Article

Genome-Wide Analysis of Whole Human Glycoside Hydrolases by Data-Driven Analysis *in Silico*

Takahiro Nakamura[1]†, Muhamad Fahmi[1]†, Jun Tanaka[1], Kaito Seki[1], Yukihiro Kubota[2] and Masahiro Ito[1,2*]

[1] Advanced Life Sciences Program, Graduate School of Life Sciences, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan

[2] Department of Bioinformatics, College of Life Sciences, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan

* Correspondence: maito@sk.ritsumei.ac.jp; Tel.: +81-77-561-5301 (M.I.)

† These authors have equal contribution to this work.

Abstract: Glycans are involved in various metabolic processes via the functions of glycosyltransferases and glycoside hydrolases. Analysing the evolution of these enzymes is essential for improving the understanding of glycan metabolism and function. Based on our previous study of glycosyltransferases, we performed a genome-wide analysis of whole human glycoside hydrolases using the UniProt, BRENDA, CAZy, and KEGG databases. Using cluster analysis, 319 human glycoside hydrolases were classified into four clusters based on their similarity to enzymes conserved in chordates or metazoans (Class 1), metazoans (Class 2), metazoans and plants (Class 3), and eukaryotes (Class 4). The eukaryote and metazoan clusters included *N*- and *O*-glycoside hydrolases, respectively. The significant abundance of disordered regions within the most conserved cluster indicated a role for disordered regions in the evolution of glycoside hydrolases. These results suggest that the biological diversity of multicellular organisms is related to the acquisition of *N*- and *O*-linked glycans.

Keywords: glycoside hydrolase; glycan; phylogenetic profiling

1. Introduction

Glycans are present in various biological molecules including glycoproteins, glycolipids, and proteoglycans and in more than half of all human proteins. Glycans are widely distributed in eukaryotes, bacteria, and archaea [1] and have similar structures in different organisms, including yeasts, plants, insects, and chordates [2]. The high conservation of glycans in different species is biologically meaningful [3]. Human glycans can be classified into four major categories: *O*-linked (mucin-type) glycans, *N*-linked glycans, glycosphingolipids, and glycosaminoglycans. These glycans play important roles *in vivo*, including in cell membrane/extracellular matrix (ECM) construction, cell adhesion, protein stabilisation, and transmission of information [4,5]. Abnormalities in glycan structures are closely related to certain diseases such as neurological disorders, cancer metastasis, Alzheimer’s disease, and diabetes [6]. The diversity of glycan functions depends on the diversity of glycan structures, i.e. the combination of monosaccharides constituting a glycan, differences in binding sites, and differences in branching modes. However, the mechanisms mediating the acquisition of various glycan categories, balance between biosynthesis and degradation, and essential biological significance of glycans are unclear.

The biosynthesis and degradation of various glycan structures are mainly catalysed by glycosyltransferases and glycoside hydrolases, respectively. Glycosyltransferases function to regulate the elongation of glycans, and variations in glycosyltransferases result in diverse substrate specificities such as the type of sugar to be transferred and specific binding position of the sugar. For genome-wide evolutionary analysis of glycosyltransferases, we previously performed lineage profile analysis of 173 human glycosyltransferases [3]. The results indicate that human glycosyltransferases can be roughly divided into four categories based on their similarity to enzymes conserved in deuterostomes, metazoans, eukaryotes, and eukaryotes, bacteria, and archaea. Two glycosyltransferase groups, synthesise *O*- and *N*-linked glycans, are present in the Golgi apparatus in deuterostomes and metazoans and in the endoplasmic reticulum of eukaryotes. Thus, we found that the localisation and function of glycosyltransferases conserved among deuterostomes, metazoans, and other eukaryotes were distinctly different. Furthermore, our findings suggested that *N*-linked glycan structures existed before *O*-linked glycans during the evolution of these molecules in humans [3].

Glycoside hydrolases have substrate specificities similar to those of glycosyltransferases; however, many of glycosyltransferases have a strict substrate specificity, whereas glycoside hydrolases show a looser substrate specificity. Glycoside hydrolases function to cleave glycosidic bonds in glycans, and many are in lysosomes [7]. In addition to glycan degradation in lysosomes, glycoside hydrolases are closely associated with *in vivo* functions, such as the quality control of proteins by the processing of high-mannose-type (*N*-linked-type) glycans and remodelling of ECM comprising *O*-linked glycans and glycosaminoglycans [8]. Notably, glycoside hydrolases have been shown to play roles in lysosomal storage diseases. The lysosome is an intracellular organelle that decomposes waste products via the functions of various hydrolytic enzymes. Lysosomal storage diseases are caused by the accumulation of undegraded substances because of genetic abnormalities affecting the expression of glycoside hydrolases in lysosomes [9]. Symptoms of lysosomal storage diseases are diverse and severe. Additionally, both the synthesis of glycans and decomposition of glycans are involved in biological functions; however, the detailed functions of sugar hydrolases in lysosomes have not been determined. Particularly, the roles of glycoside hydrolases for *O*-linked glycans are unclear in lysosomal storage diseases [8].

Protein evolution is driven by function, which critically depends on the structure. This is supported by comparison of evolutionary rates between ordered and disordered structured proteins. Disordered regions commonly evolve faster than ordered structures [10–13] because of differences in the relative constraints that maintain folding interactions [14]. However, there are exceptions to this rule. For instance, specific functional binding and modification regions of a disordered structure are constrained [13,15], thus introducing heterogeneity into evolutionary rates.

In this study, we evaluated glycan degradation by performing a genome-wide analysis of 319 human glycoside hydrolases. By comparing the results of analysis of glycosyltransferases [3] and their protein structures, we clarified the acquisition process of each glycan category during evolution.

2. Results

2.1. Human glycoside hydrolase dataset

In this study, 319 human glycoside hydrolases (Table S2) were retrieved from the UniProt [16], CAZy [17], and BRENDA [18] databases. The dataset was verified using the Gene Ontology (GO) term GO:0016798. Of the 319 human glycoside hydrolases in the dataset, 251 overlapped with glycoside hydrolases in the GO database (Table S1); among the 251 genes involved in the GO, 178 genes overlapped with glycoside hydrolases in the InterPro database [19]. Most data extracted using GO were related to nucleic acid-related glycoside hydrolases.

2.2 Human glycoside hydrolases belong to four evolutionary classes

The 319 human glycoside hydrolases in the dataset were classified into four clusters by phylogenetic profiling (Figure 1, Tables S3 and S4) and cluster analysis. The four clusters included enzymes with orthologs primarily conserved in chordates or metazoans (class 1), metazoans (class 2), metazoans and plants (class 3), and eukaryotes (class 4). The molluscs *Octopus bimaculoides*, *Crassostrea gigas*, and *Lottia gigantea* and cnidarian *Nematostella vectensis* were classified in the same cluster as the Chordata. Additionally, two deuterostome taxa, i.e. Choanoflagellatea and *Dictyostelium*, showed greater conservation relative to all fungi (Figure 1).

