SCIENTIFIC REPORTS
natureresearch

There are amendments to this paper

OPEN

Mutually exclusive locales for N-linked glycans and disorder in human glycoproteins

Shyamili Goutham[1], Indu Kumari[2], Dharma Pally[1], Alvina Singh[1], Sujasha Ghosh[1], Yusuf Akhter[3] & Ramray Bhat[1]*

Several post-translational protein modifications lie predominantly within regions of disorder: the biased localization has been proposed to expand the binding versatility of disordered regions. However, investigating a representative dataset of 500 human N-glycoproteins, we observed the sites of N-linked glycosylations or N-glycosites, to be predominantly present in the regions of predicted order. When compared with disordered stretches, ordered regions were not found to be enriched for asparagines, serines and threonines, residues that constitute the sequon signature for conjugation of N-glycans. We then investigated the basis of mutual exclusivity between disorder and N-glycosites on the basis of amino acid distribution: when compared with control ordered residue stretches without any N-glycosites, residue neighborhoods surrounding N-glycosites showed a depletion of bulky, hydrophobic and disorder-promoting amino acids and an enrichment for flexible and accessible residues that are frequently found in coiled structures. When compared with control disordered residue stretches without any N-glycosites, N-glycosite neighborhoods were depleted of charged, polar, hydrophobic and flexible residues and enriched for aromatic, accessible and order-promoting residues with a tendency to be part of coiled and β structures. N-glycosite neighborhoods also showed greater phylogenetic conservation among amniotes, compared with control ordered regions, which in turn were more conserved than disordered control regions. Our results lead us to propose that unique primary structural compositions and differential propensities for evolvability allowed for the mutual spatial exclusion of N-glycosite neighborhoods and disordered stretches.

One of the common co- and post-translational modifications of polypeptides is the conjugation of branched glycosylations to asparagines (known as N-linked glycosylations)[1]. N-linked glycosylation begins with the assembly of an oligosaccharide on dolichol pyrophosphate and the subsequent transfer of the oligosaccharide to the asparagine residues of polypeptides in the lumen of the endoplasmic reticulum; the oligosaccharide is further remodeled in the Golgi complex. Several proteins that end up in the extracellular milieu or as transmembrane proteins are N-linked glycoconjugates. The establishment of organismal morphologies has been sought to be understood through the interactions of a highly conserved set of proteins known as the developmental genetic toolkit[2]. Most toolkit proteins, which are involved in tissue-scale processes, such as cell-cell and cell-matrix adhesion, diffusion-driven signaling and cell movement, are extracellular- or membrane-bound glycoproteins[3,4]. The fundamental role of such toolkit proteins in mechanisms of organ- and organismal-development across diverse clades suggests evolutionary constraints on their structures and folds, while they may have continued to evolve to perform newer functions as organisms occupied and constructed newer niches.

A large number of eukaryotic proteins show flexible tertiary structures and are known as intrinsically disordered proteins (IDPs)[5–7]. Other proteins, while not being entirely disordered can possess variable lengths of disorder. The inherently flexible nature of disordered regions enhances the proteins' repertoire of interacting or binding partners, and also plays an important role in the catalytic functions of the proteins[8,9]. Multiple cytoplasmic and nuclear proteins (such as those involved in signal transduction, and transcription factors) are highly disordered, although extracellular proteins have been found to also have disordered stretches[10–13]. In fact, more

[1]Department of Molecular Reproduction, Development and Genetics, Indian Institute of Sciences, Bangalore, 560012, India. [2]School of Earth and Environmental Sciences, Central University of Himachal Pradesh, District-Kangra, Shahpur, Himachal Pradesh, 176206, India. [3]Department of Biotechnology, Babasaheb Bhimrao Ambedkar University, Vidya Vihar, Raebareli Road, Lucknow, Uttar Pradesh, 226025, India. *email: ramray@iisc.ac.in

than 20% of extracellular proteins consist of more than 50% average disordered residues, with approximately 57% extracellular proteins containing at least one single disordered stretch of 30 continuous residues[14]. Several post-translational modifications (PTMs) have been reported to be located within disordered regions e.g., phosphorylations of serines and threonines are found within unstructured regions both in eukaryotes[15–17] and in prokaryotes[18]. Serine/threonine phosphorylations within disordered stretches play crucial roles in the interactions between the proteins and their ligands. They can stabilize the tertiary structural organization of the disordered region in order to enhance its binding to the protein's cognate ligand(s)[19] besides further stabilizing the bound state of the protein-ligand complex[20]. In addition to phosphorylation, O-linked glycosylations have also been predicted to be conjugated to the disordered regions of proteins[21] and mediate resistance to proteolysis leading to their evolutionary selection[22].

An emerging body of literature links the contribution of N-linked glycosylation to the structure, tertiary fold and stabilization of proteins[23]. Intermediate steps in N-glycan biosynthesis act as checkpoints to ensure that only correctly folded proteins traffic along the ER-Golgi axis[24]. Thermodynamic and molecular dynamics studies show that glycans enhance the stability of the tertiary structure of proteins[22,25,26]. Keeping the above observations in mind, we asked whether the sites of N-glycan conjugation (N-glycosites) showed any bias in their locations vis-a-vis (dis)ordered regions of N-glycoproteins. Investigating a dataset of 500 representative human N-glycoproteins (out of a total of 1124 proteins with experimentally elucidated N-glycosites), wherein we mapped and predicted ordered and disordered regions, we found N-glycosites to be enriched predominantly within ordered regions relative to disordered stretches. We present biochemical (proximate) and evolutionary (distal) reasons for this enrichment and the mutual exclusion of N-glycosites and disordered regions.

Materials and Methods

Sequence retrieval and determination of N-glycosites.

Proteins were selected from Uniprot by sequentially applying the following criteria: 1. Encoded by the human genome and 2. Proteins having at least one experimentally elucidated N-linked glycosylation. Subsequently, the disorder prediction was performed using GenesilicoMetadisorder (the best predictor of disorder based on CASP8 and CASP9), which is a meta-predictor based on the following predictors of disorder: POODLE-I, IPDA, IUPRED-I, DISPRO, POODLE-S, IUPRED-S, SPRITZ-I, PRDOS, RONN, DISOPRED2, DISEMBL AND SPRITZ-S[27]. The sequences with annotated N-glycosites and predicted disordered regions are given in Supplementary File 1. Uniprot gave the total number of human N-glycoproteins at 1124, of which the first 500 were chosen for further analysis. Amino acid enrichment was mapped using the ProtParam tool from Expasy (See Supplementary File 2)[28]. In addition, another list of proteins that have been experimentally demonstrated to have disorder was curated from the database DISPROT[29,30] and annotated for both their N-glycosites and disordered regions (Supplementary File 3).

Gene Ontology (GO) analysis of the proteins.

For prediction of gene ontology terms, sequences of the selected proteins were scanned with the GOanna server[31]. GOanna performs a BLAST search against protein sequences that have a GO number. Gene ontological diversity was measured for the 500-protein sequence set and the complete sequence set of human proteins with experimentally elucidated N-linked glycoconjugation. The GO details of both sets are given in Supplementary File 4, wherein both sequence sets are shown to belong to an identical set of ontological categories with reasonably similar patterns of ontological representation.

Phylogenetic analysis.

The protein sequences were aligned using ClustalΩ and a phylogenetic tree was constructed using 1000 bootstrap value using MEGA5[32]. Maximum likelihood was used to construct the phylogenetic tree of the proteins as it is used for the analysis of sequences of diverse origins, while maximum parsimony method was used to construct the phylogenetic tree of the proteins which have similar functions with a comparatively higher sequence homology than the former set of sequences. The phylogenetic tree was viewed using the program FigTree (http://tree.bio.ed.ac.uk/software/figtree/). The subsets of proteins used for tree construction were constituted based on results obtained from GO analysis. Supplementary Fig. 1 represents a flow chart for the construction of phylogenetic trees.

