PhosIDP: a web tool to visualize the location of phosphorylation sites in disordered regions

Sonia T. Nicolaou[1,2], Max Hebditch[1], Owen J. Jonathan[1], Chandra S. Verma[2,3,4], Jim Warwicker[1,*]

[1]School of Biological Sciences, Faculty of Biology, Medicine and Health, Manchester Institute of Biotechnology, University of Manchester, Manchester M1 7DN, UK
[2]Bioinformatics Institute, Agency for Science, Technology, and Research (A*STAR), Singapore 138671, Singapore [3]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551 [4]Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543

* jim.warwicker@manchester.ac.uk

Abstract

Charge is a key determinant of intrinsically disordered protein (IDP) and intrinsically disordered region (IDR) properties. IDPs and IDRs are enriched in sites of phosphorylation, which alters charge. Visualizing the degree to which phosphorylation modulates the charge profile of a sequence would assist in the functional interpretation of IDPs and IDRs. PhosIDP is a web tool that shows variation of charge and fold propensity upon phosphorylation. In combination with the displayed location of protein domains, the information provided by the web tool can lead to functional inferences for the consequences of phosphorylation. IDRs are components of many proteins that form biological condensates. It is shown that IDR charge, and its modulation by phosphorylation, is more tightly controlled for proteins that are essential for condensate formation than for those present in condensates but inessential.

Introduction

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) of proteins have been challenging the traditional structure-function paradigm over the last two decades. They exhibit a range of conformations, from molten globules to random coils [1], with their net charge correlated to their conformational preference [2,3]. The absence of distinct structure in IDPs/IDRs can be attributed to their high net charge and low hydrophobicity [4]. The flexible nature of IDPs/IDRs allows them to be easily regulated by post-translational modifications (PTMs) [5]. Phosphorylation is a post-translational modification enriched in IDPs/IDRs [6], that alters the charge of serine, threonine and tyrosine amino acids by substituting a hydroxy functional group with a negatively charged phosphate group. Phosphorylation plays a crucial role in many biological processes, where it can modify charge and hydrophobicity, and modulate interactions with partners [7].

According to the polyelectrostatic model, interconverting conformations of charged IDPs and/or IDRs bind to their partners through an average electrostatic field caused by long-range electrostatic interactions [8-10]. Phosphorylation can modify the electrostatic interactions of disordered regions by altering their charge which can either enhance or reduce their binding affinity for binding partners [10]. Conformational ensembles of IDPs/IDRs depend on the balance of all charge interactions in the protein, rather than individual short-range interactions between the binding site of the protein and a binding partner. Therefore, they can interact with partners through several distinct binding motifs and/or conformations, and various functional elements will switch between availability for interaction and burial [10].

It is becoming clear that phase-separating proteins and other molecules, such as mRNA, mediate a variety of biological functions [11]. Proteins containing IDRs are commonly found in membraneless organelles, formed by liquid-liquid phase separation (LLPS) [12]. Messenger RNAs are also commonly sequestered during LLPS, typically alongside proteins with IDRs and specific RNA-binding motifs [12], so specific and non-specific charge-charge interactions are likely to be involved. Phosphorylation (and therefore charge variation) is a known control mechanism in LLPS formation [13]. Databases of proteins observed to associate with membraneless organelles are being collated. Notably, the DrLLPS resource [14] classifies proteins undergoing LLPS as core (localised proteins that have been verified as essential for granule assembly/maintenance), client (localised but non-essential), and regulator (contribute to modulation of LLPS but are not located in the condensate).

PhosIDP is an addition to the protein-sol software suite [15], created for visualizing the effects of phosphorylation on protein sequence charge profiles, and thereby facilitating functional hypotheses. In order to assist in understanding the behaviour of IDPs/IDRs, the web tool visualizes changes resulting from phosphorylation across a sequence, rather than focusing on the consequences of individual phosphorylation sites. Further, offline calculations for datasets of proteins are presented that demonstrate the role of phosphorylation in modulating charge for IDRs in core LLPS proteins, but not client proteins.