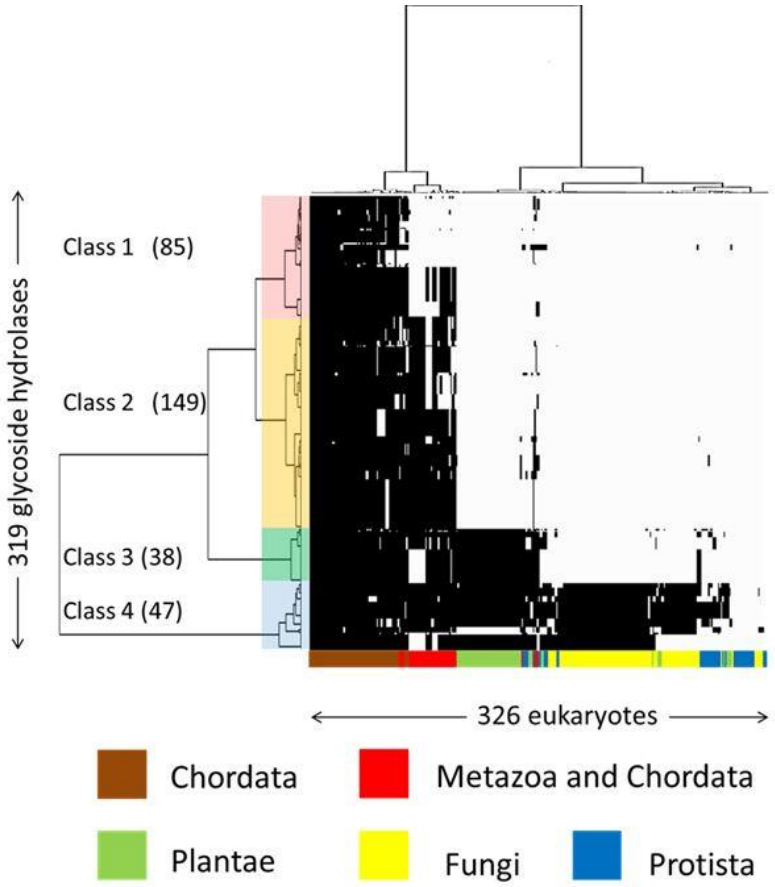

Figure 1. Phylogenetic profiling of human glycoside hydrolases. The X-axis shows 326 organisms (Table S3) that underwent genome sequencing, and the Y-axis shows the 319 human glycoside hydrolases (Table S2). Based on the phylogenetic tree, human glycoside hydrolases were classified into four characteristic clusters, defined as classes 1–4, which included 85, 149, 38, and 47 human glycoside hydrolases, respectively. The black regions indicate the presence of human glycoside hydrolase orthologs in specific groups of organisms, shown in different colours on the X-axis. In class 1, chordates or metazoans are included. In the metazoans in class 1, some metazoan animals such as some fly species were excluded.

2.3 Functions of human glycoside hydrolases differ among classes

Next, we characterised the types of glycans degraded by each class of glycoside hydrolases. The results showed that classes 1 and 2 (Figures 2a, b) contained glycoside hydrolases such as hyaluronidase, lysozyme, and chitinase which degraded glycosaminoglycans, and glycoside hydrolases such as glucosyl ceramidase and sialidase that degraded glycolipids. Class 4 contained glycoside hydrolases that only degraded *N*-linked glycans (Figure 2d). Our analysis showed that glycoside hydrolases degrading galactose and *N*-acetylgalactosamine, which are commonly found in *O*-linked glycans, were unevenly distributed in classes 1 and 2 (Figure 2). These results suggest that *O*-linked glycans were obtained after acquisition of *N*-linked glycans in the evolution of glycosyl hydrolases (GHs) as shown in the analysis of glycosyltransferases (GTs) [3].

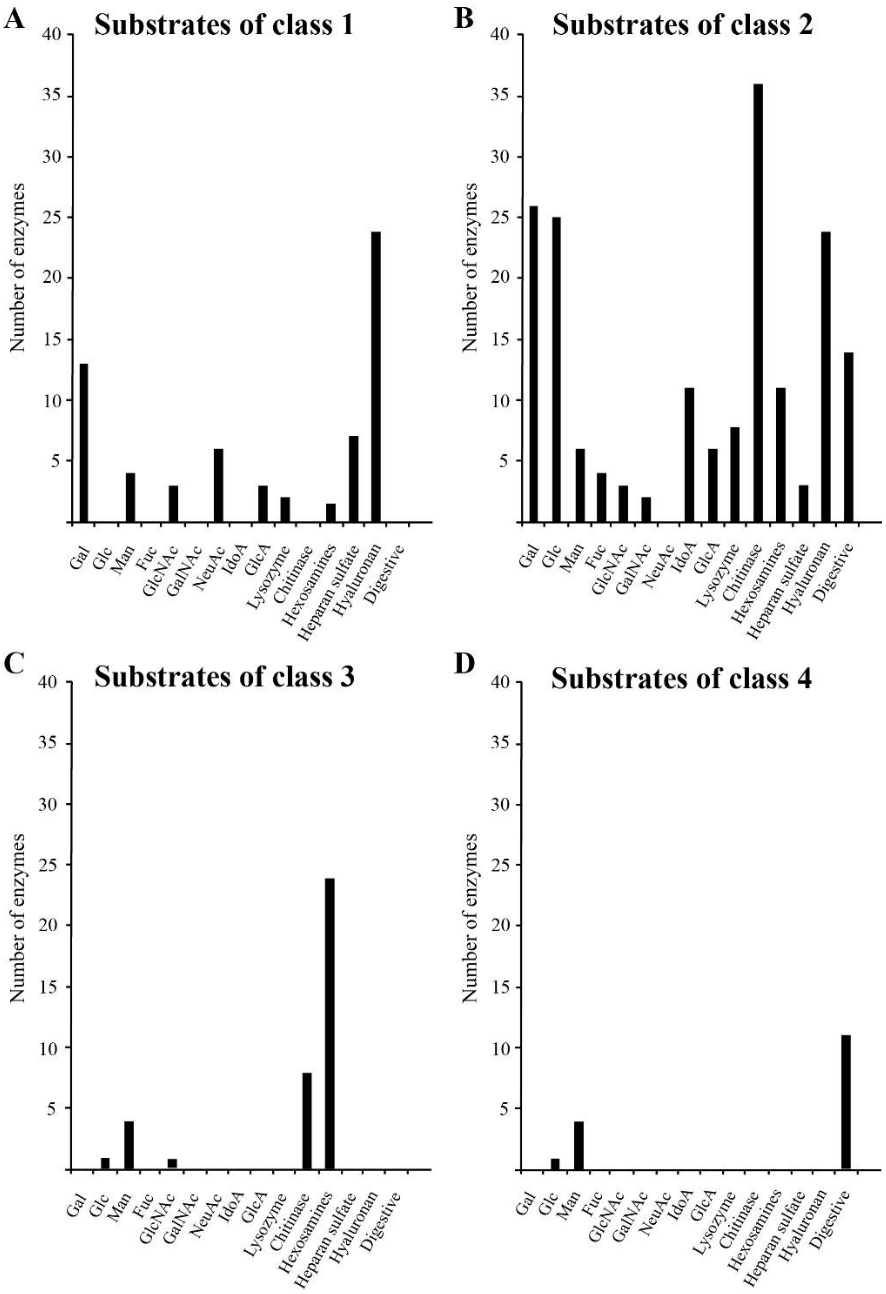

Figure 2. Degradation substrates of human glycoside hydrolases from each class. The X- and Y-axes show the degradation substrates and number of human glycoside hydrolases, respectively. Degradation substrates are shown for class 1 (**a**), class 2 (**b**), class 3 (**c**), and class 4 (**d**).

