Molecular Biology and Evolution

Genome-Wide Evolutionary Conservation of *N*-Glycosylation Sites

Chungoo Park[1] and Jianzhi Zhang*[,1]

[1]Department of Ecology and Evolutionary Biology, University of Michigan

***Corresponding author:** E-mail: jianzhi@umich.edu.

Associate editor: Claudia Schmidt-Dannert

Abstract

Although posttranslational protein modifications are generally thought to perform important cellular functions, recent studies showed that a large fraction of phosphorylation sites are not evolutionarily conserved. Whether the same is true for other protein modifications, such as *N*-glycosylation is an open question. *N*-glycosylation is a form of cotranslational and posttranslational modification that occurs by enzymatic addition of a polysaccharide, or glycan, to an asparagine (N) residue of a protein. Examining a large set of experimentally determined mouse *N*-glycosylation sites, we find that the evolutionary rate of glycosylated asparagines is significantly lower than that of nonglycosylated asparagines of the same proteins. We further confirm that the conservation of glycosylated asparagines is accompanied by the conservation of the canonical motif sequence for glycosylation, suggesting that the above substitution rate difference is related to glycosylation. Interestingly, when solvent accessibility is considered, the substitution rate disparity between glycosylated and nonglycosylated asparagines is highly significant at solvent accessible sites but not at solvent inaccessible sites. Thus, although the solvent inaccessible glycosylation sites were experimentally identified, they are unlikely to be genuine or physiologically important. For solvent accessible asparagines, our analysis reveals a widespread and strong functional constraint on glycosylation, unlike what has been observed for phosphorylation sites in most studies, including our own analysis. Because the majority of *N*-glycosylation occurs at solvent accessible sites, our results show an overall functional importance for *N*-glycosylation.

Key words: glycosylation, phosphorylation, evolution, asparagine, solvent accessibility.

Research article

Introduction

Posttranslational protein modification increases the diversity of protein function and is involved in a wide range of cellular processes, such as regulation and signaling. *N*-glycosylation is a form of cotranslational and posttranslational modification that occurs by enzymatic addition of a polysaccharide, or glycan, to an asparagine (N) residue of a protein, via complex biochemical pathways (reviewed in Freeze 2006 and Varki 2009). Most proteins that travel through the endoplasmic reticulum–Golgi conduit are *N*-glycosylated. Case studies have shown that *N*-glycosylation plays important roles in protein folding and stability, cell–cell recognition, signaling, and other cellular processes (Freeze 2006; Ohtsubo and Marth 2006; Varki 2009), and the positions glycosylated in a glycoprotein appear conserved (Freeze 2006). Further, a recent analysis found enzymes responsible for glycan synthesis to be conserved in sequence evolution (Montanucci et al. 2010). Defects in glycan synthesis or structure have been implicated in over 30 human diseases, affecting almost all organ systems (Freeze 2006; Lau and Dennis 2008). For example, congenital disorders of glycosylation (CDG), previously known as carbohydrate deficient glycoprotein syndromes (Hagberg et al. 1993), are a group of rare but severe genetic diseases (Jaeken and Matthijs 2001; Schachter and Freeze 2009). Due to serious infections and organ failures in the first several years of life, childhood mortality of CDG is as high as 20% (Matthijs et al. 2000; Kjaergaard et al. 2001). Furthermore, the deficiency of *N*-glycosylation pathways causes autoimmune diseases (Ohtsubo and Marth 2006; Green et al. 2007). Conversely, misglycosylation can be harmful. For example, a point mutation creating an extra *N*-glycosylation site in fibrillin 1 affects the processing and proper assembly of monomers of fibrillin-1, resulting in neonatal Marfan syndrome (Lonnqvist et al. 1996). Although *N*-glycosylation occurs in thousands of proteins, it remains unclear what fraction of it is functionally important. This question is relevant because not all protein modifications are functionally important. For example, it was reported that a sizable fraction of phosphorylation sites are evolutionary unconserved and therefore are unlikely to have important physiological functions (Lienhard 2008; Holt et al. 2009; Landry et al. 2009; Chen et al. 2010).

A recent large-scale proteomic study identified *N*-glycosylation sites in vivo from four organs (brain, liver, kidney, and heart) and the blood plasma of the laboratory mouse (Zielinska et al. 2010). They identified 6,367 *N*-glycosylated sites (referred to as N^+ sites) on 2,352 proteins (referred to as glycoproteins) and confirmed that the vast majority (96.5%) of N^+ sites are followed by the canonical motif of !P-[S|T]-!P, where !P is not proline and [S|T] is serine or threonine. These authors also found that mouse glycoproteins are more likely than nonglycoproteins to have one-to-one orthologs in other vertebrates (Zielinska et al. 2010). Nevertheless, it is unknown whether glycosylated

asparagines are more conserved than nonglycosylated asparagines, which would be expected if a large proportion of glycosylation is functionally important. In this study, we test the above hypothesis by a comparative genomic analysis of *N*-glycosylation sites.