For estimation of the conservation of disorder sites and N-linked glycosylation neighborhoods, the orthologs of each protein from three species *Pan troglodytes, Mus musculus and Gallus gallus* were obtained from Ensembl and aligned with human proteins using ClustalΩ (Supplementary File 5). The proteins that were analyzed represent a smaller subset of the complete protein list (where the separation of N-glycosites and disordered regions was established), that had 1. orthologs across all the three above-mentioned species and 2. showed a non-overlapping separation between the N-glycosite neighborhoods and disordered regions. Shannon information entropy was measured for randomly chosen 11-amino acid ordered and disordered stretches (1 each for every protein) with no N-glycosites, using Protein Variability Server (http://molbiol.edu.ru/eng/index.html) (Supplementary File 6).

Amino acid residue frequency analysis.

The comparison of frequency of occurrence of amino acids within N-glycosite neighborhoods, N-glycosite-less ordered and disordered regions was performed using Composition Profiler (http://www.cprofiler.org/)[33]. Bootstrap iterations were kept at 10,000, significance value was kept at 0.05 and Bonferroni's multiple-comparison correction was applied.

Protein functional site prediction.

In order to compute the probability that the residue combinations constituting disordered patches or N-linked glycosylation neighborhoods likely represent clusters of function-determining residues, full length protein sequences were input into the Universal Evolutionary Trace server (http://lichtargelab.org/software/uet)[34]. The evolutionary trace of residues that varies among evolutionarily distant protein homologs is ranked higher than for residues that vary among evolutionarily proximal protein homologs[35]. The program searches for all homologous sequences within the Uniref90 database (with a minimum e-value cutoff of 1e-250) up to 500 sequences.The mean of residue ranks was computed and compared between residues from the disordered control regions, ordered control regions and residues flanking N-glycosite neighborhoods (Supplementary File 7).

Statistics. We used both descriptive statistics and significance tests to ascertain the differences in residue characteristics between N-glycosite neighborhoods, ordered and disordered regions. For each protein, the mean value of a given characteristic was measured across a residue stretch and when comparing across protein sets, the mean, median and standard error of the mean values were calculated. In order to measure significance, the non-parametric Wilcoxon's matched-pairs signed rank test was performed. Significance was measured through P value using GraphPad Prism software. For representation, box and whisker plots were selected in order to show the interquartile range that provides a comprehensive understanding of the difference in values between two large sets of data with wide valuedistributions.

Results

N-glycosites are predominantly enriched in ordered regions of N-glycoproteins. We used UniProt to collate a list of all human glycoproteins with at least one experimentally determined N-glycosite. Among 1124 such N-glycoproteins, 500 proteins representing the diverse biological function of the complete dataset were processed through in Silico Metadisorder to predict the amino acids that constituted disordered sequences. We then annotated each N-glycoprotein for (dis)order and N-glycosite location. N-linked glycoconjugated asparagines (or N-glycosites) were found to be significantly enriched within ordered regions. (Fig. 1; Supplementary File 1). Most prediction and meta-prediction algorithms tend to 'overpredict' disorder, i.e., the possibility of false positives is higher: therefore, the enrichment of N-glycan-conjugated asparagines within ordered stretches could be even higher than our observation. We then took a complementary approach: upon searching DISPROT, a database of proteins with experimentally annotated disordered regions[30], we identified a set of eighteen human proteins with varying extents of disorder and also with at least one N-glycosite. In these eighteen proteins, N-glycosites were also found to be enriched significantly in the ordered regions (Fig. 1; Supplementary File 3).

In order to test whether the exclusion (or the less common inclusion) of N-glycosites from (and within) disordered stretches is a characteristic of one or more phylogenetically conserved protein families, we classified our protein set on the basis of gene ontologies. We observed that our analyzed set of N-glycoproteins perform a diverse range of predicted biological functions (Supplementary File 4). Subsets of proteins within each ontological classifier varied in their phylogenetic relatedness: for example, proteins related to immune system function were closely related (Supplementary Fig. 2 top), however, those that clustered together under the ontological heading of metabolic functions were very distantly related (Supplementary Fig. 2 bottom). Even a comparatively limited set of proteins with N-glycosites in their disordered regions were distributed across a wide set of ontological categories. This suggested that N-glycoproteins with disordered stretches represented a very diverse set of phylogenetically distant proteins.

The residue signature for N-linked glycosylation has been established to be asparagine, followed by any amino acid except proline (as it introduces a kink in the protein structure), followed in turn by a serine (Ser) or threonine (Thr) (i.e., N(X-P)S/T)[36]. Therefore, we asked whether the differential enrichment for N-glycosites could simply be reconciled by enrichment of asparagine, serine and/or threonine within ordered stretches of proteins due to their inherent biochemical difference from disordered stretches.

N-linked glycoconjugation sequon but not its individual amino acid constituents are enriched within ordered regions. There are several descriptors of protein disorder; one of the dominantly used descriptors is that of a consecutive stretch of polar and/or charged amino acids[37]. Since asparagine, serine and threonine are polar in nature, we estimated the abundance of these three amino acids within ordered and disordered segments of our protein dataset. The percentage of asparagine was found to be insignificantly different between disordered and ordered regions of N-glycoproteins. Serine and threonine were surprisingly found to be significantly enriched within disordered- rather than ordered- stretches (Fig. 2; Supplementary File 2). This discounted the hypothesis that the relative enrichment in N-linked glycosylation in ordered regions could be the result of a specific enrichment of asparagine, serine and/or threonine.

We then asked what fractions of asparagines participate in constituting a N(X-P)S/T sequon within ordered and disordered regions. Despite an unbiased distribution of asparagine between ordered and disordered regions, and relative enrichment of serine and threonine within the disordered region, the probability of constitution of the minimal signature required for N-glycan conjugation (N(X-P)S/T) was found to be significantly higher within ordered regions (Fig. 2).

We then probed for the relative proportion of individual amino acids that occupy the "X" of the sequon (N(X-P)S/T). Empirical and bioinformatic analysis with model and natural proteins, respectively, have suggested that the identity of "X" in the sequon may determine the extent of its glycosylation: bulky, hydrophobic or acidic amino acids may reduce the efficiency of glycosylation. Conversely, small, hydrophilic or basic amino acids being present as X may increase the propensity of the sequon to be glycosylated[38,39]. We analyzed the identity of X in the sequons that are present in predicted regions of order and disorder: we found higher proportion of hydrophobic and bulky residues as X in the sequons inhabiting the ordered regions compared with the disordered region (Supplementary File 9). The proportion of acidic and basic amino acids as X was higher in the sequons from the disordered region. These findings are consistent with the amino acid compositions of the ordered and disordered regions. Therefore, these properties of the sequon glycosylation do not explain why N-linked glycosylation predominates in ordered regions. We then asked if the residues surrounding N-linked glycosylation sequons were biochemically distinct from those constituting the disordered regions of proteins.

