A study of the structural properties of sites modified by the *O*-linked 6-N-acetylglucosamine transferase

Short title: Structural characterisation of *O*-GlcNAc sites

keywords: post-translational modification, computational biology, structural analysis, O-GlcNAc, protein disorder

Thiago Britto-Borges and Geoffrey J. Barton*

Division of Computational Biology
School of Life Sciences
University of Dundee
Dow Street, Dundee, DD1 5EH
UK`

*To whom correspondence should be addressed: g.j.barton@dundee.ac.uk

Abstract

Protein O-GlcNAcylation (O-GlcNAc) is an essential post-translational modification (PTM) in higher eukaryotes. The O-linked β-N-acetylglucosamine transferase (OGT), targets specific Serines and Threonines (S/T) in intracellular proteins. However, unlike phosphorylation, fewer than 25% of known O-GlcNAc sites match a clear sequence pattern. Accordingly, the three-dimensional structures of O-GlcNAc sites were characterised to investigate the role of structure in molecular recognition. Of the 143/1,584 O-GlcNAc sites in 620 proteins were mapped to protein X-ray structures. The modified S/T were 1.7x more likely to be annotated in the REM465 field which defines missing residues in a protein structure, while 7 O-GlcNAc sites were solvent inaccessible and unlikely to be targeted by OGT. The 132/143 sites with complete backbone atoms clustered into 10 groups, but these were indistinguishable from clusters from unmodified S/T. This suggests there is no prevalent three-dimensional motif for OGT recognition. Predicted features from the 620 proteins were compared to unmodified S/T in O-GlcNAcylated proteins and globular proteins. The Jpred4 predicted secondary structure shows that modified S/T were more likely to be coils. 5/6 methods to predict intrinsic disorder indicated O-GlcNAcylated S/T to be significantly more disordered than unmodified S/T. Although the analysis did not find a pattern in the site three-dimensional structure, it revealed the residues around the modification site are likely to be disordered and suggests a potential role of secondary structure elements in OGT site recognition.

I. Introduction

Protein *O*-GlcNAcylation, or *O*-GlcNAc, is a dynamic, intracellular glycosylation essential to mammalian development[1,2]. In animals, two enzymes mediate this post-translational modification: the glycosyltransferase O-linked 6-N-acetylglucosamine transferase (OGT), which adds a single, non-extensible *O*-GlcNAc moiety to serine/threonine (S/T) in the target protein; and the hexosaminidase *O*-GlcNAcase (OGA) that removes it. UDP-GlcNAc, the sugar donor to the protein *O*-GlcNAcylation, is a product of the hexosamine pathway, hence the concentration of intracellular glucose and the degree of protein *O*-GlcNAcylation levels are associated[3, 4]. At the physiological level, dysfunction of OGT activity has been linked to disease of the cardiovascular system, diabetes, impaired development, cancer and neurodegeneration[5–9]. At the cellular level, protein *O*-GlcNAcylation acts with phosphorylation, ubiquitylation and other reversible post-translational modifications in a network of cell signalling events that promote cellular adaptation to the viral infection process[10], regulation of transcription[11] and metabolism[12, 13].

Technical advances in mass spectrometry have led to an increase in the number of experimentally determined *O*-GlcNAc sites from 50 in the year 2000 to more than 1,000 today[14]. However, there are still obstacles to mapping *O*-GlcNAc sites reliably. The modification has a low abundance[15] and is ten times less common than protein phosphorylation[16]. Thus, the unmodified version of the peptide can suppress the *O*-GlcNAcylated peptide mass/charge signal. In addition, methods to enrich *O*-GlcNAcylated peptides in samples have limited specificity[16, 17], and the β-glycosidic bond is labile under the peptide fragmentation step which determines the modification's position within the peptide fragment.

Two machine learning methods have been used to detect patterns in the sequence of *O*-GlcNAc sites[18, 19] with limited success[20]. One of the limiting factors was that, unlike phosphorylation sites, O-GlcNAc sites lack a clear pattern in the primary structure. This is illustrated in Figure 1 which compares the relative sequence entropy for sites modified by OGT and three protein kinases in the PhosphoSitePlus database[14]. The observed relative entropy for OGT sites shows small signal, in contrast to protein kinase A (PKA; peak in -3 and +2), protein kinase C (PKC; peak in -3) and casein kinase 2 (CK2, peak in +3) sites. This implies that the sequence in the sites recognised by OGT carries less information than those recognised by PKA, PKS or CK2 and so are harder to distinguish from unmodified sites by sequence alone.

The crystal structure of OGT in a ternary complex with UDP–GlcNAc and a peptide substrate revealed that the OGT and the peptides' residues predominantly make contact via the peptide backbone[21, 22]. This fact reduces the importance of the peptide side chain in the enzyme active site, the cleft where the reaction occurs. A short structural motif, instead of sequence motif, could work as a point of molecular recognition even with a degenerate sequence. Accordingly, in this paper, the three-dimensional structures of S/T OGT substrates were examined to determine if they have distinct structural motifs and patterns of secondary structure or solvent accessibility. In addition, the predicted secondary structure and disorder were compared for known OGT substrates and S/T unlikely to be modified.

II. Methods

A. Data sources

The data selection process is summarised in Figure 2. A total of 1,533 modified sites from 676 proteins were selected by combining proteins curated from the literature up until 2011[18] and from 2011-2013[20]. The sites were filtered to keep 7-residue long motifs with unique sequences. The

resulting dataset contained 1,385 sites in 620 proteins. This dataset is referred to hereafter as the "modified sequence sites" (MSS). For comparison, 100,329 S/T from the same proteins, but not annotated as OGT-modified, were selected as a background and are referred to here as the "unmodified sequence sites" (USS).

B. Mapping O-GlcNAc sites to protein structures

Protein chains > 30 residues long from structures determined by X-ray crystallography to ≤ 2.50 Å resolution were selected from the Protein Data Bank[23] (PDB: 2nd August of 2015). Mapping the 1,385 OGT sites from 620 proteins to PDB structures by SIFTS[24] located 45 OGT sites in 24 proteins of known structure. The structures of a further 107 sites were identified by searching the sequences of O-GlcNAcylated proteins against the PDB chains with BLAST and filtering by a conservative E-value $\leq 10^{-25}$ to minimise erroneous matches. The cutoff of $\leq 10^{-25}$ was found empirically to ensure the reliability of the match in the region of each site by inspecting all alignments between query and PDB sequence at different thresholds. Selecting the protein chain with highest coverage (SIFTS) or E-value (BLAST) left 143 sites in 107 proteins for further analysis, referred to hereafter as the "143 Structural Sites" (SS143).