Materials and Methods

Mouse *N*-glycosylation sites were determined by Zielinska et al. (2010), who developed a filter aided sample preparation method in which glycopeptides are enriched by binding to lectins on the top of a filter and mapped 6,367 *N*-glycosylation sites on 2,352 proteins in four mouse tissues and blood plasma using high-accuracy mass spectrometry. This data set was downloaded from the supplementary materials in Zielinska et al. (2010).

For comparison, we also analyzed a recently published data set of phosphorylation sites experimentally determined in a large-scale proteomic study from nine mouse tissues (brain, brown fat, heart, liver, lung, kidney, pancreas, spleen, and testis) (Huttlin et al. 2010). The data set included nearly 36,000 phosphorylation sites (83% serines, 15% threonines, and 2% tyrosines) from 6,296 proteins.

One-to-one orthologous proteins between mouse and rat, mouse and human, and mouse and chicken were obtained from ENSEMBL (http://www.ensembl.org/index.html) (Release 59). The orthologous protein sequences were aligned using MUSCLE (Edgar 2004) with the default option. Protein pairs with sequence identities lower than 50% in the aligned regions were excluded from subsequent analyses. SABLE 2.0 (Wagner et al. 2005) was used to predict the solvent accessibility of *N*-glycosylation sites in mouse.

Results and Discussion

Glycosylated Asparagines Evolve More Slowly Than Nonglycosylated Asparagines

To investigate the evolutionary rate of glycosylated N residues, we obtained one-to-one orthologous protein sequences from mouse and rat. We first focused on the mouse–rat comparison because using closely related species reduces the uncertainty in ortholog identification and sequence alignment, the latter being particularly important in the subsequent analysis of motif conservation (Balla et al. 2006). There are 3,543 experimentally identified mouse N^+ sites in the set of 1,322 mouse–rat orthologous proteins. We compared the N^+ sites with the remaining 34,215 N residues in the same glycoproteins, hereafter, named nonglycosylated N sites or N^0 sites. This comparison naturally controls the impact of protein expression level on both glycoprotein identification and the rate of protein sequence evolution (Liao et al. 2006). We computed the fractions (r) of mouse N^+ and N^0 sites that are no longer occupied by N in rat, which are approximate amino acid substitution rates (uncorrected for multiple hits) at N^+ and N^0 sites. We found r to be 29.6% lower for N^+ sites (0.062) than for N^0 sites (0.088) ($P < 0.0001$, χ^2 test; fig. 1*A*), indicating that a glycosylated N is on average more conserved than a nonglycosylated N. This difference was further confirmed by randomly picking the same number of N^0 sites as the number of N^+ sites in each glycoprotein and then comparing the evolutionary rates of N^+ and N^0 sites (0.066 vs. 0.117; $P < 0.0001$).

In the above analysis, we assumed no false negative in the experimental detection of N^+ sites from the identified glycoproteins. To reduce the impact of potential false negatives, we define N^- sites as those N^0 sites in the identified glycoproteins that are not followed by a canonical motif. In other words, N^- sites are more likely than N^0 sites to be truly nonglycosylated. A total of 30,240 N^- sites were identified from 34,215 N^0 sites. We found r to be slightly greater for N^- sites (0.090) than for N^0 sites (0.088), but the difference is not statistically significant ($P = 0.38$; fig. 1*A*). Thus, the potential false-negative errors in the experiment have no significant impact on our analysis. For this reason, we use N^0 rather than N^- sites in all subsequent analyses.

To further confirm our finding of a reduced r for N^+ compared with N^0 sites, we conducted similar analyses by comparing mouse with human and chicken (fig. 1*A*). In the mouse–human comparison of 1,403 N^+ sites and 13,980 N^0 sites, r is 40.7% lower for N^+ sites (0.095) than N^0 sites (0.160) ($P < 0.0001$). In the mouse–chicken comparison of 415 N^+ sites and 4,649 N^0 sites, r is 36.7% lower for N^+ sites (0.149) than N^0 sites (0.236) ($P < 0.0001$). Thus, N^+ sites are generally more conserved than N^0 sites in vertebrate evolution.

In the mouse–rat comparison, we also compared r among N^+, N^0, and N^- sites for glycoproteins identified from each of the four tissues and blood plasma and found similar results (fig. 1*B*). Thus, N^+ sites are more conserved than N^0 sites in glycoproteins regardless of their tissue expressions.