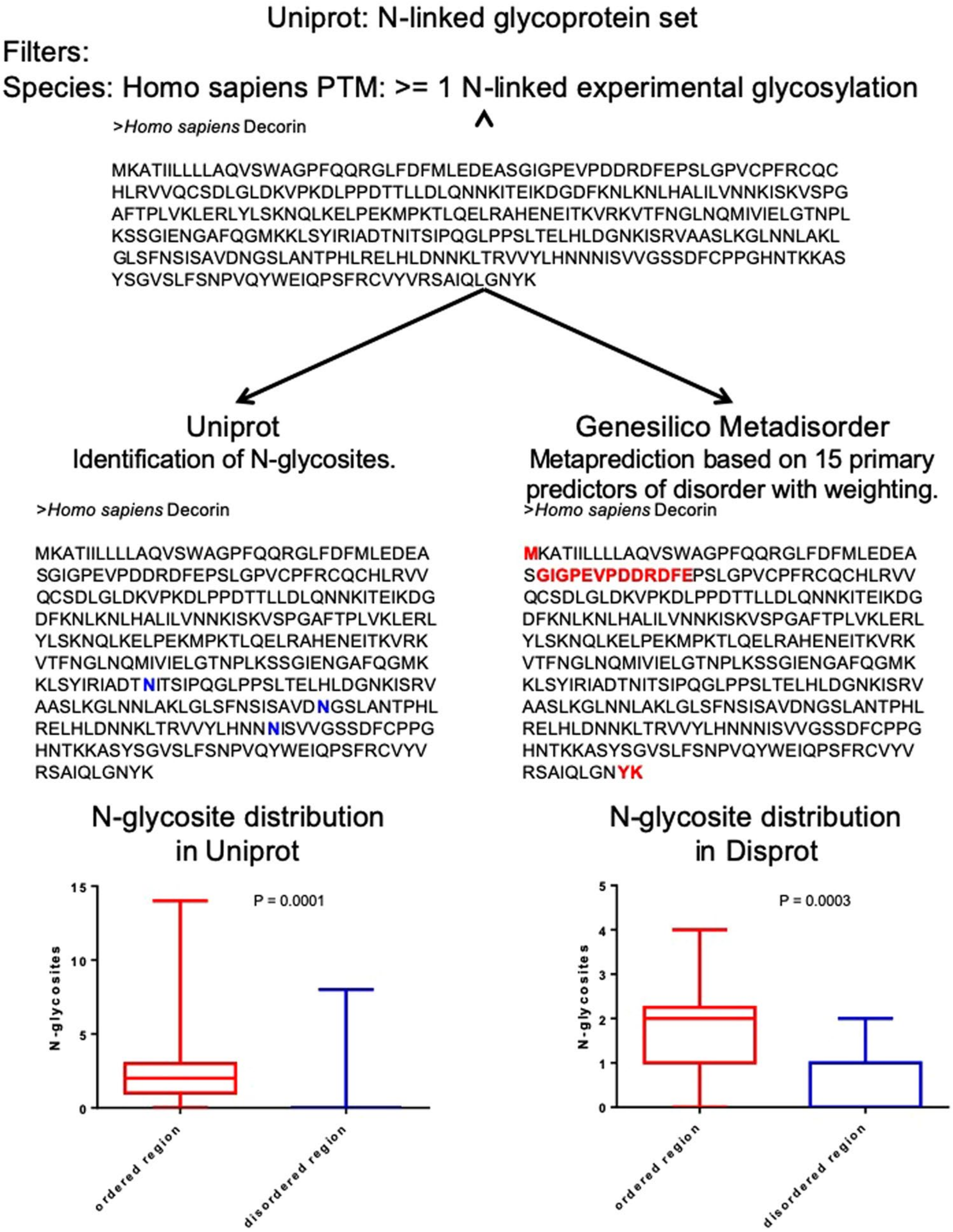
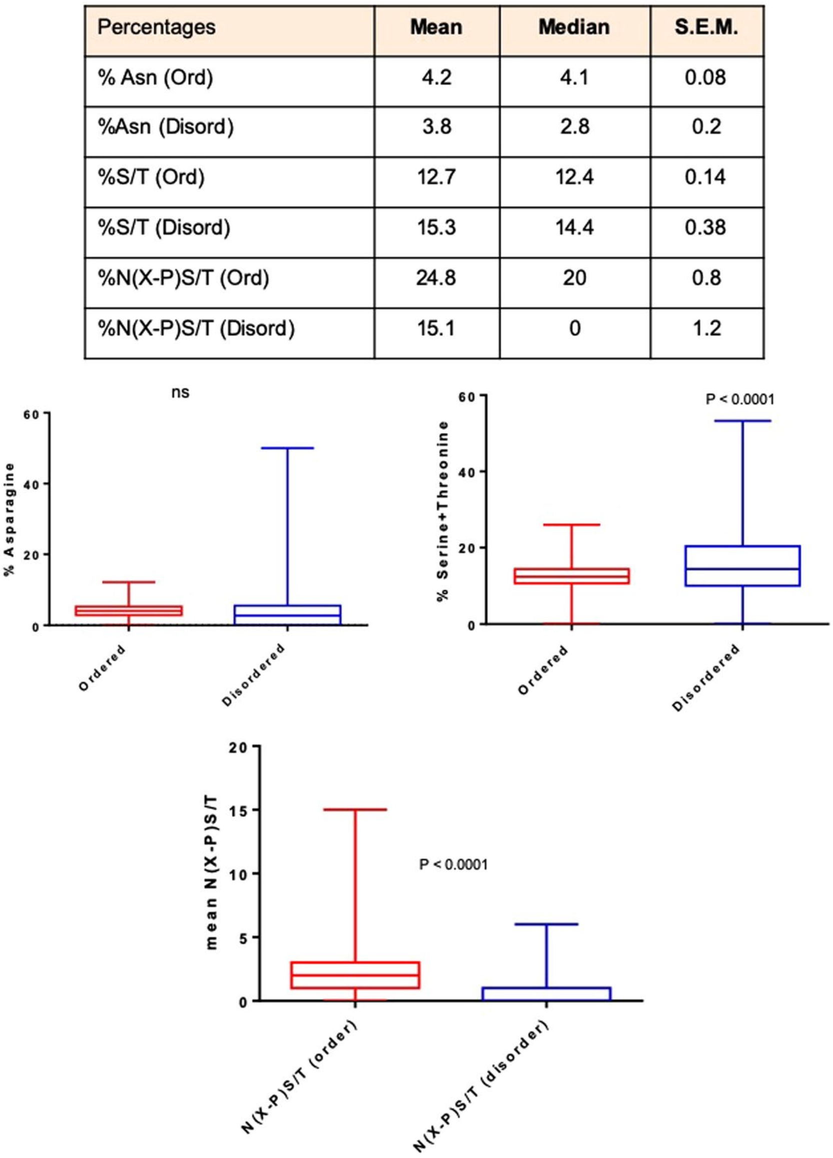

