18]. At low oxygen levels, hydroxylation is reduced and HIF-1α accumulates to dimerize with HIF-1α, enters the nucleus, and activates expression of new genes appropriate for response to hypoxia. The finding that a HyPro residue in a cytoplasmic protein like Skp1 can be capped by a glycan, as was previously known to occur in the secretory pathway of plants [19], provides a potential new mechanism for regulating the level of proteins in the cytoplasm.

In this article, characteristics of the *Dictyostelium* Skp1 GTs and prolyl 4-hydroxylase (P4H) will be reviewed. In addition, results of a search for related sequences in the genomes of other organisms are presented which suggest that (1) the pathway has its evolutionary origin in prokaryotes, (2) a related glycosylation pathway exists in the cytoplasm of other lower eukaryotes, and (3) that early steps of the pathway have counterparts in the secretory pathway of eukaryotes.

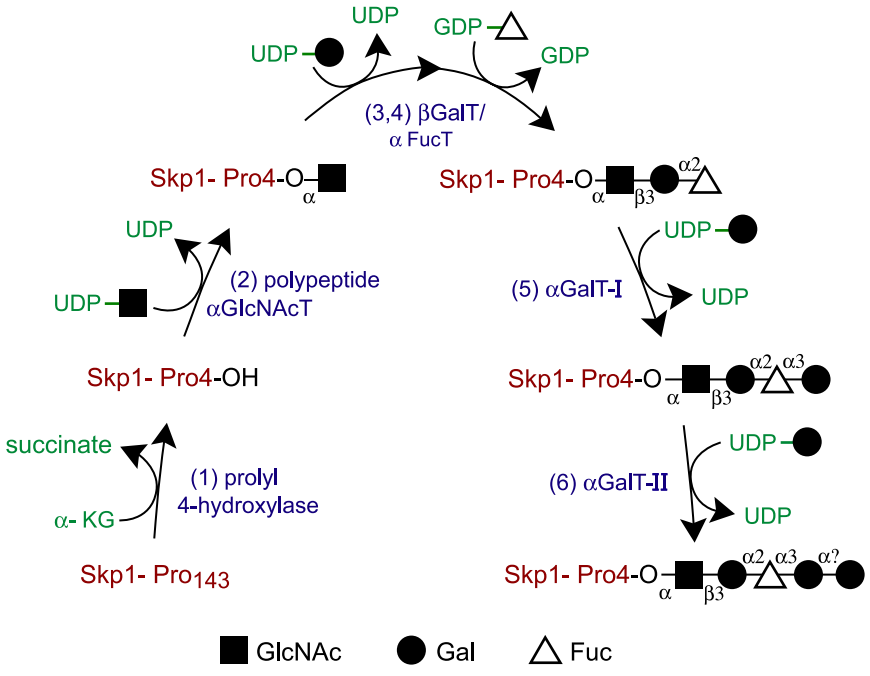


Fig. 1. The *Dictyostelium* Skp1 modification pathway. Skp1 Pro-143 is sequentially modified by a soluble P4H and 5 soluble, sugar nucleotide-dependent glycosyltransferase activities. Enzymes are in blue; substrates are in red, green and black.

2. Initiation of the glycan: the Skp1-HyPro GlcNAcT

2.1. Properties of the enzyme

The first GT in the pathway (Fig. 1), Skp1 polypeptide *N*-acetyl-α-glucosaminyltransferase (pp αGlcNAcT), was purified as an activity that transfers [^{3}H]GlcNAc from UDP-[^{3}H]GlcNAc to a recombinant mutant Skp1 that is poorly modified when expressed in *Dictyostelium* [20]. Activity toward mutant Skp1 and a 23-amino acid synthetic peptide, which includes the 4-hydroxyproline glycosylation site of Skp1, is associated with a single soluble protein GnT51. GnT51 appears to be fully active when expressed recombinantly in *E. coli* [21]. Activity is critically dependent on a divalent cation such as Mn^{2+} and a reducing agent such as dithiothreitol, and exhibits a strong preference for Skp1 over the synthetic HyPro-peptide. The sequence of GnT51, referred to as Dd-ppGnT1, is distantly related to those of enzymes that initiate mucin-type *O*-glycosylation in the Golgi of animals (Section 2.2). Like the animal polypeptide *N*-acetyl-α-D-galactosaminyltransferases (pp αGalNAcT), Dd-ppGnT1 has a neutral pH optimum, prefers low salt, and requires a divalent metal ion for activity. However, the signal anchor and spacer domain present in the pp αGalNAcTs (Fig. 2B.2) and other Golgi enzymes [22] are absent from Dd-ppGnT1 (Fig. 2A.2), suggesting that this enzyme is not targeted to organelles. Indeed, as a soluble enzyme whose activity depends on a reducing agent and is stimulated maximally by submicromolar rather than submillimolar concentrations of substrates, Dd-ppGnT1 behaves like a cytoplasmic protein. It is unrelated to the Thr/Ser-*O*-βGlcNAcT also present in the cytoplasm as well as the nucleus [5]. *O*-βGlcNAcT modifies many targets in the cytoplasm and nucleus, inverts the anomeric configuration of GlcNAc during transfer from UDP-GlcNAc, is not metal-dependent, and belongs to a different GT superfamily [23]. Therefore, the initial step in Skp1 glycosylation is novel and occurs in the cytoplasm, i.e., Skp1 does not translocate to vesicles of the secretory pathway to begin its glycosylation.

2.2. Relationship to Golgi pp αGalNAcTs

Traditional mucin-type *O*-glycosylation, a major and seemingly distinct modification pathway of the secretory apparatus, occurs in vertebrate and invertebrate animals [2,24] and, as recently documented, in the apicomplexan protozoan *Toxoplasma gondii* [25]. This modification is

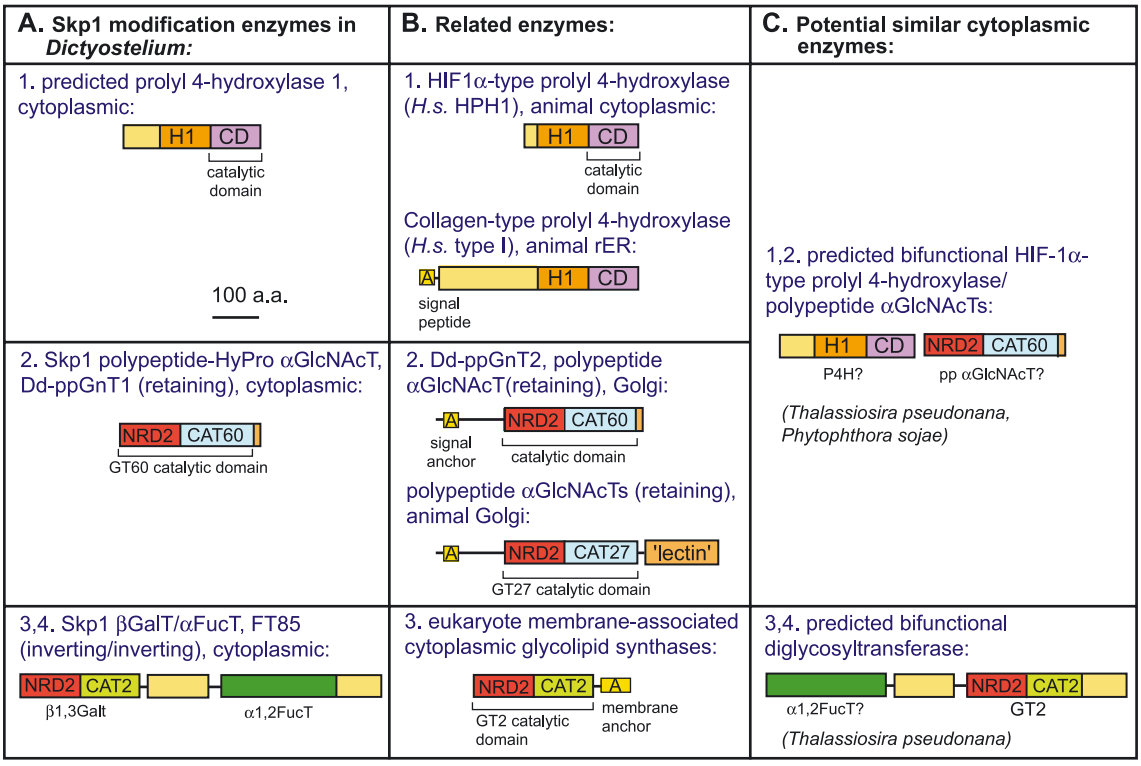


Fig. 2. Domain architecture of modifying enzymes. A: Predicted and known Skp1-modification enzymes, including (1) P4H-1, (2) Dd-ppGnT1, and (3,4) the bifunctional β3GalT/α2FucT. B: Comparison with known enzymes from eukaryotes, including (1) HIF-1α- and collagen-type P4Hs, (2) Golgi-localized pp αGlcNAcTs from unicellular eukaryotes and pp αGalNAcTs from animals, and (3) an example of a glycolipid synthase, such as Dol-P-Glc synthase, which is membrane-associated but oriented toward the cytoplasm. C: Comparison of new cytoplasmic enzymes predicted from the sequence analyses in Figs. 3–5, including (1,2) a bifunctional P4H/pp αGlcNAcT, and (3,4) a predicted bifunctional GT. H1, conserved P4H domain; CD, P4H catalytic domain; NRD2, GT nucleotide recognition domain 2; CAT60, family GT60 catalytic subdomain; CAT27, family GT27 catalytic subdomain; lectin, lectin-like domain; CAT2, family GT2 catalytic subdomain; A, Signal or membrane anchor. Scale bar is shown in panel A.1.

initiated by the attachment of GalNAc in α-linkage to the hydroxyl side chain of Thr or Ser, catalyzed by the action of a pp αGalNAcT. The pp αGalNAcTs, classified as family GT27 [26], are encoded by a family of similar genes numbering up to possibly 24 in higher animals and 4–5 in the protozoan apicomplexans (see below), although less than half have been documented experimentally.

The pp αGalNAcT catalytic domain consists of an N-terminal, 125-residue nucleotide recognition domain-2 (NRD2)-type subdomain, and a second subdomain of about the same length that includes the so-called Gal/GalNAc-type motif [24,27], referred to as CAT27 in Fig. 2B.2. The NRD2-subdomain is a shared feature of several, mostly cytoplasmic GT families including GT2, GT21, GT27 and GT60 [6,26,27]. Homology between Dd-ppGnT1, classified as family GT60 [26], and Mm-ppGalNAcT1 is supported by a sequence alignment that extends through both of these subdomains [28] (Fig. 3A). Consensus amino acids associated with the DxD-like DxH-motif characteristic of the pp αGalNAcT NRD2 subdomain, and with the Gal/GalNAc-motif, are conserved in Dd-ppGnT1. Conservative amino acid substitutions of D or H in the DxH-motif abrogate the enzyme activities of both Dd-ppGnT1 and Mm-ppGalNAcT1 [21,29].