C. Site definition and clustering

The three-dimensional structure of OGT with its substrates suggests the region of contact between OGT and a modifiable S/T includes the residues and +/- 3 amino acids either side[21, 22]. From the structural sites returned in Mapping O-GlcNAc-sites to protein structures, "132 Structural Sites" (hereafter SS132) had at least one match with all backbone atoms for the 7-residue long site and were retained for further analysis. Cα atoms of each residue and the Cα and

112 the Cβ for the central S/T were superimposed for all pairs of sites. Hierarchical clustering by
113 complete linkage was applied on the resulting matrix of root-mean-square deviation (RMSD)
114 values and clusters selected where all pairs of peptides were within 3 Å RMSD of each other.
115

116 ### D. Structural properties of sites

117 Protein secondary structure assignments were obtained from DSSP[25]. DSSP annotates 7 different
118 secondary structure states: 3_{10} helix (G), α helix (H), π helix (I), bends (S), turns (T), isolated (B)
119 and extended (E) β-bridge. These assignments were reduced to three states: G and H to helices
120 (H); I, B and E to strands (E); and all other, including residues with no assignment, to coils (C)[26].
121 The solvent accessible area from DSSP was normalised by the residue's maximum accessible
122 area[27]. A S/T was considered exposed if its relative solvent accessibility (RSA) was $> 25\%$;
123 partially buried if the RSA $> 5\%$ and $\leq 25\%$, and buried if RSA $\leq 5\%$. Cα B-factors were
124 standardised (Z-score normalised) over the B-factors for all Cα in the same chain.
125

126 ### E. Prediction of protein disorder and secondary structure

127 Protein secondary structure predictions for the proteins in the MSS dataset were performed by
128 JPred4[28]. Since JPred4 limits sequence longer than 800 residues, 300 of the sequences in the
129 MSS dataset sequences were trimmed while ensuring the modified S/T was at least 100 residues
130 away from the N- and C-termini to avoid edge effects. The intrinsic disorder was predicted by
131 JRonn (Java implementation of Ronn[29]), IUPred[30] and DisEMBL[31] through the JABAWS[32]
132 command line application. Between them, these methods provide 6 different disorder prediction
133 scores: DisEMBL-REM465 (0.6), DisEMBL-COILS (0.516), DisEMBL-HOTLOOPS (0.1204),
134 IUPred-Long (0.5), IUPred–Short (0.5) and JRonn (0.5). The score ordered/disordered classes

were defined by the cut-offs (in parenthesis) defined by the methods' authors. Disorder predictions were also performed on a background set of 1,164 S/T selected at random from globular proteins in the Astral dataset[33] version 2.04, referred to hereafter as the "Globular Set" (GS).

F. Statistical analysis and code

The data collection, processing, analysis and the Cα clustering steps, were written in the Python programming language (Python Software Foundation, version 2.7 http://www.python.org) with the libraries Pandas (version 0.17)[34] and Biopython (version 1.65)[35]. Statistical tests were performed with the StatsModels (version 0.6) and Scipy (version 0.16) libraries. A p value (p) threshold was set to 0.05.

III. Results and Discussion

A. Analysis of O-GlcNAc sites in proteins of known three-dimensional structure

Previous reports have suggested that O-GlcNAc sites, like phosphorylation sites, are predominantly present in disordered regions of proteins[36]. One indication of structural disorder is the crystallographic B-factor which indicates regions of the protein that lack crystallographic contacts. However, the standardised B-factor distribution on the SS143 dataset is the same for modified and unmodified S/T (Kruskal-Wallis two-sample test $p = 0.12$).

In X-ray crystal structures, the REM465 residue annotation indicates residues that are missing from the protein structure model and has previously been used as an indicator of structural

157 disorder[31]. Of the 143 S/T in the SS143 dataset, 26 are in regions of the protein structure labelled
158 as REM465. In comparison, 553 of 4,811 unmodified S/T from the same protein structures are
159 also found in REM465 regions. Accordingly, O-GlcNAcylated S/T in these proteins are 1.7
160 times more likely to be in REM465 regions (Fisher's exact test $p = 0.02$). This finding is
161 consistent with O-GlcNAcylated S/T occurring more frequently in disordered or highly flexible
162 regions.
163 Table 2 summarises the DSSP assigned secondary structure for the SS143 compared to the 4,811
164 unmodified S/T in the same proteins. The proportions of H, E and C are equivalent for the two
165 groups implying that there is no preference in the secondary structure for modified S/T in this
166 dataset.
167 Residues that are buried in the protein structure are not thought to be targeted by protein kinases,
168 due to structural constraints. Figure 3 illustrates that there is no difference between modified and
169 unmodified S/T with respect to relative solvent accessibility (RSA). 45% (65) of S/T in the O-
170 GlcNAcylated proteins are exposed to solvent (RSA > 25%). Surprisingly, 7 *O*-GlcNAc sites,
171 listed in Table 2, have an RSA < 5%, suggesting they are inaccessible to OGT in the natively
172 folded protein.
173

174 **B. Groups of sites with similar local structure**

175 Since the secondary structure and relative accessibility of modified S/T were indistinguishable
176 from unmodified S/T, the local structure of the 7 residue peptides centred on S/T was
177 investigated by pairwise superposition and clustering (see Methods). 36 sites produce singlet
178 clusters, while the remaining 96 sites fall into 10 clusters. Sites in clusters had less than 3 Å
179 RMSD from each other. Figure 4 illustrates the superimposed structures for sites in clusters,

where green, yellow and grey represent residues in H, E, C secondary structures, respectively. The clusters show that sites are found in a wide range of secondary structure states as summarised in Appendix 1. The sites in Clusters E, G and J, have consistent consensus secondary structures. Clusters A–D, F, H and I are all variants on coil-helix or coil-strand transitions.