Conservation of Canonical Motifs Following Conserved Glycosylated Sites

The evolutionary conservation of N^+ sites relative to N^0 sites is presumably caused by the widespread importance of *N*-glycosylation. Under the assumption that the canonical motif for *N*-glycosylation is the same across vertebrates, we predict that the canonical motif following an N^+ should also be conserved if the N^+ site is evolutionary conserved; this would not be the case for an unconserved N^+. Indeed, 94% of mouse N^+ sites that are both conserved in rat and followed by canonical motifs in mouse are also followed by canonical motifs in rat (fig. 2). By contrast, the conservation of canonical motifs is only 51% when unconserved N^+ sites are examined (fig. 2). This difference ($P < 0.0001$) confirms that the conservation of N^+ is most likely caused by the importance of glycosylation. Similar results were obtained from mouse–human and mouse–chicken comparisons (fig. 2).

Regional Variation in Substitution Rate

To confirm that the lower r at N^+ sites than N^0 sites is not due to preferential localization of N^+ sites to more conserved regions of protein sequences, we compared substitution rates of neighboring residues of N^+ and N^0 sites. Interestingly, r is generally higher in the neighborhood of

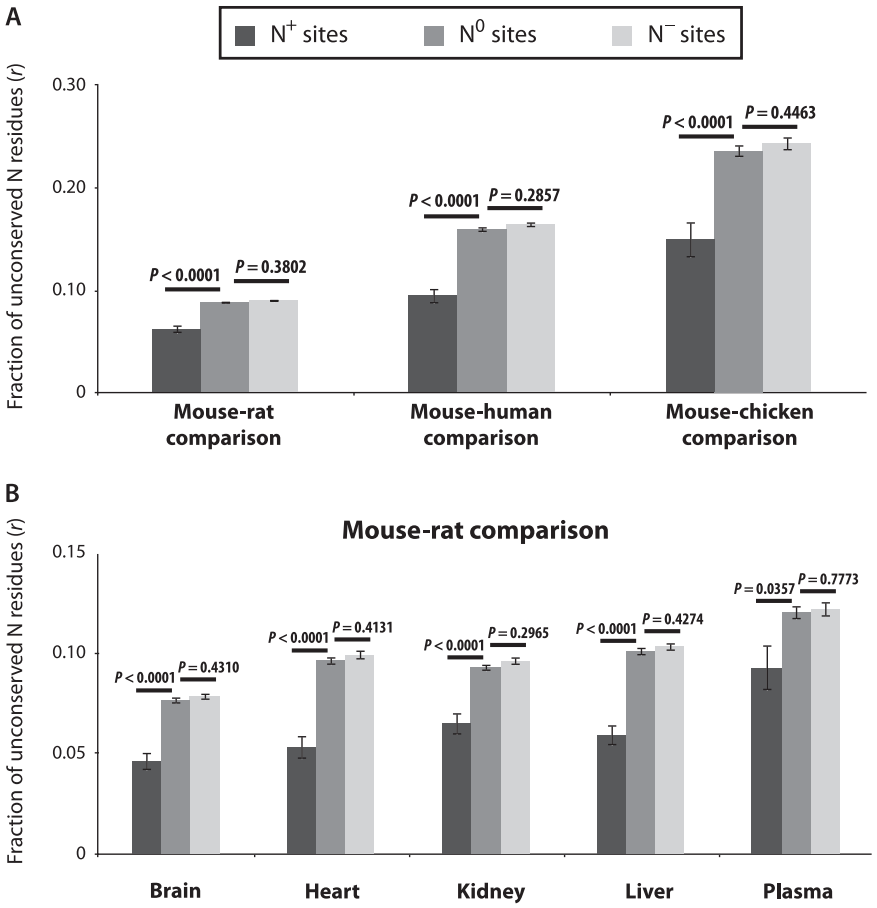

Fig. 1. Fraction of unconserved asparagine (N) residues in glycoproteins identified from (*A*) any of five mouse tissues and (*B*) each of the five mouse tissues. Error bars indicate standard errors. *P* values are from χ^2 tests. N^+ sites are experimentally identified glycosylated sites, N^0 sites are all other N sites in the same proteins, and N^- sites are those N^0 sites that are not followed by canonical glycosylation motifs.

N^+ sites than in the neighborhood of N^0 sites, with the exception of the second position downstream of N^+ (fig. 3), which is part of the canonical motif. Thus, our finding of the conservation of N^+ residues is not due to preferential localization of N^+ sites to more conserved regions of proteins. Rather, N^+ sites tend to be located in regions with relatively high substitution rates, which is likely due to preferential localization of N^+ on protein surfaces (see below).