Uniprot: N-linked glycoprotein set
Filters:
Species: Homo sapiens PTM: >= 1 N-linked experimental glycosylation
>Homo sapiens Decorin
MKATIILLLLAQVSWAGPFQQRGLFDFMLEDEASGIGPEVPDDRDFEPSLGPVCPFRCQC
HLRVVQCSDLGLDKVPKDLPPDTTLLDLQNNKITEIKDGDFKNLKNLHALILVNNKISKVSPG
AFTPLVKLERLYLSKNQLKELPEKMPKTLQELRAHENEITKVRKVTFNGLNQMIVIELGTNPL
KSSGIENGAFQGMKKLSYIRIADTNITSIPQGLPPSLTELHLDGNKISRVAASLKGLNNLAKL
GLSFNSISAVDNGSLANTPHLRELHLDNNKLTRVVYLHNNNISVVGSSDFCPPGHNTKKAS
YSGVSLFSNPVQYWEIQPSFRCVYVRSAIQLGNYK
Uniprot
Identification of N-glycosites.
>Homo sapiens Decorin
MKATIILLLLAQVSWAGPFQQRGLFDFMLEDEA
SGIGPEVPDDRDFEPSLGPVCPFRCQCHLRVV
QCSDLGLDKVPKDLPPDTTLLDLQNNKITEIKDG
DFKNLKNLHALILVNNKISKVSPGAFTPLVKLERL
YLSKNQLKELPEKMPKTLQELRAHENEITKVRK
VTFNGLNQMIVIELGTNPLKSSGIENGAFQGMK
KLSYIRIADTNITSIPQGLPPSLTELHLDGNKISRV
AASLKGLNNLAKLGLSFNSISAVDNGSLANTPHL
RELHLDNNKLTRVVYLHNNNISVVGSSDFCPPG
HNTKKASYSGVSLFSNPVQYWEIQPSFRCVYV
RSAIQLGNYK
Genesilico Metadisorder
Metaprediction based on 15 primary predictors of disorder with weighting.
>Homo sapiens Decorin
MKATIILLLLAQVSWAGPFQQRGLFDFMLEDEA
SGIGPEVPDDRDFEPSLGPVCPFRCQCHLRVV
QCSDLGLDKVPKDLPPDTTLLDLQNNKITEIKDG
DFKNLKNLHALILVNNKISKVSPGAFTPLVKLERL
YLSKNQLKELPEKMPKTLQELRAHENEITKVRK
VTFNGLNQMIVIELGTNPLKSSGIENGAFQGMK
KLSYIRIADTNITSIPQGLPPSLTELHLDGNKISRV
AASLKGLNNLAKLGLSFNSISAVDNGSLANTPHL
RELHLDNNKLTRVVYLHNNNISVVGSSDFCPPG
HNTKKASYSGVSLFSNPVQYWEIQPSFRCVYV
RSAIQLGNYK
N-glycosite distribution in Uniprot
P = 0.0001
N-glycosites
0
5
10
15
ordered region
disordered region
N-glycosite distribution in Disprot
P = 0.0003
N-glycosites
0
1
2
3
4
5
ordered region
disordered region

Figure 1. Workflow for the determination of the exclusion between N-glycosites and predicted disordered stretches A list of proteins was compiled from Uniprot based on the criteria that they were encoded by the human genome and at the same time had at least one N-linked glycosylation established through experimental elucidation. Subsequently, one of or more N-glycosites were annotated in its amino acid sequence. The sequence was also submitted to Genesilico Metadisorder in order to annotate the predicted disordered residue stretches. Bottom left: Box and whisker plot representation of the relative occurrence of experimentally established N-glycosites in ordered versus disordered regions of the 500 human N-glycoprotein set from Uniprot. Bottom right: Box and whisker plot representation of the relative occurrence of experimentally established N-glycosites in ordered versus disordered regions of the 18-protein dataset from Disprot. For both graphs, boxes represent interquartile range and whiskers extend from minimum to maximum values. Statistical significance is given by P-value measured using Wilcoxon's matched-pairs signed rank test.

N-glycosite neighborhoods have unique residue compositions based on polarity, length of side chain and aromaticity. We began by annotating five residues upstream- and downstream of the glycoconjugated asparagines as N-glycosite neighborhoods of proteins in our dataset. Additionally 11-residue stretches that did not contain N-glycosites were also identified in the disordered and ordered regions and were defined as disordered and ordered control regions respectively. We estimated the relative frequencies of amino acids between these three 11-residue subsets across our 500-protein dataset and estimated the significance in enrichment or deletion of residue properties such as aromaticity, polarity[40], charge, hydrophobicity[41], flexibility[42],



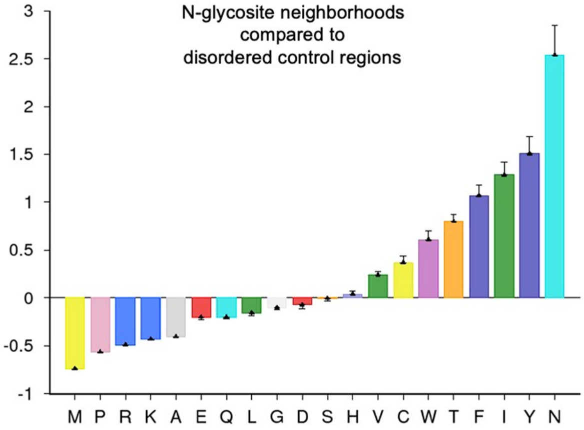
Percentages	Mean	Median	S.E.M.
% Asn (Ord)	4.2	4.1	0.08
%Asn (Disord)	3.8	2.8	0.2
%S/T (Ord)	12.7	12.4	0.14
%S/T (Disord)	15.3	14.4	0.38
%N(X-P)S/T (Ord)	24.8	20	0.8
%N(X-P)S/T (Disord)	15.1	0	1.2

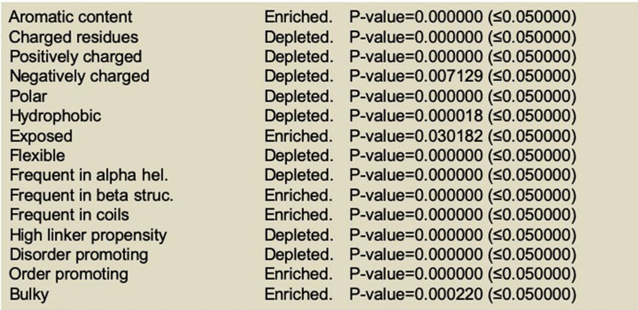
Figure 2. The N-linked glycosylation sequon but not the individual sequon residues show enrichment in the ordered regions (Top) Table showing the mean, median and standard error of mean in the percentage of Asparagines (Asn), Serines (Ser) and Threonines (Thr) and the N-linked glycosylation sequon (N(X-P)S/T) within the ordered and disordered regions of the 500 human N-glycoprotein set. The percentage of asparagine, serine and threonine and N(X-P)S/T for each protein is provided in Supplementary File 3. (Bottom) Box and whisker plot representation of the relative proportion of asparagine, serine and threonine, and N(X-P) S/T within the 500 human N-glycoprotein subset. For the graphs, boxes represent interquartile range and whiskers extend from minimum to maximum values. Statistical significance is given by P-value measured using Wilcoxon's matched-pairs signed rank test.

surface exposure[43], normalized frequency of occurrence in β structures and coils[44], propensities to contribute to order[45], and bulkiness[40].

We first compared the set of ordered control regions with disordered control regions (Supplementary Figs. 3 and 4). The former showed an enrichment of aromatic, bulky, hydrophobic, order-promoting residues with propensity to participate in β structures. It also showed a depletion for charged, polar and flexible residues that are exposed, promote disorder and participate in coil secondary structures. These observations served to confirm the veracity of our controls.