The buried sites, which are listed in Table 3, group in clusters D and G. The 3 sites in cluster D are unlikely to be targeted by OGT because they are buried in the protein core. In contrast, the 2/4 sites in cluster G (structures 3abm and 4y7y) might be modified since are located at a dimer interface, and so the monomer could be modified. The remaining two sites in cluster G (structures 2zxe and 4l3j) lie on a loop that could potentially move to expose them to OGT.

To see if the clusters found for the SS132 dataset are features of *O*-GlcNAc modification or just reflect the composition of the protein structures, 132 sites, centred on unmodified S/T, were randomly sampled with replacement from the same proteins and clustered. The process was repeated 1,000 times and the resulting clusters compared to those clusters in the SS132 dataset. The number of clusters identified in each sample ranged from 10-14 (95% CI), which is consistent with the SS132 dataset. Furthermore, the structural clusters identified for the random sampling included structural clusters similar to those for the modified sites, suggesting there are no dominant secondary structural or conformational patterns indicative of *O*-GlcNAc modified sites in the SS132 dataset.

C. Analysis of features predicted for the "Modified Sequence Sites" dataset (MSS)

Since the structural analysis of *O*-GlcNAc sites is limited by the number of sites in proteins of known three-dimensional structure, prediction algorithms were applied to the sequences in the MSS and USS datasets, as detailed in Methods. The proportions of S/T in the levels of solvent accessibility predicted by JPred are equivalent in the MSS and USS datasets, as shown in Table 4. 1% of the S/T are predicted to be buried in the MSS and USS datasets. Again, the result is unexpected, since sites modified by PTM are thought to be accessible in the protein native fold.

While the structural sites in the SS143 dataset have equal proportions of the secondary structure states, the result from secondary structure predictions on the MSS set showed that *O*-GlcNAc sites are likely to reside in coils, if compared to the USS dataset.

Table 5 shows an increase of the proportion of modified S/T in C ($p < 0.01$) and a corresponding reduction in H ($p < 0.01$), but no change in E ($p = 0.6$). The enrichment of sites in C is consistent with the need to place modified S/T in loops that are more likely to be mobile and so more accessible to OGT. The proportions of secondary structure assigned by DSSP and predicted by JPred4 differ. While secondary structure prediction has limited accuracy, the number of samples in the SS143 dataset is limited and potentially biased toward structured regions in proteins. Also, clustering sites in the SS132 dataset highlight groups that are more likely to occur near to the transition between a secondary structure element and C, as observed in several members of clusters A–D, F and H. The regions of transition between C and H/E are harder to predict than contiguous secondary structure elements, and this may also contribute to the observed enrichment in C.

The analysis of SS143 dataset showed an enrichment of S/T in REM465 regions likely to be disordered or highly mobile. To explore this further, 3 disorder prediction algorithms, giving a total of 6 disorder scores, were run on the MSS and USS datasets as detailed in Methods. Table 6 shows that, with the exception of DisEMBL-HOTLOOPS which is trained structural B-factors, all methods report a small but significant increase in mean predicted disorder for the modified S/T. To confirm this result, the MSS dataset was compared to the GS dataset, which was selected from proteins known to be predominantly globular, and hence an ordered background. In Figure 5, DisEMBL-HOTLOOPS shows an increase in the ratio of disordered residues around the modified S/T. DisEMBL-COILS and JRonn also indicate a small increase, not in a specific region, but rather for 40 residues around the S/T. IUpred-Long, IUPred-Short and DisEMBL-REM465 show a bigger increase of the ratio of disordered residues in the MSS dataset and IUpred-Short and REM465 have a clearer peak within -15 to 15 residues from the modified S/T. Overall, all methods indicate an increased proportion of predicted disorder in the MSS dataset when compared to the GS dataset.

IV. Conclusions and final remarks

Despite the substantial evidence of protein structural disorder in the MSS and the SS143 datasets, the SS132 dataset clearly indicates that some of the examined sites appear within ordered regions of the protein structure. Furthermore, InterproScan[37] analysis of *O*-GlcNAc sites assigned 19 % of the sites to protein domains, this is similar to with the 25 % phosphoserines and phosphothreonines in PFAM domains[14, 38], which are thought to be mostly ordered by definition. So, like protein phosphorylation, *O*-GlcNAcylated S/T are found in both ordered and disordered regions.

The local tertiary structure of *O*-GlcNAc sites is indistinguishable from unmodified sites. So, how does OGT recognise the site it modifies? OGT may force the unfolding of the targeted substrate[39]. Moreover, OGT participates in macromolecular assemblies[40], and so adaptor proteins should be important. Non-local interactions might also act in OGT substrate recognition. In protein kinase C (PKC) substrate recognition, residues distant in the protein sequence but close in its three-dimensional structure are critical[41]. Other components, such as UDP–GlcNAc concentration and subcellular location-dependent interactions, modulate OGT activity[42], but their part in substrate recognition is still unknown. In conclusion, although no three-dimensional fingerprint was detected during the structural characterisation of OGT-modified sites, the work confirmed that S/T and surrounding residues are more disordered than the backgrounds tested and that sites in transition between C to H/E might be involved, suggesting that the structural flexibility has a role on OGT site recognition.

Acknowledgements

We would like to thank Dr. Tom Walsh and the University of Dundee IT department for computing support; Prof. Daan van Aalten and DVA group for advice and discussions.

Funding

This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES process 1529/12-9; studentship to T.B.B).

Conflict of Interest:

None declared.