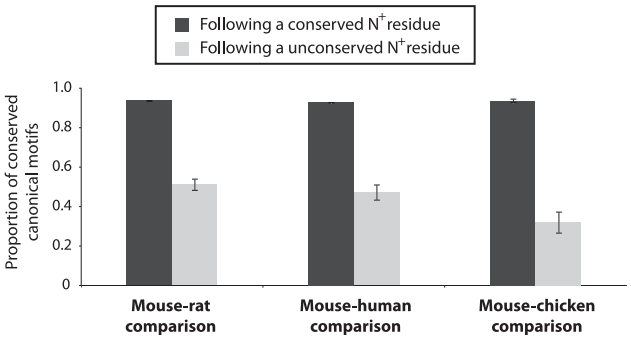

Fig. 2. Proportion of conserved canonical glycosylation motifs. Error bars indicate standard errors.

Substitution Patterns of Unconserved Asparagines

Our inference of functional constraint on glycosylation from the disparity in substitution rate between N^+ and N^0 sites is based on the assumption that the only factor causing the substitution rate difference between N^+ and N^0 sites is glycosylation. To validate this assumption, we compared the substitution patterns of N^+ and N^0 sites, because, under our assumption, the substitution patterns should be the same when an N is substituted. This is indeed the case for the mouse–rat ($P = 0.33$, χ^2 test), mouse–human ($P = 0.92$), and mouse–chicken ($P = 0.19$) comparisons (fig. 4). This observation verifies our assumption and thus supports our inference of selection on glycosylation from the difference in substitution rate between N^+ and N^0 sites.

No Redundancy Among Multiple Glycosylated Sites Within a Protein

It was reported that, although the presence of a phosphorylation site may be important for a protein, the exact position of the phosphorylation site in the protein is unimportant and can shift in evolution (Holt et al. 2009; Moses and Landry 2010). The position shift presumably

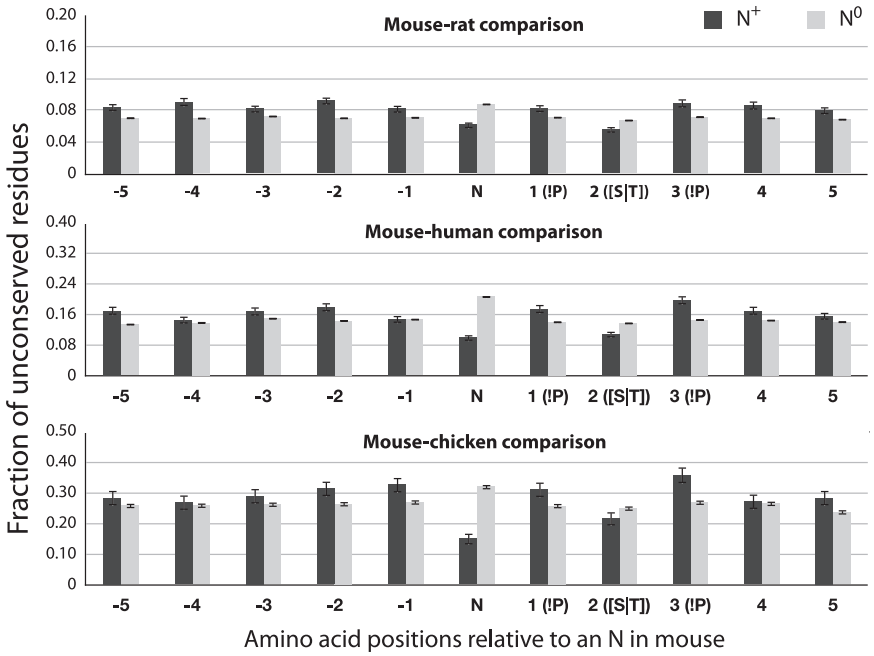

Fig. 3. Variation in substitution rate in the neighborhood of mouse N sites. Error bars indicate standard errors.

occurs in two steps: creation of a new phosphorylation site and the subsequent loss of the old site. The protein in the intermediate stage of this scenario contains two phosphorylation sites, although only one is necessary. We wonder whether this type of functional redundancy also exists in glycoproteins that contain multiple N^+ sites. To test this hypothesis, we grouped mouse glycoproteins by the number of N^+ sites in a protein and calculated the ratio between the r of N^+ sites and that of N^0 sites for each group using the mouse–rat comparison (fig. 5). The redundancy hypothesis predicts that the ratio should increase with the number of N^+ sites per protein, if all