Most pp αGalNAcTs modify peptides with amino acid repeats rich in Thr and Ser but unique amino acid sequences can also be targeted [24]. Some pp αGalNAcTs appear to be specialized for unmodified peptides, whereas others depend on previously added GalNAc residues. This specificity has been correlated with a 150-residue C-terminal domain (Fig. 2B.2) with homology to the plant lectin ricin [24]. Dd-ppGnT1 does not appear to contain this lectin-like C-terminal domain, and is thus more similar to a putative pp αGalNAcT from *C. elegans*, Gly8, which also lacks this domain.

Therefore, it appears that Dd-ppGnT1 is homologous to the pp αGalNAcTs, but transfers a distinct HexNAc to a different acceptor hydroxyamino acid in an alternative compartment of the cell, and lacks the lectin domain found at the C-terminus of most pp αGalNAcTs. The differential compartmentalization is associated with absence of the N-terminal signal anchor and spacer region typically found in Golgi GTs.

2.3. Identification of other potential pp αGlcNAcTs

To investigate whether enzymes similar to Dd-ppGnT1 might occur in other organisms, tBLASTn-, PSI-BLAST- and PHI-BLAST-based searches for related sequences were performed in publicly accessible databases. Numerous similar genes in lower eukaryotes and prokaryotes were predicted based on weak sequence similarity throughout the approximately 275-amino acid catalytic domain, and conservation of key motifs such as DxH- and Gal/GalNAc-like sequences (Fig. 3A). They all lack the ricin-like C-terminal domain of the pp αGalNAcTs. With the exception of a prokaryotic sequence from the Gram-positive bacterium *Carboxydothermus hydrogenoformans*, these are more similar to Dd-ppGnT1 than to the Golgi pp αGalNAcTs.

They can be divided into two groups: those with N-terminal rER/Golgi targeting sequences as seen in the pp αGalNAcTs (Fig. 3A, blue names), and those without N-terminal targeting sequences as for Dd-ppGnT1 (green and orange names).

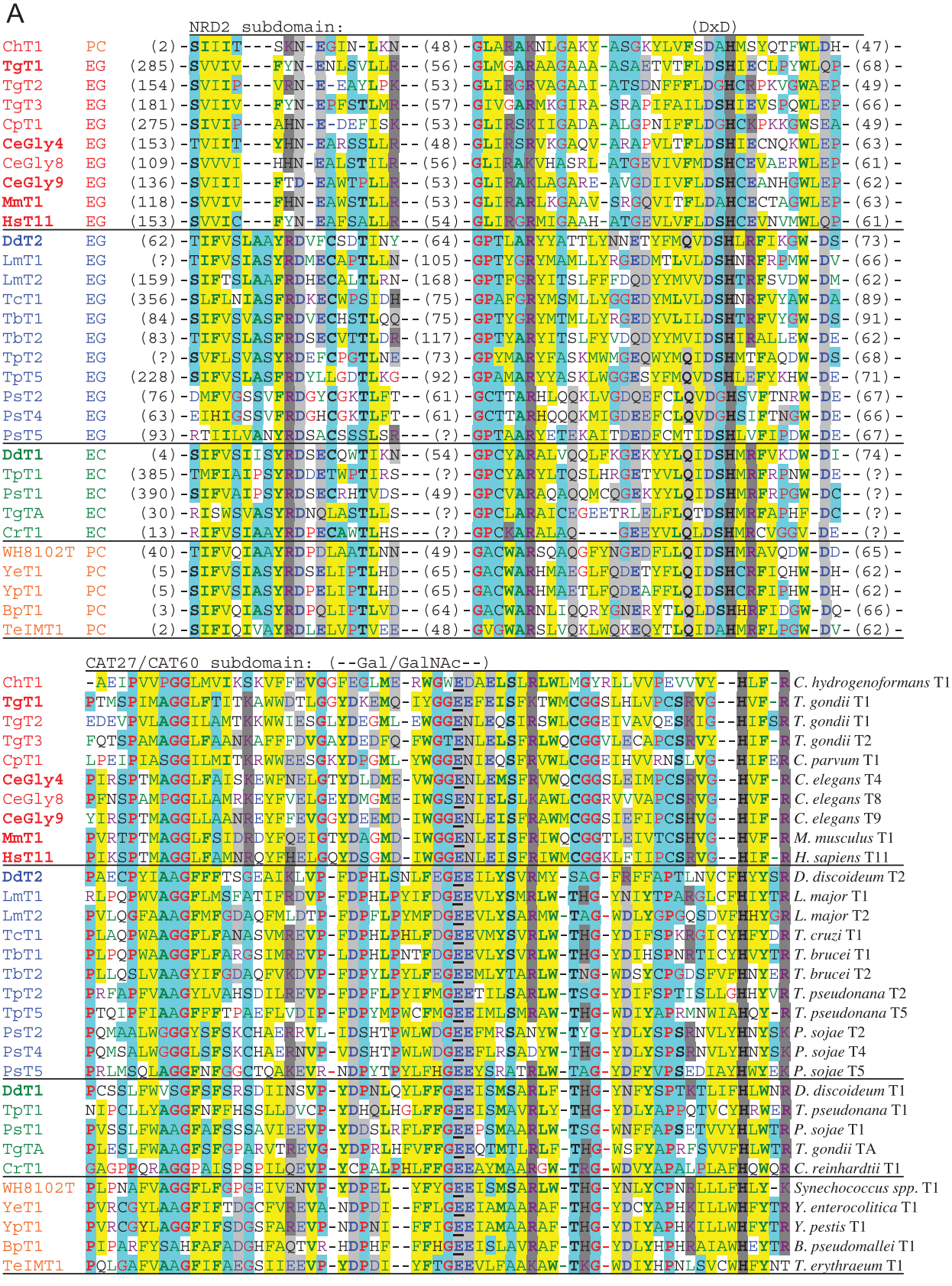
2.3.1. Golgi pp αGlcNAcTs

Predicted genes in the first group consist of a Dd-ppGnT1-like catalytic domain within a putative type 2 membrane protein, including an N-terminal signal anchor followed by a spacer-like sequence. They generally have six conserved Cys-residues consistent with a role in disulfide bonding as do the pp αGalNAcTs, but at distinct locations.

To test the prediction that these sequences encode pp αGlcNAcTs, a predicted protein from this group found in *Dictyostelium* was examined further. *Dictyostelium* is known to form GlcNAcα1-Thr and GlcNAcα1-Ser linkages in mucin-type peptide repeats in the GPI-anchored cell surface protein SP29 (PsA) [30,31] and, based on serological cross-reactivity with anti-glycan monoclonal antibodies, probably also on the spore coat protein SP85 (PsB) [32]. The gene predicted to encode this enzyme had been previously cloned as *cis4c* in a screen for genes which, when disrupted, resulted in resistance of clonal growth to the chemical cisplatin, and was known to be expressed throughout the life cycle [33]. When expressed as a recombinant protein substituted with an N-terminal cleavable signal peptide, the *cis4c* gene product was recovered from the growth medium and robustly catalyzed transfer of [^{3}H]GlcNAc from UDP-[^{3}H]GlcNAc to synthetic peptides corresponding to mucin-type repeats of SP29 and SP85 [34]. Thr residues in each of the sequence repeats in these peptides were modified with αGlcNAc, based on mass spectrometry and Edman degradation. *O*-glycosylation of SP29, SP85, and many other glycoproteins were dependent on the expression of the *cis4c(gntB)* gene product, Dd-ppGnT2, in vivo. Dd-ppGnT2 also complemented mutations in the *modB* locus, a previously described gene required for normal *O*-glycosylation of cell surface and secretory proteins, cell–cell adhesion, slug migration, cell sorting and spore coat formation [34]. These findings demonstrate multiple functions for mucin-type *O*-glycosylation in *Dictyostelium*.

Based on these observations, it is proposed that similar sequences in the other lower eukaryotes, including the trypanosomatids *Trypanosoma cruzi*, *T. brucei*, and *Leishmania major*, the diatom *Thalassiosira pseudonana*, and the algal plant pathogen *Phytophthora sojae* (Fig. 3A), are also pp-Thr αGlcNAcTs. A pp αGlcNAcT activity has been described in Golgi extracts of *T. cruzi* [35,36]. This activity, as well as Dd-ppGnT2 but not another retaining enzyme, Skp1 αGalT-I, is sensitive to inhibition in vitro by two nucleotide conjugates (A. Ercan, N. Heise, C.M. West, unpublished data) that potently and selectively inhibit ani-