References

1. Shafi R, Iyer SP, Ellies LG, O'Donnell N, Marek KW, Chui D, Hart GW, Marth JD. The O-GlcNAc transferase gene resides on the X chromosome and is essential for embryonic stem cell viability and mouse ontogeny. Proc Natl Acad Sci U S A 2000;97(11):5735–5739.
2. O'Donnell N, Zachara NE, Hart GW, Marth JD. Ogt-Dependent X-Chromosome-Linked Protein Glycosylation Is a Requisite Modification in Somatic Cell Function and Embryo Viability. Mol Cell Biol 2004;24(4):1680–1690.
3. Buse MG. Hexosamines, insulin resistance, and the complications of diabetes: current status. Am J Physiol Endocrinol Metab 2006;290(1):E1–E8.
4. Abdel Rahman AM, Ryczko M, Pawling J, Dennis JW. Probing the hexosamine biosynthetic pathway in human tumor cells by multitargeted tandem mass spectrometry. ACS Chem Biol 2013;8(9):2053–2062.
5. Liu J, Marchase RB, Chatham JC. Increased O-GlcNAc levels during reperfusion lead to improved functional recovery and reduced calpain proteolysis. Am J Physiol Heart Circ Physiol 2007;293(3):H1391-9.
6. McClain DA, Lubas WA, Cooksey RC, Hazel M, Parker GJ, Love DC, Hanover JA. Altered glycan-dependent signaling induces insulin resistance and hyperleptinemia. Proc Natl Acad Sci U S A 2002;99(16):10695–10699.
7. Mariappa D, Zheng X, Schimpl M, Raimi O, Ferenbach AT, Müller H-AJ, Aalten DMF van. Dual functionality of O -GlcNAc transferase is required for Drosophila development. Open Biol 2015;5(12):150234.
8. Lynch TP, Ferrer CM, Jackson SR, Shahriari KS, Vosseller K, Reginato MJ. Critical Role of O-Linked -N-Acetylglucosamine Transferase in Prostate Cancer Invasion, Angiogenesis, and Metastasis. J Biol Chem 2012;287(14):11070–11081.
9. Liu F, Iqbal K, Grundke-Iqbal I, Hart GW, Gong C-X. O-GlcNAcylation regulates phosphorylation of tau: A mechanism involved in Alzheimer's disease. Proc Natl Acad Sci 2004;101(29):10804–10809.
10. Chen D, Juárez S, Hartweck L, Alamillo JM, Simón-Mateo C, Pérez JJ, Fernández-Fernández MR, Olszewski NE, García JA. Identification of secret agent as the O-GlcNAc transferase that participates in Plum pox virus infection. J Virol 2005;79(15):9381–9387.
11. Kuo M, Zilberfarb V, Gangneux N, Christeff N, Issad T. O-glycosylation of FoxO1 increases its transcriptional activity towards the glucose 6-phosphatase gene. FEBS Lett 2008;582(5):829–834.
12. Wells L. Mapping Sites of O-GlcNAc Modification Using Affinity Tags for Serine and Threonine Post-translational Modifications. Mol Cell Proteomics 2002;1(10):791–804.
13. Parker GJ, Lund KC, Taylor RP, McClain DA. Insulin resistance of glycogen synthase mediated by o-linked N-acetylglucosamine. J Biol Chem 2003;278(12):10022–10027.
14. Hornbeck P V, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res 2015;43(Database issue):D512-20.
15. Roquemore EP, Dell A, Morris HR, Panico M, Reason AJ, Savoy LA, Wistow GJ, Zigler JS, Earles BJ, Hart GW. Vertebrate lens alpha-crystallins are modified by O-linked N-acetylglucosamine. J Biol Chem 1992;267(1):555–563.
16. Hahne H, Gholami AM, Kuster B, Moghaddas Gholami A, Kuster B. Discovery of O-

GlcNAc-modified proteins in published large-scale proteome data. Mol Cell Proteomics 2012;11(10):843–850.

17. Ma J, Hart GW. O-GlcNAc profiling: from proteins to proteomes. Clin Proteomics 2014;11(1):8.
18. Wang J, Torii M, Liu H, Hart GW, Hu Z. dbOGAP - An Integrated Bioinformatics Resource for Protein O-GlcNAcylation. BMC Bioinformatics 2011;12(1):91.
19. Jia C-Z, Liu T, Wang Z-P. O-GlcNAcPRED: a sensitive predictor to capture protein O-GlcNAcylation sites. Mol Biosyst 2013;9(11):2909–2913.
20. Jochmann R, Holz P, Sticht H, Stürzl M. Validation of the reliability of computational O-GlcNAc prediction. Biochim Biophys Acta - Proteins Proteomics 2014;1844(2):416–421.
21. Lazarus MB, Nam Y, Jiang J, Sliz P, Walker S. Structure of human O-GlcNAc transferase and its complex with a peptide substrate. In: Nature. Volume 469. Nature Publishing Group; 2011. p 564–567.
22. Schimpl M, Zheng X, Borodkin VS, Blair DE, Ferenbach AT, Schüttelkopf AW, Navratilova I, Aristotelous T, Albarbarawi O, Robinson D a, Macnaughtan M a, Aalten DMF van. O-GlcNAc transferase invokes nucleotide sugar pyrophosphate participation in catalysis. Nat Chem Biol 2012;8(12):969–974.
23. Velankar S, Alhroub Y, Alili A, Best C, Boutselakis HC, Caboche S, Conroy MJ, Dana JM, Ginkel G van, Golovin A, Gore SP, Gutmanas A, Haslam P, Hirshberg M, John M, Lagerstedt I, Mir S, Newman LE, Oldfield TJ, Penkett CJ, Pineda-Castillo J, Rinaldi L, Sahni G, Sawka G, Sen S, Slowley R, Sousa da Silva AW, Suarez-Uruena A, Swaminathan GJ, Symmons MF, Vranken WF, Wainwright M, Kleywegt GJ. PDBe: Protein Data Bank in Europe. Nucleic Acids Res 2011;39(Database issue):D402-10.
24. Velankar S, Dana JM, Jacobsen J, Ginkel G van, Gane PJ, Luo J, Oldfield TJ, O'Donovan C, Martin M-J, Kleywegt GJ. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res 2013;41(Database issue):D483-9.
25. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–2637.
26. Cuff J a, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. In: Proteins. Volume 34. 1999. p 508–519.
27. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. In: Proteins. Volume 40. John Wiley & Sons, Inc.; 2000. p 502–511.
28. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: A protein secondary structure prediction server. Nucleic Acids Res 2015;43(W1):W389-94.
29. Yang ZR, Thomson R, McNeil P, Esnouf RM. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 2005;21(16):3369–3376.
30. Dosztányi Z, Csizmók V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 2005;347(4):827–839.
31. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein Disorder Prediction. Structure 2003;11(11):1453–1459.
32. Troshin P V, Procter JB, Barton GJ. Java bioinformatics analysis web services for multiple sequence alignment--JABAWS:MSA. Bioinformatics 2011;27(14):2001–2002.
33. Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins--

extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 2014;42(Database issue):D304-9.