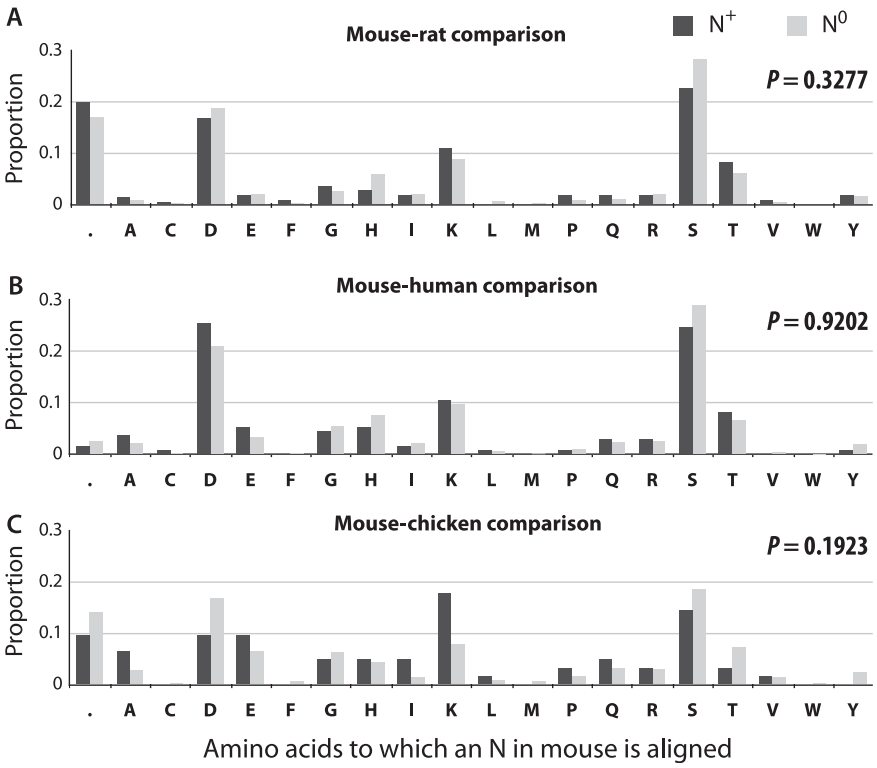

Fig. 4. Substitution patterns of mouse N^+ and N^0 sites in (*A*) mouse–rat, (*B*) mouse–human, and (*C*) mouse–chicken comparisons. The dot indicates a deletion. *P* values are from χ^2 tests.

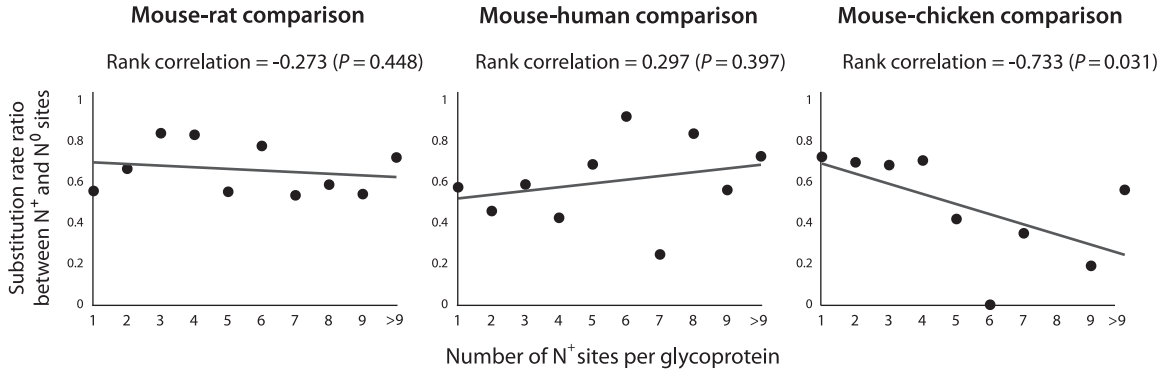

Fig. 5. Correlation between the number of N^+ sites per glycoprotein and the substitution rate ratio between N^+ and N^0 sites.

glycoproteins have the same number of functional N^+ sites. However, there is no significant correlation between the number of N^+ sites in a protein and the ratio (rank correlation $= -0.27$, two tail $P = 0.45$), providing no support for functional redundancy of multiple N^+ sites within a protein. We also confirmed this result in the mouse–human comparison (fig. 5). In the mouse–chicken comparison, however, a significantly negative correlation was observed (fig. 5), suggesting that each N^+ site becomes more important relative to each N^0 site as the number of N^+ sites per glycoprotein increases. Although the cause of this phenomenon is unclear, we note that the significance of the correlation is weak ($P = 0.031$), and it might have been an artifact of multiple testing because three tests were conducted here (fig. 5). The inconsistency in the direction of the correlation among the three comparisons (fig. 5) is consistent with the interpretation that the significance in the mouse–chicken comparison is spurious. In sum, there is no indication of functional redundancy of multiple N^+ sites within a glycoprotein.