2.4 Comparison of decomposition substrates among classes of glycoside hydrolases

Next, we investigated other differences among the classes of sugar hydrolases. Substrates and products of human glycoside hydrolases were referenced according to the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database [20], and the relationships between the degradation of glycan structures and glycoside hydrolases were mapped (Figure 3). Human glycoside hydrolases of the high-mannose-type *N*-linked glycans, particularly those with processing function, were widely conserved in eukaryotes (Figure 3a). Glycoside hydrolases classified into class 2 or 3 were involved in the degradation of complex *N*-linked glycans. This result suggests changes in substrates from complex glycans to functional substances in human glycoside hydrolases that originated from multicellular organisms (Figure 3b). Glycans of *N*-linked glycoproteins and glycolipids were degraded by specific glycoside hydrolases at the nonreducing end (Figure 3c). Many glycoside hydrolases had exo-type functions allowing for the decomposition of monosaccharides at nonreducing ends. In contrast, glycoside hydrolases were classified into class 1 had endo-type functions and acted to decompose the interior region of carbohydrate chains. An endo-type glycoside hydrolase was shown to enhance the efficiency of endoplasmic reticulum-associated degradation (ERAD) of folding-deficient proteins in the protein quality control process [21]. These findings suggest that the acquisition of a mechanism involved in alleviating endoplasmic reticulum stress contributed to chordate evolution. The structure of human glycosaminoglycans was largely degraded by glycoside hydrolases obtained from chordates, except keratan sulphate, which was decomposed by glycoside hydrolases from classes 2 and 3 (Figure 3d).

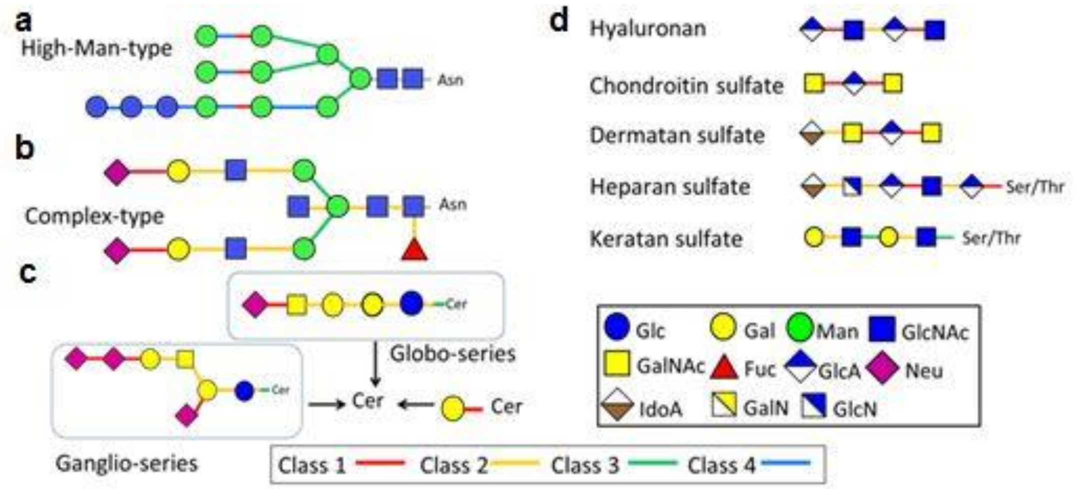

Figure 3. Mapping of evolutionary information to metabolic pathways. Different classes are shown by different colours, as indicated. The links connecting human glycoside hydrolase classes and degradation substrates show bonds degraded by each class of enzymes. The single glycan is shown based on the Consortium for Functional Glycomics symbol. The figure shows high-mannose-type *N*-glycans (**a**), complex-type *N*-glycans (**b**), glycolipids (**c**), and glycosaminoglycans (**d**).

2.5 Identification of glycoside hydrolases important for the evolution to mammals

Molecular phylogenetic analysis was conducted to investigate how human glycoside hydrolases evolved in the process of evolution from chordates to mammals (Figure 4). Sialidase, which is involved in neuronal and muscle differentiation, and lysozyme, which plays an important role in mammalian embryos, were acquired before the emergence of cartilaginous fish and of the common ancestor of birds and mammals, respectively. Glucosylceramidase, a class 2 glycolipid-metabolising enzyme, was conserved in most Chordata but was lost during evolution in some chordates including *Gallus* spp. and *Xenopus laevis*. Accordingly, we hypothesised that sialidase, lysozyme, and glucosylceramidase were necessary for the evolution to mammals.

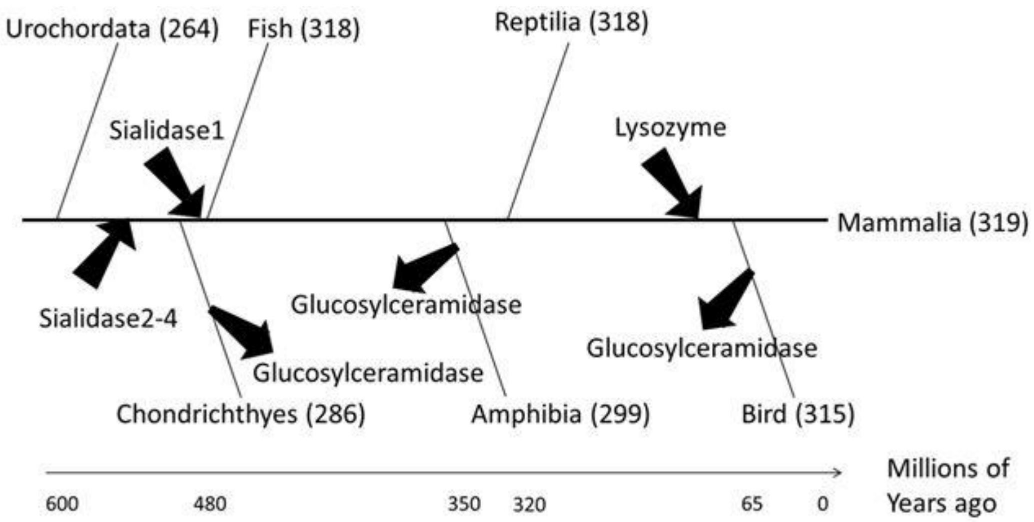

Figure 4. Evolution of human glycoside hydrolases. The time divergence analysis refers to Rowe's tree of life [31]. The X-axis shows the time of evolution, and arrows indicate the acquisition or loss of human glycoside hydrolases.

