Prediction of quality-control degradation signals in yeast proteins

Kristoffer E. Johansson[1], Bayan Mashahreh[2], Rasmus Hartmann-Petersen[1], Tommer Ravid[2,*], and Kresten Lindorff-Larsen[1,*]

1: Linderstrøm-Lang Centre for Protein Science, Section for Biomolecular Sciences, Department of Biology, University for Copenhagen, Copenhagen, Denmark

2: Department of Biological Chemistry, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

* To whom correspondence may be addressed: tommer.ravid@mail.huji.ac.il, lindorff@bio.ku.dk

Abstract

Effective proteome homeostasis is key to cellular and organismal survival, and cells therefore contain efficient quality control systems to monitor and remove potentially toxic misfolded proteins. Such general protein quality control to a large extent relies on the efficient and robust delivery of misfolded or unfolded proteins to the ubiquitin-proteasome system. This is achieved via recognition of so-called degradation motifs—degrons—that are assumed to become exposed as a result of protein misfolding. Despite their importance, the nature and sequence properties of quality-control degrons remain elusive. Here, we have used data from a yeast-based screen of 23,600 17-residue peptides to build a predictor of quality-control degrons. The resulting model, QCDPred (Quality Control Degron Prediction), achieves good accuracy using only the sequence composition of the peptides as input. Our analysis reveals that strong degrons are enriched in hydrophobic amino acids and depleted in negatively charged amino acids, in line with the expectation that they are buried in natively folded proteins. We applied QCDPred to the entire yeast proteome, enabling us to analyse more widely the potential effects of degrons. As an example, we show a correlation between cellular abundance and degron potential in disordered regions of proteins. Together with recent results on membrane proteins, our work suggest that the recognition of exposed hydrophobic residues is a key and generic mechanism for proteome homeostasis.

Introduction

Most proteins are required to fold into their native three-dimensional structures to carry out their cellular functions. Mutations or environmental stress conditions may, however, lead to partial unfolding or misfolding of these natively folded proteins, or to disassembly of complexes. Within the cell, such structurally unstable or misfolded proteins may form aggregates and often display proteotoxicity via poorly understood mechanisms. Cells are therefore equipped with an effective protein quality control (PQC) system that catalyses the refolding or degradation of aberrant proteins to sustain proteostasis (Clausen et al., 2019; Hipp et al., 2019; Rosenzweig et al., 2019). The degradative route of the eukaryotic PQC system often relies on the ubiquitin-proteasome system to clear the cell of non-native proteins. Here, E3 ubiquitin-protein ligases, frequently assisted by molecular chaperones, first conjugate the non-native proteins to ubiquitin-chains, which in turn target the proteins for proteasomal degradation (Abildgaard et al., 2020; Alberti et al., 2003; Enam et al., 2018; Jones et al., 2020; Kriegenburg et al., 2014).

Many E3s are involved in specific cellular functions and are highly specific enzymes (Deshaies and Joazeiro, 2009; Zheng and Shabek, 2017). As virtually all cellular proteins may become misfolded, it is, however, essential that the PQC-linked degradation system is not too specific so that it can degrade a wide range of misfolded substrates, but simultaneously leaves native proteins intact. The exact substrate preferences of the PQC E3s are largely unknown, and accordingly, the discriminating feature(s) of misfolded PQC targets are unknown and presently not possible to predict. E3s recognize regions in target proteins termed degradation signals or degrons (Guharoy et al., 2016; Ravid and Hochstrasser, 2008). The so-called primary degron interacts—directly or indirectly via adaptor proteins—with the E3, and thus determines specificity. In addition, a nearby lysine residue that acts as the acceptor site for ubiquitylation, the so-called secondary degron, is also required. Finally, a partially unfolded or disordered region in the substrate protein, termed the tertiary degron, is required for the proteasome to commence threading the protein into the particle for degradation (Prakash et al., 2004). We here focus on the properties of primary PQC degrons that determine the specificity in PQC-linked proteasomal degradation. Despite several efforts (Geffen et al., 2016; Maurer et al., 2016), the discerning features of the PQC degrons remain rather enigmatic. These degrons are, however, expected to include hydrophobic areas that are buried in the native protein but become exposed upon unfolding or misfolding.

Previously, Geffen et al. used a Ura3-GFP reporter system to map large numbers of degradation signals in *Saccharomyces cerevisiae* (Geffen et al., 2016). These experiments revealed a large breadth in the sequences that cause degradation without any obvious individual motifs. More recently, we used the same reporter system to show that Hsp70-binding motifs may function as PQC degrons (Abildgaard et al., 2021). Specifically, we found that inserting known Hsp70 binding motifs into the reporter system led to Hsp70-dependent degradation.

In a new set of experiments, we extended these ideas to map more systematically degrons in *Saccharomyces cerevisiae* (Mashahreh et al., 2022). Specifically, we used a fluorescence-based reporter system coupled to fluorescence-activated cell sorting (FACS) and sequencing to study the degradation of 23,600 17-residue long peptides. Here, we use this data to study the properties of the degrons and build a prediction method for PQC degrons. Our analysis confirms our previous observation that groups of hydrophobic residues—often buried inside the core of folded proteins—can act as strong degrons when exposed to the PQC system. We also show that negatively charged residues counteract the degradation signal, potentially explaining why these are common in many disordered proteins. Finally, we show that the

presence of PQC degrons in disordered regions of proteins may influence their cellular abundance.

Results and Discussion

We aimed to develop a prediction algorithm to determine which protein sequences, when exposed to the cytosolic PQC system, can be recognized and targeted for degradation. As input to our model, we used data from an assay in which 23,600 17-residues long peptides from the yeast proteome had been fused to GFP in a reporter construct (Mashahreh et al., 2022). In this construct, peptides with strong degron properties cause degradation of GFP which, when normalized by an expression control, can be distinguished from non-degron peptides using FACS and quantified by sequencing. Details and validation of the experiments are described elsewhere (Mashahreh et al., 2022).

