Feature analysis and *in silico* prediction of lower solubility proteins in three eukaryotic model systems

by

Gerard Chan

B.Sc. (Hons.) in Life Sciences, National University of Singapore, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Genome Science and Technology)

The University of British Columbia
(Vancouver)

April 2015

Abstract

Regulation of protein solubility, or the ability of proteins to remain soluble within the cell, is an important part of protein homeostasis. This is highlighted with the disruption of protein homeostasis and dysregulation of solubility being associated with various neurodegenerative diseases. Using quantitative mass spectrometry and computational analyses, we identify low solubility proteins under unstressed conditions in three eukaryotic model systems: yeast cells, human neuroblastoma cells, and mouse brain tissue. Using an internal reference, we account for protein abundance, and allow for the analysis of proteins based on their partitioning between the soluble and insoluble fractions, rather than purely on their abundance within the insoluble fraction. We identified several intrinsic traits such as length, disorder, abundance, molecular recognition features, and low complexity regions which are correlated with protein solubility. These features have been previously shown to be associated with protein-protein interactions. This suggests that, under unstressed conditions, lower solubility in proteins may be linked to functional aggregation, rather than aberrant aggregation. We then present two predictors which may be used to predict the *in vivo* solubility of proteins, built using the many traits examined in this work. The linear regression model is able to give estimates of protein solubility, although proteins near the threshold between low and normal solubility may be misclassified. The Support Vector Machine is able to reliably distinguish between low and high solubility proteins, but is unable to reliably distinguish low and normal solubility proteins. We have identified several traits that distinguish low solubility proteins from other proteins, as well as developed two models that are able to estimate the solubility of proteins.

Preface

The majority of the work presented in this thesis has been published in Albu et al., 2014. Mice used in this study were grown by members of the Johnson lab, with brain tissues harvested by Taghizadeh, Hu, and Mehran. Preparation of the human and mouse biological samples and GO analysis were done by Dr. Razvan Albu. Preparation of the yeast biological samples, RNase experiments, Western Blots, and generation of the amino acid compass were performed by Mang Zhu. Source code for the amino acid compass was written by Alex Ng and previously published in Ng et al., 2013. Processing of data for CAI, secondary structure, ANCHOR MoRFs, ELMs, IUPred,Pfam, and disulfide bond prediction were performed by Eric Wong. Analysis of all other protein properties, generation of boxplots, and generation of models were carried out by Gerard Chan.

Table of Contents

List of Tables

List of Figures

Glossary

AAindex amino acid index

AD alzheimer's disease

AUC area under curve

CAI codon adaptation index

ELMs eukarotic linear motifs

FDR false discovery rate

GO gene ontology

GRAVY grand average of hydrophobicity

HA hemagglutinin

HPLC high performance liquid chromatography

HS higher solubility

IPOD insoluble protein deposit

JUNQ juxtanuclear quality control

LC liquid chromatography

LCRs low complexity regions

LPS lowest-probability subsequences

LOWESS locally weighted scatterplot smoothing

LS lower solubility

LTQ linear-trapping quadrupole

MCC matthews correlation coefficient

MoRFs molecular recognition features

MS mass spectrometry

MSE mean squared error

NS normal solubility

PD parkinson's disease

RSPG random sampling of proteins within groups

SGD saccharomyces genome database

SILAC stable isotope labeling by amino acids in cell culture

SVM support vector machine

UPS ubiquitin proteasome system

Acknowledgments

First and foremost, I would like to thank my supervisor Professor Thibault Mayor as well as my committee members Professor Jörg Gsponer and Professor Christopher Loewen. You all have given me a great deal of guidance on technical matters as well as soft skills such as keeping track of the big picture while not losing sight of small details. Without your patience and encouragement, this project would not have been possible.

Professor Jörg Gsponer, your guidance and supervision on the bioinformatics of this project have been crucial to the development of the models in this project. The feedback and advice have played an important role in guiding me in this endeavor. It would not have been possible without you.

Much thanks go out to Dr Razvan Albu and Mang Zhu. Your preparation of the mass spectrometry (MS) samples used in this study, upstream data processing of MS spectra, as well as carrying out the biochemical assays form an indispensable part of this project. Numerous discussions with you have helped shape this project into the work it currently is.

Eric Wong and Alex Cumberworth deserve many thanks for the invaluable advice with regard to bioinformatic tools and best practices for programming. Prior to my rotation in the Gsponer lab, I had no experience with programming. With your help and guidance, I was able to understand the fundamentals of programming and gain the confidence to pursue a project that heavily utilised this skill set.

Nawar Malhis has been instrumental in giving a lot of helpful advice in the building of the model. Your guidance has helped develop the model beyond the simple linear model and helped to refine it.

All the members of the Mayor, Gsponer and Loewen labs have also provided

numerous informal discussions. The constant feedback and examination of ideas and approaches has been crucial in bringing the project from its nascent state to what it is today.

The Genome Science and Technology program has provided me with a stable yet flexible platform to explore the many options available to me. The strong culture of collaboration and the numerous opportunities to build a network of contacts with a variety of areas of expertise is excellent. The wide breadth it provides has set it apart and proven to be valuable in my development.

I would like to thank all the administrators and support staff in the GSAT program, the NCE, as well as the MSL. Your contributions and dedication have provided a stable and coducive environment without which all of the work in this thesis would not have been possible.

My friends, old and new, have been a huge boon to me these past years. From old friends I have known for over a decade, to new friends I have gotten to know in Canada, you have all played a huge part in my life and made it that much richer.

I would also like to thank my parents, Richard and Lucy, as well as my brother Alvin, and my fiancée Rui Qing, for their support. Moving to a different continent to pursue my graduate studies was a huge step for me, and your encouragement and support were invaluable to me as I progressed through this phase of my life. Without you, I would not be the person I am today.

Chapter 1

Introduction

The introduction of this thesis will cover several areas.

- Section 1.1: Protein homeostasis
- Section 1.2: Aggregation and disease
- Section 1.3: Functional aggregation
- Section 1.4: Predicting aggregation
- Section 1.5: Aims and scope of project

1.1 Protein homeostasis

Protein homeostasis, also known as proteostasis, is crucial to the well being of cells. Given the high concentration of biological molecules such as proteins within cells, misfolded or damaged proteins present considerable risks. The folding of a protein can be disrupted during synthesis or even after it has attained its native conformation. Factors such as mutations, translation errors and stresses, including but not limited to extreme temperatures, pressure, and pH, can cause a protein to misfold and potentially form amyloid or amorphous aggregates [18, 41]. This is detrimental to the cell due to the loss-of-function [76] as well as potentially toxic nature of amyloid and amorphous aggregates [92], which have the ability to form non-native interactions with cellular machinery and impair their functions [14].

Cells rely on the protein quality control network to prevent the accumulation of aberrant protein species, either through refolding them via the use of chaperones [66] or disposing of them via proteolysis. The ubiquitin proteasome system (UPS) plays a major role in clearing aberrant proteins in the cell [22], targetting them for degradation to the proteasome via covalent attachment of ubiquitin [67, 73]. Failure of the UPS to effectively clear these proteins can lead to detrimental outcomes brought about by the accumulation of aberrant proteins [10, 59, 116].

Another means by which cells address the issue of aberrant proteins is by sequestering them within quality control compartments [114] such as aggresomes [62, 69], Q-bodies [39], the juxtanuclear quality control (JUNQ) and insoluble protein deposit (IPOD) compartments [82]. These compartments may then be cleared by macroautophagy [61, 69] or by asymmetrical partitioning of these structures upon cell division [2, 15, 115, 134].

Macroautophagy is one mechanism by which cells can dispose of aberrant proteins sequestered in quality control compartments. The body to be disposed of is engulfed in a double membrane to form the autophagosome, which then fuses with the lysosome, resulting in the degradation of autophagosomal contents by lysozomal enzymes [74]. The ability of processes such as macroautophagy to maintain homeostasis is known to decline with age, contributing to age-related neurodegenerative diseases [124].

1.2 Aggregation and diseases

Misfolded proteins have the potential to assemble into large, insoluble structures held together by hydrophobic intermolecular interactions. Such structures can be classified into amyloid or amorphous aggregates. Amyloids display a characteristic fibrillar structure consisting of β-sheets running perpendicular to the axis of the fibrils [34]. Studies have shown that short protofibrils in the early stages of fibril formation may in fact be more toxic than mature fibrils [14]. In contrast to the ordered structure of amyloid aggregates, amorphous aggregates are assemblies that do not contain such ordered intermolecular bonds [133].

When the numerous quality control mechanisms designed to dispose of and mitigate the damage caused by aberrant proteins are overcome, various pathologies can arise. Protein aggregation has been associated with more than 40 diseases in humans [19, 106]. Of these, neurodegenerative diseases display among the most crippling symptoms, leading to them being the focus of intense research efforts. α synuclein has been associated with parkinson's disease (PD) [85] and amyloid-beta fibrils with alzheimer's disease (AD) [130], with recent studies highlighting other inclusions and their associations with various pathologies [129]. Many protein deposits in various disease contexts contain ubiquitin [79][6], suggesting that they were targeted for degradation, but somehow managed to evade the quality control pathways in the cell. Studies have shown that [80] disease associated protein aggregation may be able to act as a nucleus for the aggregation of endogenous proteins, potentially allowing for the propagation of the disease state to otherwise healthy cells [7, 8, 46, 52, 56, 104]. Some have proposed a model whereby while all proteins are theoretically able to form amorphous or amyloid aggregates, certain proteins simply possess a higher propensity to form them under a given set of conditions [20, 68]. Certain inherent traits, such as stretches of high hydrophobicity, high beta-sheet propensity, and low charge, are associated with a higher propensity to form amorhous or amyloid aggregates [21]. Transfer of these stretches from an amylogenic protein domain to a non-amylogenic protein has been shown to induce aggregation [123]. Improving our understanding of protein aggregation and solubility will be important for the development of therapies for proteopathies.