2.6 Evolution of glycosyltransferases and glycoside hydrolases

To compare the acquisition processes of human glycosyltransferases and human glycoside hydrolases (Table S5), phylogenetic profiling analysis of human sugar hydrolases and human glycosyltransferases was performed (Figure 6a, Table S6). The results showed that human glycosyltransferases and human glycoside hydrolases were classified into four characteristic clusters, defined as classes I–IV, based on their similarity to enzymes conserved in chordates or metazoans (Class I), metazoans (Class II), metazoans and plants (Class III), and eukaryotes (Class IV). In this analysis, *Strongylocentrotus* were classified together with Class I, whereas enzymes of the chordates *Ciona intestinalis* and *Branchiostoma floridae* were classified together with Class II. However, degradation enzymes for the core structure of *N*-linked glycans had a lower degree of conservation in other organisms than that of human glycosyltransferases. Additionally, β-1,4-mannosyl-glycoprotein 4-β-*N*-acetylglucosaminyltransferase (MGAT3), a human glycosyltransferase of a bisecting GlcNAc, was found to be conserved in the complex type (*N*-glycan). Because few studies have evaluated human glycoside hydrolases, it was difficult to map *O*-glycans to a metabolic pathway; however, this *O*-glycan hydrolase was classified in the same cluster as a human glycosyltransferase. Additionally, when we focused on sialic acid modifications, which were shown to be required for protein stabilization and neuronal differentiation, the human sialic acid glycoside hydrolase and human sialic acid glycosyltransferase were found to have be acquired in the same period during evolution (Figure 5b, c).

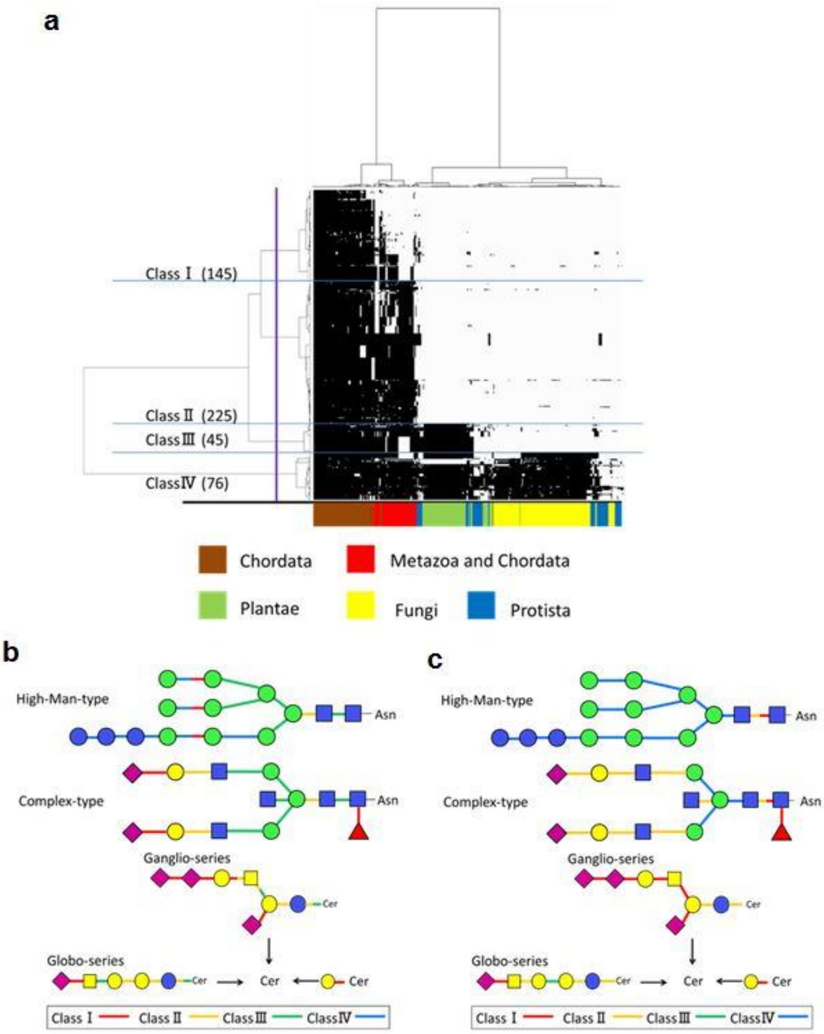

Figure 5. Phylogenetic profiling of human glycoside hydrolases and human glycosyltransferases. In (**a**), the X-axis indicates 326 organisms that underwent genome sequencing (Table S3), and the Y-axis indicates 319 human glycoside hydrolases and 172 glycosyltransferases (Tables S5 and S6). Based on the phylogenetic tree, the enzymes were classified into four characteristic clusters, defined as classes I–IV, which included 145, 225, 45, and 76 enzymes, respectively. Classes are indicated by different colours. Links between human glycosyltransferases and glycoside hydrolases classes (**a**) and degradation (**b**) or synthesis (**c**) substrates are shown. The single glycan is shown based on the Consortium for Functional Glycomics symbol. Addition or removal of Neu (magenta triangle) occurs during sialic acid modification.

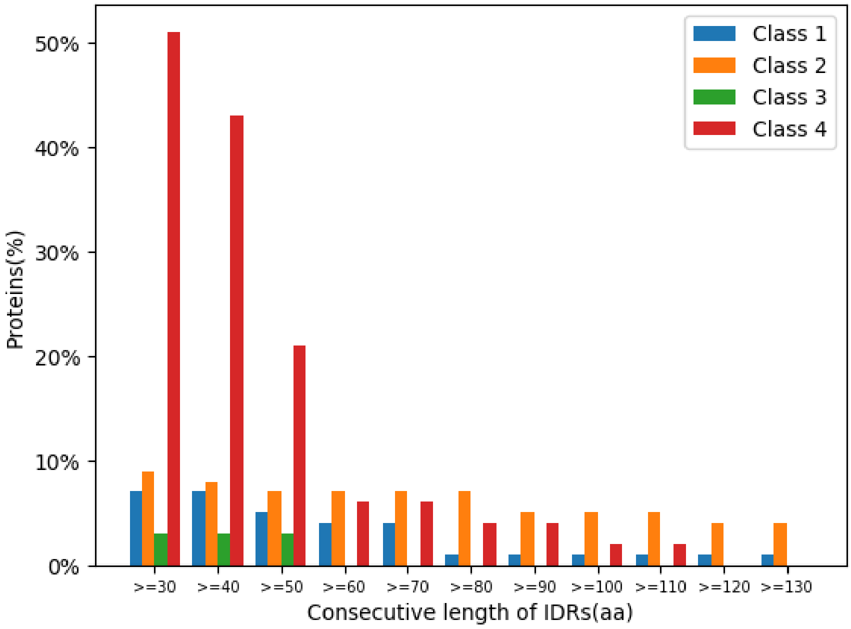

Figure 6. Percentages of glycoside hydrolases with specific consecutive lengths of IDRs in each class. Classes are indicated by different colours.

2.7 Intrinsically disordered regions of glycoside hydrolases

The ratios of the lengths of intrinsically disordered regions (IDRs) to the total amino acid protein sequences [22] were analysed for the 319 human glycoside hydrolases (Figures. 6 and 7). The presence of continuous stretches of IDRs was predominant within class 4, which showed significantly higher ratios than those in the other classes (Figure 6). More than 50% of class 4 members had a continuous stretch of IDR of more than or equal to 30 amino acids, whereas only 10% or less members of the other classes had this continuous stretch of IDR. These results are consistent with the distribution of protein lengths, which were commonly longer among class 4 members than among those from the other classes (Figure 7).