A



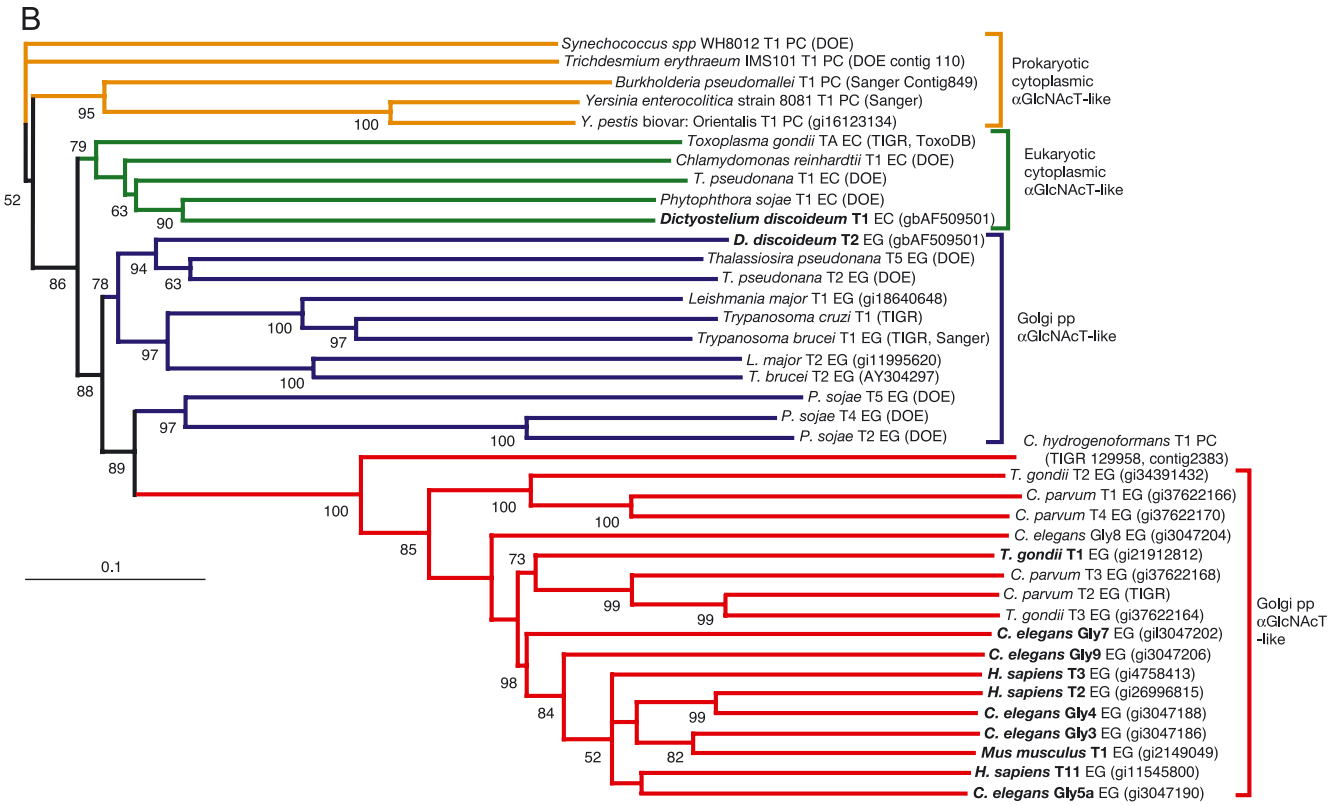


Fig. 3. Analysis of Skp1 αGlcNAcT-like sequences. Sequences similar to the catalytic domain of DdppGnT1 were identified using tBLASTn, PSI-BLAST with inclusion of the most similar Dd-ppGnT1-like sequences, and PHI-BLAST based on the DxD-like DxH-motif. A: Sequences were aligned manually according to the chemical character of the residues as described previously [6,28], and selected examples are shown in this panel. Because the sequences are highly divergent with few amino acid identities, amino acids are color-coded with respect to chemical similarities which were used as basis for the alignment, with hydrophobicity receiving the greatest weight. To emphasize related amino acids, positions occupied by residues of similar character are highlighted with colors coded to broader chemical similarities, and positions occupied by identical amino acids across all or nearly all of a subgroup are in bold. Locations of the NRD2 and CAT27/60 subdomains are defined by the top bars, positions of the DxD- and Gal/GalNAc-like motifs are indicated parenthetically above the top bars, and the putative catalytic Glu-residue in the Gal/GalNAc-like motif is underlined. The number of preceding (to the N-terminus) and intervening amino acids are in parentheses. Sequences are named by an acronym corresponding to the genus and species (spelled out in panel B), and are classified, grouped and color-coded as EG (eukaryote Golgi, i.e., predicted signal anchor), EC (eukaryote cytoplasmic, i.e., no signal anchor), or PC (prokaryote cytoplasmic). Names of sequences encoding established activities are in bold. B: A more extended sequence alignment, corresponding to residues 5–50, 66–126, 137–153, 158–183, 186–261, 272–282 and 362–386 (of 423 total) of Dd-ppGnT1, and 63–110, 135–192, 201–218, 225–249, 252–328, 337–347 and 361–387 (of 403 total) of Dd-ppGnT2, was subjected to a phylogenetic analysis using a distance algorithm. Trees were generated using PAUP* [42] under the minimum evolution criterion with tree-bisection reconnection branch swapping. Heuristic searches were performed with 10,000 replicates. Bootstrap values are given when they exceed 50%. The best tree, rooted with a cyanobacterial sequence, is shown. Clades are color-coded according to sequence groups in panel A. Sequences lacking identifiers were assembled from and confirmed by overlapping raw genomic and EST sequence reads from the indicated databases (see Acknowledgments for database URLs).

mal pp αGalNAcTs [37]. This provides further evidence for the homology of these two groups of enzymes. These pp αGlcNAcT-like sequences may be the microbial counterparts of the animal/apicomplexan pp αGalNAcTs, since no genomes have been found to contain both sequence types.

2.3.2. *Cytoplasmic pp αGlcNAcTs*

The second group of sequences is predicted to encode enzymes in the cytoplasmic compartment, based on the absence of apparent N-terminal signal anchor sequences. These sequences have few conserved Cys-residues consistent with the instability of disulfide bonds in the cytoplasm. They are found in both eubacteria (Fig. 3A, orange names) and eukaryotes (green names), but not archeae. The prokaryote sequences belong to single ORFs that are similar to Dd-ppGnT1 throughout the length of the protein, and share most sequence motifs. Their functions are undefined but the sequences from some organisms are contiguous to structural genes which in other bacteria are known to encode fimbria-like glycoproteins [38,39]. The eubacteria in which they are found are diverse representatives of cyanobacteria and proteobacteria. Included in the proteobacterial group are the pathogenic *Yersinia* species (including a partial sequence from *Y. pseudotuberculosis*, data not shown) and *Burkholderia pseudomallei*. Skp1-like and P4H-like sequences are not detectable in their genomes, so these putative enzymes might modify other hydroxyamino acids in distinct proteins.

Dd-ppGnT1-like sequences are also present in the genomes of the lower eukaryotes *T. pseudonana* (a diatom),

P. sojae (soybean stem and root rot; an oomycete), *T. gondii* (an apicomplexan) and *Chlamydomonas reinhardtii* (green alga). These genomes encode Skp1-like sequences (not shown) which conserve the glycosylation site and neighboring sequence context of *Dictyostelium* Skp1; these sequences are therefore candidates for modification by these predicted pp αGlcNAcTs. Interestingly, the predicted gene products from *T. pseudonana* and *P. sojae* appear to be multidomain proteins unlike Dd-ppGnT1, with P4H-like domains at their N-termini and GnT1-like domains at their C-termini (Fig. 2C.1,2). As discussed below (Section 5.2), these N-terminal P4H-like domains may act in concert with the Dd-ppGnT1-like domains, to modify Skp1 analogous to the *Dictyostelium* pathway which, however, utilizes two separate proteins for this purpose.

2.3.3. Model for the evolution of the initial step of mucin-type O-glycosylation

To examine the evolutionary relationships among the 21 predicted Dd-ppGnT1-like and Dd-ppGnT2-like proteins, an alignment of the sequences of the catalytic domains from these and selected pp αGalNAcTs was subjected to a phylogenetic analysis. The best tree, shown in Fig. 3B, is rooted with a sequence from an ancient group, the cyanobacteria, which have three known Dd-ppGnT1-like sequences (two are shown). The other cyanobacterial sequence is similar and they are not polytomic when the tree is rooted with other sequences. With the exception of a sequence from a Gram-positive bacterium, *C. hydrogenoformans*, the gene tree resembles a species tree.

The earliest branch (in orange) includes three cytoplasmic sequences from enterobacteria including the bubonic plague agent *Yersinia pestis*. This position is consistent with phylogenetic studies representing enterobacteria as derivative to the proteobacteria [40].

In the eukaryotic realm, the tree branches with predicted cytoplasmic and Golgi-associated (i.e., type 2 membrane proteins) sequences occupying distinct clades (green and blue/red). Because one organism, *T. gondii*, is represented by expressed sequences in both clades, this bifurcation probably represents an ancient gene duplication associated with eukaryotic compartmentalization. The cytoplasmic clade (in green) includes Dd-ppGnT1 suggesting that these other sequences also encode active pp αGlcNAcTs.

The type 2 membrane protein (Golgi-like) sequences branch into two subclades. The upper subclade includes eight sequences (in blue) from lower eukaryotes including Dd-ppGnT2 [34], suggesting that these sequences may encode active pp αGlcNAcTs. Sequences from *Dictyostelium* and the diatom *T. pseudonana* are resolved from the trypanosomatid sequences, including representatives from *T. cruzi*, *T. brucei*, and *L. major*. The sequence from *T. cruzi* is an excellent candidate to encode the known pp-Thr αGlcNAcT activity in this organism [35,36]. Recent studies show that this sequence is expressed in epimastigotes (Heise, West and Previato, unpublished data) and a sequence from *T. brucei* appears to encode a Golgi protein [67]. The *Dictyostelium* and *T. cruzi* sequences appear to be single copy genes, but *T. pseudonana*, *L. major*, and *T. brucei* each have at least two sequences. The *T. brucei* and *L. major* sequences are related as pairs that are more similar across species than within, suggesting a gene duplication and functional specialization which preceded speciation in the trypanosomatid lineage. These predicted orthologs are likely to have conserved specialized functions, as previously suggested for certain pairs of pp αGalNAcTs [41].

The lower Golgi subclade (blue/red) includes a cluster of three deeply branched sequences (in blue), from the algal plant pathogen *P. sojae*, that most resemble the pp αGlcNAcT-like sequences, and a long-stemmed branch which includes the pp αGalNAcTs (in red). In the red subclade, the 10 animal pp αGalNAcTs (*C. elegans*, *H. sapiens*, *M. musculus*) show an evolutionary relationship similar but not identical to the relationship derived in Ref. [41]; this may be due to differences in the range of amino acids or species selected. Seven sequences from the apicomplexan group of protozoa (*T. gondii*, *C. parvum*) are also found in this group and branch most deeply, which correlates with their more ancient speciation. One of these, Tg-ppGnT1 from *T. gondii*, is a documented pp-Ser/Thr αGalNAcT [25]. The long stem of the red subclade therefore correlates with a change in function of the encoded protein from a pp-Ser/Thr αGlcNAcT to a pp-Ser/Thr αGalNAcT. Two sequences from *T. gondii* are more similar to sequences from *C. parvum* than to each other, suggesting that they represent orthologs resulting from gene duplications that preceded speciation. The grouping of the putative Gly8 pp αGalNAcT from *C. elegans* with the apicomplexan sequences (*T. gondii, C. parvum*) predicts that Gly8 encodes a bona fide pp αGalNAcT despite the absence of a ricin-like C-terminal domain.