Using the sequencing counts to quantify which peptides lead to low or high levels of GFP, we separate the assayed peptides into six bins based on their protein stability index (PSI) (Fig. 1A). Here a high PSI score indicates peptides associated with high GFP fluorescence (non-degrons), and a low PSI score with peptides that cause decreased GFP fluorescence (degrons). The number between one and four reflects the average position of a peptide among the four bins of the fluorescence distribution of sorted cells. As discussed below and elsewhere (Mashahreh et al., 2022), the decreased GFP signal is in most cases due to proteasomal degradation resulting from the fusion with the peptide. We find that the amino acid composition of the peptides in each of the six bins shows a clear trend towards increased hydrophobicity in peptides that are associated with low GFP signals (likely degrons) and negatively charged side chains in peptides that do not act as degrons (Fig. 1A and Fig. S1).

We explore the predictive power of this observation using a logistic regression model that takes the amino acid composition as input (see Methods). This choice of model is motivated by a bimodal distribution of the PSI scores (Fig. 1A), which we simplify while building the prediction model by dividing peptides into those with degron activity (positive label) and those without (non-degrons or negative label). This model captures the trend well with an area under the curve (AUC) of 0.79 in a receiver operating characteristic (ROC) curve analysis. To further explore which part of the data is best described by this model, we trained three models that each excluded a slice of peptides with intermediate PSI. These showed an increasingly accurate fit to the data with an AUC up to 0.88 and without a decrease in the AUC when the models are evaluated on all data (Fig. S2). Thus, we conclude that peptides with intermediate PSI measurements do not contribute to the model accuracy and retrain a final model without these (Fig. 1B and Methods section). AUC values were retained in 10-fold cross validations. We term the final model QCDPred because the peptides were selected to represent the signals recognized by the PQC system (Mashahreh et al., 2022). Python and R code to run QCDPred is available at https://github.com/KULL-Centre/papers/tree/main/2022/degron-predict-Johansson-et-al.

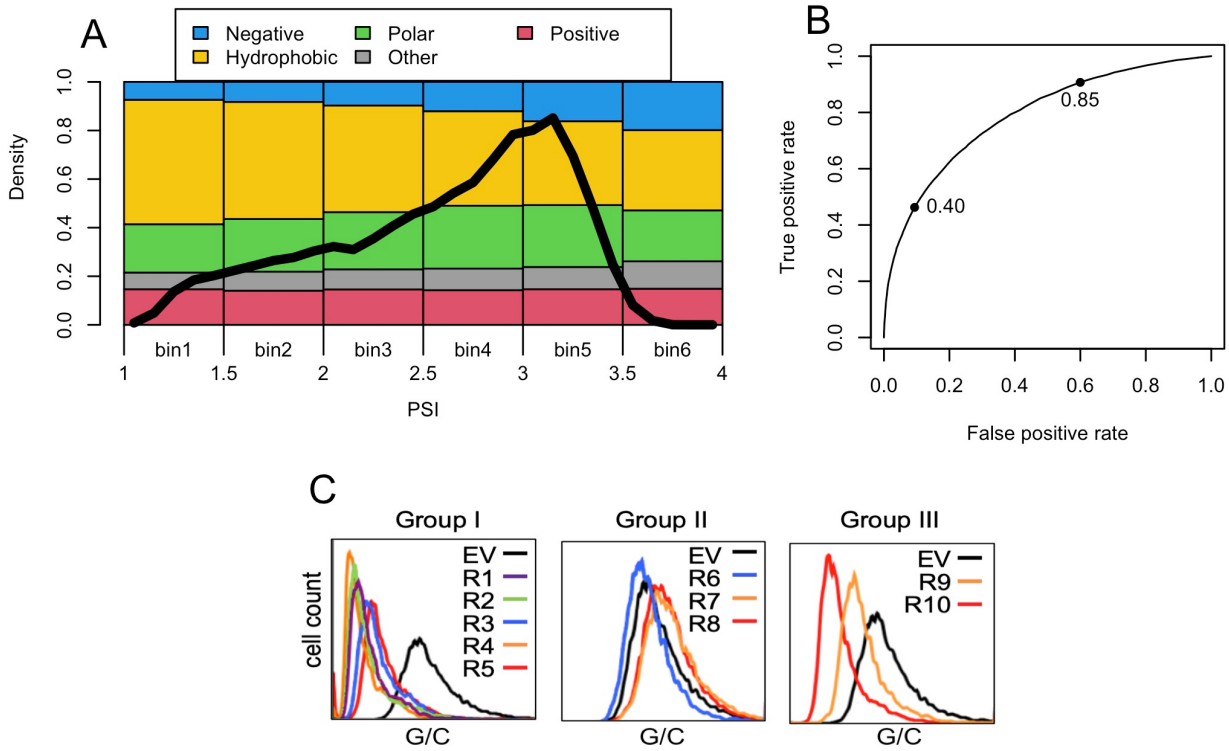


Figure 1. *A) The amino acid composition in each bin shows a trend towards fewer hydrophobic (CAVMLIFYW) and more negatively charged side chains (DE) in stable proteins (high PSI) compared to degrons (low PSI). Polar (NQST), positively charged (HKR) and other amino acids (GP) change less across the PSI distribution. B) A logistic regression model that takes amino acid composition as input captures these trends well with an AUC of 0.79. The two points indicate that applying threshold values of 0.85 or 0.40 give a true-positive rate of 91% or a false positive rate of 9% respectively. C) Flow cytometry distributions for ten peptides predicted to be degrons (group 1) or stable (groups 2 and 3). The green fluorescence is normalized to red fluorescence to correct for plasmid levels (G/C).*

While the AUC evaluations above suggest that most of the information in the experimental data is contained within the overall amino acid composition, we also explored whether a more complex model that use the di-peptide composition as input, could achieve a better fit. Each 17-residue peptide was thus represented by 16 dipeptides, with the idea that these would hold some context information for each amino acid. The accuracy of this model was, however, only slightly better with and AUC of 0.80 for the best model. Furthermore, the parameters of the dipeptide model (Fig. S3) were strongly correlated (Pearson correlation coefficient of 0.88) with the average of the two corresponding single amino acid parameters, suggesting that approximately the same information is contained in the simpler model (Fig. S4). The high accuracy of these amino acid composition models supports the observation that the ability of a peptide to act as a PQC degron is to a large extent determined solely by its amino acid composition and that hydrophobic peptides are strong degrons.