34. McKinney W, Team PD. Pandas - Powerful Python Data Analysis Toolkit. In: Pandas - Powerful Python Data Analysis Toolkit. 2015. p 1625.

35. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, Hoon MJL de. Biopython: freely available Python tools for computational molecular biology and bioinformatics. In: Bioinformatics. Volume 25. 2009. p 1422–1423.

36. Trinidad JC, Barkan DT, Gulledge BF, Thalhammer a., Sali a., Schoepfer R, Burlingame a. L. Global Identification and Characterization of Both O-GlcNAcylation and Phosphorylation at the Murine Synapse. Mol Cell Proteomics 2012;11(8):215–229.

37. Zdobnov EM, Apweiler R. InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics 2001;17(9):847–848.

38. Beltrao P, Albanèse V, Kenner LR, Swaney DL, Burlingame A, Villén J, Lim WA, Fraser JS, Frydman J, Krogan NJ. Systematic Functional Prioritization of Protein Posttranslational Modifications. Cell 2012;150(2):413–425.

39. Pathak S, Alonso J, Schimpl M, Rafie K, Blair DE, Borodkin VS, Schüttelkopf AW, Albarbarawi O, Aalten DMF van. The active site of O-GlcNAc transferase imposes constraints on substrate sequence. Nat Struct Mol Biol 2015;22(9):744–750.

40. Wells L, Kreppel LK, Comer FI, Wadzinski BE, Hart GW. O-GlcNAc Transferase Is in a Functional Complex with Protein Phosphatase 1 Catalytic Subunits. J Biol Chem 2004;279(37):38466–38470.

41. Duarte ML, Pena DA, Nunes Ferraz FA, Berti DA, Paschoal Sobreira TJ, Costa-Junior HM, Abdel Baqui MM, Disatnik MM-H, Xavier-neto J, Lopes de Oliveira PS, Schechtman D, Augusto F, Ferraz N, Berti DA, José T, Sobreira P, Costa-Junior HM, Muhammad M, Baqui A, Disatnik MM-H, Xavier-neto J. Protein folding creates structure-based, noncontiguous consensus phosphorylation motifs recognized by kinases. Sci Signal 2014;7(350):ra105-ra105.

42. Nagel AK, Ball LE. O-GlcNAc transferase and O-GlcNAcase: achieving target substrate specificity. Amino Acids 2014;46(10):2305–2316.

43. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004;14(6):1188–1190.

Tables

Table 1 Dataset summary. See Methods for details.

Dataset name	Number of sites	Number of proteins	Short name
Modified Sequence Sites	1,385	620	MSS
Unmodified Sequence Sites	100,329	620	USS
Structural Sites	143	106	SS143
Structural Sites with backbone	132	93	SS132
Globular Set	1,164	1,164	GS

Table 2 DSSP assigned secondary structure proportion of S/T in the SS143 dataset compared to unmodified S/T in same protein chains. 95% CI – 95% confidence interval; n – number of S/T. The *p* value refers to the two-tailed z-score test between the proportions of modified and unmodified groups.

	Modified		Unmodified		*p* value
Secondary structure	Proportion (n)	95% CI [lower, upper]	Proportion (n)	95% CI [lower, upper]	
C	0.55 (78)	[0.46, 0.63]	0.51 (2475)	[0.50, 0.53]	0.36
H	0.25 (36)	[0.18, 0.32]	0.32 (1525)	[0.31, 0.33]	0.06
E	0.20 (29)	[0.13, 0.27]	0.17 (811)	[0.16, 0.18]	0.27
Total	143		4,811		

Table 3 Structural evidence of buried *O*-GlcNAc sites in the SS143 dataset. RSA – site mean relative solvent accessibility; Cluster id – Clusters in Figure 4.

PDB id	Chain	Position	Cluster id	RSA
1f4j	B	114	D	0.05
3cb2	B	170	D	0.02
4qvp	T	131	D	0.01
2zxe	A	366	G	0.02
3abm	R	63	G	0.01
4l3j	A	180	G	0.01
4y7y	Z	190	G	0.04

Table 4 JPred4 predicted solvent accessibility for S/T in the MSS and USS datasets. The proportions of buried S/T as predicted by the Jnetsol method in JPred4. The proportions of buried S/T are significantly smaller for modified group. 95% CI – 95% confidence interval; n – number of S/T predicted to be buried. The *p* value refers to the two-tailed z-score test between the modified and unmodified groups.

	Modified (MSS)		**Unmodified (USS)**		***p* value**
Buried at	Proportion (n)	95% CI [lower, upper]	Proportion (n)	95% CI [lower, upper]	
0%	0.01 (7)	[0.00, 0.01]	0.01 (836)	[0.008, 0.009]	0.18
5%	0.04 (55)	[0.03, 0.05]	0.04 (3,917)	[0.038, 0.040]	0.86
25%	0.29 (403)	[0.27, 0.31]	0.35 (28,044)	[0.27, 0.28]	0.31

Table 5 JPred4 predicted secondary structure proportions for S/T in the MSS and USS datasets. 95% CI – 95% confidence interval; n – the number of S/T; the *p* value refers to the two-tailed z-score test between the modified and unmodified groups.

	Modified (MSS)		**Unmodified (USS)**		***p* value**
Secondary structure	Proportion (n)	95% CI [lower, upper]	Proportion (n)	95% CI [lower, upper]	
C	0.88 (1,205)	[0.86, 0.90]	0.829 (83,150)	[0.826, 0.831]	<0.01
H	0.08 (106)	[0.07, 0.09]	0.126 (12,684)	[0.124, 0.128]	<0.01
E	0.05 (66)	[0.04, 0.06]	0.045 (4,495)	[0.044, 0.046]	0.6

Table 6 Predicted disorder between modified and unmodified S/T. All disorder prediction methods, excepting DisEMBL-HOTLOOPS, reveal a small but significant increase of mean disorder score for modified S/T over unmodified ones. The *p* value refers to the two-tailed t-test between the modified and unmodified groups. SE – standard error.