Results

Phosphorylation mediates substantial charge alteration in the IDR of CIRBP

As an exemplar of proteins of interest, human cold-inducible RNA binding protein (CIRBP, UniProt [16] ID Q14011) is used, in particular as a representative of a core set of proteins that are crucial for the formation of stress granules (SGs), examples of liquid-liquid phase separated membraneless organelles [17]. CIRBP possesses a structured N-terminal RNA recognition motif (RRM) that mediates specific RNA interactions, and a C-terminal disordered arginine/glycine-rich (RGG) region that is involved in weak multivalent RNA interactions [17]. Both the positive charge and disorder of the unphosphorylated RGG region are apparent in Fig. 1 (panels A and B). It is known that phosphorylation can regulate the phase transitions of stress granules [13]. Here, the incorporation of phosphorylation (as recorded in UniProt) shows that the C-terminal end of the disordered region of CIRBP becomes substantially more negatively charged, noting that the N-terminal end of the disordered region remains positively charged.

Nucleophosmin, a protein with order-disorder transitions coupled to phosphorylation

The potential of our server to highlight the role phosphorylation could play in structural transitions is demonstrated with human nucleophosmin, which (in common with CIRBP) is classified as a core protein in regard to formation of membraneless organelles [14]. Nucleophosmin shuttles between the nucleus and cytoplasm, undertaking multiple roles [18]. The current focus is on how prediction of structured and unstructured regions in the phosIDP server, and the presence of phosphorylation sites, correlates with segments for which structural flexibility is known to correlate with function. The core domain is structurally polymorphic, with pentamer formation dependent on phosphorylation status [19]. Phosphorylation biases towards monomer over pentamer, and this balance is also influenced by ligand binding, contributing to the rich functional properties of nucleophosmin

[20]. Only a part of the folded core domain is predicted as structured, using sequence-based fold propensity (Fig. 2). Colour-coding is common between the predicted structured/unstructured plot (Fig. 2B) and the core domain structure (Fig. 2E, 4n8m [19]), with the predicted structured and unstructured regions forming the two parts of the monomer. Extensive interactions of the monomer within the pentamer (Fig. 2E) are consistent with the 3D structural stability being dependent upon oligomerisation. Furthermore, the localisation of phosphorylation sites at the monomer interface likely reflects their role in the monomer – pentamer equilibrium [19]. Therefore, if a user were studying the phosIDP server results, in the absence of known 3D structure, a reasonable prediction would be that conformation predicts only as weakly folded, and that phosphorylation alters the structured/unstructured balance, thereby potentially mediating function. Such a hypothesis could then be subject to the types of experimental analysis that have been applied to investigate the nucleophosmin core domain.

The other domain known to be structured for nucleophosmin, and also detected with the phosIDP sequence analysis, lies at the C-terminus (CTD, Fig. 2 panels B, D, F). The CTD is predicted as relatively weakly folded, with colour-coding transferred from sequence prediction to the NMR structure (2vxd [21]), consistent with observation that its folding stability is significantly lower than that of the pentamer unit formed by the core domain, as measured by thermal and chemical denaturation [22]. Indeed, mutation of the CTD has been associated with loss of structure and amyloid formation, possibly related to pathogenesis [23,24]. These results again indicate the utility of the server for identifying regions that are predicted to be of relatively low folded state stability. With regard to phosphorylation, notable amongst CTD sites (Fig. 2F) is the location of Ser243 at the start of the first helix in the CTD, and lying in the portion predicted (from sequence, coloured blue) to have only marginal folded state stability. Since favourable interactions arising from a phosphorylated helix N-cap are greater than any amino acid N-cap [25], it is possible that phosphorylation at Ser243 is coupled with CTD folding stability and thereby with biological activity.

Charge distributions and phosphorylation in stress granule proteins

Whilst proteins termed core, such as CIRBP and nucleophosmin, are central to SG formation, client proteins can be found in SGs but do not themselves cause formation of SGs [14]. Having found in the two examples studied that SG core protein phosphorylation is extensive, we wondered how core protein charge compares with that of client proteins. In order to study protein regions rather than the overall charge of the protein, and examine local effects, charge in overlapping 30 amino acid windows was summed. Further this was divided into regions predicted to be structured or intrinsically disordered, according to the

fold propensity results of the server. Distributions of the windowed net charge values are shown for IDRs of core and client proteins, and their phosphorylated counterparts are compared (Fig. 3). Interestingly, unphosphorylated core proteins tend towards overall charge neutrality more than client proteins, but with a distinct enrichment for positive charge. Upon phosphorylation the enrichment for positive charge is largely removed, perhaps indicative (on average) of reduced RNA interaction and a tendency towards SG dissolution more so than formation, although both behaviors have been observed experimentally, dependent on the system [13]. Although the calculations presented in Fig. 3 are offline, they demonstrate the relevance of studying charge distributions and how they change with phosphorylation.