To validate the high-throughput experiment and test QCDPred experimentally, we selected ten peptide sequences that we studied—one at a time—using the same reporter system and flow cytometry readout (Fig. 1C) (Mashahreh et al., 2022). Since these ten peptides originated from yeast proteins, we first retrained our model on a set of peptides without any sequence overlap

to these (see Methods). Five peptides (R1–R5), which we predicted to have degron activity (group I), and three peptides that we predicted not to have strong degrons (R6–R8; group II) were confirmed in this experiment. In contrast, two peptides (R9 and R10) were found to have some degron activity even though they were not predicted as degrons by QCDPred (group 3). We also found that treating the cells with a proteasome inhibitor (Bortezomib) restored GFP fluorescence for the degron sequences, demonstrating that the lowered fluorescence is due to proteasomal degradation, lending further support that QCDPred predicts substrates for PQC degradation (Fig. S5).

The simple logistic regression model used in QCDPred has the advantage of being easily interpretable with the estimated parameters giving the degron contribution of the individual amino acid types—in the context of a 17 amino acid peptide (Fig. 2). The values in our model confirm our initial observation that the hydrophobic side chains in Met, Val, Leu, Ile and Cys, together with the aromatic side chains in Phe, Tyr and Trp, all promote degron activity. In contrast, the negatively charged side chains in Asp and Glu counteract degradation and have a strongly stabilising effect, whereas the positively charged side chains in Lys and—in particular Arg—have a much smaller effects. Asn, Gln, Gly and Pro have intermediate stabilising effects.

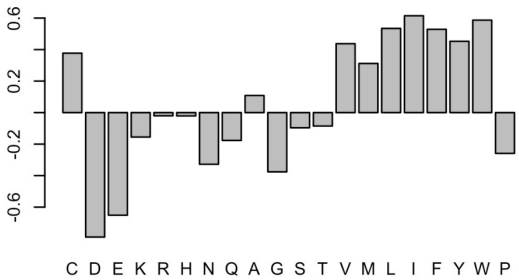


Figure 2. *Estimated parameters for the logistic regression model may be interpreted as the degron propensity of individual amino acid types.*

QCDPred also has the advantage of being fast to evaluate, and we thus calculated a degron probability score for 2.2 million residues in the yeast proteome (Dubreuil et al., 2019) in less than a minute (data available online at https://github.com/KULL-Centre/papers/tree/main/2022/degron-predict-Johansson-et-al). As an initial analysis of this dataset, we examined whether the presence of degrons in proteins might be related to their cellular abundance. We exclude proteins containing trans-membrane regions from this analysis because trans-membrane regions are enriched in hydrophobic amino acids that QCDPred would recognise, although these are unlikely to get exposed to the PQC system of the cytosolic environment (Dubreuil et al., 2019). Indeed, as we show elsewhere aberrant exposure of trans-membrane regions to the PQC leads to degradation (Mashahreh et al., 2022). The disordered regions of the remaining proteins are expected to be in contact with the PQC system whose effect is captured by QCDPred, and we thus hypothesized that the presence of degrons in these sequences would lead to decreased abundance. Indeed, we find that the average degron probability of these disordered residues correlates with protein abundance (Dubreuil et al., 2019; Wang et al., 2015) (Fig. 3a). While other mechanisms have also been suggested to lead to the observed relationship between amino acid composition and abundance (Dubreuil et al., 2019; Macossay-Castillo et al., 2019), our results suggest that QCDPred captures a significant degradation signal and thus a general mechanism of proteolytic stability in yeast cells. In line with this hypothesis that decreased abundance is due to more rapid degradation, we also find a correlation between QCDPred scores and protein half-life (Christiano et al., 2014) (Fig. 3b).

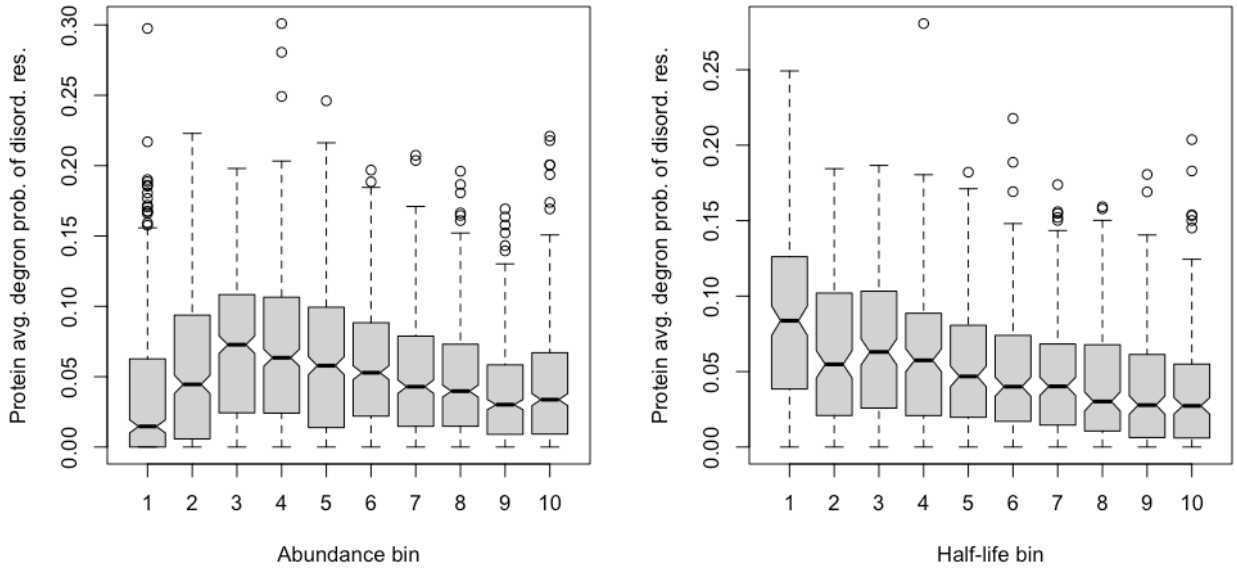


Figure 3. *The average degron probability of disordered residues decreases for proteins of the yeast proteome that are more abundant and has longer half-life.*

We note that the average degron probabilities of disordered regions are in general low, which is expected since strong degradation signals in these regions should cause immediate degradation. We also note that in our analysis of protein abundance the two lowest-abundance bins are exceptions to the overall correlation which might be explained by the low concentrations of these proteins. The protein concentration corresponds to approximately 5-30 nM which is lower than many typical binding affinities (Dubreuil et al., 2019), thus perhaps bypassing the degradation mechanism.