Method	**Mean score modified (MSS) ± SE**	**Mean score unmodified (USS) ± SE**	***p* value**
DisEMBL-REM465	0.48 ± 0.004	0.47 ± 0.001	0.01
DisEMBL-COILS	0.60 ± 0.004	0.58 ± 0.001	<0.01

DisEMBL-HOTLOOPS	0.10 ± 0.001	0.10 ± 0.001	0.45
IUpred-Long	0.59 ± 0.006	0.55 ± 0.001	<0.01
IUpred-Short	0.48 ± 0.005	0.45 ± 0.001	<0.01
JRonn	0.62 ± 0.004	0.61 ± 0.001	0.02

427

428

429

Legend to figures

Figure 1: Sequence relative entropy of sites (+/- 7 residues) from 4 post-translational modifications. Three kinases with most sites in PhosphoSitePlus database[14]: protein kinase A (PKA with 1285 sites), protein kinase C (PKC with 930 sites) and casein kinase 2 (CK2 with 742 sites). 1530 OGT sites were compiled from the same database. The sequence relative entropy was calculated with the WebLogo library[43]. Lines show mean relative entropy and the semi-transparent area represents 95% confidence intervals. Unlikely the PKA, PKC and CK2, known OGT sites have no clear sequence consensus.

Figure 2: Diagram of the relationships among the 5 datasets used in this work. See Methods for details.

Figure 3: RSA of modified S/T in the SS143 dataset and unmodified S/T in same proteins. DSSP calculated solvent accessibility was normalised by the residue theoretical maximum accessibility, and the derived scores were reduced to three levels: buried ($RSA \leq 0.05$), partially buried ($0.05 < RSA \leq 0.25$) and exposed ($RSA > 0.25$). The y-axis and x-axis carry the RSA levels and the RSA distribution for each level, respectively. The mean RSA is equivalent between modified and unmodified residues, at all three levels.

Figure 4: Structural superimpositions for the 10 clusters comprising 96 sites in the SS132 dataset. Pairs of sites were superimposed on their 7 Cα atoms and the Cβ of the central S/T. Their pairwise RMSD were clustered with complete linkage and Euclidean distance. Clusters were defined by a 3 Å RMSD threshold. Green, yellow and grey represent residues in H, E, C secondary structures elements, respectively.

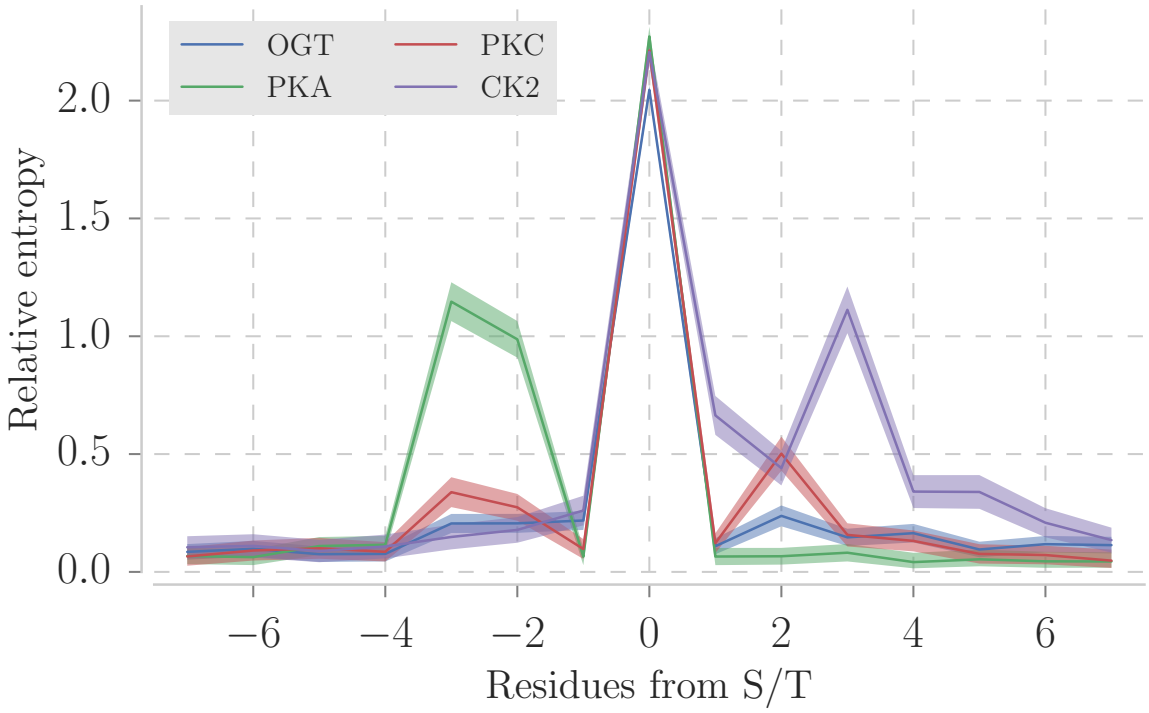
Figure 5: Predicted disorder around *O*-GlcNAc sites in the MSS compared randomly selected S/T in the GS-dataset. The y-axis shows the $\log_{10}$ odds ratio of the between the proportion of

453 disordered residues in the MSS dataset and the proportion of disordered residues in the GS
454 dataset. The semi-transparent area represents 95% confidence intervals. A residue was defined as
455 disordered given a method threshold. The x-axis represents the distance in residues of the central
456 residue, always a S/T. DisEMBL-REM465, IUpred-short predict protein structural disorder
457 specifically around the modification site, while the other methods detect intrinsic disorder over
458 O-GlcNAcylated proteins. DisEMBL-HOTLOOPS shows a less pronounced increase compared
459 to the other methods.