Discussion

Previous work has demonstrated the importance of net charge and also the balance between positive and negative charge in the analysis of IDR conformation and function [26]. Here, the crucial role that phosphorylation plays in many systems, not only through site-specific interactions, but also with global switching of charge profiles, can be quickly and conveniently studied with phosIDP, subject to data collation in the UniProt database. Two examples were taken from a dataset of 'core' proteins that are integral for membraneless organelle formation. For CIRBP (Fig. 1), phosphorylation of a disordered RGG domain, summed over several sites, reduces the net positive charge close to neutrality, similar to the effect seen overall in IDR regions of the core protein set (Fig. 3). Given the mRNA localisation in membraneless organelles, it is apparent that charge modulation could fine tune stability. In the second example, nucleophosmin (Fig. 2), phosphorylation is enriched close to the boundaries of predicted structured and disordered regions, which for the oligomerisation domain relates to structural transitions that are known to be modulated by phosphorylation. We suggest that the phosIDP web tool will be particularly useful when looking for these regions of predicted weak disorder or structure, and the effect of phosphorylation, which can in turn be matched to available experimental data or lead to experiment design.

Methods

Using the PhosIDP web tool

As part of the protein-sol software suite [15], PhosIDP is freely available at https://protein-sol.manchester.ac.uk/phosidp without requiring a license or registration. In order for the

software to execute the user needs to provide a valid UniProt ID [16]. Results are displayed graphically with additional information provided below the plots. Data is also available to download as text. The custom URL is available for 7 days from the day of creation.

The phosIDP algorithm

Upon receiving a UniProt protein ID [16], phosIDP calculates charge at neutral pH as a sequence-based profile, averaged over a sliding window of 21 amino acids [15], with a second charge profile added that corresponds to the addition of a double negative charge for each phosphorylation site recorded in UniProt (Fig. 1A). A hydropathy scale [27] is combined with net charge to predict IDRs [4,28], in a scheme repurposed from the existing implementation on protein-sol [15]. The scale for fold propensity (Fig. 1B) extends from positive/folded to negative/unfolded. An additional feature of the folded state prediction is a sawtooth window that smooths structured and intrinsically disordered regions. Short stretches of IDR within ordered regions may legitimately describe the behavior of loops, but could also obscure the overall domain structure. The sawtooth envelope is created from the predicted profile with the caveat that disordered regions with fewer than 10 amino acids in length take on the average prediction for a window of up to 41 amino acids centered on that region. Additionally, predicted IDRs separated by structured regions of fewer than 10 amino acids are combined into one longer IDR. For reference, phosphorylation sites, as curated by UniProt (Fig. 1C), and Pfam domains [29] (Fig. 1D) are displayed.

Datasets of proteins in membraneless organelles

Three publically available databases were studied in order to generate a dataset of proteins that have been located in membraneless organelles. MSGP [30] is a database of proteins in mammalian stress granules, and PhaSepDB [31] contains proteins that undergo LLPS in various organelles. Each of these have overlap with DrLLPS [14], with the advantage of DrLLPS recording proteins as core, client, or regulator. Since our aim was to compare core and client proteins, DrLLPS was used to generate sets of 56 core proteins and 728 client proteins.

Data availability

The reported web tool is freely available online. The data associated with LLPS proteins can be provided upon request.

References

1 Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6**, 197-208, doi:10.1038/nrm1589 (2005).

2 Yamada, J. *et al.* A bimodal distribution of two distinct categories of intrinsically disordered structures with separate functions in FG nucleoporins. *Mol Cell Proteomics* **9**, 2205-2224, doi:10.1074/mcp.M000035-MCP201 (2010).

3 Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci U S A* **110**, 13392-13397, doi:10.1073/pnas.1304749110 (2013).

4 Uversky, V. N., Gillespie, J. R. & Fink, A. L. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41**, 415-427 (2000).

5 Forman-Kay, J. D. & Mittag, T. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure* **21**, 1492-1499, doi:10.1016/j.str.2013.08.001 (2013).

6 Iakoucheva, L. M. *et al.* The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* **32**, 1037-1049, doi:10.1093/nar/gkh253 32/3/1037 (2004).

7 Bah, A. & Forman-Kay, J. D. Modulation of Intrinsically Disordered Protein Function by Post-translational Modifications. *J Biol Chem* **291**, 6696-6705, doi:10.1074/jbc.R115.695056 (2016).

8 Borg, M. *et al.* Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proc Natl Acad Sci U S A* **104**, 9650-9655 (2007).