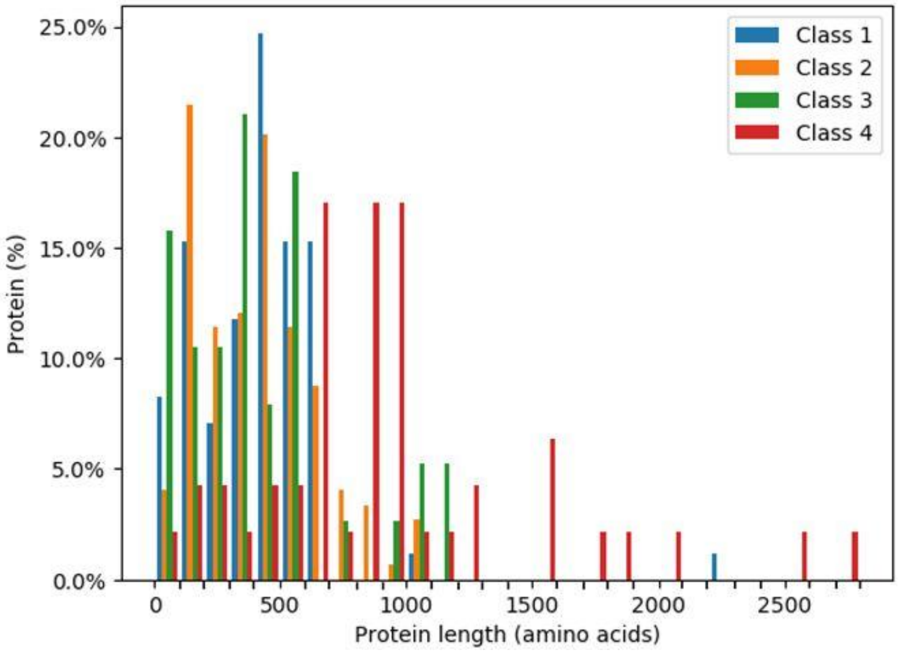

Figure 7. Percentages of glycoside hydrolases with the indicated lengths in each class. Classes are indicated by different colours.

3. Discussion

In this study, we performed phylogenetic analysis of human glycoside hydrolases to evaluate the evolution of glycan-mediated biological systems. We found that 319 human glycoside hydrolases were classified into four clusters, including enzymes with orthologs in chordates, metazoans, metazoans and plants, and eukaryotes. We also compared the dataset in this study to enzymes annotated by GO and found that 78.7% enzymes overlapped. Thus, most enzymes in the dataset of this study have already been annotated by using GO. Based on these findings, we propose that the acquisition of each human-type glycoside hydrolase gene was associated with the development of an intracellular protein-producing system and extracellular glycan-dependent biological interactions, as well as with the development and diversification of neuronal and neuromuscular functions. Consistent with data from a previous study showing that *N*-linked glycosyltransferases were widely conserved from the ancestral species of eukaryotes [3], the acquisition of high-mannose-type *N*-glycan-degrading enzymes occurred from ancestral species of eukaryotes. Among these enzymes, endo-β-*N*-acetylglucosaminidase and α-mannosidase 2C1 are localised in the endoplasmic reticulum, with the ERAD machinery facilitating accurate quality control of glycosylated proteins [23]. Similarly, the acquisition of high-mannose-type *N*-glycan-degrading enzymes was closely correlated with lectin-mediated glycoprotein folding [24]. Thus, precise regulation of *N*-glycan synthesis and degradation may play a central role in ensuring the integrity of *N*-glycan-mediated biological processes in eukaryotes. Our results showed that human-type *N*-glycan-degrading enzymes and the intracellular ERAD-related quality control of the protein-producing system were conserved throughout eukaryotes.

During the evolution of metazoans, polysaccharide-degrading enzymes such as lysozyme and chitinase, glycosaminoglycan-degrading enzymes, and hyaluronidase were acquired. Because these molecules are essential in the defence against bacterial infections, as well as for fertilisation and ECM remodelling [25–27], acquisition of these degrading enzymes may play important roles in regulating glycan-mediated biological functions. Among the glycosaminoglycan-degrading enzymes, keratan sulphate-processing enzymes are involved in many biological processes, whereas other degrading enzymes such as hyaluronidases, chondroitinases, heparitinases, and dermatan sulphate-degrading enzymes are mainly involved in neuronal functions [28]. Thus, during the evolutionary development of neuronal tissues, regulation of *O*-glycan modifications by *O*-glycan-degrading hydrolases may have played important roles in both plasma membrane-mediated and ECM-dependent biological functions. In terms of glycan degradation of *N*-glycans and glycolipids, ancestrally acquired human glycoside hydrolases can show degradation activity for the nonreducing end, whereas the sialic acid-degrading enzyme sialidase is essential for degrading the reducing end. Thus, complex-type *N*-glycans and glycolipids may have evolved by the addition of new sugars at the nonreducing end of ancestrally acquired glycans in multicellular organisms.

During evolution to chordates, an endo-α-mannosidase, MANEMA, was acquired. As described above, most exo-type mannosidases were acquired from ancestral eukaryotes, and the acquisition of the endo-type mannosidase MANEMA conferred organisms with the ability to efficiently degrade misfolded proteins. During evolution to chordates, genomic gains of sialidase genes occurred twice before the ancestral chordates evolved into teleosts. Because sialic acid-mediated modification of proteins is essential for muscle, neuronal, and lysosomal functions [28], the acquisition of sialidases may have been essential for the development of neuronal and neuromuscular structures, and lysosome-mediated protein degradation systems during evolution to chordates.

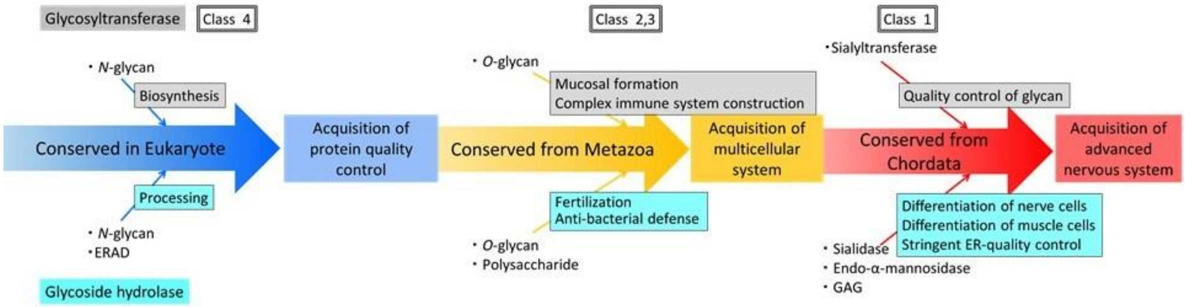

Figure 8. Schematic representation of a model for acquisition processes of human glycosyltransferases and human glycoside hydrolases using the results of a previous study [3] and this study. Blue arrows indicate enzymes that are widely conserved in eukaryotes; yellow arrows indicate enzymes that are conserved in metazoans; and red arrows indicate enzymes acquired from chordates.

In Rowe’s phylogenetic tree (Figure 8), sialidases and lysozyme were acquired during the evolution of mammals. Because sialidases regulate higher cerebral functions, the acquisition of these enzymes may have yielded more highly organised neuronal and muscular structures, facilitating the evolution of neuromuscular development. Similarly, the acquisition of lysozyme by ancestral species of mammals may have facilitated the development of the viviparous system in these organisms.