Conclusions

Here we present the—to our knowledge—first prediction model for PQC degrons. This work complements previous work on mapping regulatory degrons, which are typically more specific. Our results confirm previous findings that hydrophobic sequences are recognized and targeted for degradation (Fredrickson et al., 2011; Hickey et al., 2021). Surprisingly, however, we find clear differences between how different polar and charged residues contribute to degrons. Most notably, we find that the presence of the negatively charged side chains in Asp and Glu counteract degradation whereas the positively charged side chains in Lys and Arg have a more neutral effect. These results complement our recent observation that Hsp70 binding motifs, which are often hydrophobic and positively charged, can act as degrons (Abildgaard et al., 2021; Kampmeyer et al., 2022). The effects of hydrophobic, positively and negatively charged residues are also validated by experiments that show that replacing hydrophobic residues with Glu in a degron can prevent degradation, whereas replacing the same amino acids with Arg cannot (Mashahreh et al., 2022).

In the future, it will be interesting to explore the effects of these differences further in the context of cellular functions and evolution. Here we note for example that transactivation domains of transcription factors are enriched in hydrophobic and negatively charged residues that are known to be key for function (Sigler, 1988; Staller et al., 2022). We speculate that the presence and necessity of the negatively charged residues in these may not only be due to direct

effects on function, but also to counteract the higher degradation propensity afforded by the presence of for example many aromatic residues. In addition, it has been found that negatively charged proteins only rarely engage with molecular chaperones (Houben et al., 2020; Koopman and Rüdiger, 2020; Rousseau et al., 2006). More generally, we show a clear relationship between the presence of degrons in disordered regions and the abundance and degradation rates of yeast proteins. In this context we note that human disease mutations in disordered regions are substantially enriched in Glu-to-Lys substitutions, as well as Arg-to-Trp and Arg-to-Cys substitutions (Vacic et al., 2012; Vacic and Iakoucheva, 2012), that we suggest could lead to more rapid degradation of these proteins. Finally, we note that the sequence properties that we have mapped also resemble those of transmembrane regions, and indeed we elsewhere show that when transmembrane regions are mislocalized to the cytosol they are also targeted for degradation (Mashahreh et al., 2022).

In many cases, the effects discussed above can be traced back to differences in hydrophobic content of the amino acid sequences. For example, the overall amino acid compositions of disordered and folded proteins are different (Tompa, 2002), and the depletion of hydrophobic amino acids in disordered proteins underlies their inability to fold into specific three-dimensional structures. Thus, by recognizing hydrophobic sequences, the PQC system can distinguish misfolded globular or transmembrane proteins from for example disordered regions or those on surfaces of folded proteins, and in this way display functional specificity without substantial sequence specificity.

More than half of disease-causing missense variants are thought to lead to protein degradation (Cagiada et al., 2021), but the molecular mechanisms of how these variants are recognized by the cell is still incompletely understood. Our simple, but accurate prediction algorithm to identify degrons from sequence may help understand the complex relationship between sequence variation and abundance. We and others have previously shown that many disease-causing missense variants are thermodynamically destabilized (Abildgaard et al., 2019; Casadio et al., 2011; Gersing et al., 2021; Kampmeyer et al., 2022; Matreyek et al., 2018; Pey et al., 2003; Stein et al., 2019; Yue et al., 2005). The results presented here provide a key missing part of the puzzle, namely a quantitative model for which exposed sequences are PQC targets, and we hope our work may help develop quantitative models that combine protein stability and recognition via the PQC.

While our work is a substantial step forward, there are some limitations. First, in the library we have screened we only vary 17-residue long peptides, and so our model only predicts degrons that fit within this length. More importantly, an unfolded or misfolded protein may have multiple degrons and we have previously shown that an increased number of exposed degrons lead to progressively lower abundance (Abildgaard et al., 2021). Thus, future work should quantify the relationship between the length and number of degrons and protein abundance. Our model also does not provide additional insight into the presence of potentially specific sequence motifs. We believe this is mostly due to two, related effects. First, the PQC system needs to be able to target a wide range of proteins, and thus the PQC may have evolved to recognize generic properties of misfolded proteins such as exposed hydrophobic cores. Second, there is likely extensive redundancy in the components of the PQC system including E3 enzymes, chaperones and adaptor proteins (Breckel and Hochstrasser, 2021; Hickey et al., 2021; Maurer et al., 2016; Samant et al., 2018; Theodoraki et al., 2012). Thus, a more extensive set of experiments using different genetic perturbations are likely needed to disentangle the sequence preferences of different components. Finally, we note that our experiments and analyses are focused on detecting linear degrons. While we expect that these are important in the general PQC system, future experiments and computational analyses are needed to discover

potential PQC degrons that are perhaps localized in a three-dimensional structure but not in the linear polypeptide chain.

In summary, we conclude that the approach we have developed provides a first, key step in understanding and predicting the sequence properties of PQC degrons. We hope that the computational model, QCDPred, the experimental data (Mashahreh et al., 2022) and the application to the yeast proteome will be useful to understand better the determinants of protein abundance and quality control.

Methods

Logistic regression model

For all models, we use the 18,616 most confident peptide measurements with more than 50 sequencing reads across all gates in the flow cytometry experiment (Mashahreh et al., 2022). Here, 4790 peptides with PSI < 2.2 were labelled as degron-active (positive) and 8769 peptides with PSI > 2.8 as degron inactive or stable (negative). Based on a 10-fold cross validation, the strength of the ridge regularization penalty in the logistic model was set to 0.001 and the model was retrained on all the labelled data resulting in the same AUC of 0.85 as during cross validation. To evaluate the model on data that includes unlabelled peptides with intermediate PSI, we labelled these as degrons if PSI > 2.5 and non-degrons otherwise. The same procedure was followed for the dipeptide model using a regularization strength of 1.0. Models were made in *The R project for statistical computing* [R] (R Core Team, 2020) using the *glmnet* package (Friedman et al., 2010). Code and data are available at https://github.com/KULL-Centre/papers/tree/main/2022/degron-predict-Johansson-et-al.