This analysis suggests the following simple model for the evolution of the initiating step in mucin-type *O*-glycosylation. The original GT may have emerged in a cyanobacterium via domain shuffling, with its NRD2 domain contributed by an inverting family GT2 gene and its CAT60-domain contributed by a retaining GT gene. This enzyme may have been a pp-Ser/Thr αGlcNAcT, as P4H-like genes that might generate a 4-hydroxyproline acceptor are not detectable in the bacterial genomes with pp αGlcNAcT-like sequences (see Section 5.2). This gene persists today in selected Gram-negative cyanobacteria and proteobacteria, which may have derived from cyanobacteria [40]. A copy of this gene was retained with the appearance of eukaryotes, and persists today in many unicellular creatures where it appears to have specialized as a pp-HyPro αGlcNAcT devoted to the modification of Skp1-like proteins in the cytoplasm. A duplication of this gene occurred early during eukaryotic radiation with the new gene, still a pp-Ser/Thr αGlcNAcT, acquiring an additional N-terminal sequence for compartmentalizing the catalytic domain within the lumen of the Golgi apparatus. In *Dictyostelium*, both of these genes encode pp αGlcNAcTs.

Prior to evolution of the apicomplexan protozoa the Golgi-type copy transmuted into a pp-Ser/Thr αGalNAcT, as represented by *T. gondii*, and was distributed to both invertebrate and vertebrate animals. This gene underwent multiple duplications, possibly preceding the animal radiation based on the close association of the putative *C. elegans* pp αGlcNAcT Gly8 with the protozoan sequences. The sequence in the Gram-positive bacterium *C. hydrogenoformans* most likely occurred as a lateral gene transfer. Additional gene duplications resulted in possibly 24 predicted pp αGalNAcT enzymes in humans. These enzymes are specialized for the primary and secondary modifications of distinct target sequences in different cell types [24].

The implications of this model are that *O*-glycosylation evolved anciently in the cytoplasm of Gram-negative bacteria, and likely targeted the side chains of Thr or Ser residues. A similarly ancient origin for the *N*-glycosylation pathway has recently been uncovered in the periplasmic space of bacteria [43]. This pathway persists today in the cytoplasms of selected lower eukaryotes, where it is specialized for the modification of HyPro residues formed on target proteins by cytoplasmic P4Hs (see Section 5). In the two unicellular eukaryotic genomes that have not yet yielded cytoplasmic P4H-like sequences (*T. gondii* and *C. reinhardtii*), which are also the most deeply branched in the eukaryotic cytoplasmic (green) clade, the pathway may target Ser or Thr residues. The pathway may have been deleted from fungi and other lower organisms, a phenomenon that has been described for other genes [44]. The cytoplasmic enzymes may target unique sequences in a small range of proteins. A novel role for this enzyme blossomed with the eukaryotic radiation, when a new copy of the gene was modified and amplified to encode a Golgi-targeted enzyme that modifies both unique and mucin-type repeat sequences in a large number of different proteins. Prior to metazoan evolution, this Golgi gene appears to have converted from a pp αGlcNAcT to a pp αGalNAcT, and this gene subsequently underwent duplications within the lineages of their respective organisms. These genes initiate formation of the bulk of *O*-glycans produced by mammalian cells today. Though it is not known why the apicomplexan/animal lineage adopted the use of GalNAc in place of GlcNAc, it correlates with the ability of UDP-4-epimerase to also form UDP-GalNAc from UDP-GlcNAc, a capability not possessed by the UDP-4-Glc-epimerase from *T. cruzi* [45]. This substitution might have been allowed as both *N*-acetylated amino sugars have similar effects on the conformation of the peptide backbone [46], and would presumably have provided greater opportunity for diversification of subsequent glycosyl modifications between *O*-linked and *N*-linked (which usually lack GalNAc) glycans after expression of the duplicate gene product in the Golgi. The proposed evolutionary relationship between protozoan cytoplasmic pp αGlcNAcTs and animal Golgi pp αGalNAcTs, while speculative, represents a useful simplification that might stimulate transfer of knowledge between these two classes of enzymes.

3. Extension of the glycan: the Skp1 βGalT/α*FucT*

3.1. A bifunctional diglycosyltransferase

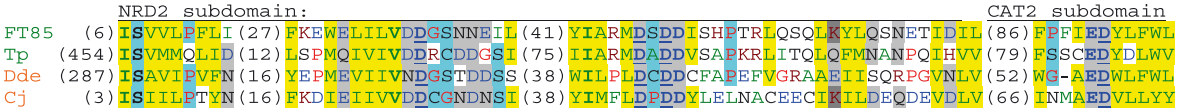
In *Dictyostelium* Skp1, the αGlcNAc is extended by the action of a β1,3GalT and an α1,2FucT (Fig. 1), resulting in the formation of a trisaccharide corresponding to the human type 1 blood group H antigen. The enzymes that form this structure, however, are essentially unrelated to those that form it in mammalian cells. The question of why evolution has convergently arrived at the same complex structure in *Dictyostelium* and mammals remains unanswered.

The two sugars are added by distinct catalytic domains of the same enzyme protein [17,47,48], FT85 (Fig. 2A.3), whether it is purified from *Dictyostelium* or expressed recombinantly in *E. coli*. FT85 is responsible for Skp1 modification in vivo based on gene disruption analysis. Kinetic studies show that this bifunctional diGT is able to modify Skp1 processively, but that each catalytic domain can also act independently of the other either in the intact protein or when expressed as separate polypeptides. It is therefore likely that FT85 is designed to efficiently extend the Skp1 glycan from the monosaccharide to the trisaccharide state, which would avoid intracellular accumulation of the disaccharide glycoform. This protein architecture is also employed in the biosynthesis of glycosaminoglycan polymers in certain bacteria [49] and probably some eukaryotes [26], though a processive mechanism has not been established.

3.2. Evolution of the β3GalT-domain

The βGalT domain belongs to the large GT2 family of GT-A superfamily domains [6,26], typified by SpsA, that share sequence motifs including the NRD2 subdomain, with its DxD-motif, and other conserved Asp-residues (underlined in Fig. 4A), invert the anomeric linkage of the sugar to the nucleotide, and exhibit metal-dependence. GT-2 catalytic domains, of approximately 250–300 residues, are characteristically associated with membranes but oriented toward the cytoplasmic compartment of the cell (Fig. 2B.3) [6,27]. Interestingly, the NRD2 subdomain is also present in the GT60 family enzyme Dd-ppGnT1 (see above), consistent with the proposed evolutionary origin of this domain in the cytoplasm. In bacteria and archeae, these enzymes contribute to the biosynthesis of lipopolysaccharide and capsular precursor glycolipids, cellulose, and hyaluronic acid and other glycosaminoglycans, all of which are co-synthetically translocated to the cell surface [6,50]. In eukaryotes, the GT2 domain has been adapted for the biosynthesis of Dol-P-Man, Dol-P-Glc, and glucosylceramide, in addition to its continued use in the formation of the polysaccharides cellulose, hyaluronate, and chitin [6]. The βGalT-domain of FT85 is a specialized example that is not membrane-associated, and modifies a glycoprotein that remains in the cytoplasm (or nucleus), rather than a glyco-

A. β3GalT-like sequences:



B. α2FucT-like sequences:

C. Identities and similarities:

	β3GalT-domain (270 aa)		α2FucT-domain (285 aa)		combined	
	identical	similar	identical	similar	identical	similar
FT85 vs. Tp	21%	53%	25%	60%	23%	57%
FT85 vs. Dde	19%	51%	28%	59%	24%	55%
Tp vs. Dde	17%	50%	19%	52%	18%	51%
FT85 vs. Cj	22%	60%	-		-	-

D. Origin of sequences:

FT85	EC	*Dictyostelium discoideum* (gi9022426, gbAAF82378)
Tp	EC	*Thalassiosira pseudonana*, Contig assembled from multiple overlapping raw genomic fragment sequences accessed at DOE Joint Genome Institute (http://genome.jgi-psf.org/thaps0/thaps0.home.html)
Dde	PC	*Desulfovibrio desulfuricans*, hypothetical protein G20 (giZP00129692, gi23474398)
Cj	PC	*Campylobacter jejuni*, β3GalT (gi6940833)

Fig. 4. Alignment of diglycosyltransferase-like sequences. A: Selected sequences with similarity to FT85 β3GalT catalytic domain. NRD2-like and CAT2-like subdomains are labeled above the bars. Conserved Asp (D)-residues are underlined. Underlined DxD-residues represent DxD motif; underlined ED-residues represent the predicted catalytic base. Numbers of amino acids upstream or skipped are represented in parentheses. Dashes represent empty positions. Sequence names are color-coded as in Fig. 3A. B: Sequences with similarity to FT85 α2FucT domain. C: Percentage of identical and similar residues between sequence pairs for each entire predicted catalytic domain. D: Origin of the sequences.

lipid or polysaccharide substrate that is subsequently transferred across a membrane.

A screen of eukaryotic genomes reveals a small number of family GT2-like sequences [6] with characteristic motifs that are predicted to be soluble proteins, i.e., not membrane bound as would be expected for glycolipid or polysaccharide synthesis. These are therefore candidates for mediating cytoplasmic glycosylation as seen for *Dictyostelium* Skp1. For example, a conserved ORF in *C. elegans* [6], zebrafish, *Xenopus*, mouse, and humans is expressed and up-regulated in human cancer cell lines. This predicted enzyme may have other targets as currently known human Skp1 isoforms lack the equivalent of Pro143 and do not appear to be modified in the same fashion as *Dictyostelium* Skp1 (C.M. West et al., unpublished data). In a second, intriguing example found in several lower organisms, a GT2-domain appears to be fused to an αFucT-like domain as seen for *Dictyostelium* FT85 (see below). Since domains of family GT2 can mediate the transfer of many sugar types, it is difficult a priori to predict the linkage formed from these putative enzyme sequences.

3.3. Evolution of the α2FucT-domain

The FT85 αFucT domain is distinct in sequence and metal-dependence from other known α1,2FucTs [17,48], though the three-dimensional structures of these enzymes remain to be determined. A BLAST search revealed similar FT85 αFucT-like sequences in the genomes of the proteobacterium *Desulfovibrio desulfuricans* and the diatom *T. pseudonana*. A comparison of three highly conserved regions is shown in Fig. 4B. These sequences are 25–28% identical, and about 60% similar, to a 285-residue interval of the FT85 α-2-L-fucosyltransferase (α2FucT)-domain, and are slightly less similar to each other (Fig. 4C). The α2FucT-domain was suggested previously to be a family GT2 member [6], but this alignment shows that the sequence motifs in FT85 on which this suggestion was based are not conserved for this group consistent with earlier site-directed mutagenesis studies [48]. Interestingly, these α2FucT-like sequences are adjacent to β3GalT-like sequences (Fig. 4A) predicting that these genes encode soluble, two-domain proteins as for *Dictyostelium* FT85, except that the order of the domains is reversed (Fig. 2C.3). The degree of identity and similarity among the β3GalT-like sequences is slightly less than that of the associated α2FucT-like sequences and a bacterial β3GalT (Fig. 4C), so the enzymatic activity of this domain cannot be predicted with certainty. Since *T. pseudonana* also possesses a genomic DNA sequence predicted to encode a protein with both P4H and pp αGlcNAcT activity (see above and below), and a Skp1-like gene with the equivalent of Pro143 (not shown), the proteins encoded by these genes may comprise a

cytoplasmic glycosylation pathway in diatoms that is related to the Skp1 modification pathway defined in *Dictyostelium*.