1.3 Functional aggregation

While many amorphous and amyloid aggregates have negative consequences for cells, functional aggregates are a class of aggregates that are part of normal cellular processes. Amyloid fibrils, characterized by their fibrillar cross β-sheet structure, have commonly been thought to be detrimental. However, it has been shown to be utilized by bacteria and fungi as a structural component, due to the high yield-strength and protease resistant nature of amyloids [44]. p53 is an example of a well known protein that can form functional aggregates as part of its normal function, existing as a homotetramer in its active form [94, 95]. Some peptide and secretory hormones have utlized the optimized packing of amyloid-like cross β-sheet rich conformations for their storage [81]. Other proteins such as TIA-1 in yeast [48], ataxin-1 in humans [94], and Pumilio in flies [110] have also been shown to be able to form functional aggregates. Functional aggregates have also been associate with other functions such as epigenetic inheritance [111] and formation of stress granules [48]. This highlights that although aggregation can be a detrimental scenario that cells need to manage, it can also serve a functional role in cells. Interestingly, recent studies have suggested that functional and dysfunctional aggregation are indeed promoted via similar forces, and that regulation of these forces is crucial for maintaining the balance between these two competing pathways [94, 96].

Several traits have been associated with the ability of proteins to form functional aggregates. Low complexity regions in proteins such as TIA1, FUS, CIRBP, RBM3, hnRNPA1, hnRNPA2 and SUP35 have been shown to be necessary and sufficient to cause aggregation of the proteins [64]. The work of Kato et al.showed that truncations of RNA-binding proteins that removed the RNA binding domains, and only contained their low complexity domains were capable of forming hydrogels, networks of interacting proteins with an aqueous phase contained within [3]. Truncations lacking the low complexity regions, in contrast, did not display the ability to form hydrogels. The work by Salazar et al. highlights how Q/N rich regions are important for the regulation of Pumilo function. In the absence of the Q/N rich region, the suppression of toxicity caused by Pumilo expression was not observed. Disordered proteins have been associated with the formation of functional assemblies known as Woronin bodies in plants [72]. The family of proteins known as septin pore-associated proteins that are part of Woronin bodies are highly charged and enriched in amino acids typically found in disordered proteins. Low complexity regions as well as disorder have thus been associated with lower solubility and functional aggregation.

1.4 Predicting aggregation

Due to the pathological association of amyloid aggregates with disease, many amyloid aggregation predictors have been developed [93][119][40][60][135][23][118][83][27][89]. TANGO[40] makes predictions on the aggregation propensity of proteins by calculating the partitioning of the segments of the protein between the aggregation state and the non-aggregation state. AGGRESCAN[23] utilizes experimentally derived aggregation propensities [30] and considers local stretches in proteins to determine aggregation propensity. Both were shown to accurately identify known aggregation prone proteins reliably. Given that not all aggregates are amyloid in nature, we wanted to explore and characterize a wider range of potentially aggregating proteins. While many studies have characterized aggregation under stress condition such as heat shock[87, 88, 131] and proteosomal inhibition[128], we were interested in studying protein aggregation under steady state conditions. Using insights gleaned from the analysis of protein solubility, we decided to build a model that would then be able to predict solubility under unstressed conditions *in silico*.

1.5 Aims and scope of project

We hypothesized that even under unstressed conditions in cells, some proteins are more prone to lower solubility than others, and that there are specific traits that distinguish lower solubility proteins from other proteins. Using quantitative mass spectrometry (MS), we have identified low solubility proteins in three eukaryotic model organisms: budding yeast, human neuroblastoma tissue culture cells, and mouse brain tissue. Analysis of these low solubility proteins highlights traits that draw a link to functional aggregation and macro-molecular assemblies. Using these traits, two models (a linear model and a support vector machine) were built to enable the *in silico* identification of low solubility proteins.

The aims of the project are as follows

- To identify proteins more prone to low solubility
- To identify traits that distinguish low solubility proteins form other proteins
- To use traits identified to build a model capable of predicting low solubility propensity

Chapter 2

Methods

Samples from three model systems were prepared and analyzed by quantitative proteomic MS. Computational and bioinformatic analysis of proteins identified allowed the identification of certain traits correlating to protein solubility. Fitting the data obtained to supervised learning models allowed for the prediction of the solubility of a protein based on its properties.

The methods section of this thesis will cover several areas.

- Section 2.1: Quantitative proteomic mass spectrometry
- Section 2.2: Biochemical assays
- Section 2.3: Computational and bioinformatic analysis
- Section 2.4: Models for prediction of lower solubility propensity

2.1 Quantitative proteomic mass spectrometry

Work in the following section was carried out by Dr Razvan F. Albu and Mang Zhu. Full details of the methods used can be found in [4]

2.1.1 Biological sample preparation

The biological samples from each of the three model organisms were prepared using native lysis to allow for the identification of proteins within the lower solubility fraction. Denaturing lysis would prevent the identification of proteins within the lower solubility fraction.

Yeast

Yeast stable isotope labeling by amino acids in cell culture (SILAC) strains were labeled with light or heavy arginine and lysine residues in order to carry out a quantitative mass spectrometry analysis. Cells were grown for at least 7 generations at 25°C. Cultures were grown to mid log phase (OD_{600} 0.8-1) before harvesting and lysis in lysis buffer (100 mM HEPES, 250 mM KCl, 1 mM PMSF, 1PIC, 1 mM phenanthroline and 10 mM chloroacetamide, 1% Triton X-100). To verify that detergent choice did not significantly influence results, the experiment was repeated with Triton X-100 substituted for 1% Igepal CA-630 with 0.5% deoxycholate. Lysate was pre-cleared before detergent-insoluble proteins were pelleted by centrifugation at 16000 rcf at 4°C for 15 min. The pellet was washed twice before resuspension. The protein concentration of both the supernatant and pellet fraction were measured using the DC Protein Assay (BioRad). For the RNase treatment, a final concentration of 20μg/ml RNase (Roche). RNA extraction was carried out using TRIzol® Reagent (Life Science Technologies) according to the manufacturer's protocol. Effectiveness of Rnase treatment was validated by agarose gel electrophoresis.

Human cells

Human neuroblastoma tissue culture cells (SH-SY5Y) were labeled with light or heavy arginine or lysine residues for a SILAC analysis. Cells were grown at 37°C for at least 11 divisions (determined by cell counting). Confluent cells were harvested and lysed in lysis buffer (50 mM TrisHCl, pH 8.5, 150 mM NaCl, 0.5% Na-deoxycholate, 1% Igepal CA-630, 1 PIC, 1 mM PMSF, 1 mM phenanthroline, 0.5 mM DTT) containing Igepal CA-630. Lysate was pre-cleared before detergent insoluble proteins were pelleted by centrifugation at 50,000 rcf for 1h at 4°C. The pellet was washed twice before resuspension. The supernatant was subjected to methanol chloroform precipitation and proteins extracted were resuspended in HU buffer (8 M urea, 100 mM HEPES, pH 8.0). Protein concentrations were measured using the DC Protein Assay (BioRad).

Mouse brain tissue

Female C57BL/6 mice (non-littermates) were grown to 11 weeks of age, after which brain tissue was harvested. Harvested tissue was flash frozen in liquid N_2 and then lysed by cryogrinding. Brain samples from three mice were pooled for analysis. After resuspension in lysis buffer, lysate was pre-cleared by centrifugation and detergent insoluble proteins were pelleted by centrifugation at 50,000 rcf for 1h at 4°C. The pellet was washed twice before resuspension. The supernatant was subjected to methanol chloroform precipitation and proteins extracted were resuspended in HU buffer. Protein concentrations were measured using the DC Protein Assay (BioRad).

2.1.2 Sample preparation and offline fractionation

Yeast samples were subjected to in-gel trypsin digestion, while human and mouse samples were subjected to in-solution trypsin digestion. The heavy-labeled insoluble pellet fraction and light-labeled soluble supernatant fraction from each of the organisms was mixed in a 1:1 ratio by mass. For experiments involving RNase, light-labeled soluble supernatant from cells untreated with RNase was mixed with medium and heavy-labeled insoluble pellets (without and with RNase treatment respectively) in a 1:1:1 ratio by mass (as shown in Figure 3.18a). Mouse brain samples were labeled after tryptic digestion using formaldehyde-cyanoborohydride, attaching a 28Da (light) or a 32Da (heavy) moiety to primary amines as described in [113]. Consistent with the labeling scheme for human and yeast, the insoluble pellet fraction was labeled heavy and the soluble supernatant fraction was labeled light.

Approximately 200μg of tryptic peptides were fractionated by offline high pH reverse-phase chromatography. 96 fractions of 40 seconds each were collected and pooled in a non-contiguous manner [132][120], with 9 pooled fractions for yeast and 10 pooled fractions for human and mouse. Yeast label swap experiments and experiments comparing light- and heavy-labeled insoluble pellets did not undergo an offline fractionation step.

2.1.3 Liquid chromatography- tandem mass spectrometry

Each of the fractions prepared in 2.1.1 was analyzed using liquid chromatography (LC)-MS/MS (Tandem Mass Spectrometry) on a linear-trapping quadrupole (LTQ) Orbitrap Velos (Thermo) coupled to an Agilent 1100 Series Nanoflow high performance liquid chromatography (HPLC) as described in [4].

Spectra were searched by the ANDROMEDA algorithm, in the MaxQuant environment (version 1.5.0.0) against the saccharomyces genome database (SGD)[17] for yeast (Feb 3, 2011), and the Uniprot human [117] (Apr 16, 2014) and mouse (Feb 19, 2014) databases. The search was configured largely using the default MaxQuant parameters. We allowed for a 1% false discovery rate at both the peptide and protein level.

2.2 Biochemical assays

Work in the following section was carried out by Mang Zhu. Full details of the methods can be found in [4].

Western blotting

Two proteins each from the lower solubility (LS) and higher solubility (HS) bin were selected. The endogenous copies of these proteins were tagged with a triple-hemagglutinin (HA) tag amplified from parent vector pFA6a-3HA-His3MX6. The pellet and supernatant fraction were separated as detailed in 2.1.1, normalized in a 10:1 ratio. Samples were resolved on 4%-20% gradient gels (BioRad), and transferred onto 0.45 µm nitrocellulose membranes (BioRad). Immunodetection was carried out using anti-HA primary antibodies(1:2000 dilution,12CA5, AbLab) and LI-COR secondary antibodies (1:10000 dilution, LI-COR Biosciences). Images were acquired using the CLx Odyssey (LI-COR Biosciences) and quantification was carried out using Image Studio v3.1 (LI-COR Biosciences)

2.3 Computational analysis

Processing of data for CAI, secondary structure, ANCHOR MoRFs, ELMs, IUPred, Pfam, and disulfide bond prediction were performed by Eric Wong.