Validation experiment

Ten peptides, R1–R10 (Table S1), were selected for experimental validation using FACS sorting profiles as readout rather than sequencing. These peptides were re-cloned into the parental vector and were tested and analysed as described (Mashahreh et al., 2022). While these peptides were not part of the training data for QCDPred, these peptides have some sequence overlap with the training data. Therefore, we retrained a model excluding the 24 peptides in the original dataset that contain either the first or the last eight residues of R1–R10.

Proteome analysis

For the proteome half-life and abundance analysis, we used a collection of proteins from Dubreuil *et al.* (Dubreuil et al., 2019) annotated with cellular abundance levels from the Pax database (Wang et al., 2015) and residues predicted to be disordered by IUPRED (Dosztanyi et al., 2005). The 30% of residues with highest IUPRED score were assumed to be disordered. This set is particularly relevant here because it excludes proteins with any trans-membrane regions. For 1993 of the 4533 proteins in this set we were able to assign a half-life (Christiano et al., 2014). For each protein we calculated the QCDPred degron probability of each consecutive 17-mer and assigned the value to the central residue position. Finally, for each protein we averaged the degron probability of the positions that are predicted to be disordered. These predictions are available at https://github.com/KULL-Centre/papers/tree/main/2022/degron-predict-Johansson-et-al and predictions for the full yeast proteome are available at https://erda.ku.dk/archives/3f5bf0ada384be330541a3c1b04168d9/uniprot_2022_01_proteome_UP000002311_qcdpred.txt.gz.

Acknowledgments

This work is a contribution from the Novo Nordisk Foundation centre PRISM (NNF18OC0033950; to R.H.-P. and K.L.-L.), and also acknowledges support from a NSF-BSF grant (2016722 ; to R.G.G and T.R). B.M. acknowledges the support of the Neubauer doctoral fellowship fund.

References

1. Abildgaard, A.B., Gersing, S.K., Larsen-Ledet, S., Nielsen, S.V., Stein, A., Lindorff-Larsen, K., and Hartmann-Petersen, R. (2020). Co-Chaperones in Targeting and Delivery of Misfolded Proteins to the 26S Proteasome. Biomolecules. 10.

2. Abildgaard, A.B., Petersen, S.D., Larsen, F.B., Kampmeyer, C., Johansson, K.E., Stein, A., Ravid, T., Andreasson, C., Jensen, M.K., Lindorff-Larsen, K., and Hartmann-Petersen, R. (2021). HSP70-binding motifs function as protein quality control degrons. bioRxiv.

3. Abildgaard, A.B., Stein, A., Nielsen, S.V., Schultz-Knudsen, K., Papaleo, E., Shrikhande, A., Hoffmann, E.R., Bernstein, I., Gerdes, A.M., Takahashi, M., Ishioka, C., Lindorff-Larsen, K., and Hartmann-Petersen, R. (2019). Computational and cellular studies reveal structural destabilization and degradation of MLH1 variants in Lynch syndrome. Elife. 8.

4. Alberti, S., Esser, C., and Höhfeld, J. (2003). BAG-1--a nucleotide exchange factor of Hsc70 with multiple cellular functions. Cell Stress. Chaperones. 8, 225-231.

5. Breckel, C.A., and Hochstrasser, M. (2021). Ubiquitin Ligase Redundancy and Nuclear-Cytoplasmic Localization in Yeast Protein Quality Control. Biomolecules. 11.

6. Cagiada, M., Johansson, K.E., Valanciute, A., Nielsen, S.V., Hartmann-Petersen, R., Yang, J.J., Fowler, D.M., Stein, A., and Lindorff-Larsen, K. (2021). Understanding the Origins of Loss of Protein Function by Analyzing the Effects of Thousands of Variants on Activity and Abundance. Mol. Biol. Evol. 38, 3235-3246.

7. Casadio, R., Vassura, M., Tiwari, S., Fariselli, P., and Luigi, M.P. (2011). Correlating disease-related mutations to their effect on protein stability: a large-scale analysis of the human proteome. Hum. Mutat. 32, 1161-1170.

8. Christiano, R., Nagaraj, N., Fröhlich, F., and Walther, T.C. (2014). Global proteome turnover analyses of the Yeasts S. cerevisiae and S. pombe. Cell Rep. 9, 1959-1965.

9. Clausen, L., Abildgaard, A.B., Gersing, S.K., Stein, A., Lindorff-Larsen, K., and Hartmann-Petersen, R. (2019). Protein stability and degradation in health and disease. Adv. Protein Chem. Struct. Biol. 114, 61-83.

10. Deshaies, R.J., and Joazeiro, C.A. (2009). RING domain E3 ubiquitin ligases. Annu. Rev. Biochem. 78, 399-434.

11. Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics. 21, 3433-3434.

12. Dubreuil, B., Matalon, O., and Levy, E.D. (2019). Protein Abundance Biases the Amino Acid Composition of Disordered Regions to Minimize Non-functional Interactions. J. Mol. Biol. 431, 4978-4992.

13. Enam, C., Geffen, Y., Ravid, T., and Gardner, R.G. (2018). Protein Quality Control Degradation in the Nucleus. Annu. Rev. Biochem. 87, 725-749.

14. Fredrickson, E.K., Rosenbaum, J.C., Locke, M.N., Milac, T.I., and Gardner, R.G. (2011). Exposed hydrophobicity is a key determinant of nuclear quality control degradation. Mol. Biol. Cell 22, 2384-2395.

15. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33, 1-22.

16. Geffen, Y., Appleboim, A., Gardner, R.G., Friedman, N., Sadeh, R., and Ravid, T. (2016). Mapping the Landscape of a Eukaryotic Degronome. Mol. Cell 63, 1055-1065.

17. Gersing, S.K., Wang, Y., Grønbæk-Thygesen, M., Kampmeyer, C., Clausen, L., Willemoës, M., Andréasson, C., Stein, A., Lindorff-Larsen, K., and Hartmann-Petersen, R. (2021). Mapping the degradation pathway of a disease-linked aspartoacylase variant. PLoS. Genet. 17, e1009539.

18. Guharoy, M., Bhowmick, P., and Tompa, P. (2016). Design Principles Involving Protein Disorder Facilitate Specific Substrate Selection and Degradation by the Ubiquitin-Proteasome System. J. Biol. Chem. 291, 6723-6731.