4. Peripheral α-galactosyl modifications

The Skp1 pentasaccharide contains two peripheral α-linked Gal residues attached to the core trisaccharide Fucα1,2Galβ1,3GlcNAc- (Fig. 1). Their addition in vivo is strictly dependent on the presence of Fuc based on analysis of a GDP-Fuc synthesis mutant, and substrates lacking Fuc are not αGalT acceptors in extracts [51]. Mild acid hydrolysis of the glycopeptide cleaves off three sugars simultaneously suggesting that the αGal-residues are linked via Fuc [9]. A UDP-Gal-dependent Skp1 αGalT activity was recently purified 2400-fold, and shown to form a Galα1,3Fuc linkage based on co-chromatography with synthetic Gal-Fuc-Bn standards [51]. Dd-αGalT1 is a soluble protein and requires Mn^{2+} and reducing agent for activity. This enzyme therefore has biochemical properties of a cytoplasmic protein like the earlier enzymes in the pathway. The attachment position of the other αGal is not certain. Dd-αGalT1 displays marked preference for native Skp1 compared to denatured Skp1, and for the nonreducing terminal disaccharide relative to the native trisaccharide [51]. Therefore, it is suggested that Dd-αGalT1 requires a proper three-dimensional conformation of Skp1 and its acceptor glycan prior to addition of αGal. Since polypeptide recognition would seem unnecessary for specific targeting of Skp1 considering the relative uniqueness of the target trisaccharide, it may constitute a quality control mechanism for proper Skp1 folding. Consistent with the quality control model, an isoform of Skp1 with point mutations in its N-terminal region is inefficiently processed by the αGalTs in vivo [9].

5. Cytoplasmic prolyl hydroxylation

5.1. Hydroxylation of Skp1 Pro143

Hydroxylation of Pro143 of Skp1 is a prerequisite for its glycosylation (Fig. 1) and has been demonstrated by MS–MS studies on the purified Skp1 glycopeptide [9]. The hydroxyl group is assumed to be located at the 4-position because a synthetic 4-HyPro-peptide is a good substrate for the subsequently acting Skp1 pp αGlcNAcT (Dd-ppGnT1) [20], but this remains to be established directly. Modification of Skp1 by a conventional P4H is supported by the finding that two inhibitors of P4Hs, α,α′ -dipyridyl and ethyl-2,3-dihydroxybenzoate, effectively suppress glycosylation of Skp1 in cells [10]. A coupled assay for Skp1 P4H based on transfer of [^{3}H]GlcNAc from UDP-[^{3}H]GlcNAc to the modified Pro of Skp1 in the presence of added Dd-ppGnT1 has been developed (H. van der Wel, C.M. West, unpublished data). The Skp1 P4H activity appears to be cytoplasmic, depends on Fe^{2+}, α-ketoglutarate, and ascorbic acid, which are co-substrates or co-factors for other known P4Hs, and is inhibited by α,α′ -dipyridyl and 2,3-dihydroxybenzoate.

Two classes of P4Hs are now known: the familiar rER P4Hs which are essential for collagen triple helix formation in animals and glycosylation of extensins and arabinogalactans in plants, and the more recently discovered P4Hs which modify HIF-1α and a number of other proteins in the cytoplasm and nucleus [52,53]. Hydroxylation of either of two distinct Pro residues in HIF-1α results in recognition by pVHL E3 ubiquitin-ligase [54], polyubiquitination, and subsequent degradation in the 26S-proteasome. In mammals, there are several HIF-1α-type P4Hs which vary in cell type expression and cytoplasmic or nuclear compartmentalization [55,56]. These P4Hs are dependent on physiological levels of molecular oxygen and, in hypoxic conditions, HIF-1α is not modified and therefore escapes degradation [57]. The sequences of the catalytic domains and upstream regions of the collagen- and HIF-1α-classes are distinct but can be aligned (Fig. 5A), and regions further upstream that are involved in acceptor substrate recognition [58], association with other subunits, and organelle targeting are divergent (Fig. 2B.1).

The Skp1 P4H might be most similar to the HIF-1α-class based on similar cytoplasmic localization. The sequence context of Skp1Pro143 is KNDFT**P**EEEQIRK, which appears distinctive from the canonical sequence context of HIF-1α-P4Hs, $LxxLAPx_{3-4}D_{2-3}$, and from the $(xPG)_n$ repeat sequences of collagen-type domains. However, site-directed mutagenesis reveals that the HIF-1α-P4H sequence is tolerant of substitutions [59,60], and a P4H from the vascular plant *A. thaliana* modifies both collagen and HIF-1α peptides [61], suggesting that these enzymes might also recognize other targets and that it is not yet possible to predict target specificity based on enzyme amino acid sequence.

5.2. Candidate Skp1 P4H genes

The *Dictyostelium* genome has been sequenced to a 10-fold level of redundancy indicating that coding sequences for >95% of the genes are accessible [62]. To search for candidate P4H genes in *Dictyostelium* and other organisms, sequence databases were queried with the catalytic domains (region CD in Fig. 2B.1) of the *egl-9* (HIF-1α) P4H from *C. elegans* and the collagen-type P4H from the protozoan virus PBCV-1 (*Paramecium bursaria* Chlorella virus-1), using the tBLASTn program. This sequence consists of about 120 amino acids and is typically located near the C-termini of proteins [52]. The search yielded five high-scoring sequences from *Dictyostelium*, and additional sequences from bacteria and other lower eukaryotes, including two that also have Dd-ppGnT1-like sequences, *T. pseudonana* and *P. sojae*. These sequences are predicted to reside in cytoplasmic proteins based on the apparent absence of N-terminal signal peptides. They each possess the two His- and one Asp-residues implicated in coordi-

nating Fe^{++}, and the basic residue thought to bind the C-5 hydroxyl of α-ketoglutarate in all known P4H's; these are asterisked in the alignment of selected sequences shown in Fig. 5A. They are distinct from other α-ketoglutarate-dependent dioxygenases such as HIF-asparaginyl hydroxylase and lysyl oxidases, based on the distance between the second His and the basic residue and overall sequence differences. To investigate their relatedness, they were

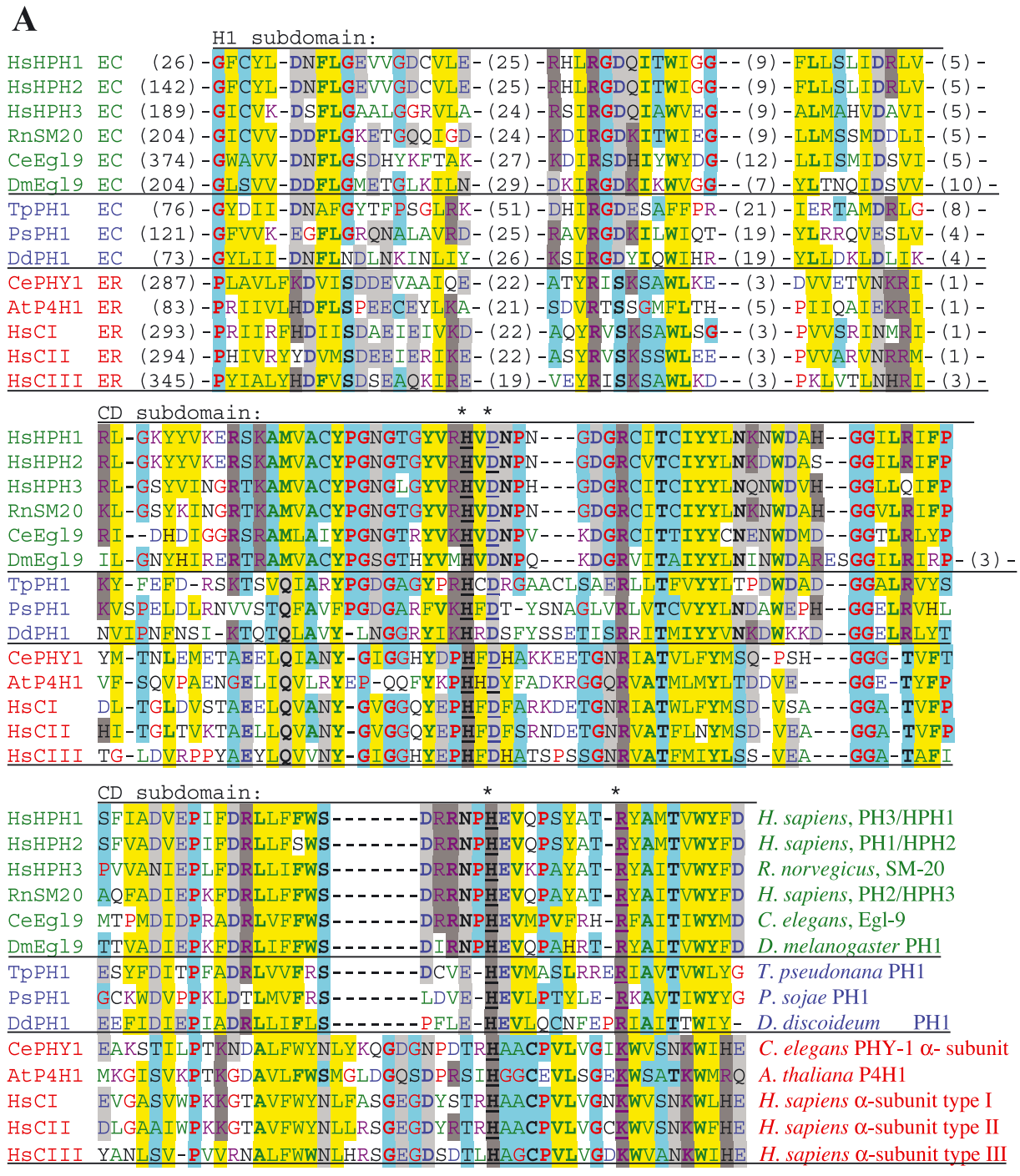


Fig. 5. Analysis of P4H-like sequences. A: Sequences with P4H-like catalytic domains were identified by BLAST in publicly accessible databases, color-coded, and aligned as in Fig. 3A. Sequences associated with four motifs, which span most of the length of HsHPH1 (Fig. 2B.1), are referred to as H1 (hydroxylase-1) and CD (putative catalytic domain) on the top bar. Known catalytically important residues for these enzymes are asterisked above the bar. Predicted compartmentalization, species names, and color-coding are as in Fig. 3. The collagen-type P4Hs are localized in the rER (in red), whereas the HIF-type P4Hs are localized in the cytoplasm (in green). Predicted P4Hs from Protist genomes appear to be cytoplasmic and are in blue. B: The CD-region (residues 108–216 of HsHPH1 of 244 total) of a larger alignment that included sequences from additional genomes was subjected to a phylogenetic analysis as described in Fig. 3B, and the best, unrooted tree is shown. Names of sequences encoding known P4H activities are in bold.