2.3.1 Identification of lower solubility proteins

From the proteins identified via MaxQuant search results in 2.1.3, proteins flagged as contaminants or reverse hits were removed, as well as proteins identified by fewer than two peptides. Proteins that did not have a reported quantification value due to inconsistent labeling or orphaned analyte issues were also removed. Duplicate database entries with identical amino acid sequences were removed. The quantification ratios were ranked and a plot of $\log_2$ ratios against rank (highest to lowest) was generated (Figure 3.2a-c). To separate proteins into different categories based on their solubility, we first devised a method to determine appropriate cutoffs. A smoothed curve was generated by application of a locally weighted scatterplot smoothing (LOWESS) function (python module statsmodels, lowss function, 3 iterations, no linear interpolation, 0.667 of data used for each y-value estimate) and 10% of points with the lowest gradients were chosen. The R^2 value of all points between the first and last unsmoothed equivalent of these points was calculated. Data points of increasing $\log_2$ ratio were added until the corresponding R^2 value dropped below 0.99. The process was repeated with points of decreasing $\log_2$ ratio. Points that were thus defined were considered to correspond to proteins of normal solubility (NS). A trendline was fitted to these points and the intercepts of that trendline at the first and last ranked point on the plot were taken as the cutoff for the lower solubility (LS), i.e. high ratio, and higher solubility (HS), i.e. low ratio, bins respectively. If the LS or HS cutoff was less extreme than the most extreme point in the NS bin, the most extreme point in the NS bin was used as the cutoff instead. In most cases, there were points that lay between the LS and NS as well as NS and HS bins. These points were excluded from further analysis as their solubility was deemed ambivalent. Next, protein groups containing proteins flagged as having one or more transmembrane domains (as determined by uniprot annotations) were removed. The CYC2008 [97] (for yeast) and Quorum [108] (for mouse and human) databases were scanned to determine which proteins were part of complexes.

2.3.2 Plotting and statistical analyses

Figures were generated using R (www.r-project.org) and the python library Matplotlib [58]. Statistical tests were carried out using NumPy [122] and SciPy [91] and R. For boxplots, boxes show the middle 50% of data points, the whiskers represent 1.5 the inter-quartile range. n corresponds to the number of proteins in each bin, o to the number of outliers, which are not shown. p-values calculated by the MannWhitney U test and listed in Table A.10, Table A.11, and Table A.12 are displayed above the figures. A dotted gray line represents a p-value lower than 0.05 but which was not significant after multiple testing correction, while solid lines denoting significant values after correction. Statistical comparisons were made between the LS and NS bins, as well as the NS and HS bins. The multiple testing correction used was the Bonferroni correction. The Amino Acid Compass plotted using an in-house R-script as described previously [87].

2.3.3 Protein properties

Mouse and human proteomes were omitted during the analysis due to the large number of isoforms as well as the tissue specific nature of many proteins. Protein length and amino acid composition were calculated based on sequences in the databases used in 2.1.3. The random sampling of proteins within groups (RSPG) analysis was carried out to determine the length variation of proteins within protein groups. For each of the 1000 iterations, one protein was picked randomly from each protein group and the median value of each bin (LS, HS, or NS) was computed. Yeast was omitted from the RSPG analysis due to the small number of protein groups observed.

Matching of mouse and human orthologs was carried out using Roundup [31]. gene ontology (GO) analysis was performed using DAVID [32]. Pfam data files were downloaded from the Pfam database [43] on June 6, 2014. pfam_scan.pl version 1.5 (used with HMMER-3.1b1, downloaded from hmmer.org [42]) from the Pfam database was used at default settings to search for matches to each protein sequence. For protein groups containing multiple entries, the first protein in the group was used for Pfam and GO analysis. Protein abundance data from previous studies [47][29] was utilized. Aggregation propensity of proteins was predicted using TANGO [40](aggregation prone: at least 1 stretch of 7 residues with aggregation tendency above 50%, non-aggregation prone: no residue in entire protein with aggregation tendency above 50%) as well as AGGRESCAN [23] (aggregation prone: at least 1 stretch of 8 residues with a4v value greater than 0.5, non-aggregation prone: no stretch longer than 5 residues with a4v value greater than 0.5 using previously defined cut-offs [50]. low complexity regions (LCRs) were retrieved from lowest-probability subsequences (LPS)-annotate, which has defined LCRs in a number of organisms as previously published [55]. Disulphide bond prediction was performed using DIpro v2.0 [16] at default settings. Phosphorylation sites were retrieved from PhosphoSitePlus [57] for mouse and human, and from PhosphoGRID for yeast [109]. PSIPRED [63] was used to obtain secondary structure predictions. DISOPRED2 [126] and IUPred [35] were used for the prediction of protein disorder. Stretches of disordered resides longer than 5 as predicted by DISOPRED were taken to be disordered patches. molecular recognition features (MoRFs), intrinsically disordered stretches on proteins that assume an ordered conformation upon protein-protein interaction, were retrieved from ANCHOR [36]. eukarotic linear motifs (ELMs) were retrieved from the ELM database on the ANCHOR server [99]. When normalizing against disorder, proteins with less than 10% disorder were excluded from the analysis. The codon adaptation index (CAI) of proteins were calculated using CAIcal [98]. Reference tables were obtained from the CAIcal server on Jun 10, 2014 (mouse and yeast) and June 16, 2014. Coding sequences were obtained on June 6,2014 from SGD [17] (for yeast) and the UniParc database [117] (for mouse and human). Yeast SGD accession numbers were mapped to Uniprot accession numbers using the Uniprot mapping tool. Coding sequences were also used to calculate GC content, as well as the number of codons one substitution away from a stop codon (close stop). Localization information for yeast was retrieved from previously published data by [77]. The number of transmembrane helices was predicted using TMHMM v2 [70] (yeast only). Hydrophobicity was calculated using the grand average of hydrophobicity (GRAVY) index [71]. The number of codons per protein was calculated using coding sequences previously mentioned as published in [1]. Amino acid abundance by percentage, as well as the percentage of positive (H,R,K), negative (D,E), polar (S,T,Y,C,N,Q), hydrophobic (G,A,V,L,I,F,M), aromatic (F,Y,W), and rare (C,W,H,M) amino acids, as well as net charge per protein were calculated using the sequences used for searching in 2.1.3. Patches of aromatic, hydrophobic, positive, negative, polar (Q and N, as well as S and T) residues was calculated from the database sequences in 2.1.3 using the same methods described in [55] with categories of amino acids rather than individual species.

2.4 Models for prediction of lower solubility propensity

The protein traits derived from Section 2.3.3 were used to build machine learning models to predict lower solubility propensity.

2.4.1 Multiple regression model

Multiple linear regressions were carried out in R with the glmnet library [45]. The model was built using stepwise linear regression in the forward direction with elastic net regularization. 5-fold cross-validation was carried out for the training set, with the lambda values chosen to optimize the mean squared error (MSE) on the training set. 5-fold cross validation was carried out, with the data divided randomly into 5 equal parts, and one part per iteration used as the test set, and the remaining ones used for training. MSE values and correlation coefficients were obtained from each iteration of the test set.

2.4.2 Support vector machine

The support vector machine (SVM) model was built in R using the e1071 library [84]. One fifth of the data was randomly chosen as the test set in turn, and the remaining data was used for training. The model was built (cost = 10, method=”C-classification”, kernel = ”linear”) using parameters determined by the built in tuning function. 5-fold cross validation was carried out.

Chapter 3

Results

This section of the thesis will describe the results of the mass spectrometry experiments which were used to identify low solubility proteins, as well as the bioformatic analyses carried out. The results of those analyses were then used to build two machine learning models to predict protein solubility.

Most of the results presented here were published in [4]. Sample processing and mass spectrometry were performed by Dr. Razvan Albu (human and mouse tissues) and Mang Zhu (yeast cells). RNase experiments, Western Blots, and generation of the amino acid compass were performed by Mang Zhu. Processing of data for CAI, secondary structure, ANCHOR MoRFs, ELMs, IUPred,Pfam, and disulfide bond prediction were performed by Eric Wong. All other computational analyses were performed by Gerard Chan (myself).

- Section 3.1: Identification and feature analysis of low solubility proteins
  - Section 3.1.1: Isolation of lower solubility proteins
  - Section 3.1.2: Identification of lower solubility proteins
  - Section 3.1.3: Lower solubility (LS) proteins are longer than higher solubility (HS) proteins
  - Section 3.1.4: LS proteins are more aggregation prone in yeast but not in human or mouse
  - Section 3.1.5: LS proteins contain biases for particular amino acids

  - Section 3.1.6: LS proteins are more highly charged and are less hydrophobic
  - Section 3.1.7: Choice of detergent does not significantly affect solubility of LS proteins
  - Section 3.1.8: LS proteins are more disordered and contain more molecular recognition features (MoRFs) and eukarotic linear motifs (ELMs) than higher solubility (HS) proteins
  - Section 3.1.9: LS RNase treatment increases the solubility of RNA associated proteins, but does not affect the overall properties of low solubility proteins
  - Section 3.1.10: Coding sequences for LS proteins contain a lower GC content
  - Section 3.1.11: LS proteins possess numerous traits that distinguish them from HS proteins

- Section 3.2: Models to predict protein solubility
  - Section 3.2.1: Regularized linear model
  - Section 3.2.2: Support vector machine

3.1 Identification and feature analysis of low solubility proteins

In this section, we outline how we distinguished lower solubility proteins from other proteins in our experiments, and then examined certain features that distinguished these lower solubility proteins from other proteins.