19. Hickey, C.M., Breckel, C., Zhang, M., Theune, W.C., and Hochstrasser, M. (2021). Protein quality control degron-containing substrates are differentially targeted in the cytoplasm and nucleus by ubiquitin ligases. Genetics 217, 1-19.

20. Hipp, M.S., Kasturi, P., and Hartl, F.U. (2019). The proteostasis network and its decline in ageing. Nat. Rev. Mol. Cell Biol. 20, 421-435.

21. Houben, B., Michiels, E., Ramakers, M., Konstantoulea, K., Louros, N., Verniers, J., van der Kant, R., De, V.M., Chicória, N., Vanpoucke, T., Gallardo, R., Schymkowitz, J., and Rousseau, F. (2020). Autonomous aggregation suppression by acidic residues explains why chaperones favour basic residues. EMBO J. 39, e102864.

22. Jones, R.D., Enam, C., Ibarra, R., Borror, H.R., Mostoller, K.E., Fredrickson, E.K., Lin, J., Chuang, E., March, Z., Shorter, J., Ravid, T., Kleiger, G., and Gardner, R.G. (2020). The extent of Ssa1/Ssa2 Hsp70 chaperone involvement in nuclear protein quality control degradation varies with the substrate. Mol. Biol. Cell 31, 221-233.

23. Kampmeyer, C., Larsen-Ledet, S., Wagnkilde, M.R., Michelsen, M., Iversen, H.K.M., Nielsen, S.V., Lindemose, S., Caregnato, A., Ravid, T., Stein, A., Teilum, K., Lindorff-Larsen, K., and Hartmann-Petersen, R. (2022). Disease-linked mutations trigger exposure of a protein quality control degron in the DHFR protein. bioRxiv.

24. Koopman, M.B., and Rüdiger, S.G. (2020). Behind closed gates - chaperones and charged residues determine protein fate. EMBO J. 39, e104939.

25. Kriegenburg, F., Jakopec, V., Poulsen, E.G., Nielsen, S.V., Roguev, A., Krogan, N., Gordon, C., Fleig, U., and Hartmann-Petersen, R. (2014). A chaperone-assisted

382 degradation pathway targets kinetochore proteins to ensure genome stability.
383 PLoS. Genet. 10, e1004140.

384 26. Macossay-Castillo, M., Marvelli, G., Guharoy, M., Jain, A., Kihara, D., Tompa, P., and
385 Wodak, S.J. (2019). The Balancing Act of Intrinsically Disordered Proteins:
386 Enabling Functional Diversity while Minimizing Promiscuity. J. Mol. Biol. 431,
387 1650-1670.

388 27. Mashahreh, B., Armony, S., Johansson, K.E., Chapelbaum, A., Friedman, N., Gardner,
389 R.G., Hartmann-Petersen, R., Lindorff-Larsen, K., and Ravid, T. (2022).
390 Conserved degronome features governing quality control-associated proteolysis.
391 bioRxiv https://doi.org/10.1101/2022.04.06.487275.

392 28. Matreyek, K.A., Starita, L.M., Stephany, J.J., Martin, B., Chiasson, M.A., Gray, V.E.,
393 Kircher, M., Khechaduri, A., Dines, J.N., Hause, R.J., Bhatia, S., Evans, W.E.,
394 Relling, M.V., Yang, W., Shendure, J., and Fowler, D.M. (2018). Multiplex
395 assessment of protein variant abundance by massively parallel sequencing. Nat.
396 Genet. 50, 874-882.

397 29. Maurer, M.J., Spear, E.D., Yu, A.T., Lee, E.J., Shahzad, S., and Michaelis, S. (2016).
398 Degradation Signals for Ubiquitin-Proteasome Dependent Cytosolic Protein
399 Quality Control (CytoQC) in Yeast. G3. (Bethesda.) 6, 1853-1866.

400 30. Pey, A.L., Desviat, L.R., Gámez, A., Ugarte, M., and Pérez, B. (2003). Phenylketonuria:
401 genotype-phenotype correlations based on expression analysis of structural and
402 functional mutations in PAH. Hum. Mutat. 21, 370-378.

403 31. Prakash, S., Tian, L., Ratliff, K.S., Lehotzky, R.E., and Matouschek, A. (2004). An
404 unstructured initiation site is required for efficient proteasome-mediated
405 degradation. Nat. Struct. Mol. Biol. 11, 830-837.

406 32. R Core Team. (2020). R: A language and environment for statistical computing. R
407 Foundation for Statistical Computing, Vienna, Austria. https://www.R-
408 project.org/.

409 33. Ravid, T., and Hochstrasser, M. (2008). Diversity of degradation signals in the ubiquitin-
410 proteasome system. Nat. Rev. Mol. Cell Biol. 9, 679-690.

411 34. Rosenzweig, R., Nillegoda, N.B., Mayer, M.P., and Bukau, B. (2019). The Hsp70
412 chaperone network. Nat. Rev. Mol. Cell Biol. 20, 665-680.

413 35. Rousseau, F., Serrano, L., and Schymkowitz, J.W. (2006). How evolutionary pressure
414 against protein aggregation shaped chaperone specificity. J. Mol. Biol. 355, 1037-
415 1047.

416 36. Samant, R.S., Livingston, C.M., Sontag, E.M., and Frydman, J. (2018). Distinct
417 proteostasis circuits cooperate in nuclear and cytoplasmic protein quality control.
418 Nature 563, 407-411.

419 37. Sigler, P.B. (1988). Transcriptional activation. Acid blobs and negative noodles. Nature
420 333, 210-212.

421 38. Staller, M.V., Ramirez, E., Kotha, S.R., Holehouse, A.S., Pappu, R.V., and Cohen, B.A.
422 (2022). Directed mutational scanning reveals a balance between acidic and
423 hydrophobic residues in strong human activation domains. Cell Syst.

424 39. Stein, A., Fowler, D.M., Hartmann-Petersen, R., and Lindorff-Larsen, K. (2019).
425 Biophysical and Mechanistic Models for Disease-Causing Protein Variants.
426 Trends Biochem. Sci. 44, 575-588.