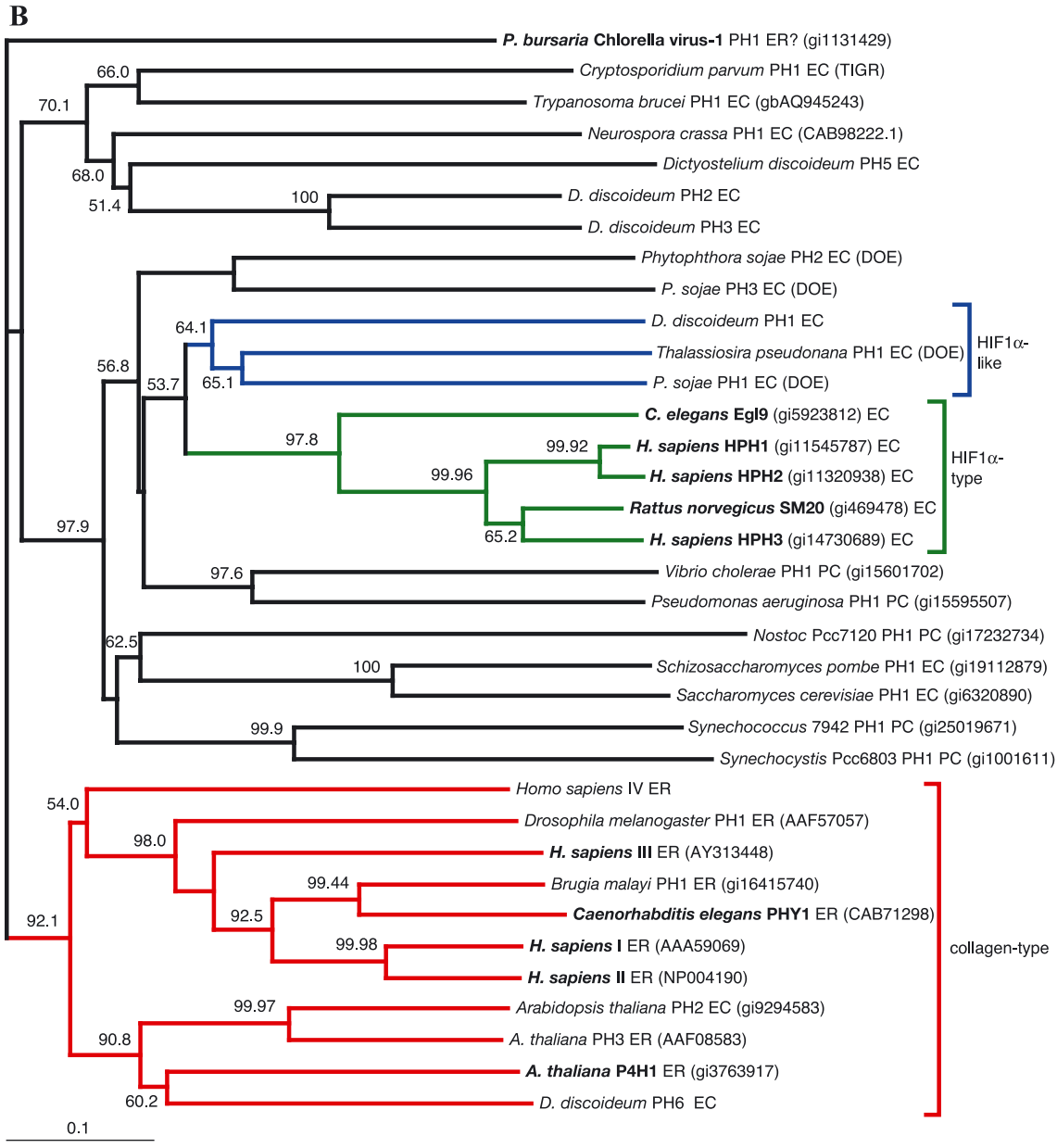


Fig. 5 (*continued*).

subjected to a phylogenetic analysis as described above for the pp αGlcNAcT-like sequences, and the best, unrooted tree is shown in Fig. 5B. The sequences resolved into two major clades. The lower clade (in red) includes all the known collagen-type rER-localized (ER) P4H sequences (in bold). The other sequences in this clade with predicted signal peptides (denoted ER) are expected to encode proteins with similar activity. Two sequences lack canonical signal peptides (denoted EC) and are predicted to be cytoplasmic but have collagen-type P4H sequences.

The known HIF-1α-type P4H's (in bold) occupy a discrete subclade (in green) of the upper clade (Fig. 5B). A closely related subclade (in blue) branches from the green subclade. This contains a sequence from *Dictyostelium* (DdPH1) and two other lower eukaryotes, *T. pseudonana* (TpPH1) and *P. sojae* (PsPH1). All the sequences in this large clade are predicted to lie at the C-terminus of cytoplasmic proteins (based on the absence of detectable N-terminal signal peptides), except for TpPH1 and PsPH1, which have Dd-ppGnT1-like C-terminal domains (see below). The predicted proteins from eukaryotes are denoted EC and those from prokaryotes are denoted PC. They each most resemble the HIF-1α-class, though the upper subclade has poor bootstrap support and lies inter-

mediate between the two classes when the tree is rooted with any of the bacterial sequences. The expression and identity of these sequences remain to be investigated. The bacterial sequences lie in multiple subclades and may have evolved by horizontal transfer, as seen for certain metabolic genes from bacteria to eukaryotes (e.g., Ref. [63]). The Chlorella virus (PBCV-1) sequence, which represents a third branch, has been shown to be a P4H possibly of the collagen type [64], but lacks a clearly defined signal peptide/anchor. This sequence is more related to the collagen-type clade when the tree is rooted with any of the sequences in the upper clade.

DdPH1, TpPH1 and PsPH1 share additional similarity with both the HIF-1α- and collagen-type sequences for an additional 80 upstream amino acids extending almost to the N-terminus of the shortest P4H of the group, the HIF-1α-type HPH1 (region H1 of Fig. 2A.1 and B.1, and Fig. 5A). Within the HIF-1α-type (green) clade, human HPH-1 is 53% identical and 80% similar to *C. elegans* Egl-9 over the entire extended region shown in Fig. 5A (H1 and CD regions). In comparison, DdPH1, TpPH1 and PsPH1 are 33–39% identical and 58–67% similar to Egl-9 and each other. These sequences are only 17–18% identical and 49% similar to the collagen-type P4H from *C. elegans* (PHY1 from the red subclade), providing additional support that these lower eukaryotic sequences are of the HIF-1α-type. In addition, their conserved catalytic domain basic residue is Arg, like the HIF-type, rather than Lys like the collagen-type. DdPH1, TpPH1 and PsPH1 lack the N-terminal Zn-finger sequence seen in some but not all HIF-1α-type P4Hs.

Strikingly, TpPH1 and PsPH1 appear to reside in the same coding sequence as the pp-αGnT1-like sequences in these organisms (as depicted in Fig. 2C.1). These predicted genes would therefore encode bifunctional enzymes with the ability to both 4-hydroxylate and subsequently add an αGlcNAc to the same amino acid, as occurs on *Dictyostelium* Skp1. In the rER of mammals, a trifunctional protein (LH3) with a related architecture mediates the hydroxylation of lysine and its subsequent modification by Gal and Glc [65]. A bifunctional mechanism is also employed later in the *Dictyostelium* Skp1 pathway (Fig. 1), where the FT85 protein processively catalyzes the addition of βGal and αFuc and, similarly, an FT85-like gene also occurs in *T. pseudonana* (Section 3.3). The association of HIF-1α-P4H-like domains with predicted pp αGlcNAcT domains suggests that these proteins are enzymatically active in a common pathway.

This analysis reveals candidate genes that encode bifunctional enzymes which perform initial steps of prolyl hydroxylation and glycosylation in the cytoplasm of two lower eukaryotes in addition to *Dictyostelium*. In turn, the similarity of DdPH1 to TpPH1 and PsPH1 makes it a candidate gene for mediating Skp1 prolyl hydroxylation in *Dictyostelium*. It is intriguing that the hypothesized hydroxyproline residues may subsequently be capped by a sugar moiety, which would provide another level of regulation of the signaling function mediated by the P4Hs (Section 1).

6. Prospects for glycosylation of cytoplasmic protein-hydroxyproline in other organisms

The equivalent of Pro143 in *Dictyostelium* Skp1 is present in predicted Skp1 proteins of all known unicellular eukaryotes including yeast, fungi, trypanosomes, apicomplexans, oomycetes and algae, and in vascular plants and invertebrate (but not vertebrate) animals. Therefore, these Skp1 proteins are each candidates for hydroxylation as observed in *Dictyostelium*. Consistent with this possibility, the genomes of most of these organisms contain nucleotide sequences in putative genes predicted to encode proteins in the cytoplasm that contain P4H catalytic domains (Fig. 5B). The resemblance of many of these gene sequences to the HIF-1α-class of P4Hs suggests that hydroxylation may be O_2-dependent as for the HIF-1α-class enzymes (Section 1), but the activity, regulation and actual targets of these proteins, whose expression have not even been verified, remain to be examined.