3.1.1 Isolation of lower solubility proteins

Our aim was to identify features that distinguished lower solubility proteins in eukaryotic cells under steady state conditions. While many studies have studied systems under non-steady state conditions, we felt that the solubility landscape under steady state conditions would allow us to examine features that would be intrinsic to the proteins and proteomes, free from the influence of stresses and chemical

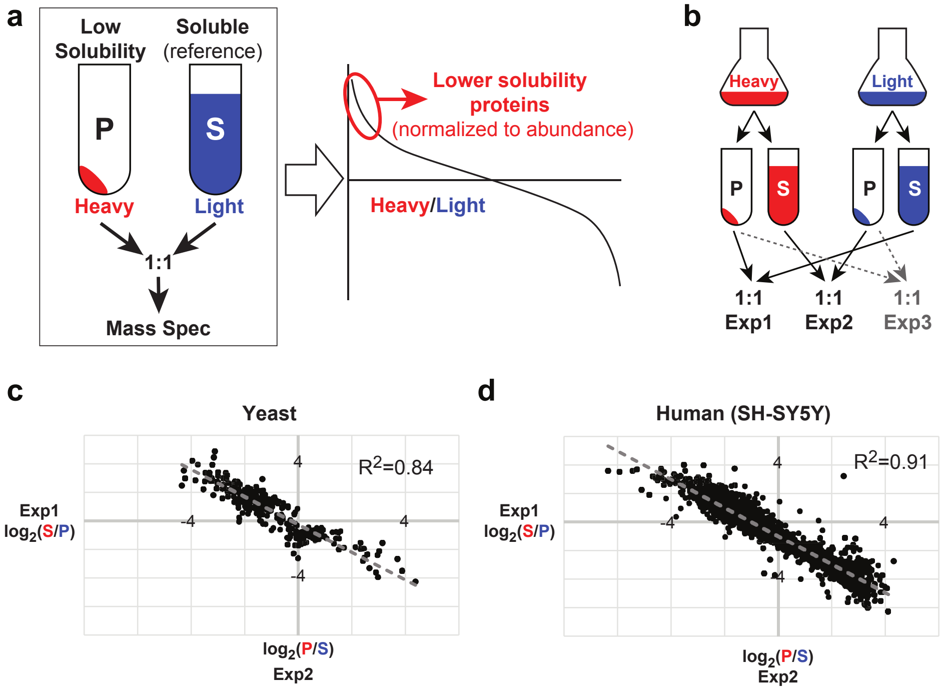

Figure 3.1: Overview of the approach used to identify lower solubility proteins. (a) Proteins in the low solubility fraction (P:pellet) are mixed with soluble proteins (S:supernatant) in equal ratios by mass and analyzed by quantitative LC-MS/MS. Normalization for abundance is achieved by comparison between the two fractions. (b) Various combinations of soluble and lower solubility fractions from light- or heavy- labeled samples. The results of experiment 1 and 2 are shown in panels c and d respectively. (c, d) Results of the label-swap experiments carried out for yeast (c) and human (SH-SY5Y) cells (d). The $\log_2$ ratio values of the proteins quantified in both experiments were plotted, and the coefficient of determination (R^2) is shown in each figure. Figure reproduced from [4].

inhibitors. The yeast model organism *Saccharomyces cerevisiae* was chosen for its relative simplicity and well-characterized nature. Mouse brain and human neuroblastoma SH-SY5Y cells [12] were used as disruption of protein homeostasis in neuronal tissue is associated with various pathologies. To account for the variation between individual mice, brain tissue from three young adult (approximately 11 weeks old) mice was combined.

Cells were grown and harvested as detailed in Section 2.1.1. Harvested cells were lysed by cryogrinding under native conditions. Low solubility proteins were defined to be those that formed a pellet after centrifugation of native lysate, while those that remained in the supernatant were defined as soluble proteins. The pellet and supernatant fractions of the lysate were mixed as shown in Figure 3.1a then analysed by quantitative mass spectrometry (see Section 2.1.2 and Section 2.1.3).

The quantification ratio of the lower solubility vs soluble proteins was obtained by directly comparing the lower solubility and soluble fractions. Without taking the soluble fraction into account, abundant proteins with a modest proportion present in the lower solubility fraction would be overrepresented in relation to less abundant proteins with a higher proportion present in the low solubility fraction. Direct quantification of proteins in the lower solubility fraction would be reflective of the abundance of proteins in the lower solubility fraction, rather than the solubility of proteins. SILAC was used to label yeast and human cells, and dimethylation via isotopically-tagged formaldehyde was used to label mouse sample proteins. Equal amounts of light-labeled soluble and heavy-labeled low solubility proteins from each organism were mixed and analyzed as detailed in Section 2.1.2 and Section 2.1.3. Given that the soluble and low solubility fractions in the yeast and human samples were derived from distinct populations of cells, a label swap experiment(Figure 3.1b) was performed in order to ascertain the degree of variation between populations, as well as the variation due to handling and labeling efficiency. For the label swap experiment, the heavy-labeled soluble fraction was mixed with the light-labeled low solubility fraction. The yeast and human samples had R^2 values of 0.84 and 0.91 respectively(Figure 3.1c and d), compared to the non-label swapped experiments, indicating a large amount of similarity. This reassured us that there were minimal artefacts attributable to handling, labeling, and variation between populations.

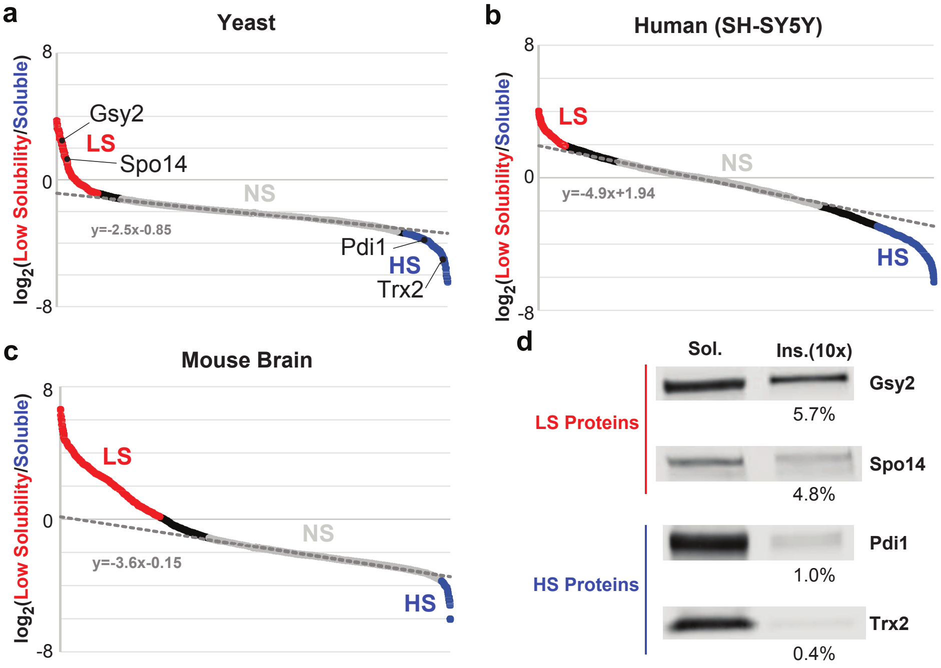

Figure 3.2: Analysis of raw LC-MS/MS quantification data and validation of low solubility delimitation. (ac) Quantitative mass spectrometry data comparing the low solubility to the soluble fraction for the yeast (a), human (b) and mouse (c) samples, using a base-2 logarithmic scale. The calculated trendline is shown as a dashed dark gray line with indicated equation. Individual data points are colored according to their assigned bin: red for lower solubility (LS), gray for normal solubility (NS), blue for higher solubility (HS), and black for data points not included in any category. (d) Western blots depicting the relative amounts of proteins in the supernatant and low solubility fraction, with ratio of low solubility to soluble expressed as a percentage. Proteins are marked on the plot in (a). Figure reproduced from [4].

3.1.2 Identification of lower solubility proteins

Using quantitative mass spectrometry, 1738, 2584, and 2326 protein groups were quantified in yeast, human and mouse respectively. Based on the peptides identified during the MS experiment, it is sometimes not possible to distinguish which of several proteins one or more peptides was derived from. In such cases, all possible proteins are represented as a protein group, comprised of two or more protein candidates.

We next divided the proteins into three bins for further analysis. As shown in Figure 3.2a-c, for each of the three organisms, a trendline was fitted to as many points as possible while maintaining a high R^2 value of 0.99 or higher, as detailed in Section 2.3.1. The intercepts of the trendline were used as cutoffs to delimit the lower solubility (LS) and higher solubility (HS) bins. The points used for the generation of the trendline were categorized as normal solubility (NS). These proteins were considered to be of non-extreme solubility. NS proteins had fairly similar solubilities, as seen by the gentler slope in that part of the plot. We thus decided to use the NS bin as a point of reference when making comparisons to the LS and HS bins. As in the human and yeast samples, the majority of the proteins fell inside the NS bin, with a small portion of proteins in the lower solubility (LS) and higher solubility (HS) bins. In the mouse sample, a larger proportion of the protein groups fell inside the LS bin than human and yeast. This could be due to the increased complexity of the brain tissue used in the mouse sample, compared to the simpler cultured yeast cells as well as undifferentiated human neuronal cells. Unlike the yeast and human tissue culture cells, neuronal cells in the mouse brain are less capable of reducing the amount of aberrantly aggregated protein per cell via asymmetrical division [2, 15].

After binning, proteins annotated as having one or more transmembrane domains were removed, as their solubility would be dependent on the choice of detergent used for lysis [9]. 257 yeast proteins, 187 human proteins, and 243 mouse proteins were thusly removed. Proteins in complexes were not removed as no trend was observed between complex size (number of different proteins in the complex) and solubility (average solubility of all partners in the complex) Figure 3.3. Removal of transmembrane proteins did not substantially affect the distribution of solubilities. From the filtered data, the lower solubility (LS) bin contained 96, 170, 530 proteins from yeast, human and mouse respectively. The higher solubility (HS) bin contained 180, 343, 51 proteins from yeast, human, mouse respectively. The NS bin contained approximately 2/3 of the quantified proteins, with 1095, 1200, and 1254 proteins in yeast, human, mouse respectively. Proteins that were not included in the NS bin, and did not meet the LS or HS cutoff, were excluded from

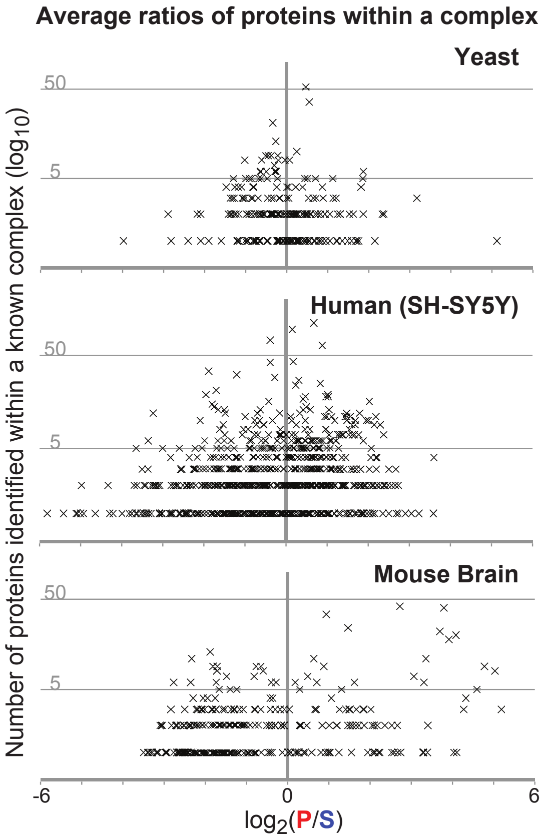

Figure 3.3: Average solubilities of proteins within a complex. The number of components of known complexes in our dataset was plotted against the average solubility (based on components identified) for each protein complex in yeast, human, and mouse. Figure reproduced from [4].

further analysis due to ambivalence regarding their solubility.