9 Serber, Z. & Ferrell, J. E., Jr. Tuning bulk electrostatics to regulate protein function. *Cell* **128**, 441-444, doi:10.1016/j.cell.2007.01.018 (2007).

10 Mittag, T., Kay, L. E. & Forman-Kay, J. D. Protein dynamics and conformational disorder in molecular recognition. *J Mol Recognit* **23**, 105-116 (2010).

11 Sehgal, P. B., Westley, J., Lerea, K. M., DiSenso-Browne, S. & Etlinger, J. D. Biomolecular condensates in cell biology and virology: Phase-separated membraneless organelles (MLOs). *Anal. Biochem.* **597**, 113691, doi:10.1016/j.ab.2020.113691 (2020).

12 Zhou, H. X., Nguemaha, V., Mazarakos, K. & Qin, S. Why Do Disordered and Structured Proteins Behave Differently in Phase Separation? *Trends Biochem Sci* **43**, 499-516, doi:10.1016/j.tibs.2018.03.007 (2018).

13 Hofweber, M. & Dormann, D. Friend or foe-Post-translational modifications as regulators of phase separation and RNP granule dynamics. *J Biol Chem* **294**, 7137-7150, doi:10.1074/jbc.TM118.001189 (2019).

14 Ning, W. *et al.* DrLLPS: a data resource of liquid-liquid phase separation in eukaryotes. *Nucleic Acids Res* **48**, D288-D295, doi:10.1093/nar/gkz1027 (2020).

15 Hebditch, M., Carballo-Amador, M. A., Charonis, S., Curtis, R. & Warwicker, J. Protein-Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics* **33**, 3098-3100, doi:10.1093/bioinformatics/btx345 (2017).

16 UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506-D515, doi:10.1093/nar/gky1049 (2019).

17 De Leeuw, F. *et al.* The cold-inducible RNA-binding protein migrates from the nucleus to cytoplasmic stress granules by a methylation-dependent mechanism and acts as a translational repressor. *Exp. Cell Res.* **313**, 4130-4144, doi:10.1016/j.yexcr.2007.09.017 (2007).

18 Cela, I., Di Matteo, A. & Federici, L. Nucleophosmin in Its Interaction with Ligands. *Int J Mol Sci* **21**, doi:10.3390/ijms21144885 (2020).
19 Mitrea, D. M. *et al.* Structural polymorphism in the N-terminal oligomerization domain of NPM1. *Proc Natl Acad Sci U S A* **111**, 4466-4471, doi:10.1073/pnas.1321007111 (2014).
20 Banerjee, P. R., Mitrea, D. M., Kriwacki, R. W. & Deniz, A. A. Asymmetric Modulation of Protein Order-Disorder Transitions by Phosphorylation and Partner Binding. *Angew Chem Int Ed Engl* **55**, 1675-1679, doi:10.1002/anie.201507728 (2016).
21 Grummitt, C. G., Townsley, F. M., Johnson, C. M., Warren, A. J. & Bycroft, M. Structural consequences of nucleophosmin mutations in acute myeloid leukemia. *J Biol Chem* **283**, 23326-23332, doi:10.1074/jbc.M801706200 (2008).
22 Marasco, D. *et al.* Role of mutual interactions in the chemical and thermal stability of nucleophosmin NPM1 domains. *Biochem Biophys Res Commun* **430**, 523-528, doi:10.1016/j.bbrc.2012.12.002 (2013).
23 Di Natale, C. *et al.* Nucleophosmin contains amyloidogenic regions that are able to form toxic aggregates under physiological conditions. *FASEB J.* **29**, 3689-3701, doi:10.1096/fj.14-269522 (2015).
24 Russo, A. *et al.* Insights into amyloid-like aggregation of H2 region of the C-terminal domain of nucleophosmin. *Biochim Biophys Acta Proteins Proteom* **1865**, 176-185, doi:10.1016/j.bbapap.2016.11.006 (2017).
25 Andrew, C. D., Warwicker, J., Jones, G. R. & Doig, A. J. Effect of phosphorylation on alpha-helix stability as a function of position. *Biochemistry* **41**, 1897-1905 (2002).
26 Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. & Pappu, R. V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys J* **112**, 16-21, doi:10.1016/j.bpj.2016.11.3200 (2017).
27 Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105-132 (1982).
28 Prilusky, J. *et al.* FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21**, 3435-3438 (2005).
29 El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427-D432, doi:10.1093/nar/gky995 (2019).
30 Nunes, C. *et al.* MSGP: the first database of the protein components of the mammalian stress granules. *Database (Oxford)* **2019**, doi:10.1093/database/baz031 (2019).
31 You, K. *et al.* PhaSepDB: a database of liquid-liquid phase separation related proteins. *Nucleic Acids Res* **48**, D354-D359, doi:10.1093/nar/gkz847 (2020).
32 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).