Although glucosylceramidase genes are conserved both in ancestral chordates and mammals, these genes disappeared during the diversification of Chondrichthyes, amphibians, and birds. These results indicate that glucosylceramidases are essential enzymes regulating the mammalian-specific functions of glycolipids. Alternatively, glucosylceramidases may not have been essential but were continuously maintained during the evolution of mammals. Because glucosylceramidases are highly regulated in higher vertebrates, glucosylceramidase activity may have been essential for nervous system development in mammals. Further studies are required to confirm these hypotheses.

Because most high-mannose-type human *N*-linked glycosyltransferases and *N*-glycoside hydrolases co-evolved in eukaryotes, high-mannose-type human *N*-glycan-dependent ERAD is thought to be essential for the precise regulation of *N*-glycan-mediated biological processes. Similarly, both glycosyltransferases and glycoside hydrolases for glycolipids and *O*-glycans were acquired at nearly the same time and co-evolved together. Thus, the development and diversity of glycolipid- and *O*-glycan-mediated biological systems were likely essential for multiple functions, including formation of the mucous membrane system and highly organised immune system, in the evolution to metazoans and vertebrates. Because complex-type bisecting GlcNAcs inhibit elongation of the β-1,6-GlcNAc branch at the nonreducing end of the core mannose of an *N*-glycan to stabilise the structure of the glycan, we focused on the timing of the acquisition of bisecting GlcNAc hydrolase and transferase. In contrast to bisecting GlcNAc hydrolases, which were acquired from more distant ancestral species, the complex-type bisecting GlcNAc transferase MGAT3 was acquired later and is conserved in most metazoans. Because complex-type bisecting GlcNAcs stabilise various biological functions including the E-cadherin-dependent cell adhesion system, the acquisition of bisecting GlcNAc elongation enzymes may have been involved in the evolution of metazoans [29].

However, the best approach to the direct evolution of these glycoside hydrolases remains unclear. Previously, we suggested that the evolutionary origin and functional acquisition of proteins are closely related to their IDRs [22]. Our results showed that the most conserved class also contained the greatest number of consecutive stretches of IDRs. Additionally, class 4 proteins commonly contain *N*-glycan-degrading enzymes and intracellular ERAD-related quality control proteins, such as ER degradation-enhancing mannosidase-like proteins (EDEMs), which are ER-resident members of the glycoside hydrolase 47 family, recruiting terminally misfolded polypeptides present in the ER lumen to the downstream ERAD pathway [30, 31]. In this study, all EDEMs, including EDEM1–3, were predicted to have disordered regions. The presence of disordered regions at the N-terminus of EDEM1 has been reported previously based on modelling and prediction studies. These regions have been shown to be important for recognising glycosylated and non-glycosylated misfolded proteins, even when the carbohydrate-binding domain is highly impaired [30]. Long consecutively disordered residues (>30) may function as entropic chains or can be involved in interactions using combinations of recognition motifs or domains [32]. We previously reported that residues within disordered regions that function as entropic chains evolve quickly, whereas those involved in protein–protein interactions tend to be constrained [13]. Thus, it may be relevant for some ancient glycoside hydrolases to harbour long stretches of disordered regions because the conformational plasticity of these regions enables the recognition of or binding to multiple partners, which is beneficial for identifying misfolded proteins.

Several mechanisms may shape the evolution of GH. Despite gene duplication, acquisition of genes may occur through other processes. Some genes may be acquired *de novo* from a stretch of non-coding DNA. The acquisition of this gene may coincide with environmental conditions such as codfish antifreeze glycoprotein genes that have evolved *de novo* from non-coding DNA in the cooling time of its habitat 13–18 million years ago [33]. Another possible mechanism is horizontal gene transfer which involves the movement of transposable elements between different species; this mechanism is well-known in prokaryotes and unicellular eukaryotes and remains controversial and less established in higher organisms [33]. However, several studies have exemplified this case clearly in a complex organism such as GH genes that are found nearly exclusively and to the largest extent in western corn rootworm (*Diabrotica virgifera virgifera*) among insects and the presence of Bovine-B (BovB) retrotransposons in mammals [34, 35]. In contrast, by utilizing symbiotic relationships such as gastrointestinal tract and microbiome, the acquisition of new genes or GH may not necessary to gain a function. In this case, some bacteria in the human gastrointestinal tract utilize their GH to cleave glycans that humans are unable to process; for instance, *Bifidobacteria longum* biovar *infantis* process oligosaccharides in milk that are not digestible by human infants [36]. The acquisition of GH by horizontal gene transfer from the microbiome also appears possible, but requires further analysis.

4. Materials and Methods

4.1 Human glycoside hydrolase dataset

The glycoside hydrolase sequence data were obtained from UniProt (release 2017_03) [16] using the following queries: "glycoside hydrolase" AND "organism: human". To confirm the annotation of each retrieved sequence as glycoside hydrolase, we extracted all UniProt IDs within the glycoside hydrolase category (EC3.2) from the CAZy [17] and BRENDA [18] databases and confirmed the presence of the UniProt ID for each retrieved sequence in the CAZy [17] and BRENDA [18] databases; unannotated sequences in any of these databases were removed. This was an alternative method used to obtain more data on human glycoside hydrolase sequences than would be obtained by using GO [37] and InterPro [19] using InterPro entry glycoside hydrolase superfamily (IPR017853), and was the easiest way to obtain human glycoside hydrolases with UniProt IDs in CAZy and BRENDA. In addition, we verified our data with glycoside hydrolases obtained using GO, 553 glycoside hydrolases that have been annotated as GO: 0016798, Taxon: *Homo sapiens* were isolated and compared to the dataset. Further, to analyse the evolution of glycoside hydrolases, we categorised these enzymes based on their substrates and products into four categories including *O*-glycans (mucins), *N*-glycans (high-mannose type, complex type), glycolipids, and glycosaminoglycans based on the metabolic map in the KEGG database [20].

4.2 Phylogenetic profiling and cluster analyses

Phylogenetic profiles were generated for 326 genome-wide eukaryotic sequences using KEGG OC default parameters in the KEGG database and extracted human glycoside hydrolase data as queries. Human glycoside hydrolase conservation in eukaryotes was examined using a BLAST search (E-value: 10−3; NIH). A bit score of 1 was assigned if orthologs of the protein of interest were present in the other genome; otherwise, a bit score of 0 was assigned. Proteins with similar bit patterns were expected to have similar interactions and functions. Further, using the bit pattern as an input, cluster analysis of the 319 human GHs and 326 eukaryotes from KEGG OC were performed using Ward's method [38] based on the Manhattan distance. Computational and cluster analyses were performed using Ruby and R programming languages.

4.3 Molecular phylogenetic analysis

A phylogenetic tree of glycoside hydrolases was manually constructed, and a model for the time divergence of chordates to mammals during evolution was presented as described by Rowe [39].