To verify the identification and quantification from the mass spectrometry experiments, two proteins from the LS bin (Gys2 and Spo14) and two proteins from the HS bin (Pdi1 and Trx2) in yeast had their endogenous copy appended with a C-terminal 3x-hemagglutinin (HA) tag, and their solubility validated by Western blotting. As seen in Figure 3.2d, LS proteins tend to be about 20 times more abundant in the supernatant than the pellet, compared with HS proteins that tend to be about 100-200 times more abundant in the supernatant, making LS proteins approximately 5-10 times less soluble than their HS counterparts.

gene ontology (GO) analysis revealed that a number of LS proteins from yeast, human, and mouse were associated with RNA processing as a molecular function (Table A.2,Table A.4,Table A.6). Molecular function showed many LS proteins associated with RNA binding. Unique to the mouse sample, the LS bin also showed an enrichment of proteins associated with cytoskeletal organization and structural molecular activity. Association with structural molecular activity is consistent with the high levels of tubulin and neurofilaments in neurons [33]. We checked if there was an enrichment of Pfam domains (Table A.7,Table A.8,Table A.9) in the LS proteins compared to the NS bin, and found that RRM_1, an RNA binding domain, and filament were enriched in human and mouse. Human also showed an enrichment in septin, while mouse showed an enrichment in spectrin. However, the GO annotations found did not represent a large portion of the LS proteins we identified, suggesting that LS proteins are of a diverse nature.

Interestingly, chaperone proteins were not strongly enriched among the LS proteins. Ssa1p, a major cytosolic Hsp70 chaperone, was found within the NS bin in yeast. It is possible that, under our unstressed experimental conditions, chaperones associated to non-misfolded proteins may remain largely in the soluble supernatant fraction, while only a small portion of the chaperone population is bound to misfolded proteins.

Proteins reported in other studies to exhibit low solubility displayed a low degree of overlap with proteins in our LS bin [26, 92, 103, 128]. This can be attributed to differences in isolation methods and the solubility of proteins in stressed and unstressed conditions. Our approach also accounts for abundance of proteins by normalizing against the soluble fraction, which may be important in accounting for the bias of mass spectrometry towards more abundant proteins. As seen in Figure 3.4a, proteins identified as insoluble without normalization to the soluble proteins were of much higher abundance (ion intensity and number of molecules per cell previously published in [29] and [47] respectively). In contrast, after normalization to the soluble fraction, it can be seen that LS proteins are actually less abundant than HS proteins (Figure 3.4b). This is consistent with previous work [50] that shows proteins which are prone to forming aberrant aggregates being subject to

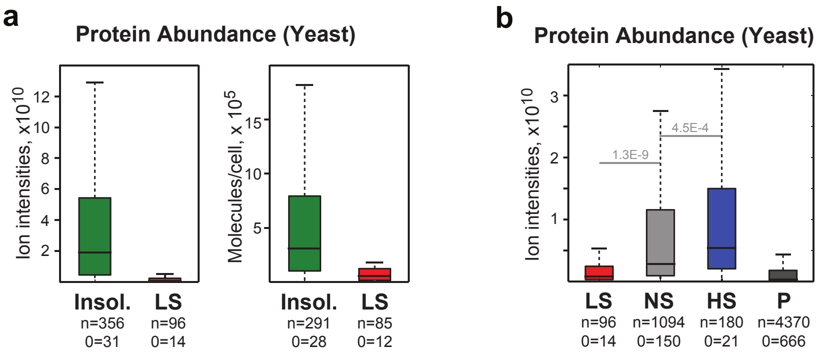

Figure 3.4: Abundance of LS proteins in this study and compared to previous studies. (a) Abundance values of proteins identified in the LS bins compared to those identified in the low solubility pellet (Insol) from a separate mass spectrometry analysis of low solubility proteins alone [4]. Protein abundances were derived based on ion intensities in mass spectrometry published by [29](left) and levels of endogenously tagged proteins as published by [47] (right). (b) Abundance of proteins in the LS, NS and HS bins, as well as in the proteome as a whole (P), determined for yeast. Figure reproduced from [4].

strict regulation at the transcriptional, translational and degradation level, keeping their concentration below the critical concentration for aggregation, while highly expressed proteins are subject to strong evolutionary pressure toward lower aggregation propensities [105, 107].

3.1.3 Lower solubility (LS) proteins are longer than higher solubility (HS) proteins

We first sought to determine whether protein length was correlated with solubility given that ubiquitinated proteins which are less soluble after heat shock have previously been shown to be longer [87]. In protein groups with more than one member, the average length was taken as representative of the whole group. In all three organisms, proteins in the LS bin tended to be longer than proteins in the NS bin, and proteins in the HS bin tended to be shorter than proteins in the NS bin Figure 3.5a, consistent with the findings in the previous study.

As mentioned earlier, peptides identified by mass spectrometry often corre-

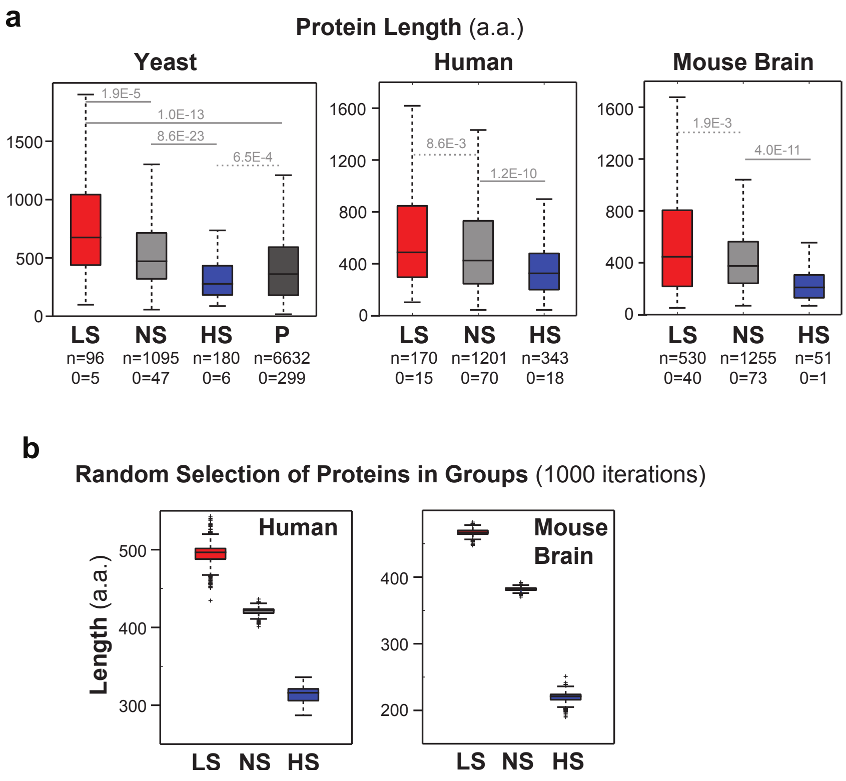

Figure 3.5: Comparison of the lengths of proteins by randomly picking versus averaging over protein groups. (a) Boxplots of the distributions of protein length (in amino acids) for yeast, human and mouse brain samples. Length of each protein group was obtained by averaging over lengths of all proteins in the group. (b) Distributions of the median protein length values after randomly selecting one protein per protein group for one thousand iterations during the random sampling of proteins within groups (RSPG) experiment. There were 115 and 298 protein groups with at least two proteins in the human and mouse datasets, respectively. Figure reproduced from [4].

spond to multiple proteins, and it is not uncommon for there to be insufficient information to allow for definitive identification of proteins. For example, in yeast one particular protein group contains two proteins: YKL156W and YHR021C. Based on the peptides identified by mass spectrometry, the peptides could have been derived from either or both of those proteins. In our experiments, 55 of 1738, 2062 of 2584, and 1296 of 2326, protein groups contained more than one protein in yeast, human, and mouse respectively. The protein groups that result from such ambiguous identification may contain proteins that are substantially different from each other. In order to gauge the amount of variability on the outcome of subsequent analysis, we carried out a random sampling of proteins within groups (RSPG). For each of the thousand iterations, one random protein from each group was selected and the median length for each of the bins defined in Section 3.1.2 was calculated. This analysis was not performed in yeast due to the small number of protein groups with two or more members observed. In human and mouse, the median values obtained in the RSPG experiment (Figure 3.5b) did not differ much from the median values obtained by averaging over the protein group, as done in Section 3.1.3 and shown in Figure 3.5a, by a large margin. The averaging of values over the protein group was thus regarded as an appropriate approximation.

3.1.4 LS proteins are predicted to be more aggregation prone in yeast but not in human or mouse

Since aggregation prone proteins should in principle display lower solubility, we next checked whether the LS bin contained more proteins predicted to be amyloid aggregation prone. In yeast, AGGRESCAN and TANGO both predicted a larger fraction of the LS to be amyloid aggregation prone than in the NS and HS bins (Figure 3.6a-b). The HS bin, was predicted to contain a higher proportion of non-amyloid aggregation prone proteins. However, this trend was not observed in human and mouse, where either no trend was observed, or amyloid aggregation prone proteins were more frequently observed in the HS bin. Given that TANGO and AGGRESCAN both look for features associated with amyloid aggregation, it is possible that human and mouse low solubility proteins possess other traits not characteristic of amyloids that contribute to their lower solubility.