Acknowledgements

We thank members of the Warwicker and Verma groups for providing feedback on the web page content and layout. The authors would like to acknowledge that this work has been supported by the University of Manchester and the Agency for Science, Technology, and

Research (A*STAR) Singapore, and by the UK EPSRC (grant EP/N024796/1). CSV thanks A*STAR for grants (grant IDs H17/01/a0/010, IAF111213C, H18/01/a0/015).

Author contributions statement

STN, MH, OJ, and JW conducted the studies and analysed the results. All authors were involved in conceiving the work and reviewing the manuscript.

Competing interests

CSV is the founder of Sinopsee Therapeutics and Aplomex. The current work has no conflict with the companies. STN, MH, OJ, and JW declare no conflict of interest.

Figure Legends

Figure 1. Results of phosIDP server for CIRBP (UniprotKB: Q14011), a stress granule-associated protein. (A) and (B) Changes to charge and fold propensity upon phosphorylation are depicted in lighter shades. The sawtooth window for smoothing disordered and ordered regions is shown in blue. (C) Phosphorylation sites from UniProt are displayed. (D) Location of known Pfam domains.

Figure 2. Sequence and structure of human nucleophosmin (UniprotKB: P06748). Windowed charge scores (A), predicted fold propensities (B), phosphorylation sites (C), and Pfam domains (D) are shown, for unmodified and phosphorylated nucleophosmin. The Pfam domains correspond with predicted structured domains, for which solved structure is available, and labelled core domain and C-terminal (C-term) domain. (E) Using the same colour-coding as fold propensity prediction (yellow/folded and blue/unfolded), a cartoon of a core domain monomer is shown against a surface of the remaining 4 monomers in a pentamer unit (PDB [32] id 4nm8 [19]). Known phosphorylation sites are indicated with green spacefill and residue numbers. (F) Also with the yellow and blue colour-coding for predicted structural order from the phosIDP server, the C-terminal domain is drawn (first model from 2vxd [21]). Known phosphorylation sites for this sequence are shown, along with an indication of the N- and C-terminal ends of the domain.

Figure 3. Core and client (and phosphorylated counterparts, denoted P) protein charge distributions, calculated from overlapping 30 amino acid windows within predicted IDRs.