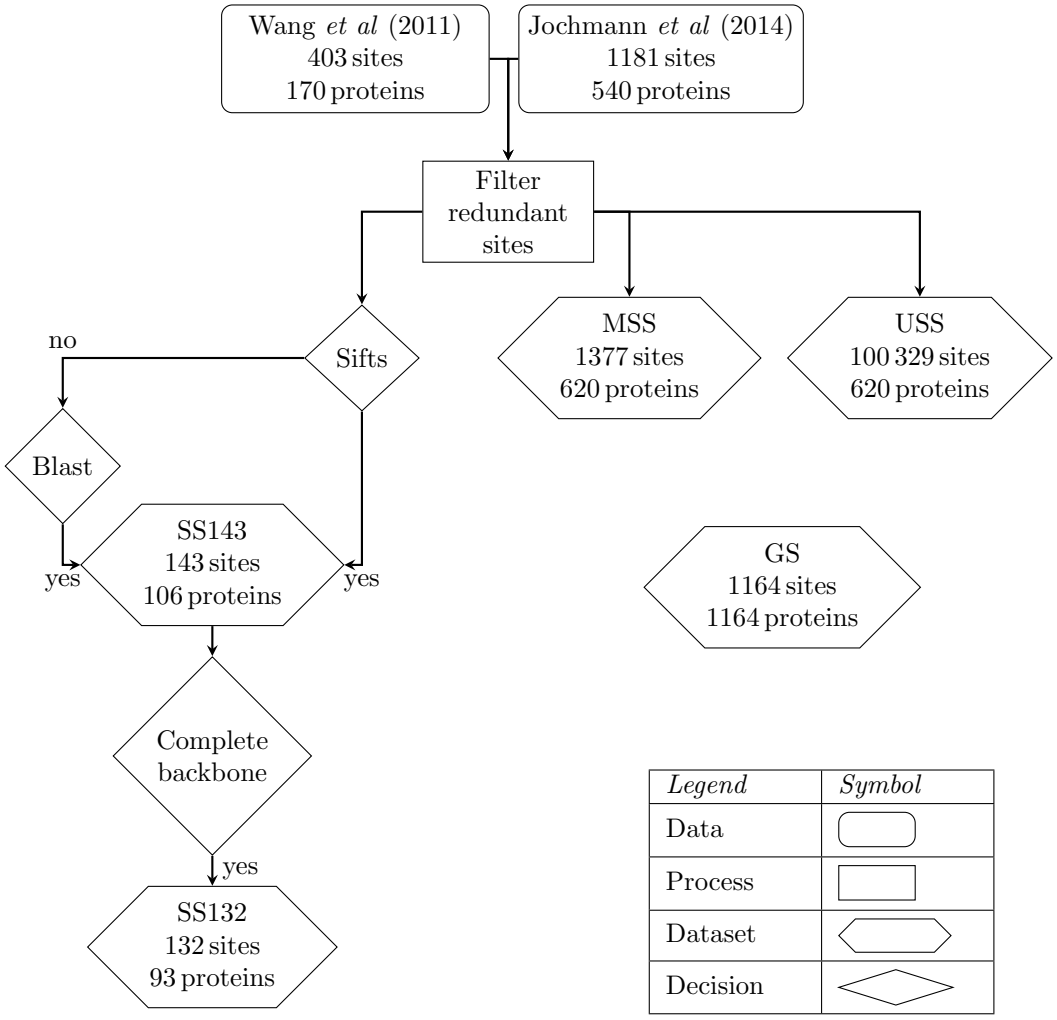
460

Figures

Figure 1



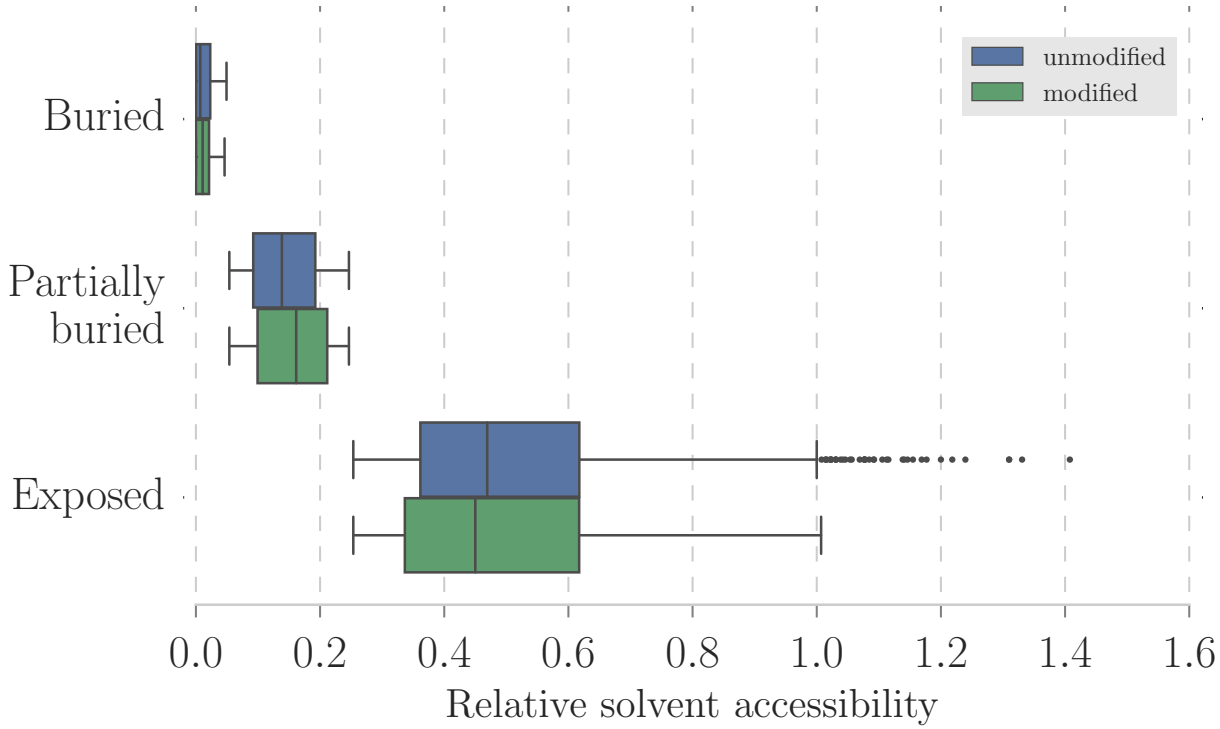
475 Figure 2



476

477