HyPro143 is modified by αGlcNAc in *Dictyostelium* Skp1. Though glycosylation of HyPro by Gal or arabinose is common in plants and certain algae (Volvocaceae), this is restricted to secretory proteins owing to the localization of the GTs to the secretory compartment [19]. The possibility of an αGlcNAc modification of Skp1 HyPro residues in the cytoplasm of other organisms is supported by the existence of gene sequences predicted to encode cytoplasmic pp αGlcNAcTs in the oomycete plant pathogen *P. sojae* and the diatom *T. pseudonana* (Fig. 3A). These sequences have not been detected in vascular plants and fungi but the possibility that they are concealed by sequence divergence cannot be excluded. It is particularly intriguing that a P4H-like catalytic domain and a pp αGlcNAcT-like domain appear to be joined within the same protein in these alga-like protozoa (Fig. 2C.1). In the rER, a similar relationship between a lysyl oxidase and two GT domains probably ensures that collagen is glycosylated at HyLys residues before folding is complete. Although it is not known if the predicted cytoplasmic enzymes modify Skp1, their apparent bifunctional nature raises the possibility that the predicted hydroxylation product of the P4H-like domain can be subsequently capped by αGlcNAc. This would almost certainly sterically block recognition of the HyPro moiety by the pVHL complex that ubiquitinates HyPro-HIF-1α [54] and other hydroxylated proteins [53], suggesting that this sugar modification would antagonize the effect of hydroxylation and provide an additional level of control on the half-life of the target protein. In addition to steric covering of the underlying HyPro, the αGlcNAc also provides a platform for extension of the glycan if requisite GTs are present as in *Dictyostelium*. The possi-

bility of extension in *T. pseudonana* is supported by the occurrence in its genome of a predicted bifunctional diGT (Fig. 2C.3,4; and Fig. 4) similar to the FT85 enzyme that acts on Skp1 in *Dictyostelium*. The *Dictyostelium* enzyme can processively extend the Skp1 glycan to the trisaccharide, which is then conditionally subject to peripheral αGal modifications (Section 4). A comparative genomics analysis of these outer modifications must await identification of the responsible genes. These complex structures might serve as ligands for lectin-like proteins known to accumulate in the cytoplasm/nucleus (e.g., Ref. [66]) or, alternatively, might serve as markers for quality control of protein folding or protein subunit assembly mediated by the processing GTs themselves [28].

Acknowledgements

Work on the Skp1 and related glycosylation pathways in the corresponding author's laboratory has been supported by grants from the NIH (GM-37539) and NSF (MCB-9730036 and -0240634). We are grateful to Bugoslaw Wojczyk (Rochester) for discussions about Apicomplexan sequences. For *Dictyostelium* EST sequence data, we acknowledge the University of Tsukuba (Japan) cDNA Sequencing Initiative. For *Dictyostelium* gDNA sequence data, we acknowledge the Institute of Biochemistry I, Cologne and the Genome Sequencing Centre Jena http://genome.imb-jena.de/dictyostelium/; supported by the Deutsche Forschungsgemeinschaft, No. 113/10-1 and 10-2), the Baylor College of Medicine (supported by the National Institute for Child Health and Human Development), and the National Biomedical Computation Resource at the San Diego Supercomputer Center (http://dicty.sdsc.edu/; supported by NIH P41-RR80605). We gratefully acknowledge the DOE Genome Institute (http://www.jgi.doe.gov/) for sequence data from *T. pseudonana*, *P. sojae*, *T. erythraeum*, *D. desulfuricans*, *Synechococcus* spp., and *C. reinhardtii*, and the Sanger Center (http://www.sanger.ac.uk/Projects/) for sequence data for *B. pseudomallei*, trypanosomatids, *T. gondii*, and *Yersinia*.

References

[1] H. Schachter, I. Brockhausen, The biosynthesis of branched *O*-glycans, Symp. Soc. Exp. Biol. 43 (1989) 1–26.
[2] G. Strous, J. Dekker, Mucin-type glycoproteins, Crit. Rev. Biochem. Mol. Biol. 27 (1992) 57–92.
[3] P. van den Steen, P.M. Rudd, R.A. Dwek, G. Opdenakker, Concepts and principles of *O*-linked glycosylation, Crit. Rev. Biochem. Mol. Biol. 33 (1998) 151–208.
[4] M. Fukuda, Roles of mucin-type *O*-glycans in cell adhesion, Biochim. Biophys. Acta 1573 (2002) 394–405.
[5] L. Wells, G.W. Hart, *O*-GlcNAc turns twenty: functional implications for post-translational modification of nuclear and cytosolic proteins with a sugar, FEBS Lett. 546 (2003) 154–158.
[6] C.M. West, H. van der Wel, E.A. Gaucher, Complex glycosylation of Skp1 in *Dictyostelium*: implications for the modification of other eukaryotic cytoplasmic and nuclear proteins, Glycobiology 12 (2002) 17R–27R.
[7] A. Varki, Metabolic radiolabeling of glycoconjugates, Methods Enzymol. 230 (1994) 16–32.
[8] A. Dell, H.R. Morris, Glycoprotein structure determination by mass spectrometry, Science 291 (2001) 2351–2356.
[9] P. Teng-umnuay, H.R. Morris, A. Dell, M. Panico, T. Paxton, C.M. West, The cytoplasmic F-box binding protein Skp1 contains a novel pentasaccharide linked to hydroxyproline in *Dictyostelium*, J. Biol. Chem. 273 (1998) 18242–18249.
[10] S. Sassi, M. Sweetinburgh, J. Erogul, P. Zhang, P. Teng-umnuay, C.M. West, Analysis of Skp1 glycosylation and nuclear enrichment in *Dictyostelium*, Glycobiology 11 (2001) 283–295.
[11] E. Kozarov, H. van der Wel, M. Field, M. Gritzali, R.D. Brown, C.M. West, Characterization of FP21, a cytosolic glycoprotein from *Dictyostelium*, J. Biol. Chem. 270 (1995) 3022–3033.
[12] B. Gonzalez-Yanes, J.M. Cicero , R.D. Brown Jr., C.M. West, Characterization of a cytosolic fucosylation pathway in *Dictyostelium*, J. Biol. Chem. 267 (1992) 9595–9605.
[13] R.J. Deshaies, SCF and cullin/RING H2-based ubiquitin ligases, Annu. Rev. Cell Dev. Biol. 15 (1999) 435–467.
[14] N. Zheng, B.A. Schulman, L. Song, J.J. Miller, P.D. Jeffrey, P. Wang, C. Chu, D.M. Koepp, S.J. Elledge, M. Pagano, R.C. Conaway, J.W. Conaway, J.W. Harper, N.P. Pavletich, Structure of the Cul1-Rbx1-Skp1-F box^{Skp2} SCF ubiquitin ligase complex, Nature 416 (2002) 703–709.
[15] K.B. Kaplan, A.A. Hyman, P.K. Sorger, Regulating the yeast kinetochore by ubiquitin-dependent degradation and Skp1p-mediated phosphorylation, Cell 91 (1997) 491–500.
[16] J.H. Seol, A. Shevchenko, A. Shevchenko, R.J. Deshaies, Skp1 forms multiple protein complexes, including RAVE, a regulator of V-ATPase assembly, Nat. Cell Biol. 3 (2001) 384–391.
[17] H. van der Wel, H.R. Morris, M. Panico, T. Paxton, A. Dell, J.M. Thomson, C.M. West, A non-Golgi α1,2-fucosyl-transferase that modifies Skp1 in the cytoplasm of *Dictyostelium*, J. Biol. Chem. 276 (2001) 33952–33963.
[18] M. Safran, W.G. Kaelin Jr., HIF hydroxylation and the mammalian oxygen-sensing pathway, J. Clin. Invest. 111 (2003) 779–783.
[19] L. Tan, J.F. Leykam, M.J. Kieliszewski, Glycosylation motifs that direct arabinogalactan addition to arabinogalactan-proteins, Plant Physiol. 132 (2003) 1362–1369.
[20] P. Teng-umnuay, H. van der Wel, C.M. West, Identification of a UDP-GlcNAc:Skp1-hydroxyproline GlcNAc-transferase in the cytoplasm of *Dictyostelium*, J. Biol. Chem. 274 (1999) 36392–36402.
[21] H. van der Wel, H.R. Morris, M. Panico, T. Paxton, A. Dell, L. Kaplan, C.M. West, Molecular cloning and expression of a UDP-GlcNAc:hydroxyproline polypeptide GlcNAc-transferase that modifies Skp1 in the cytoplasm of *Dictyostelium*, J. Biol. Chem. 277 (2002) 46328–46337.
[22] K.J. Colley, Golgi localization of glycosyltransferases: more questions than answers, Glycobiology 7 (1997) 1–13.
[23] J.O. Wrabl, N.V. Grishin, Homology between *O*-linked GlcNAc transferases and proteins of the glycogen phosphorylase superfamily, J. Mol. Biol. 314 (2001) 365–374.
[24] K.G. TenHagen, T.A. Fritz, L.A. Tabak, All in the family: the UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferases, Glycobiology 13 (2003) 1–16.
[25] B.S. Wojczyk, M.M. Stwora-Wojczyk, F.K. Hagen, B. Striepen, H.C. Hang, C.R. Bertozzi, D.S. Roos, S.L. Spitalnik, cDNA cloning and expression of UDP-*N*-acetyl-D-galactosamine:polypeptide *N*-acetylgalactosaminyltransferase T1 from *Toxoplasma gondii*, Mol. Biochem. Parasitol. 131 (2003) 93–107.
[26] P.M. Coutinho, E. Deleury, G.J. Davies, B. Henrissat, An evolving hierarchical family classification for glycosyltransferases, J. Mol. Biol. 328 (2003) 307–317.