4.4 Protein IDR analysis

Human glycoside hydrolases were classified based on structural order/disorder into three categories: structured proteins, proteins with structured domains and disordered regions, and intrinsically disordered proteins (IDPs). Allocation into these categories was performed according to the proportion of short IDRs (functional regions) of 15 residues [40]. The structured proteins were defined as proteins without any IDRs; IDPs were defined as proteins with IDRs spanning throughout the entire sequence, and the last category included proteins made up of both IDRs and structured regions [32]. The structural order/disorder propensity of the dataset was predicted using IUPred2a with 0.5 as the cut-off between order and disorder [41]. A value of 0 indicated a strong propensity for being ordered, and that of 1 indicated a strong propensity for being disordered. Continuous stretches of IDRs were plotted at n ≥ 30, 40, …, 130, as a stretch of more than 30 residues was required for categorisation as a long disordered region, with potential functions in recognition or interactions [32, 42].

4.5 Source code

The source codes used for our experiments are available at https://github.com/ritsumei-infobio/phylogenetic_profiling.

5. Conclusions

In summary, we performed genome-wide phylogenetic profiling and cluster analysis of human glycoside hydrolase proteins. Our results suggest that the acquisition of human glycoside hydrolase genes was essential for the development of the intracellular ERAD system in eukaryotes and for glycan-dependent extracellular signalling in multicellular organisms. Analysis of human glycoside hydrolase genes using Rowe's phylogenetic tree indicated that the modulation of glycan-dependent biological functions by sialidases and lysozyme and that the divergence of glucosylceramidases occurred during chordate evolution (Figure 8).

Supplementary Materials: Supplementary materials can be found at www.mdpi.com/xxx/s1.

Author Contributions: Conceptualization, T.N. and M.F.; methodology, T.N., M.F., J.T. and M.I.; software, T.N., M.F., J.T. and K.S.; validation, T.N., M.F. and K.S.; formal analysis, T.N. and M.F.; investigation, T.N., M.F., J.T., K.S., Y.K. and M.I.; resources, T.N, M.F. and M.I.; data curation, T.N., M.F., J.T. and K.S.; writing—original draft preparation, T.N. and M.F.; writing—review and editing, Y.K. and M.I.; visualization, T.N., M.F. and K.S.; supervision, M.I.; project administration, M.I.; funding acquisition, M.I.

Funding: This study was supported by the MEXT-Supported Program for the Strategic Research Foundation at Private Universities (2015–2019) and Takeda Science Foundation.

Acknowledgments: We would like to thank Information Biology Laboratory members at Ritsumeikan University.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

UniProt	The Universal Protein Resource

CAZy	Carbohydrate-Active Enzymes
KEGG	Kyoto Encyclopedia of Genes and Genomes
O-linked	Oxygen-linked-type
ECM	Extracellular matrix
N-linked	Nitrogen-linked-type
GO	Gene Ontology
GHs	Glycosyl hydrolases
ERAD	Endoplasmic reticulum-associated degradation
MGAT3	β-1,4-mannosyl-glycoprotein 4-β-N-acetylglucosaminyltransferase
GlcNAc	N-acetylglucosamine
N-glycan	N-linked glycan
O-glycan	O-linked glycan
IDRs	Intrinsically disordered regions
EDEMs	ER degradation-enhancing mannosidase-like proteins
BovB	Bovine-B
IDPs	Intrinsically disordered proteins

References

1. Schwarz, F.; Aebi, M. Mechanisms and principles of N-linked protein glycosylation. *Curr. Opin. Struct. Biol.* **2011**, *21*, 576-582, doi:10.1016/j.sbi.2011.08.005.
2. Varki, A.; Freeze, H.H.; Gagneux, P. Evolution of glycan diversity; In Essentials of Glycobiology. 2nd edition; Varki, A.; Cummings, R.D.; Esko, J.D., Freeze, H.H.; Stanley, P. et al.,; Cold Spring Harbor Laboratory Press: New York, USA, 2009; pp. 281-292.
3. Tomono, T.; Kojima, H.; Fukuchi, S.; Tohsato, Y.; Ito, M. Investigation of glycan evolution based on a comprehensive analysis of glycosyltransferases using phylogenetic profiling. *Biophys. Physicobiol.* **2015**, *12*, 57-68, doi:10.2142/biophysico.12.0_57.
4. Day, A.J.; Prestwich, G.D. Hyaluronan-binding proteins: tying up the giant. *J. Biol. Chem.* **2002**, 277, 4585-4588, doi:10.1074/jbc.R100036200.
5. Bernfield, M.; Götte, M.; Park, P.W.; Reizes, O.; Fitzgerald, M.L.; Lincecum, J.; Zako, M.Functions of cell surface heparan sulfate proteoglycans. *Ann. Rev. Biochem.* **1999**, *68*, 729-777, doi:10.1146/annurev.biochem.68.1.729.
6. Li, M.; Song, L.; Qin, X. Glycan changes: cancer metastasis and anti-cancer vaccines. *J. Biosci.* **2010**, *35*, 665-673, doi:10.1007/s12038-010-0073-8.
7. Mony, V.K.; Benjamin, S.; O'Rourke, E.J. A lysosome-centered view of nutrient homeostasis. *Autophagy* **2016**, *12*, 619-631, doi:10.1080/15548627.2016.1147671.
8. Sanderson, R.D.; Yang, Y.; Kelly, T.; MacLeod, V.; Dai, Y.; Theus, A. Enzymatic remodeling of heparan sulfate proteoglycans within the tumor microenvironment: growth regulation and the prospect of new cancer therapies. *J. Cell. Biochem.* **2005**, *96*, 897-905, doi:10.1002/jcb.20602.
9. Ballabio, A.; Gieselmann, V. Lysosomal disorders: from storage to cellular damage. *Biochim. Biophys. Acta Mol. Cell Res.* **2009**, *1793*, 684-696, doi:10.1016/j.bbamcr.2008.12.001.
10. Brown, C.J.; Johnson, A.K.; Daughdrill, G.W. Comparing models of evolution for ordered and disordered proteins. *Mol. Biol. Evol.* **2002**, 27, 609-621, doi:10.1093/molbev/msp277.
11. Chen, J.W.; Romero, P.; Uversky, V.N.; Dunker, A.K. Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J. Proteome Res.* **2006**, *5*, 879-887,doi:0.1021/pr060048x.
12. Brown, C.J.; Johnson, A.K.; Dunker, A.K.; Daughdrill, G.W. Evolution and disorder. *Curr. Opin. Struct. Biol.* **2011**, *21*, 441-446, doi:10.1016/j.sbi.2011.02.005.