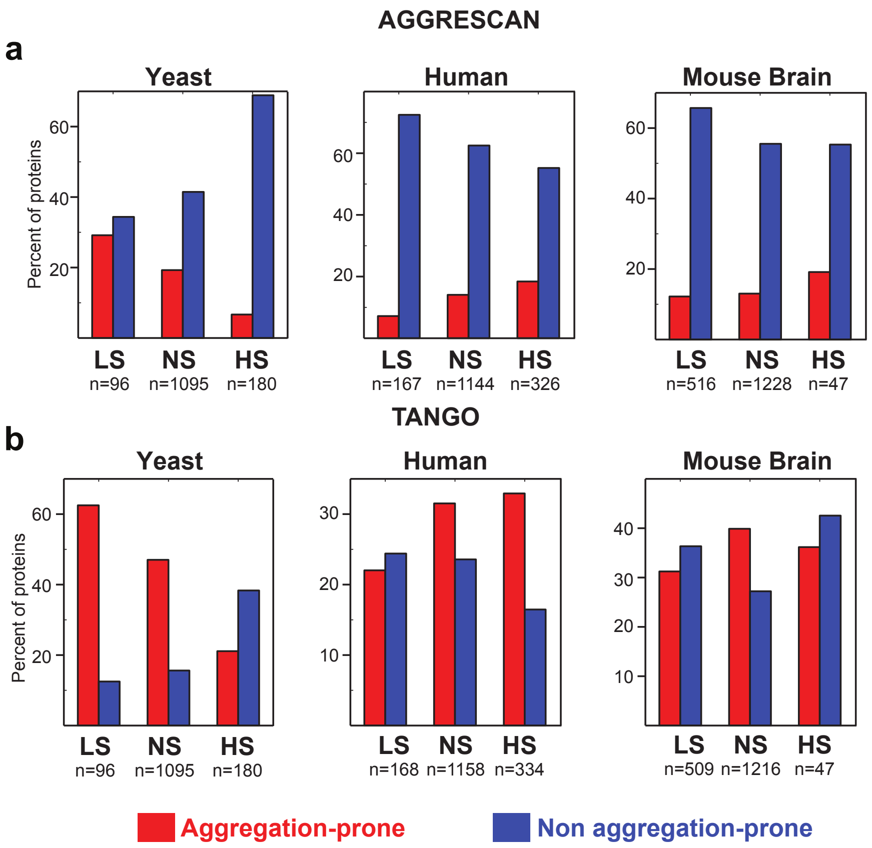

Figure 3.6: Aggregation propensity as predicted by TANGO and AGGRESCAN. Aggregation prediction (shown as percent of all proteins in each bin) for the proteins in the indicted bins in yeast, human and mouse using the AGGRESCAN (a) and TANGO (b) algorithms, as shown. Figure reproduced from [4].

3.1.5 LS proteins contain biases for particular amino acids

Since we were searching for intrinsic properties of proteins that could contribute to their low solubility, we considered the possibility that LS proteins might contain biases for certain amino acids, prompting us to examine the local and global amino acid composition of proteins. Given that certain neurodegenerative diseases have been associated with stretches of polyglutamine within proteins [54, 136], we examined LS proteins for the presence of low complexity regions (LCRs), which

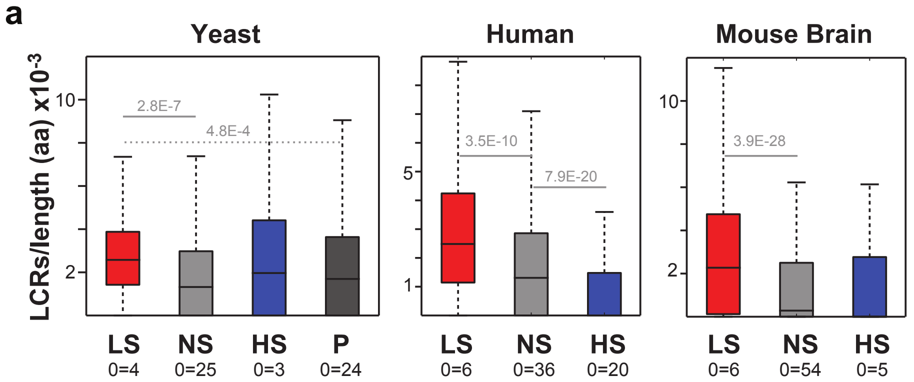

Figure 3.7: Number of low complexity regions (LCRs) per unit length of proteins in each bin, as found on LPS-annotate.

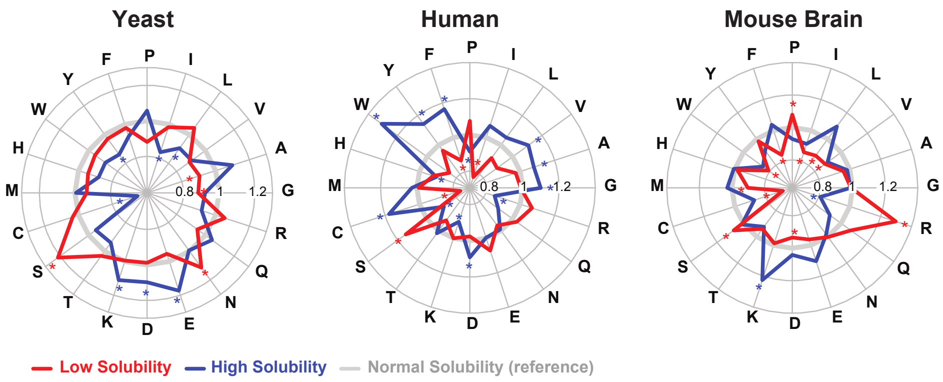

Figure 3.8: Amino acid composition of proteins in the LS and HS bins of yeast, human and mouse samples. Each data point represents the median value within the entire bin, expressed as a fold enrichment over the NS bin. Statistically significant differences are indicated by asterisks. Figure adapted from [4].

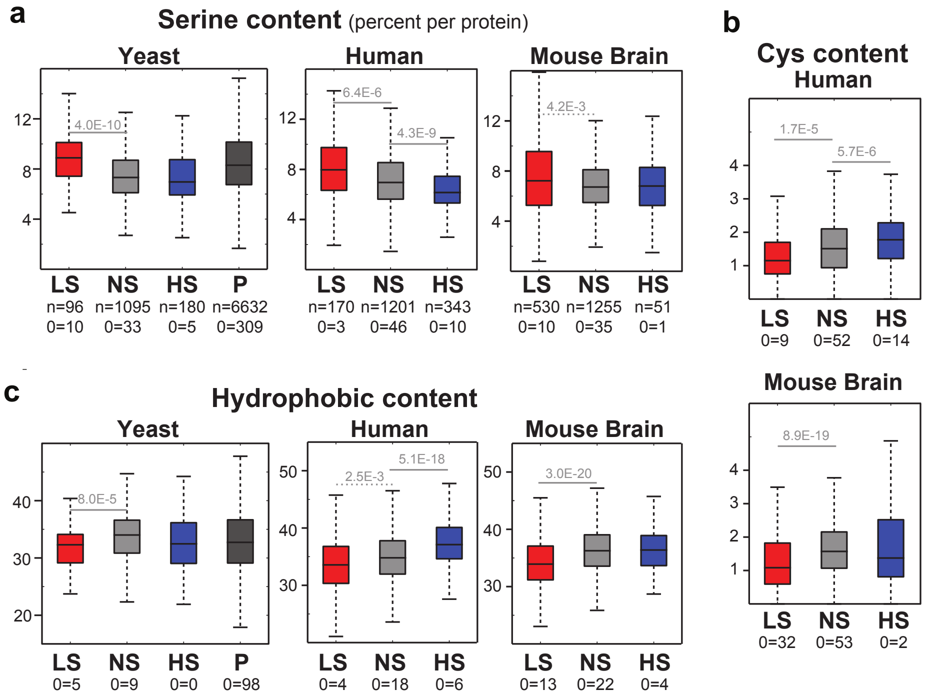

Figure 3.9: Percentage abundance of particular types of amino acids. Boxplots showing the distribution of values for the indicated amino acids (percent per protein) in the indicated sample. Shown here are the analyses for serine (a), cysteine (b), and hydrophobic residues (c). Proteins included in the hydrophobic analysis were glycine, alanine, valine, leucine, phenylalanine and methionine. Figure adapted from [4].

are local stretches of a protein that contain a bias for one or more amino acids. While proteins in the LS bin contained more low complexity regions (LCRs) per unit length (Figure 3.7), the LCRs did not show any consistent bias for any particular amino acid across all three organisms (Table A.13, Table A.14, Table A.15). With regard to general amino acid composition (Figure 3.8a), several amino acids tended to be either enriched or depleted in the LS bin vs the NS bin, with the opposite effect observed between the NS and HS bin. Interestingly, there was little overlap between yeast and the other two organisms, but the mouse and human samples tended to display more similar biases. This could be reflective of the closer evolutionary relationship between mouse and human, compared to yeast. The LS fractions tended to be enriched in glutamines, but depleted in asparagines (Figure 3.8a). Both amino acids have been previously linked to amyloid formation, with polyglutamine associated with benign amyloids and polyasparagine linked to more toxic species [53]. Serine was more enriched in the LS fraction and more depleted in the HS fraction of all three organisms (Figure 3.9a). In mouse and human, fewer cysteines were observed (Figure 3.9b), suggesting lower potential for stabilization of structure by disulphide bonds. Yeast and mouse also displayed significantly fewer hydrophobic residues in the LS bin (Figure 3.9c). The difference was observed but not significant in human. All in all, there are specific amino acid biases that set low solubility proteins apart from normal and high solubility proteins.

With certain amino acids such as serine and cysteine being associated with features such as phosphorylation and disulphide bonds, we also examined whether these features were more or less prominent in LS compared to HS proteins. Phosphorylation sites were only enriched in LS proteins for human. In yeast, it was the HS proteins that had more phosphorylation sites (Figure 3.10a), which is surprising, given the enrichment for serine observed in Figure 3.9a. LS proteins in human and mouse contained fewer predicted disulfide bonds (Figure 3.10b), consistent with the previous observation of fewer cysteines (Figure 3.9b).

3.1.6 LS proteins are more highly charged and are less hydrophobic

Charge on proteins can be used to mediate protein-protein interactions [28], we decided to examine if LS proteins might contain more net charge than their HS counterparts. In yeast and mouse, LS proteins tended to have a higher magnitude of net charge (regardless of sign) than NS and HS proteins (Figure 3.11a). When sign is taken into account, LS proteins in all three organisms had a more positive charge than HS proteins (Figure 3.11b), but not significantly more so than NS proteins. HS proteins are more negatively charged than their LS and NS counterparts. Given the ability of charged residues to mediate inter- and intramolecular interactions, it is plausible that because HS proteins tend to be less strongly charged, they might have have fewer interaction partners, as large assemblies brought about by large networks of interaction would intuitively be of lower solubility.