[27] D. Kapitonov, R.K. Yu, Conserved domains of glycosyltransferases, Glycobiology 9 (1999) 961–978.
[28] C.M. West, Evolutionary and functional implications of the complex glycosylation of Skp1, a cytoplasmic/nuclear glycoprotein associated with polyubiquitination, Cell. Mol. Life Sci. 60 (2003) 229–240.
[29] F.K. Hagen, B. Hazes, R. Raffo, D. deSa, L.A. Tabak, Structure-function analysis of the UDP-*N*-acetyl-D-galactosamine:polypeptide *N*-acetylgalactosaminyltransferase, J. Biol. Chem. 273 (1999) 8268–8277.
[30] N.E. Zachara, N.H. Packer, M.D. Temple, M.B. Slade, D.R. Jardine, P. Karuso, C.J. Moss, B.C. Mabbutt, P.M.G. Curmi, K.L. Williams, A.A. Gooley, Recombinant prespore-specific antigen from *Dictyostelium discoideum* is a β-sheet glycoprotein with a spacer peptide modified by *O*-linked *N*-acetylglucosamine, Eur. J. Biochem. 238 (1996) 511–518.
[31] E. Jung, A.A. Gooley, N.H. Packer, P. Karuso, K.L. Williams, Rules for the addition of *O*-linked *N*-acetylglucosamine to secreted proteins in *Dictyostelium discoideum*, Eur. J. Biochem. 253 (1998) 517–524.
[32] C.M. West, P. Zhang, A.C. McGlynn, L. Kaplan, Outside–In signaling of cellulose synthesis by a spore coat protein in *Dictyostelium*, Euk. Cell 1 (2002) 281–292.
[33] G. Li, H. Alexander, N. Schneider, S. Alexander, Molecular basis for resistance to the anticancer drug cisplatin in *Dictyostelium*, Microbiology 146 (2000) 2219–2227.
[34] F. Wang, T. Metcalf, H. van der Wel, C.M. West, Initiation of mucin-type *O*-glycosylation in *Dictyostelium* is homologous to the corresponding step in animals and is important for spore coat function, J. Biol. Chem. 278 (2003) 51395–51407.
[35] J.O. Previato, M. Sola-Penna, O.A. Agrellos, C. Jones, T. Oeltmann, L.R. Travassos, L. Mendonca-Previato, Biosynthesis of *O*-*N*-acetylglucosamine-linked glycans in *Trypanosoma cruzi*. Characterization of the novel uridine diphospho-*N*-acetylglucosamine:polypeptide *N*-acetylglucosaminyltransferase-catalyzing formation of *N*-acetylglucosamine alpha1 → *O*-threonine, J. Biol. Chem. 273 (1998) 14982–14988.
[36] J.A. Morgado-Diaz, C.V. Nakamura, O.A. Agrellos, W.B. Dias, J.O. Previato, L. Mendonca-Previato, W. De Souza, Isolation and characterization of the Golgi complex of the protozoan *Trypanosoma cruzi*, Parasitology 123 (2001) 33–43.
[37] H.C. Hang, C. Yu, K.G. TenHagen, E. Tian, K.A. Winans, L.A. Bertozzi, C.R. Bertozzi, Identification of polypeptide N-acetyl-α-galactosyltransferase (ppGalNAcT) inhibitors from a uridine-based library, Chem. Biol. 11 (2004) 337–345.
[38] A.E. Stephenson, H. Wu, J. Novak, M. Tomana, K. Mintz, P. Fives-Taylor, The Fap1 fimbrial adhesin is a glycoprotein: antibodies specific for the glycan moiety block the adhesion of *Streptococcus parasanguis* in an in vitro tooth model, Mol. Microbiol. 43 (1992) 147–157.
[39] C. Levesque, C. Vadeboncoeur, F. Chandad, M. Frenette, *Streptococcus salivarius* fimbriae are composed of a glycoprotein containing a repeated motif assembled into a filamentous nondissociable structure, J. Bacteriol. 183 (2001) 2724–2732.
[40] R.S. Gupta, The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins, Int. Microbiol. 4 (2001) 187–202.
[41] T. Schwientek, E.P. Bennett, C. Flores, J. Thacker, M. Hollmann, C.A. Reis, J. Behrens, U. Mandel, B. Keck, M.A. Schafer, K. Haselmann, R. Zubarev, P. Roepstorff, J.M. Burchell, J. Taylor-Papadimitriou, M.A. Hollingsworth, H. Clausen, Functional conservation of subfamilies of putative UDP-*N*-acetylgalactosamine: polypeptide *N*-acetylgalactosaminyltransferases in *Drosophila*, *Caenorhabditis elegans*, and mammals, J. Biol. Chem. 277 (2002) 22623–22638.
[42] D.L. Swofford, PAUP*. Phylogenetic Anlysis Using Parsimony (* and Other Methods). Version 4, Sinauer Associates, Sunderland, MA, 2000.
[43] M. Wacker, D. Linton, P.G. Hitchen, M. Nita-Lazar, S.M. Haslam, S.J. North, M. Panico, H.R. Morris, A. Dell, B.W. Wren, M. Aebi, N-linked glycosylation in *Campylobacter jejuni* and its functional transfer into *E. coli*, Science 298 (2002) 1790–1793.
[44] J. Roelofs, P.J. Van Haastert, Deducing the origin of soluble adenylyl cyclase, a gene lost in multiple lineages, Mol. Biol. Evol. 19 (2002) 2239–2246.
[45] J.R. Roper, M.A. Ferguson, Cloning and characterisation of the UDP-glucose 4′-epimerase of *Trypanosoma cruzi*, Mol. Biochem. Parasitol. 132 (2003) 47–53.
[46] D.M.Coltart, A.K.Royyuru, L.J.Williams, P.W.Glunz, D.Sames, Kuduk, Kuduk, S.D. Kuduk, J.B. Schwarz, X.T. Chen, S.J. Danishefsky, D.H. Live, Principles of mucin architecture: structural studies on synthetic glycopeptides bearing clustered mono-, di-, tri-, and hexasaccharide glycodomains, J. Am. Chem. Soc. 124 (2002) 9833–9844.
[47] C.M. West, T. Scott-Ward, P. Teng-umnuay, H. van der Wel, E. Kozarov, A. Huynh, Purification and characterization of an (1,2-L-fucosyltransferase, which modifies the cytosolic protein FP21, from the cytosol of *Dictyostelium*, J. Biol. Chem. 271 (1996) 12024–12035.
[48] H. van der Wel, S.Z. Fisher, C.M. West, A bifunctional diglycosyltransferase forms the Fucα1,2Galβ,3-disaccharide on Skp1 in the cytoplasm of *Dictyostelium*, J. Biol. Chem. 277 (2002) 46527–46534.
[49] P.L. de Angelis, Microbial glycosaminoglycan glycosyltransferases, Glycobiology 12 (2002) 9R–16R.
[50] C. Whitfield, I.S. Roberts, Structure, assembly and regulation of expression of capsules in *Escherichia coli*, Mol. Microbiol. 31 (1999) 1307–1319.
[51] C. Ketcham, F. Wang, S.Z. Fisher, A. Ercan, H. van der Wel, R. D. Locke, S. ud-Doulah.k, K.L. Matta, C.M. West, Specificity of a UDP-galactose:fucoside α1,3galactosyltransferase that modifies Skp1 in the cytoplasm of *Dictyostelium*, J. Biol. Chem., in press.
[52] J. Myllyharju, Prolyl 4-hydroxylases, the key enzymes of collagen biosynthesis, Matrix Biol. 22 (2003) 15–24.
[53] A.V. Kuznetsova, J. Meller, P.O. Schnell, J.A. Nash, M.L. Ignacak, Y. Sanchez, J.W. Conaway, R.C. Conaway, M.F. Czyzyk-Krzeska, von Hippel–Lindau protein binds hyperphosphorylated large subunit of RNA polymerase II through a proline hydroxylation motif and targets it for ubiquitination, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 2706–2711.
[54] J.H. Min, H. Yang, M. Ivan, F. Gertler, W.G. Kaelin, N.P. Pavletich, Structure of an HIF-1alpha–pVHL complex: hydroxyproline recognition in signaling, Science 296 (2002) 1886–1889.
[55] R.K. Bruick, S.L. McKnight, A conserved family of prolyl-4-hydroxylases that modify HIF, Science 294 (2001) 1337–1340.
[56] E. Metzen, U. Berchner-Pfannschmidt, P. Stengel, J.H. Marxsen, I. Stolze, M. Klinger, W.Q. Huang, C. Wotlaw, T. Hellwig-Burgel, W. Jelkmann, H. Acker, J. Fandrey, Intracellular localisation of human HIF-1a hydroxylases: implications for oxygen sensing, J. Cell. Sci. 116 (2003) 1319–1326.
[57] D. Lando, J.J. Gorman, M.L. Whitelaw, D.J. Peet, Oxygen-dependent regulation of hypoxia-inducible factors by prolyl and asparaginyl hydroxylation, Eur. J. Biochem. 270 (2003) 781–790.
[58] R. Hieta, L. Kukkola, P. Permi, P. Pirila, K.I. Kivirikko, I. Kilpelainen, J. Myllyharju, The peptide substrate-binding domain of human collagen prolyl 4-hydroxylases. Backbone assignments, secondary structure, and binding of proline-rich peptides, J. Biol. Chem. 278 (2003) 34966–34974.
[59] J. Huang, Q. Zhao, S.M. Mooney, F.S. Lee, Sequence determinants in hypoxia-inducible factor-1α for 26 hydroxylation by the prolyl hydroxylases PHD1, PHD2, and PHD3, J. Biol. Chem. 277 (2002) 39792–39800.
[60] T. Pereira, X. Zheng, J.L. Ruas, K. Tanimoto, L. Poellinger, Identification of residues critical for regulation of protein stability and the transactivation function of the hypoxia-inducible factor-1a by the von Hippel–Lindau tumor suppressor gene product, J. Biol. Chem. 278 (2003) 6816–6823.

[61] R. Hieta, J. Myllyharju, Cloning and characterization of a low molecular weight prolyl 4-hydroxylase from *Arabidopsis thaliana*. Effective hydroxylation of proline-rich, collagen-like, and hypoxia-inducible transcription factor alpha-like peptides, J. Biol. Chem. 277 (2002) 23965–23971.

[62] L. Kreppel, A.R. Kimmel, Genomic database resources for *Dictyostelium discoideum*, Nucleic Acids Res. 30 (2002) 84–86.

[63] J.E. Nixon, A. Wang, J. Field, H.G. Morrison, A.G. McArthur, M.L. Sogin, B.J. Loftus, J. Samuelson, Evidence for lateral transfer of genes encoding ferredoxins, nitroreductases, NADH oxidase, and alcohol dehydrogenase 3 from anaerobic prokaryotes to *Giardia lamblia* and *Entamoeba histolytica*, Euk. Cell 1 (2002) 181–190.

[64] M. Eriksson, J. Myllyharju, H. Tu, M. Hellman, K.I. Kivirikko, Evidence for 4-hydroxyproline in viral proteins. Characterization of a viral prolyl 4-hydroxylase and its peptide substrates, J. Biol. Chem. 274 (1999) 22131–22134.

[65] C. Wang, H. Luosujarvi, J. Heikkinen, M. Risteli, L. Uitto, R. Myllyla, The third activity for lysyl hydroxylase 3: galactosylation of hydroxylysyl residues in collagens in vitro, Matrix Biol. 21 (2002) 559–566.

[66] F.T. Liu, R.J. Patterson, J.L. Wang, Intracellular functions of galectins, Biochim. Biophys. Acta 1572 (2002) 263–273.

[67] C. He, J. Malsam, H. Ho, C. Chalouni, C.M. West, E. Ulla, D. Toomre, G. Warren, Golgi duplication in *Trypanosoma brucei*, J. Cell. Biol. (2004) in press.