In comparison with disordered control region residues, N-glycosite neighborhoods were found to be enriched for aromatic, bulky and order-promoting amino acids that showed a propensity for access and being part of β structures and coils (Fig. 3 and Supplementary Fig. 5). N-glycosite neighborhoods also showed a depletion for charged, polar and flexible disorder-promoting residues (Fig. 3 and Supplementary Fig. 5). N-glycosites in ordered regions when compared with the less commonly found N-glycosites in disordered regions were significantly enriched also for bulky, hydrophobic and order-promoting residues and depleted of surface-exposed,





Aromatic content	Enriched.	P-value=0.000000 (≤0.050000)
Charged residues	Depleted.	P-value=0.000000 (≤0.050000)
Positively charged	Depleted.	P-value=0.000000 (≤0.050000)
Negatively charged	Depleted.	P-value=0.007129 (≤0.050000)
Polar	Depleted.	P-value=0.000000 (≤0.050000)
Hydrophobic	Depleted.	P-value=0.000018 (≤0.050000)
Exposed	Enriched.	P-value=0.030182 (≤0.050000)
Flexible	Depleted.	P-value=0.000000 (≤0.050000)
Frequent in alpha hel.	Depleted.	P-value=0.000000 (≤0.050000)
Frequent in beta struc.	Enriched.	P-value=0.000000 (≤0.050000)
Frequent in coils	Enriched.	P-value=0.000000 (≤0.050000)
High linker propensity	Depleted.	P-value=0.000000 (≤0.050000)
Disorder promoting	Depleted.	P-value=0.000000 (≤0.050000)
Order promoting	Enriched.	P-value=0.000000 (≤0.050000)
Bulky	Enriched.	P-value=0.000220 (≤0.050000)

Figure 3. N-glycosite neighborhoods compared with disordered control regions show unique amino acid compositions. (Top) Graphical representation of fractional enrichment of amino acids in 11-residue N-glycosite-centered neighborhoods within ordered regions normalized to disordered control regions of similar length (containing no N-glycosites) (see also Supplementary Fig. 5 for tabular representation signifying enrichment and depletion of individual amino acids for the above comparison with significance). (Bottom) Tabular representation signifying enrichment and depletion of residue properties based on the above amino acid enrichment and depletion profile. Error bars represent SD. Statistical significance is given by P-value measured using two-way t-test (see[33]).

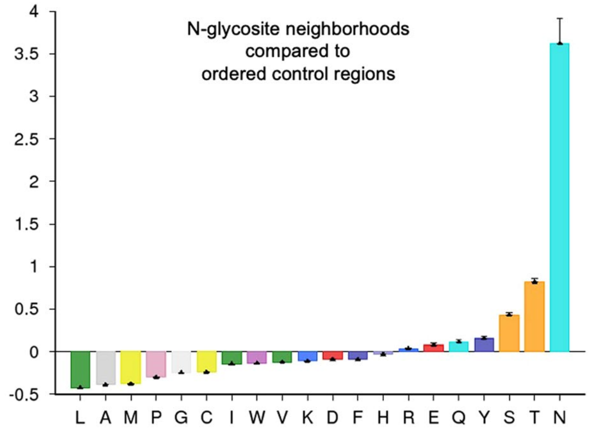
flexible, disorder-promoting amino acids (Supplementary File 8). When compared with ordered control regions, N-glycosite neighborhoods showed an enrichment for flexible and exposed residues with a tendency to be part of coils (Fig. 4 and Supplementary Fig. 6). As well, N-glycosite neighborhoods were relatively depleted of hydrophobic and bulky residues as well as those that participate in β structures (Fig. 4 and Supplementary Fig. 6).

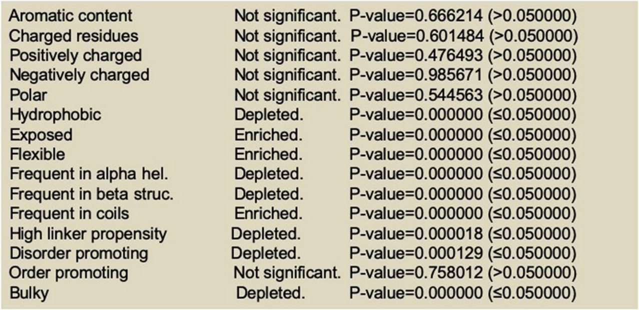
Ranking the three datasets (N-glycosite neighborhoods, ordered- and disordered- control regions) together, based on the above properties revealed two broad signatures (Supplementary Fig. 7): The first signature represented properties wherein the disordered control regions were significantly different from both the other regions (Supplementary Fig. 7 left). Aromatic and order-promoting amino acids were specifically depleted in disordered control regions. The latter also showed an enrichment for charged and polar amino acids, relative to the other two regions (between which such properties were insignificantly distinct). Properties such as bulkiness, flexibility, promotion of disorder and a tendency to be part of linker regions were also changed to the greatest extent in the disordered control regions, albeit with relative and significant differences between ordered regions and N-glycosite neighborhoods. The other signature consisted of properties for which N-glycosite neighborhoods and ordered control regions represented the ends of the comparative spectrum of differences with disordered control regions showing intermediate values (Supplementary Fig. 7 right). These were residue hydrophobicity, exposure, and the propensity to be part of coil structures.

Having ascertained the differences in primary structure between the three sequence sets, we sought to investigate their propensity for being locales where function-determining residues tend to cluster together.

Disordered regions show higher clustering of function-determining residues; N-glycosite neighborhoods show greater evolutionary conservation within the tetrapod clade.

We used universal evolutionary trace, an algorithm that identifies and ranks each amino acid within a protein that determines its function[34]. This functional importance, is therefore not centered exclusively on conservation, but on the fact that variation in evolutionarily distant homologs ranks a residue higher than variation in evolutionarily closer homologs in being able to be part of a functional interface. Such important amino acids potentially cluster together allowing prediction of putative functional motifs for proteins that have not been crystallized. We estimated that the mean evolutionary trace ranks of residues that constituted the control order regions and those that constituted N-glycosite neighborhoods were not significantly different. On the other hand, the mean trace of the disordered control sequence set was significantly higher than both the ordered subsets (Fig. 5; Supplementary





Aromatic content	Not significant.	P-value=0.666214 (>0.050000)
Charged residues	Not significant.	P-value=0.601484 (>0.050000)
Positively charged	Not significant.	P-value=0.476493 (>0.050000)
Negatively charged	Not significant.	P-value=0.985671 (>0.050000)
Polar	Not significant.	P-value=0.544563 (>0.050000)
Hydrophobic	Depleted.	P-value=0.000000 (≤0.050000)
Exposed	Enriched.	P-value=0.000000 (≤0.050000)
Flexible	Enriched.	P-value=0.000000 (≤0.050000)
Frequent in alpha hel.	Depleted.	P-value=0.000000 (≤0.050000)
Frequent in beta struc.	Depleted.	P-value=0.000000 (≤0.050000)
Frequent in coils	Enriched.	P-value=0.000000 (≤0.050000)
High linker propensity	Depleted.	P-value=0.000018 (≤0.050000)
Disorder promoting	Depleted.	P-value=0.000129 (≤0.050000)
Order promoting	Not significant.	P-value=0.758012 (>0.050000)
Bulky	Depleted.	P-value=0.000000 (≤0.050000)

Figure 4. N-glycosite neighborhoods compared with ordered control regions show unique amino acid compositions. (Top) Graphical representation of fractional enrichment of amino acids in 11-residue N-glycosite-centered neighborhoods within ordered regions normalized to ordered control regions of similar length (containing no N-glycosites) (see also Supplementary Fig. 6 for tabular representation signifying enrichment and depletion of individual amino acids for the above comparison with significance). (Bottom) Tabular representation signifying enrichment and depletion of residue properties based on the above amino acid enrichment and depletion profile. Error bars represent SD. Statistical significance is given by P-value measured using two-way t-test (see[33]).

File 5). This suggested that disordered regions scored higher than similar length stretches in the ordered regions in terms of determining protein function.