427 40. Theodoraki, M.A., Nillegoda, N.B., Saini, J., and Caplan, A.J. (2012). A network of
428 ubiquitin ligases is important for the dynamics of misfolded protein aggregates in
429 yeast. J. Biol. Chem. 287, 23911-23922.

430 41. Tompa, P. (2002). Intrinsically unstructured proteins. Trends Biochem. Sci. 27, 527-533.

431 42. Vacic, V., and Iakoucheva, L.M. (2012). Disease mutations in disordered regions--
432 exception to the rule? Mol. Biosyst. 8, 27-32.

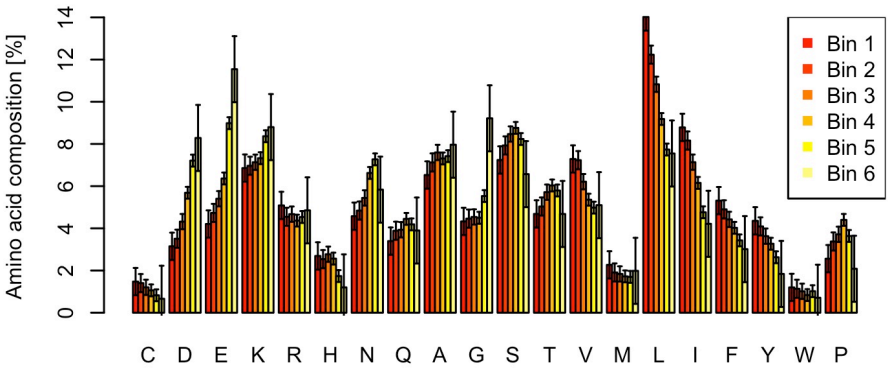
433 43. Vacic, V., Markwick, P.R., Oldfield, C.J., Zhao, X., Haynes, C., Uversky, V.N., and
434 Iakoucheva, L.M. (2012). Disease-associated mutations disrupt functionally
435 important regions of intrinsic protein disorder. PLoS. Comput. Biol. 8, e1002709.

436 44. Wang, M., Herrmann, C.J., Simonovic, M., Szklarczyk, D., and von, M.C. (2015).
437 Version 4.0 of PaxDb: Protein abundance data, integrated across model
438 organisms, tissues, and cell-lines. Proteomics. 15, 3163-3168.

439 45. Yue, P., Li, Z., and Moult, J. (2005). Loss of protein structure stability as a major
440 causative factor in monogenic disease. J. Mol. Biol. 353, 459-473.

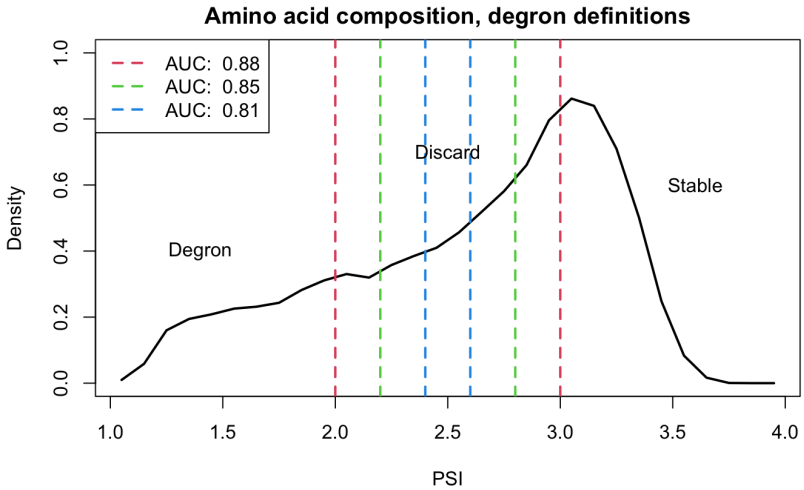
441 46. Zheng, N., and Shabek, N. (2017). Ubiquitin Ligases: Structure, Function, and
442 Regulation. Annu. Rev. Biochem. 86, 129-157.
443

444 Supplementary figures



445

446 **Figure S1.** Composition of each amino acid in the 6 bins across the PSI distribution (Fig. 1).



447

448 ***Figure S2.*** *Peptides with intermediate PSI values do not contribute to model accuracy when*
449 *included in training. We trained three models in which we excluded different sets of peptides*
450 *with intermediate PSI values. The AUCs of the resulting models increase as more peptides*
451 *of intermediate PSI are excluded. When we evaluate these three models on the full dataset*
452 *(that is when also including the intermediate PSI values) we obtain an AUC of 0.79 for all*
453 *three models.*

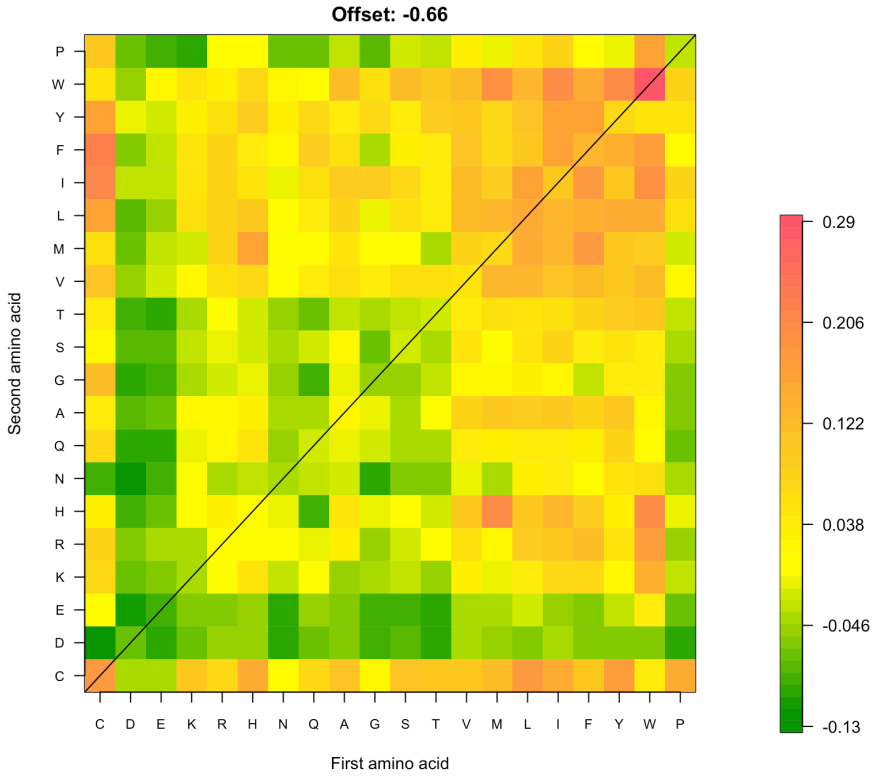


Figure S3. *Estimated parameters for a logistic regression model that uses di-peptide composition as input.*