N-Glycosylation at Solvent Accessible and Inaccessible Sites

N-glycosylation sites are known to be enriched on the surface of folded proteins (Zielinska et al. 2010). Because surface (i.e., solvent accessible) residues tend to evolve faster than buried (i.e., solvent inaccessible) residues of the same proteins (Goldman et al. 1998; Bustamante et al. 2000; Lin et al. 2007), all of the differences in substitution rate between N^+ and N^0 calculated thus far are conservative estimates. Nonetheless, it is interesting to ask whether there is a difference in the importance of glycosylation at solvent accessible sites and inaccessible sites. We computed solvent accessibility scores of all N residues of mouse glycoproteins using SABLE 2.0 (Wagner et al. 2005). As expected (Zielinska et al. 2010), the median solvent accessibility score of N^+ sites (3.8) is significantly greater than that of N^0 sites (3.4) ($P < 0.0001$, Mann–Whitney *U* test; fig. 6*A*). The enrichment of N^+ relative to N^0 residues occurs at sites with a solvent accessibility score >3 (fig. 6*A*). For simplicity, we refer to N sites with a solvent accessibility score below 3 as solvent inaccessible sites and the rest as solvent accessible sites, as was previously defined (Chen et al. 2010). Compared with an N^0, an N^+ is two times more likely to be found at a solvent accessible site than at an inaccessible site ($P < 0.0001$). Note that the transmembrane residues of membrane proteins have very low solvent accessible scores, consistent with the rare occurrence of *N*-glycosylatioin in transmembrane regions.

Interestingly, we found that the fraction of residues that are unconserved between mouse and rat is comparable between N^+ and N^0 at solvent inaccessible sites (fig. 6*B*). At solvent accessible sites, however, this fraction becomes significantly lower for N^+ sites than for N^0 sites (fig. 6*B*). These observations suggest that *N*-glycosylation at solvent inaccessible sites (24.4% of all N^+ sites) may be experimental artifacts or functionally unimportant. Among solvent accessible sites, the fraction of unconserved N^+ sites between mouse and rat is 45% lower than that of unconserved N^0 sites ($P < 0.0001$), suggesting that the functional constraint on *N*-glycosylation that occurs on solvent accessible sites is quite strong. A similar distinction between solvent accessible and inaccessible sites is observed in mouse–human and mouse–chicken comparisons (fig. 6*B*).

To verify the above finding, we conducted an additional test using the mouse–rat comparison. We identified all mouse canonical motifs following conserved N^+ sites. We then calculated the fraction (f_+) of the above motifs that are unconserved in rat. More specifically, we calculated f_+ for those mouse N^+ sites that are solvent accessible (f_{+A}) and those N^+ sites that are solvent inaccessible (f_{+I}), respectively. Similarly, we identified all mouse canonical motifs following non-N sites that are also non-N in rat. We then calculated the fraction (f_-) of the above motifs that are unconserved in rat. We further calculated f_- for those mouse non-N sites that are solvent accessible (f_{-A}) and those non-N sites that are solvent inaccessible (f_{-I}), respectively. As expected, we found f_{+A} (0.078) to be significantly lower than f_{-A} (0.100) ($P < 0.0003$; fig. 6*C*), reflecting the functional relevance and therefore the evolutionary conservation of the canonical motifs following conserved solvent accessible N^+. By contrast, f_{+I} (0.076) is not lower than f_{-I} (0.075), suggesting the lack of conservation of the canonical motifs following solvent inaccessible N^+ (fig. 6*C*). These