Figure 1

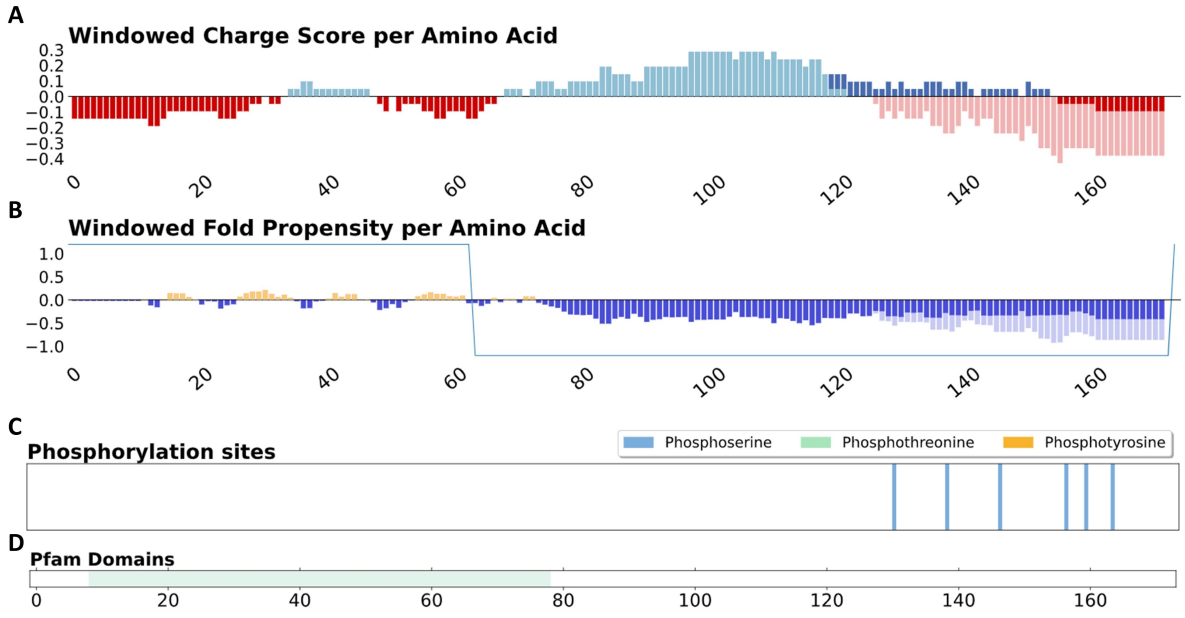

Figure 2

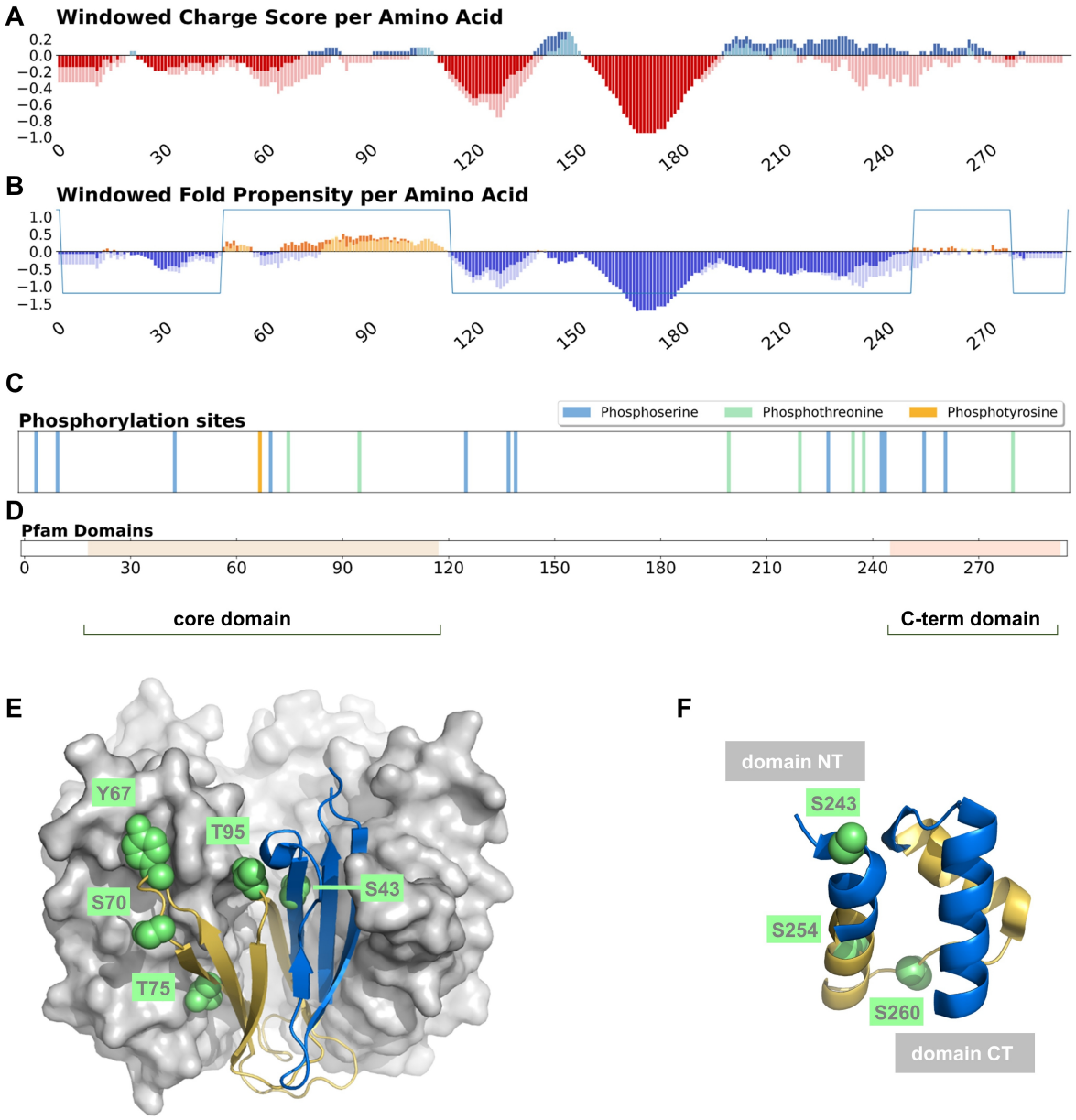

Figure 3