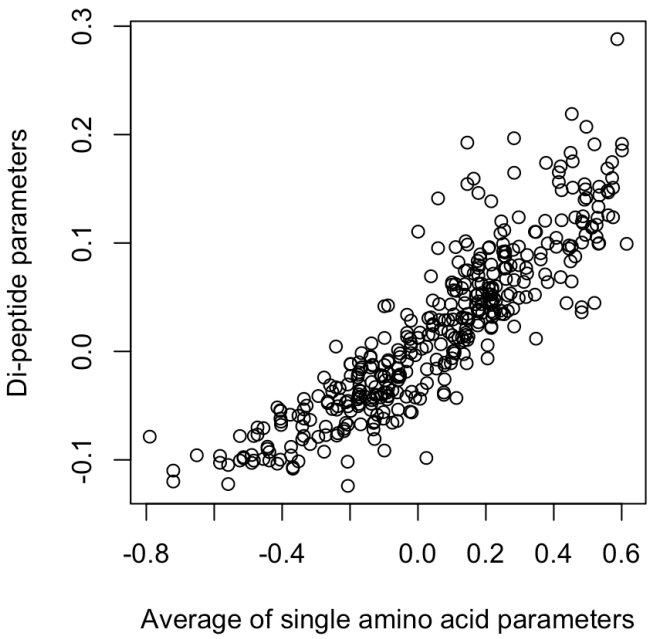
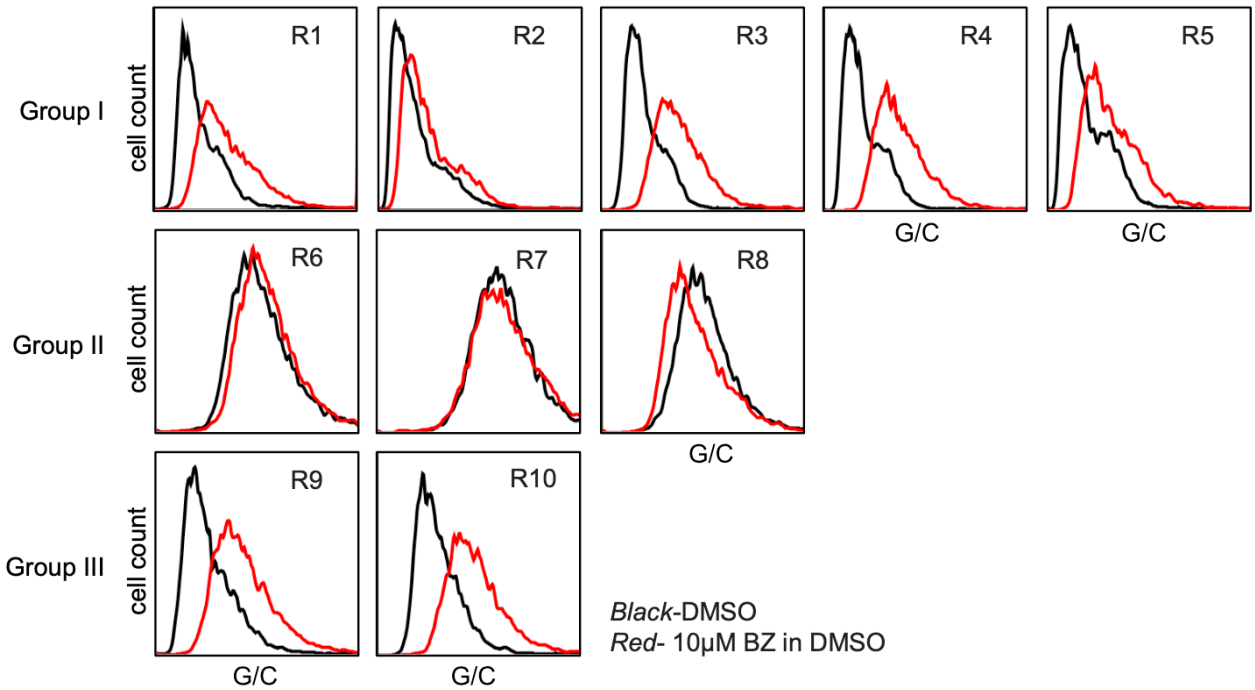


Figure S4*. The parameters estimated parameters in the dipeptide model correlate strongly (Pearson correlation coefficient 0.88) with the average of the two corresponding single amino acid parameters of the single amino acid composition model.*

461



462

463 ***Figure S5.*** *The low cellular abundance of the seven degron peptides (R1–R5, R9, R10)*
464 *increase after treatment with the proteasome inhibitor Bortezomib, confirming that they are*
465 *proteasomal targets.*

Peptide	Sequence	Gene	Start	End	QCDPred score
R1	YLVPFIIAAMVIMHLMA	COB	184	200	0.99
R2	IMRCKIYFFYLRISKKL	PCI8	304	320	0.98
R3	FKRNAVFVGTIFAGAFV	QCR9	11	27	0.89
R4	NKLHYCVSCAIHARIVR	RPS26A	69	85	0.92
R5	SLYLLSIWVKKFKWAGI	CYT1	279	295	0.97
R6	DEPDEQGNPKKRPGKLS	CYT1	136	152	0
R7	SGKSNTSNDASNKKETK	ITC1	301	317	0.01
R8	VIDKLKNENTDLKDIIQ	PCI8	383	399	0.13
R9	LNIRIQTYGETTAVDAL	RPC19	101	117	0.48
R10	KVGSIRVHKSGKLSVKI	RPC53	359	375	0.62

466

467 ***Table S1.*** *Peptides used in the validation experiments. The table shows the sequence of the*
468 *10 peptides, the yeast gene the peptides come from, the start and end amino acid position in*
469 *the protein and the QCDPred score*