13. Fahmi, M.; Ito, M. Evolutionary approach of intrinsically disordered CIP/KIP proteins. *Sci. Rep.* **2019**, *9*, p.1575, doi:10.1038/s41598-018-37917-5.
14. Goldman, N.; Thorne, J.L.; Jones, D.T. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **1998**, *149*, 445-458.
15. Bellay, J.; Michaut, M.; Kim, T.; Han, S.; Colak, R.; Myers, C.L.; Kim, P.M. An omics perspective of protein disorder. *Mol. BioSyst.* **2012**, *8*, 185-193, doi:10.1039/C1MB05235G.
16. Apweiler, R.; Martin, M.J.; O'Donovan, C.; Magrane, M.; Alam-Faruque, Y.; Antunes, R.; Barrell, D.; Bely, B.; Bingley, M.; Binns, D.; Bower, L. The universal protein resource (UniProt) in 2010. *Nucleic Acids Res.* **2010**, *38*, D142-D148, doi:10.1093/nar/gkp846.
17. Lombard, V.; Golaconda Ramulu, H.; Drula, E.; Coutinho, P.M.; Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **2013**, *42*, D490-D495, doi:10.1093/nar/gkt1178.
18. Placzek, S.; Schomburg, I.; Chang, A.; Jeske, L.; Ulbrich, M.; Tillack, J.; Schomburg, D. BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.* **2016**, p.gkw952, doi:10.1093/nar/gkw952.
19. Hunter, S.; Apweiler, R.; Attwood, T.K.; Bairoch, A.; Bateman, A.; Binns, D.; Bork, P.; Das, U.; Daugherty, L.; Duquenne, L.; Finn, R.D. InterPro: the integrative protein signature database. *Nucleic Acids Res.* **2008**, *37*,pp.D211-D215, doi:10.1093/nar/gkn785.
20. Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **2016**, *45*, D353-D361, doi:10.1093/nar/gkw1092.
21. Thompson, A.J.; Williams, R.J.; Hakki, Z.; Alonzi, D.S.; Wennekes, T.; Gloster, T.M.; Songsrirote, K.; Thomas-Oates, J.E.; Wrodnigg, T.M.; Spreitz, J.; Stütz, A.E. Structural and mechanistic insight into N-glycan processing by endo--mannosidase. *Proc. Natl. Acad. Sci.* **2012,** *109*, 781-786, doi:10.1073/pnas.1111482109.
22. Ito, M.; Tohsato, Y.; Sugisawa, H.; Kohara, S.; Fukuchi, S.; Nishikawa, I.; Nishikawa, K. Intrinsically disordered proteins in human mitochondria. *Genes Cells* **2012**, *17*, 817-825, doi:10.1111/gtc.12000.
23. Huang, C.; Harada, Y.; Hosomi, A.; Masahara-Negishi, Y.; Seino, J.; Fujihira, H.; Funakoshi, Y.; Suzuki, T.; Dohmae, N.; Suzuki, T. Endo--N-acetylglucosaminidase forms N-GlcNAc protein aggregates during ER-associated degradation in Ngly1-defective cells. *Proc. Natl. Acad. Sci.* **2015**, *112*, 1398-1403, doi:10.1073/pnas.1414593112.
24. Kornfeld, R.; Kornfeld, S. Assembly of asparagine-linked oligosaccharides. *Ann. Rev. Biochem.* **1985**, *54*, 631-664, doi:10.1146/annurev.bi.54.070185.003215.
25. Mukherjee, S.; Vaishnava, S.; Hooper, L.V. Multi-layered regulation of intestinal antimicrobial defense. *Cell. Mol. Life Sci.* **2008**, *65*, 3019-3027, doi:10.1007/s00018-008-8182-3.
26. Paoletti, M.G.; Norberto, L.; Damini, R.; Musumeci, S. Human gastric juice contains chitinase that can degrade chitin. *Ann. Nutr. Metab.* **2007**, *51*, 244-251, doi:10.1159/000104144.
27. Modelski, M.J.; Menlah, 190 G.;Wang, Y.; Dash, S.;Wu, K.; Galileo, D.S.; Martin-DeLeon, P.A. Hyaluronidase 2: a novel germ cell hyaluronidase with epididymal expression and functional roles in mammalian sperm. *Biol. Reprod.* **2014**, *91*, 109, doi:10.1095/biolreprod.113.115857.
28. Funderburgh, J.L. Keratan sulfate biosynthesis. *IUBMB Life* **2002,** *54*, 187-194, doi:10.1080/15216540214932.

Peer-reviewed version available at *Int. J. Mol. Sci.* **2019**, *20*, 6290; doi:10.3390/ijms20246290

29. Carvalho, S.; Catarino, T.A.; Dias, A.M.; Kato, M.; Almeida, A.; Hessling, B.; Figueiredo, J.; Gärtner, F.; Sanches, J.M.; Ruppert, T.; Miyoshi, E. Preventing E-cadherin aberrant N-glycosylation at Asn-554 improves its critical function in gastric cancer. *Oncogene* **2016**, *35*, 1619, doi:10.1038/onc.2015.225.
30. Marin, M.B.; Ghenea, S.; Spiridon, L.N.; Chiritoiu, G.N.; Petrescu, A.J.; Petrescu, S.M. Tyrosinase degradation is prevented when EDEM1 lacks the intrinsically disordered region. *PloS One* **2012**, 7, p.e42998, doi:10.1371/journal.pone.0042998.
31. Olivari, S.; Molinari, M. Glycoprotein folding and the role of EDEM1, EDEM2 and EDEM3 in degradation of folding-defective glycoproteins. *FEBS let.* **2007**, *581*, pp.3658-3664, doi:10.1016/j.febslet.2007.04.070.
32. Van Der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R.J.; Daughdrill, G.W.; Dunker, A.K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D.T.; Kim, P.M. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **2014**, *114*, 6589-6631, doi:10.1021/cr400525m.
33. Baalsrud, H.T.; Tørresen, O.K.; Solbakken, M.H.; Salzburger, W.; Hanel, R.; Jakobsen, K.S.; Jentoft, S. De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Mol. Biol. Evol.* **2017,** *35*, pp.593-606, doi:10.1093/molbev/msx311.
34. Walsh, A.M.; Kortschak, R.D.; Gardner, M.G.; Bertozzi, T.; Adelson, D.L.Widespread horizontal transfer of retrotransposons. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, pp.1012-1016, doi:10.1073/pnas.1205856110.
35. Eyun, S.I.,; Wang, H.; Pauchet, Y.; Benson, A.K.; Valencia-Jiménez, A.; Moriyama, E.N.; Siegfried, B.D. Molecular evolution of glycoside hydrolase genes in the western corn rootworm (Diabrotica virgifera virgifera). *PloS one* **2014**, *9*, p.e9405., doi:10.1371/journal.pone.0094052.
36. German, J.B.; Freeman, S.L.; Lebrilla, C.B.; Mills, D.A. Human milk oligosaccharides: evolution, structures and bioselectivity as substrates for intestinal bacteria. In Personalized nutrition for the diverse needs of infants and children. *Karger Publishers* **2008**, *62*, pp. 205-222, doi:10.1159/000146322.
37. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; Harris, M.A. Gene ontology: tool for the unification of biology. *Nat. Genet.* **2000**, *25*, p.25, doi:10.1038/75556.
38. Ward Jr, J.H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236-244.
39. Rowe, T. Chordate phylogeny and development; In Assembling the tree of life; Cracraft, J.; Donoghue, M.J.;Oxford University Press: Oxford, UK, 2009; pp.384-409.
40. Oates, M.E.; Romero, P.; Ishida, T.; Ghalwash, M.; Mizianty, M.J.; Xue, B.; Dosztanyi, Z.; Uversky, V.N.; Obradovic, Z.; Kurgan, L.; Dunker, A.K. D2P2: database of disordered protein predictions. *Nucleic Acids Res.* **2012,** *41*, D508-D516, doi:10.1093/nar/gks1226.
41. Mészáros, B.; Erdo˝s, G.; Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **2018**, *46*, W329-W337, doi:10.1093/nar/gky384.
42. Iakoucheva, L.M.; Brown, C.J.; Lawson, J.D.; Obradovic´, Z.; Dunker, A.K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **2002**, *323*, 573-584, doi:10.1016/s0022-2836(02)00969-5.