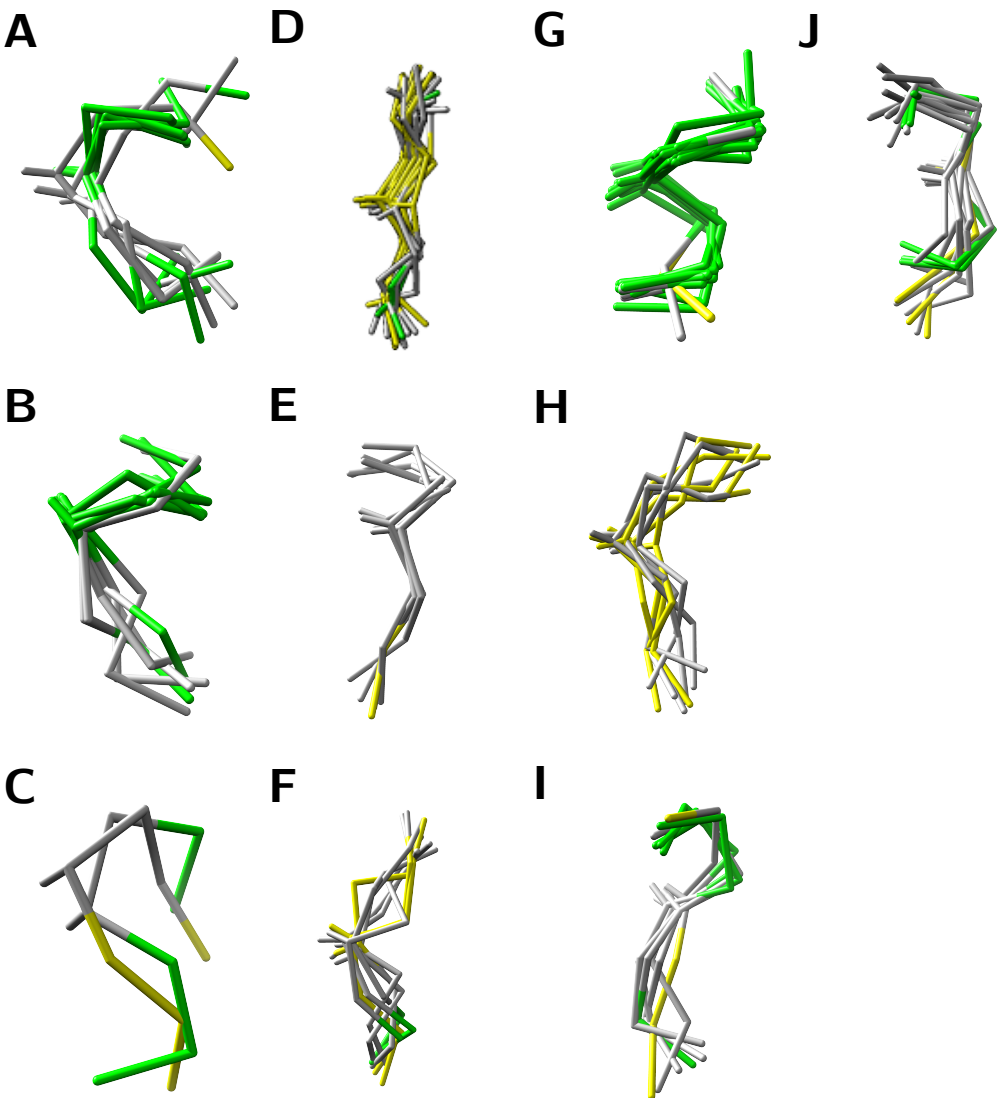
478 Figure 3



479

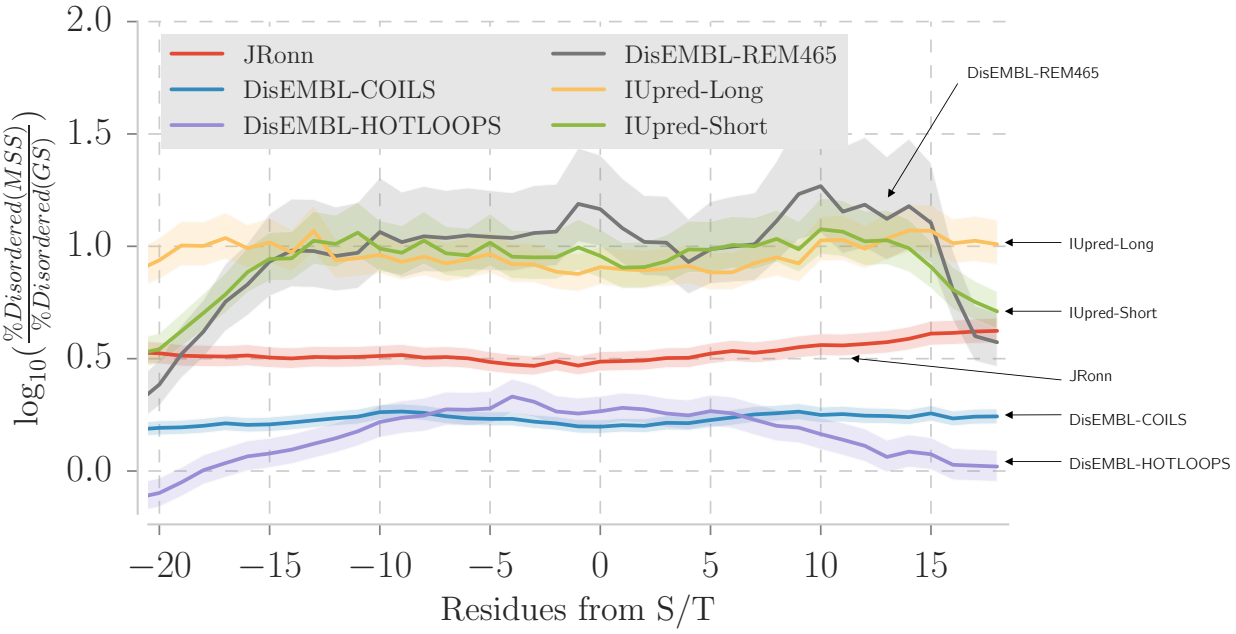
480

481 Figure 4



482

483 Figure 5



484