We then asked if N-glycosite neighborhoods represented regions of higher evolutionary conservation within ordered parts of protein primary structures. We aligned the human N-glycoprotein sequences with their orthologs from chimpanzee (*Pan troglodytes*), mouse (*Mus musculus*) and chicken (*Gallus gallus*) in order to assay a possible divergence in the extent of conservation between residues surrounding the N-glycosites (in the ordered regions) and residues in the disordered regions (Supplementary File 6). Our screen sought to probe for conservation within a wider clade (i.e., amniotes), wherein despite important differences in N-linked glycoconjugation, the primary structures of proteins shows high conservation across the vertebrate clade. We quantified conservation by measuring the mean Shannon entropy of residues constituting the N-glycosite neighborhoods and disordered regions (Fig. 6; Supplementary File 7)[46–48]. Shannon entropy is a measure of the information content within a sequence. Its measure across a given residue position within multiple aligned peptide sequences provides insight into how a residue is conserved in that position across species. When 11-residue ordered and disordered regions (with no N-glycosites) were compared, disordered sequences exhibited greater entropy. In contrast, N-glycosite neighborhoods showed lower mean entropy than full-length protein sequences suggesting residues comprising N-glycosite neighborhood show higher, and those comprising disordered regions, lower, evolutionary conservation among amniote N-glycoproteins. Previous studies have shown that the association between residue conservation and residue disorder is context-specific[21,49]. Our observations indicate that N-glycosites represent high-conservation islands within ordered regions which themselves are more conserved within the amniotic clade with respect to disordered regions.

Discussion

In this paper, we have endeavored to examine the spatial changes in the biochemical properties of amino acids that reside close to or away from the sites of N-glycan conjugation. We observe that residue neighborhoods that immediately flank the glycan conjugated sequons show a similar proportion of charged and polar residues as control ordered regions; such residues are enriched maximally within control disordered regions. On the other hand, N-glycosite neighborhoods are least enriched for hydrophobic residues, more so than even the disordered control regions. Therefore, while both N-glycosites and disordered regions are accessible to the surface, the former represents the ordered tertiary protein surface and the latter represent flexible regions.

Universal Evolutionary Trace	Mean	Median	S.E.M
N-glycosite neighborhood	27.27	20.70	1.76
Ordered control region	29.67	23.29	1.93
Disordered control region	44.46	39.00	2.76

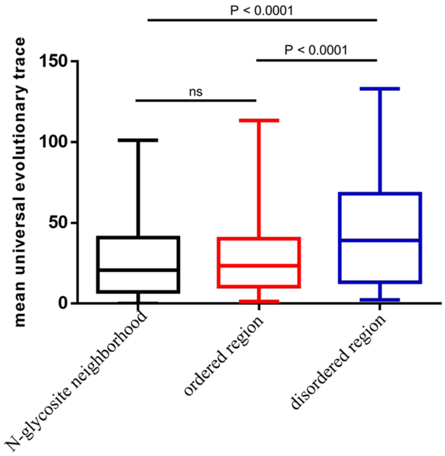
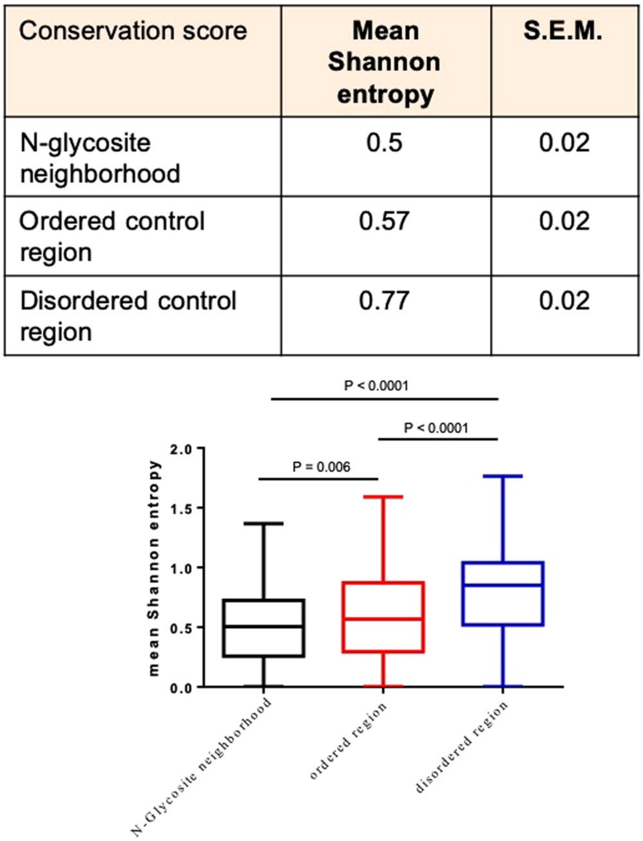


Figure 5. Disordered regions show greater ability to form functionally relevant residual clusters than ordered regions with and without N-glycosites (Top) Table showing mean and median universal evolutionary trace and the standard error of mean of N-glycosite neighborhoods, 11-residue ordered and disordered regions with N-glycosites. (Bottom) Box and whisker representation showing differences in mean universal evolutionary traces of N-glycosite neighborhoods, 11-residue ordered and disordered regions with N-glycosites. The mean evolutionary trace for each protein is provided in Supplementary File 5. For the graph, boxes represent interquartile range and whiskers extend from minimum to maximum values. Statistical significance is given by P-value measured using Wilcoxon's matched-pairs signed rank test.

We observed that while order-promoting residues are equivalently distributed between N-linked glycosites and control ordered regions, the former are relatively further depleted of disorder-promoting residues compared to the latter. N-glycans have been shown through mounting theoretical[26,50] and experimental[51] evidence to impact the folding and structure of glycoproteins (although the relationship is not always positively regulatory)[52,53]. The triose core of N-glycoproteins enhances both the kinetics and stability of tertiary glycoprotein folds[54]. In fact, N-glycosylations seems to destabilize the unfolded state more than stabilize the folded state of proteins[26]. Our results suggest that not just N-glycans but the residue neigborhoods surrounding them may help perform these functions.

Disordered regions are known to contain variable amino acid sequences and we observe that they show relatively lower phylogenetic conservation among amniotes. However, some amino acid residues which have longer side chains than others, e.g., asparagine, proline, glutamate, methionine, serine, arginine, and glutamine are significantly enriched in the disordered regions. The long side-chains of these amino acids may have some significant function in the disordered region which could be related to their ability to enhance flexibility, exposure and disorder. The abundance of pyrrolidine ring-containing proline allows it to introduce kinks or turns in the protein structure and contributes to its propensity within interdomain linkers. It is well known that proline content is generally higher within disordered stretches[22]. Therefore, N-glycosite neighborhoods are uniquely placed in terms of their biochemical properties separating them from the hydrophobic ordered core and the disordered stretches of N-linked glycoproteins.

Based on our observations of evolutionary trace and amniote phylogeny, N-glycosite neighborhoods show residue conservation but do not seem to represent hotspots for the phylogenetic clustering of amino acids that determine function. This however, does not imply that N-linked glycans or even N-glycosites do not determine protein functions. Far from it, both are crucial to fold of the protein and the function performed by it. Additionally, our results imply that disordered stretches may represent hotspots for the further accumulation of function-determining residues, which may regulate the evolvability of such N-glycoproteins.



Conservation score	Mean Shannon entropy	S.E.M.
N-glycosite neighborhood	0.5	0.02
Ordered control region	0.57	0.02
Disordered control region	0.77	0.02

Figure 6. The N-glycosite neighborhoods show greater phylogenetic conservation across amniote genomes than ordered and disordered residue stretches of similar length (Top) Table shows mean Shannon information entropy and its standard error of mean of N-glycosite neighborhoods, 11-residue ordered and disordered regions with N-glycosites. (Bottom) Box and whisker representation showing differences in mean Shannon entropy of N-glycosite neighborhoods, 11-residue ordered and disordered regions with N-glycosites. Multiple sequence alignments for each protein, for which conservation was analyzed is provided in Supplementary File 4. The mean Shannon entropy for each protein is provided in Supplementary File 5. For the graph, boxes represent interquartile range and whiskers extend from minimum to maximum values. Statistical significance is given by P-value measured using Wilcoxon's matched-pairs signed rank test.