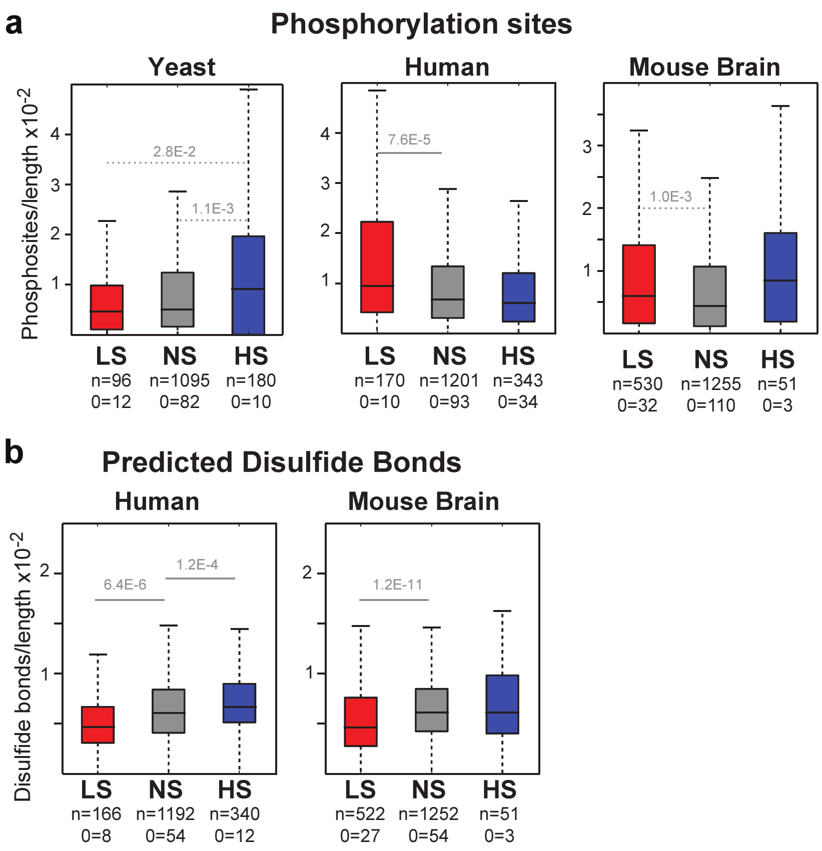

Figure 3.10: Number of phosphorylation sites and disulfide bonds in each of the three model organisms. (a) The number of phosphorylation sites normalized to length for yeast, human and mouse, as indicated. (b) Number of predicted disulfide bonds normalized to length. Figure reproduced from [4].

In light of hydrophobic interfaces being able to mediate interactions between proteins [125], we also examined whether the hydrophobicity of proteins differed between proteins in the LS and HS bins. The grand average of hydrophobicity (GRAVY) index [71] was chosen as it scores each residue with a different hydrophobicity score, allowing a more precise representation than the previous examination of what proportion of each protein was composed of hydrophobic residues (Figure 3.9c). Consistent with the observation that LS proteins were com-

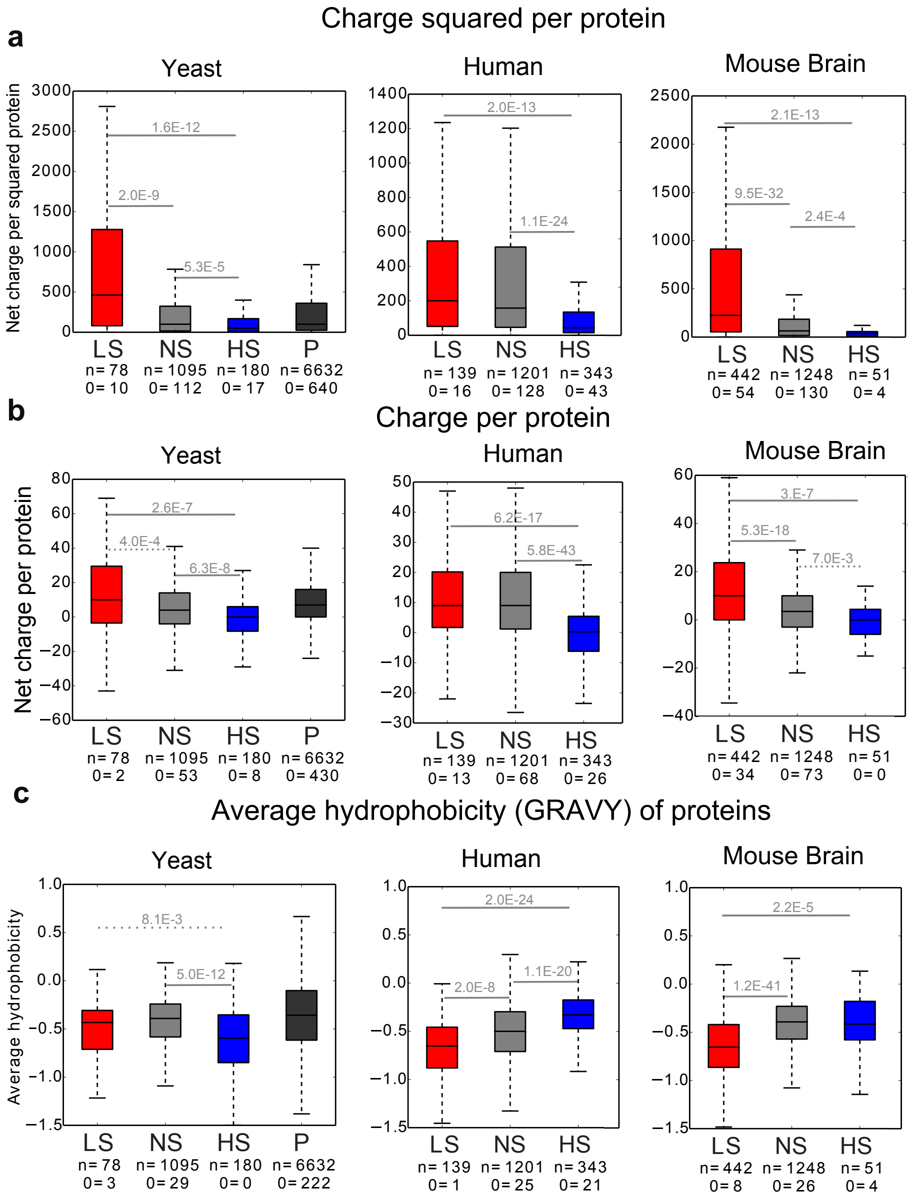

Figure 3.11: Analysis of net charge and hydrophobicity of proteins. (a-b) Net charge of proteins was calculated by assigning His, Arg, and Lys a charge of +1, Glu and Asp a charge of -1, and all other residues a charge of 0. The charges of all residues in the protein were then added together to obtain net charge. This net charge was then either squared and plotted (a) or plotted directly (b). (c) Average hydropobicity of proteins, as calculated by the grand average of hydrophobicity (GRAVY) index.

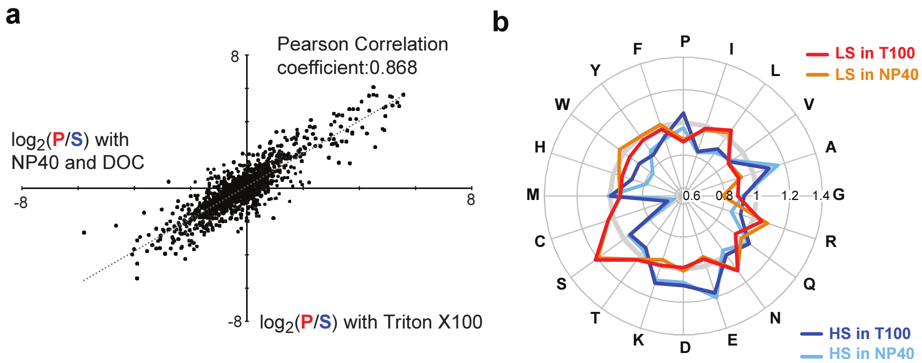

Figure 3.12: Comparison of results obtained using the two detergents NP40 and Triton X-100. (a) Comparison of ratios for proteins in yeast identified in common using different detergents for lysis. (b) Comparison of amino acid abundances in yeast for cells lysed with different detergents. Figure reproduced from [4].

prised of less hydrophobic residues, LS proteins in mouse and human had lower hydrophobicity scores (Figure 3.11c), as calculated by the GRAVY index.

3.1.7 Choice of detergent does not significantly affect solubility of LS proteins

There was a possibility that the choice of detergent might affect the perceived solubility of proteins, as well as the analyses that were based on said perceived solubility. To rule that out, we repeated the experiment in yeast (originally carried out using Triton X-100) using Igepal as the detergent, and observed little disparity between the results obtained with Triton X-100. The ratios obtained in the two experiments displayed a Pearson correlation coefficient of 0.868 (Figure 3.12a). There was also little variation observed in the amino acid enrichment between the samples treated with different detergents (Figure 3.12b). Observed effects between organisms is thus better explained by interspecies variation than by differences in detergents used.

3.1.8 LS proteins are more disordered and contain more molecular recognition features (MoRFs) and eukarotic linear motifs (ELMs) than higher solubility (HS) proteins

The study by Ng et al. also highlighted that longer, less soluble proteins that are ubiquitinated also tended to be more disordered, which would be consistent with our observation that LS proteins are more highly charged and less hydrophobic. To ascertain if this was true in our dataset, we utilized two disorder prediction algorithms, DISOPRED and IUPRED, to analyze the disorder of proteins. We found that LS proteins were more disordered than their NS and HS counterparts, as shown in Figure 3.13a and Figure 3.13b. This is consistent with the findings of the previous study [87] that proteins in the low solubility fraction after heat shock tended to be more disordered than ones that remained in the soluble fraction. LS proteins were also found to contain more coiled regions (Figure 3.13c), as predicted by PSIPRED. Given that coil regions are essentially regions that were not classified to have a fixed helical or sheet structure, this is consistent with LS proteins being more disordered.

While LS proteins contained more disorder, we also wanted to see whether this disorder resided in distinct regions of the protein or was dispersed throughout the protein, prompting us to examine how many disordered patches the proteins contained. In all three organisms, LS proteins contained more disordered patches (Figure 3.14a) of at least five residues in length. This trend was only preserved in mouse after normalization to length (Figure 3.14b), suggesting that the increased number of disordered patches in yeast and human was likely to be merely due to a correlation with length.

Disordered regions that assume an ordered conformation upon binding to partners, also known as molecular recognition features (MoRFs)[121] were enriched in the LS fraction (Figure 3.15a). Given that MoRFs must occur in regions of disorder, and that LS proteins have been shown earlier to be more disordered, it was logical to examine whether the enrichment for MoRFs was due to the increased disorder of LS proteins. However, the enrichment for MoRFs in LS proteins was still observed even after normalizing to percentage disorder of the proteins (Figure 3.15b).