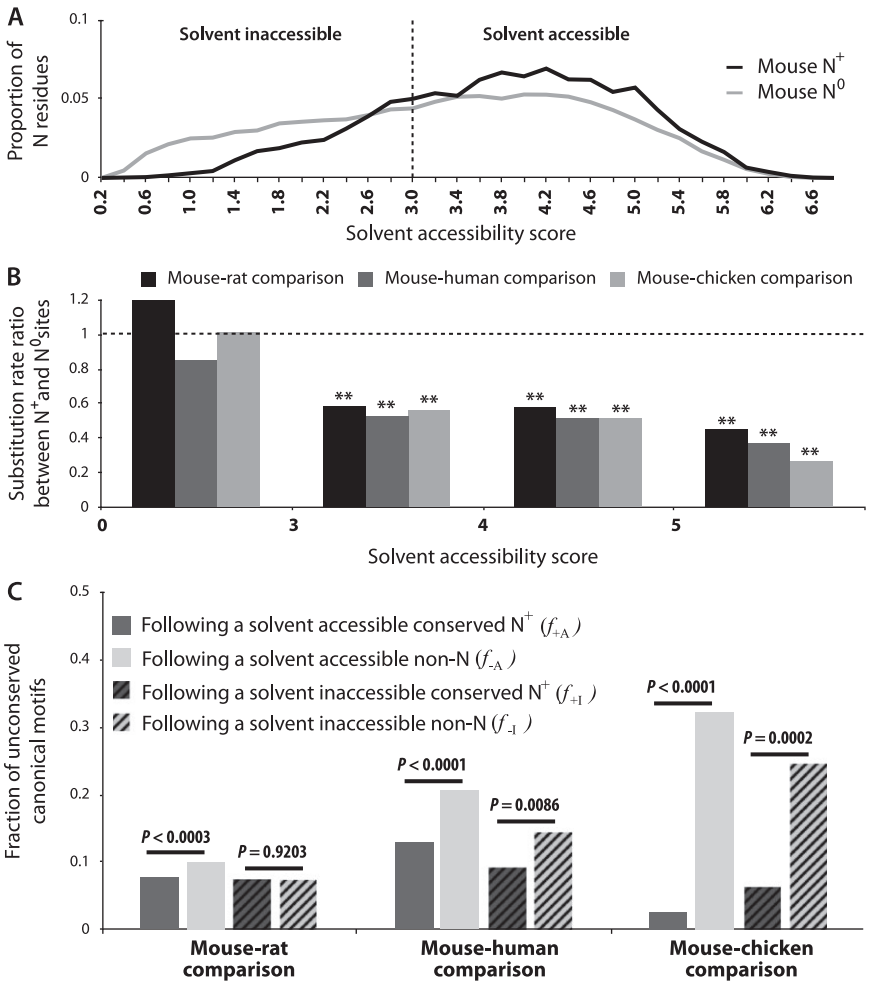

Fig. 6. *N*-glycosylation at solvent accessible and inaccessible sites. (*A*) Frequency distributions of solvent accessibility scores for N^+ and N^0 sites. (*B*) Substitution rate ratio between N^+ and N^0 sites varies with solvent accessibility. *$P < 0.05$; **$P < 0.01$. *P* values are from χ^2 tests. (*C*) Fraction of unconserved canonical motifs. *P* values are from χ^2 tests.

results further support the distinction between glycosylation at solvent accessible and inaccessible asparagines. Note that we did not compare between f_{+A} and f_{+I} or between f_{-A} and f_{-I} because they cannot be directly compared due to the generally faster evolution of solvent accessible regions than inaccessible regions. We conducted similar tests for the mouse–human comparison and mouse–chicken comparison. Although we found both $f_{+A} < f_{-A}$ and $f_{+I} < f_{-I}$ in each of these two comparisons, $(f_{-A} - f_{+A})$ is always greater than $(f_{-I} - f_{+I})$ (fig. 6*C*), reflecting stronger conservations of canonical motifs following solvent accessible N^+ than inaccessible N^+. Together, these findings suggest that *N*-glycosylation in solvent inaccessible sites represents to a certain degree experimental artifacts or physiological unimportance. The observation of a significantly lower f_{+I} than f_{-I} in the mouse–human and mouse–chicken comparisons suggests that some conserved N^+ sites at solvent inaccessible sites may still be functional. It is also possible, although less likely, that this observation is entirely due to inaccurate predictions of solvent accessibility.

Comparison Between the Evolution of *N*-Glycosylation Sites and Phosphorylation Sites

To directly compare the evolutionary rates of *N*-glycosylation sites and phosphorylation sites, we analyzed mouse proteins that have been experimentally shown to contain

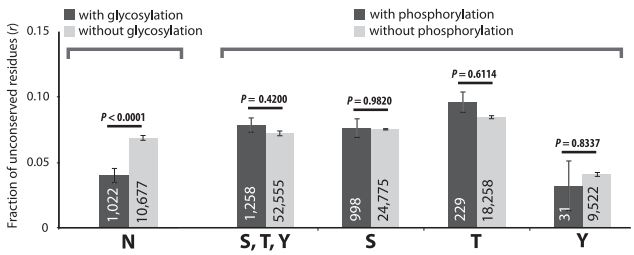

Fig. 7. Fraction of unconserved asparagine (N) residues depends on glycosylation, whereas that of unconserved serine (S), threonine (T), and tyrosine (Y) residues does not depend on phosphorylation. The comparison is made between mouse proteins that are subject to both glycosylation and phosphorylation and their rat orthologs. The number of residues considered in each sample is provided in each bar. Error bars indicate standard errors.

both types of modifications. We found r to be significantly greater for N^0 sites (0.0687) than for N^+ sites (0.0401) in *N*-glycosylation sites ($P < 0.0001$; fig. 7). By contrast, there is no significant difference in r between phosphorylated and nonphosphorylated S, T, and Y sites (fig. 7). Thus, *N*-glycosylation imposes a stronger selective constrain than phosphorylation in general.