The N-linked glycosylation of proteins also tends to act as an important quality control check point for the ability of the protein to fold itself correctly[55]. Unfolded proteins are N-glycosylated within the ER lumen and bear 3 Glc residues at the terminal end of their A-branch. N-linked glycans being bulky hydrophilic polymers can increase the thermodynamic stability of proteins, enhance their folding and allow them to evade the enzyme UGGT that would otherwise deploy unfolded proteins for proteolytic degradation[56]. However, this regulation by N-glycans of protein folding does not necessarily bias the hypothetical localization of the N-glycosites within the ordered (or disordered) regions. We would in the future, perform computational studies to examine the effect on the dynamics of disordered regions by the hypothetical presence of one or more N-glycan(s).

Our results are also consistent with an earlier approach that sought to develop algorithms to predict sites of N-linked glycosylation based on structure and pattern[57], wherein structural properties such as secondary structure were observed to contribute further to the accuracy of prediction of whether a potential N-glycosite is conjugated with a glycan structure, than local contact order. We observed, for example, that residues that are part of N-glycosite neighborhoods show greater propensity to be part of coiled structures. Whereas local contact order may not by itself be able to distinguish successfully between N-glycan-conjugated and -unconjugated glycosites according to Chuang and coworkers[57], it may contribute to constraining the localization of conjugated N-glycosites within the ordered stretches of the protein.

Our results are specific to N-glycoproteins, which are part of the extracellular milieu. Many of them being transmembrane- and secreted diffusible- proteins, function across multicellular tissue spatial scales and contribute to the developmental mechanisms by mediating discrete biophysical and biochemical functions[3]. The determinacy of their developmental roles is a function of their stability and N-glycosylation may have played an important role in the latter. Our study qualifies this by proposing that the localization of N-linked glycosylations allowed these proteins to also have unstructured interfaces in order that they interact with the various components in the extracellular milieu without compromising the global structured fold of the proteins. Our phylogenetic analysis suggests our N-glycoprotein set consists of proteins with diverse functions. However, it also shows that at least for some of the ontological categories, proteins, which shared some common function also were phylogenetically related. This clustering may help to choose specific representatives of clustered protein sets for further molecular dynamical studies. Such studies will help us better understand the role of disordered regions and N-glycosites in the evolution of protein function at the molecular level.

Data availability

All the raw data for the analyses have been provided in the Supplemenrary Information Files.

Received: 20 June 2019; Accepted: 30 January 2020;
Published online: 08 April 2020

References

1. Khoury, G. A., Baliban, R. C. & Floudas, C. A. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific reports* 1, https://doi.org/10.1038/srep00090 (2011).
2. Jaramillo, M. L. *et al*. Exploring developmental gene toolkit and associated pathways in a potential new model crustacean using transcriptomic analysis. *Development genes and evolution* **226**, 325–337, https://doi.org/10.1007/s00427-016-0551-6 (2016).
3. Newman, S. A. & Bhat, R. Dynamical patterning modules: a "pattern language" for development and evolution of multicellular form. *The International journal of developmental biology* **53**, 693–705, https://doi.org/10.1387/ijdb.072481sn (2009).
4. Engler, A. J., Humbert, P. O., Wehrle-Haller, B. & Weaver, V. M. Multiscale modeling of form and function. *Science* **324**, 208–212, https://doi.org/10.1126/science.1170107 (2009).
5. Uversky, V. N. & Dunker, A. K. Understanding protein non-folding. *Biochimica et biophysica acta* **1804**, 1231–1264, https://doi.org/10.1016/j.bbapap.2010.01.017 (2010).
6. Uversky, V. N. & Dunker, A. K. Biochemistry. Controlled chaos. *Science* **322**, 1340–1341, https://doi.org/10.1126/science.1167453 (2008).
7. Romero, P., Obradovic, Z. & Dunker, A. K. Natively disordered proteins: functions and predictions. *Applied bioinformatics* **3**, 105–113 (2004).
8. Fuxreiter, M. Fuzziness: linking regulation to protein dynamics. *Molecular bioSystems* **8**, 168–177, https://doi.org/10.1039/c1mb05234a (2012).
9. Mohan, A. *et al*. Analysis of molecular recognition features (MoRFs). *Journal of molecular biology* **362**, 1043–1059, https://doi.org/10.1016/j.jmb.2006.07.087 (2006).
10. Cortese, M. S., Uversky, V. N. & Dunker, A. K. Intrinsic disorder in scaffold proteins: getting more from less. *Progress in biophysics and molecular biology* **98**, 85–106, https://doi.org/10.1016/j.pbiomolbio.2008.05.007 (2008).
11. Dunker, A. K. & Uversky, V. N. Signal transduction via unstructured protein conduits. *Nature chemical biology* **4**, 229–230, https://doi.org/10.1038/nchembio0408-229 (2008).
12. Liu, J. *et al*. Intrinsic disorder in transcription factors. *Biochemistry* **45**, 6873–6888, https://doi.org/10.1021/bi0602718 (2006).
13. Niklas, K. J., Bondos, S. E., Dunker, A. K. & Newman, S. A. Rethinking gene regulatory networks in light of alternative splicing, intrinsically disordered protein domains, and post-translational modifications. *Frontiers in cell and developmental biology* **3**, 8, https://doi.org/10.3389/fcell.2015.00008 (2015).
14. Peysselon, F., Xue, B., Uversky, V. N. & Ricard-Blum, S. Intrinsic disorder of the extracellular matrix. *Molecular bioSystems* **7**, 3353–3365, https://doi.org/10.1039/c1mb05316g (2011).
15. Iakoucheva, L. M. *et al*. The importance of intrinsic disorder for protein phosphorylation. *Nucleic acids research* **32**, 1037–1049, https://doi.org/10.1093/nar/gkh253 (2004).
16. Marchini, F. K. *et al*. Profiling the Trypanosoma cruzi phosphoproteome. *PloS one* **6**, e25381, https://doi.org/10.1371/journal.pone.0025381 (2011).
17. Amoutzias, G. D., He, Y., Lilley, K. S., Van de Peer, Y. & Oliver, S. G. Evaluation and properties of the budding yeast phosphoproteome. *Molecular & cellular proteomics: MCP* **11**, M111 009555, https://doi.org/10.1074/mcp.M111.009555 (2012).
18. Singh, G. P. Association between intrinsic disorder and serine/threonine phosphorylation in Mycobacterium tuberculosis. *PeerJ* **3**, e724, https://doi.org/10.7717/peerj.724 (2015).
19. Gsponer, J., Futschik, M. E., Teichmann, S. A. & Babu, M. M. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science* **322**, 1365–1368, https://doi.org/10.1126/science.1163581 (2008).
20. Nishi, H., Hashimoto, K. & Panchenko, A. R. Phosphorylation in protein-protein binding: effect on stability and function. *Structure* **19**, 1807–1815, https://doi.org/10.1016/j.str.2011.09.021 (2011).
21. Nishikawa, I. *et al*. Computational prediction of O-linked glycosylation sites that preferentially map on intrinsically disordered regions of extracellular proteins. *Int. J. Mol. Sci.* **11**, 4991–5008, https://doi.org/10.3390/ijms11124991 (2010).
22. Prates, E. T. *et al*. The impact of O-glycan chemistry on the stability of intrinsically disordered proteins. *Chemical science* **9**, 3710–3715, https://doi.org/10.1039/c7sc05016j (2018).
23. Helenius, A. & Aebi, M. Intracellular functions of N-linked glycans. *Science* **291**, 2364–2369 (2001).
24. Ferris, S. P., Kodali, V. K. & Kaufman, R. J. Glycoprotein folding and quality-control mechanisms in protein-folding diseases. *Disease models & mechanisms* **7**, 331–341, https://doi.org/10.1242/dmm.014589 (2014).
25. Tams, J. W. & Welinder, K. G. Glycosylation and thermodynamic versus kinetic stability of horseradish peroxidase. *FEBS letters* **421**, 234–236 (1998).
26. Shental-Bechor, D. & Levy, Y. Effect of glycosylation on protein folding: a close look at thermodynamic stabilization. *Proc. Natl Acad. Sc.i USA* **105**, 8256–8261, https://doi.org/10.1073/pnas.0801340105 (2008).
27. Kozlowski, L. P. & Bujnicki, J. M. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC bioinformatics* **13**, 111, https://doi.org/10.1186/1471-2105-13-111 (2012).
28. Gasteiger, E. et al. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic acids research* **31**, 3784–3788 (2003).
29. Sickmeier, M. *et al*. DisProt: the Database of Disordered Proteins. *Nucleic acids research* **35**, D786–793, https://doi.org/10.1093/nar/gkl893 (2007).
30. Piovesan, D. *et al*. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic acids research* **45**, D1123–D1124, https://doi.org/10.1093/nar/gkw1279 (2017).
31. McCarthy, F. M. *et al*. AgBase: a unified resource for functional analysis in agriculture. *Nucleic acids research* **35**, D599–603, https://doi.org/10.1093/nar/gkl936 (2007).
32. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution* **10**, 512–526, https://doi.org/10.1093/oxfordjournals.molbev.a040023 (1993).
33. Vacic, V., Uversky, V. N., Dunker, A. K. & Lonardi, S. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC bioinformatics* **8**, 211, https://doi.org/10.1186/1471-2105-8-211 (2007).
34. Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology* **257**, 342–358, https://doi.org/10.1006/jmbi.1996.0167 (1996).