Conclusions

In summary, we found solvent accessible *N*-glycosylation to be generally functionally constrained and solvent inaccessible *N*-glycosylation to be much less constrained. Because the majority of *N*-glycosylation events occur at solvent accessible sites, most *N*-glycosylations are likely to be functionally important. We also found no indication of functional redundancy among multiple glycosylation sites of a glycoprotein. These features contrast recent observations on phosphorylation sites, where only a minority of sites are likely to be functionally constrained and the locations of phosphorylated residues in a protein do not seem to matter (Holt et al. 2009). These differences illustrate the complexity and diversity in the function and evolution of various posttranslational modifications. Although additional evolutionary questions about *N*-glycosylation (e.g., species specificity and tissue specificity; Varki 2009) remain to be addressed in the future, our work opens the door to genome-wide understanding of the functional importance and evolution of *N*-glycosylation.

Acknowledgments

We thank Meg Bakewell, Wenfeng Qian, Jian-Rong Yang, and two anonymous reviewers for valuable comments. Research grants (GM67030 and GM80285) from the National Institutes of Health to J.Z.

References

Balla S, Thapar V, Verma S, et al. (15 co-authors). 2006. Minimotif Miner: a tool for investigating protein function. *Nat Methods*. 3:175–177.

Bustamante CD, Townsend JP, Hartl DL. 2000. Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella enterica. *Mol Biol Evol*. 17:301–308.

Chen SC, Chen FC, Li WH. 2010. Phosphorylated and nonphosphorylated serine and threonine residues evolve at different rates in mammals. *Mol Biol Evol*. 27:2548–2554.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.

Freeze HH. 2006. Genetic defects in the human glycome. *Nat Rev Genet*. 7:537–551.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149:445–458.

Green RS, Stone EL, Tenno M, Lehtonen E, Farquhar MG, Marth JD. 2007. Mammalian N-glycan branching protects against innate immune self-recognition and inflammation in autoimmune disease pathogenesis. *Immunity* 27:308–320.

Hagberg BA, Blennow G, Kristiansson B, Stibler H. 1993. Carbohydrate-deficient glycoprotein syndromes: peculiar group of new disorders. *Pediatr Neurol*. 9:255–262.

Holt LJ, Tuch BB, Villen J, Johnson AD, Gygi SP, Morgan DO. 2009. Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* 325:1682–1686.

Huttlin EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, Villen J, Haas W, Sowa ME, Gygi SP. 2010. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* 143:1174–1189.

Jaeken J, Matthijs G. 2001. Congenital disorders of glycosylation. *Annu Rev Genomics Hum Genet*. 2:129–151.

Kjaergaard S, Schwartz M, Skovby F. 2001. Congenital disorder of glycosylation type Ia (CDG-Ia): phenotypic spectrum of the R141H/F119L genotype. *Arch Dis Child*. 85:236–239.

Landry CR, Levy ED, Michnick SW. 2009. Weak functional constraints on phosphoproteomes. *Trends Genet*. 25:193–197.

Lau KS, Dennis JW. 2008. N-Glycans in cancer progression. *Glycobiology* 18:750–760.

Liao BY, Scott NM, Zhang J. 2006. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol Biol Evol*. 23:2072–2080.

Lienhard GE. 2008. Non-functional phosphorylations? *Trends Biochem Sci*. 33:351–352.

Lin YS, Hsu WL, Hwang JK, Li WH. 2007. Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Mol Biol Evol*. 24:1005–1011.

Lonnqvist L, Karttunen L, Rantamaki T, Kielty C, Raghunath M, Peltonen L. 1996. A point mutation creating an extra N-glycosylation site in fibrillin-1 results in neonatal Marfan syndrome. *Genomics* 36:468–475.

Matthijs G, Schollen E, Bjursell C, et al. (13 co-authors). 2000. Mutations in PMM2 that cause congenital disorders of glycosylation, type Ia (CDGIa). *Hum Mutat*. 16:386–394.

Montanucci L, Laayouni H, Dall'olio GM, Bertranpetit J. 2010. Molecular evolution and network-level analysis of the N-Glycosylation metabolic pathway across primates. *Mol Biol Evol*. 28:813–823.

Moses AM, Landry CR. 2010. Moving from transcriptional to phospho-evolution: generalizing regulatory evolution? *Trends Genet*. 26:462–467.

Ohtsubo K, Marth JD. 2006. Glycosylation in cellular mechanisms of health and disease. *Cell* 126:855–867.

Schachter H, Freeze HH. 2009. Glycosylation diseases: quo vadis? *Biochim Biophys Acta*. 1792:925–930.

Varki A. 2009. Essentials of glycobiology. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.

Wagner M, Adamczak R, Porollo A, Meller J. 2005. Linear regression models for solvent accessibility prediction in proteins. *J Comput Biol*. 12:355–369.

Zielinska DF, Gnad F, Wisniewski JR, Mann M. 2010. Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell* 141:897–907.