35. Lua, R. C. *et al.* UET: a database of evolutionarily-predicted functional determinants of protein sequences that cluster as functional sites in protein structures. *Nucleic acids research* **44**, D308–312, https://doi.org/10.1093/nar/gkv1279 (2016).
36. Mellquist, J. L., Kasturi, L., Spitalnik, S. L. & Shakin-Eshleman, S. H. The amino acid following an asn-X-Ser/Thr sequon is an important determinant of N-linked core glycosylation efficiency. *Biochemistry* **37**, 6833–6837, https://doi.org/10.1021/bi972217k (1998).
37. Uversky, V. N., Gillespie, J. R. & Fink, A. L. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41**, 415–427 (2000).
38. Shakin-Eshleman, S. H., Spitalnik, S. L. & Kasturi, L. The amino acid at the X position of an Asn-X-Ser sequon is an important determinant of N-linked core-glycosylation efficiency. *The Journal of biological chemistry* **271**, 6363–6366, https://doi.org/10.1074/jbc.271.11.6363 (1996).
39. Petrescu, A. J., Milac, A. L., Petrescu, S. M., Dwek, R. A. & Wormald, M. R. Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. *Glycobiology* **14**, 103–114, https://doi.org/10.1093/glycob/cwh008 (2004).
40. Zimmerman, J. M., Eliezer, N. & Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *Journal of theoretical biology* **21**, 170–201 (1968).
41. Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *Journal of molecular biology* **179**, 125–142, https://doi.org/10.1016/0022-2836(84)90309-7 (1984).
42. Vihinen, M., Torkkila, E. & Riikonen, P. Accuracy of protein flexibility predictions. *Proteins* **19**, 141–149, https://doi.org/10.1002/prot.340190207 (1994).
43. Janin, J. Surface and inside volumes in globular proteins. *Nature* **277**, 491–492 (1979).
44. Nagano, K. Logical analysis of the mechanism of protein folding. I. Predictions of helices, loops and beta-structures from primary structure. *Journal of molecular biology* **75**, 401–420 (1973).
45. Dunker, A. K. *et al.* Intrinsically disordered protein. *Journal of molecular graphics & modelling* **19**, 26–59 (2001).
46. Pritisanac, I., Vernon, R. M., Moses, A. M. & Forman Kay, J. D. Entropy and Information within Intrinsically Disordered Protein Regions. *Entropy* **21**, 662, https://doi.org/10.3390/e21070662 (2019).
47. Bywater, R. P. Prediction of protein structural features from sequence data based on Shannon entropy and Kolmogorov complexity. *PloS one* **10**, e0119306, https://doi.org/10.1371/journal.pone.0119306 (2015).
48. Oliveira, L., Paiva, P. B., Paiva, A. C. & Vriend, G. Identification of functionally conserved residues with the use of entropy-variability plots. *Proteins* **52**, 544–552, https://doi.org/10.1002/prot.10490 (2003).
49. 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. *Journal of proteome research* **5**, 879–887, https://doi.org/10.1021/pr060048x (2006).
50. Lee, H. S., Qi, Y. & Im, W. Effects of N-glycosylation on protein conformation and dynamics: Protein Data Bank analysis and molecular dynamics simulation study. *Scientific reports* **5**, 8926, https://doi.org/10.1038/srep08926 (2015).
51. Banks, D. D. The effect of glycosylation on the folding kinetics of erythropoietin. *Journal of molecular biology* **412**, 536–550, https://doi.org/10.1016/j.jmb.2011.07.061 (2011).
52. Gavrilov, Y., Shental-Bechor, D., Greenblatt, H. M. & Levy, Y. Glycosylation May Reduce Protein Thermodynamic Stability by Inducing a Conformational Distortion. *The journal of physical chemistry letters* **6**, 3572–3577, https://doi.org/10.1021/acs.jpclett.5b01588 (2015).
53. Shental-Bechor, D. & Levy, Y. Communication: folding of glycosylated proteins under confinement. *J. Chem. Phys.* **135**, 141104, https://doi.org/10.1063/1.3650700 (2011).
54. Hanson, S. R. *et al.* The core trisaccharide of an N-linked glycoprotein intrinsically accelerates folding and enhances stability. *Proc. Natl Acad. Sci. USA* **106**, 3131–3136, https://doi.org/10.1073/pnas.0810318105 (2009).
55. Xu, C. & Ng, D. T. Glycosylation-directed quality control of protein folding. *Nature reviews. Molecular cell biology* **16**, 742–752, https://doi.org/10.1038/nrm4073 (2015).
56. Sifers, R. N. Insights into checkpoint capacity. *Nature structural & molecular biology* **11**, 108–109, https://doi.org/10.1038/nsmb0204-108 (2004).
57. Chuang, G. Y. *et al.* Computational prediction of N-linked glycosylation incorporating structural properties and patterns. *Bioinformatics* **28**, 2249–2255, https://doi.org/10.1093/bioinformatics/bts426 (2012).

Acknowledgements

RB acknowledges funding from DBT (0525) and Wellcome Trust DBT India Alliance Fellowship (WELT0041). SG (Shyamili Goutham) and DP acknowledge support from IISc for fellowship. We also thank DST-FIST [SR/FST/LS11-036/2014(C)], UGC-SAP [F.4.13/2018/DRS-III (SAP-II)] and DBT-IISc Partnership Program Phase-II (BT/PR27952-INF/22/212/2018) for infrastructure and financial support. Research in YA lab is funded by Indian Council of Medical Research (Ministry of Health & Family Welfare, Government of India) and Department of Biotechnology (Ministry of Science and Technology, Government of India).

Author contributions

S.G. (Goutham), I.K., D.P., A.S., S.G. (Ghosh), performed the analysis and contributed to writing the manuscript. Y.A. designed the in silico analyses and contributed to writing the manuscript. R.B. designed and performed the analyses and contributed to writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-61427-y.

Correspondence and requests for materials should be addressed to R.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.