While LS proteins had more MoRFs and contained more patches of disorder,

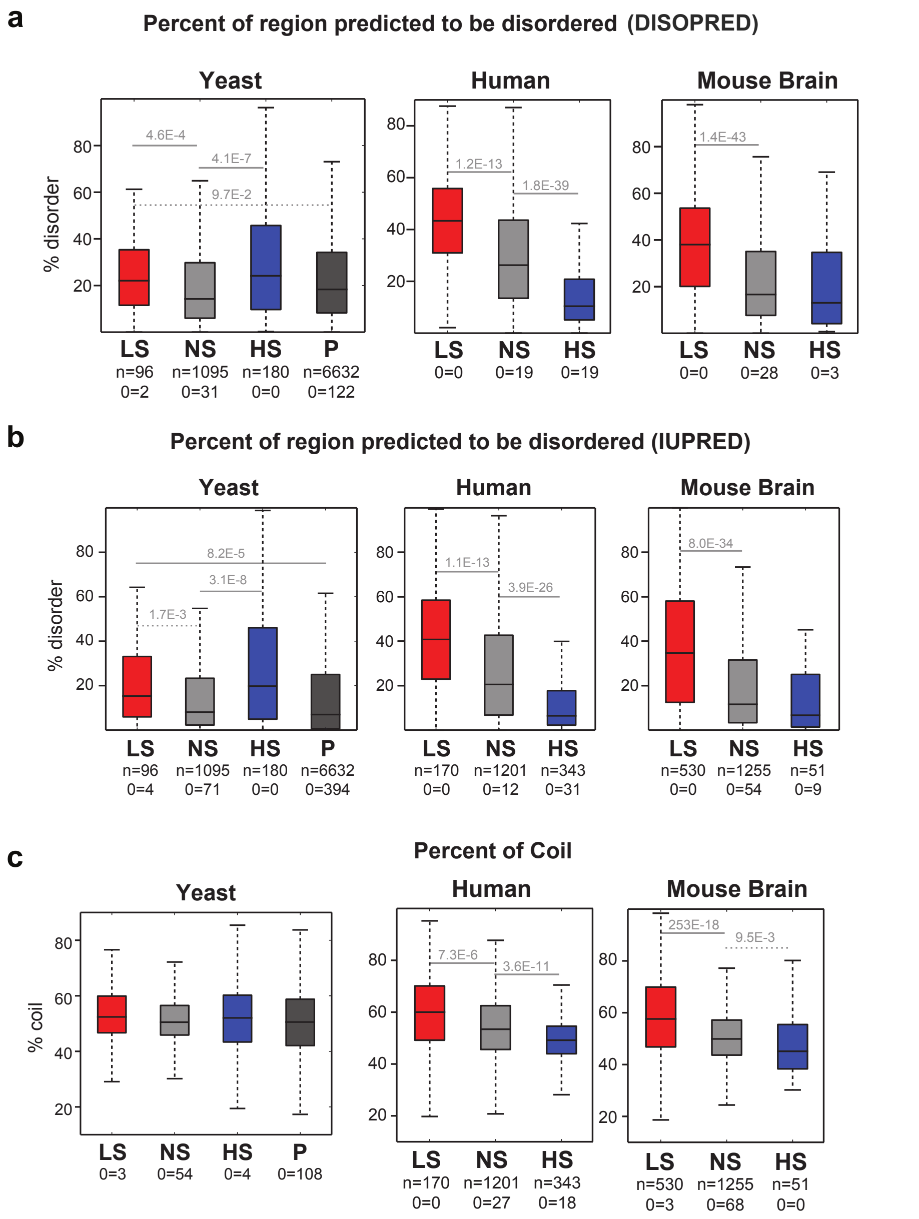

Figure 3.13: Comparison of protein percentage disorder as predicted by DISOPRED and IUPRED. (a-b) Distributions of the percent disorder of proteins as predicted by DISOPRED(a) and IUPRED (b) in the indicated bins.(c) Percent of each protein predicted by PSIPRED to be an unstructured coil. Figure reproduced from [4].

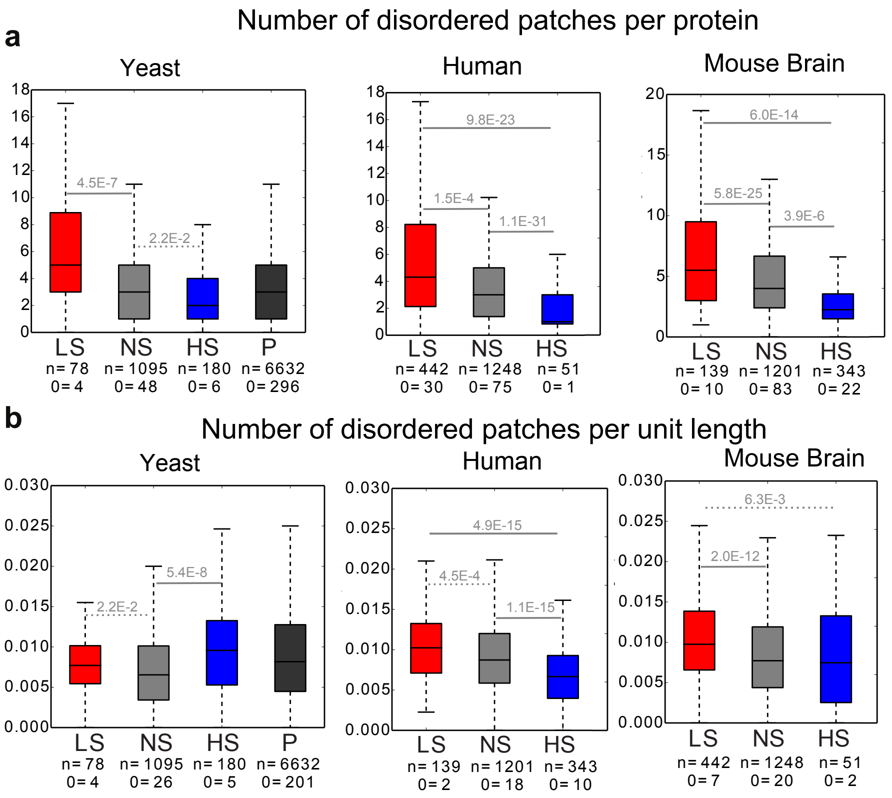

Figure 3.14: Number of disordered regions in each protein, and the number of MoRFs within each such region.(a) Number of disordered regions (at least 5 residues long) per protein. (b) Number of disordered regions (at least 5 residues long) per unit length.

it was possible that this could simply be dependent on length, and that each disordered patch would generally contain the same average number of MoRFs. In mouse and human, LS proteins contained more MoRFs per disordered patch (Figure 3.16), indicating that not only were there more MoRFs, they were also more densely clustered in disordered regions.

Short stretches of conserved amino acid sequences, known as eukarotic linear motifs (ELMs), are often involved in mediating protein-protein interactions [25]. In contrast with MoRFs, ELMs are not necessarily within disordered regions. A search of the ELM database on ANCHOR showed that yeast LS proteins contained

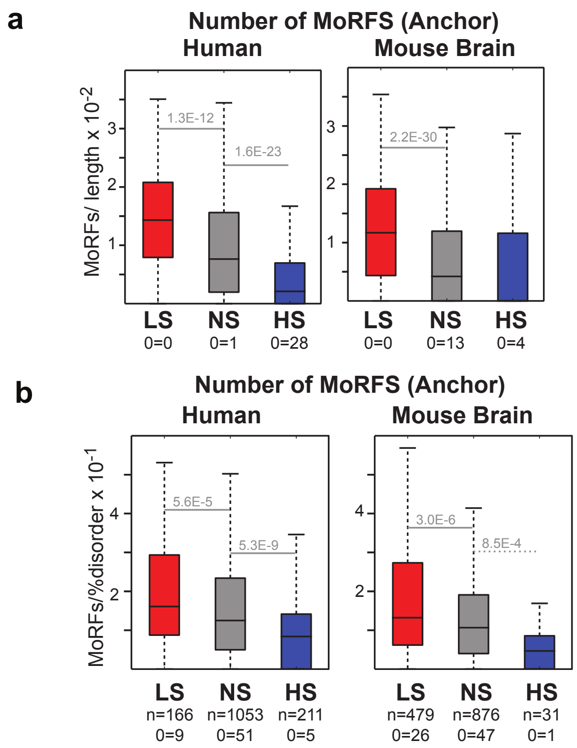

Figure 3.15: Number of molecular recognition features (MoRFs) within each protein, as predicted by the ANCHOR database. (a and b) The number of MoRFs per protein, normalized to protein length (a) and percentage disorder (b). Only proteins predicted to be at least 10% disordered were considered. Figure reproduced from [4].

more linear motifs per unit length (Figure 3.17a). However, this trend was no longer significant after normalization to disorder (Figure 3.17b).

3.1.9 RNase treatment increases the solubility of RNA associated proteins, but does not affect the overall properties of low solubility proteins

An enrichment of RNA-associated GO terms was previously observed in Section 3.1.2. We wanted to investigate the possibility that proteins binding to RNA associated assemblies might increase their propensity to precipitate during centrifugation. In order to do so, we carried out a triple-SILAC experiment with one

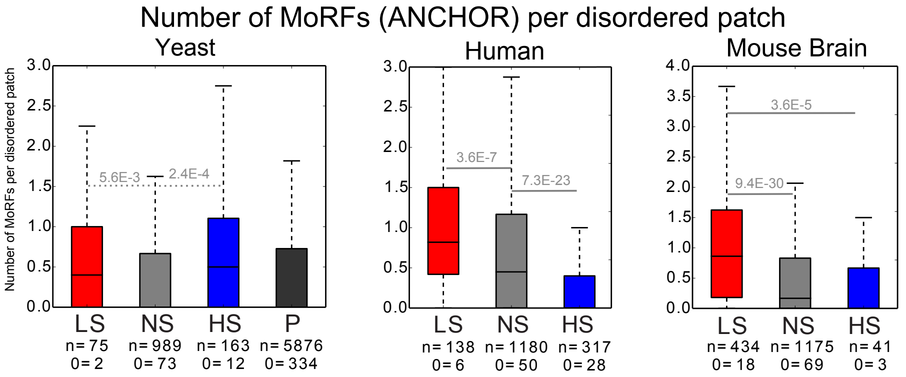

Figure 3.16: Average number of MoRFs as determined by ANCHOR within each disordered patch (of 5 residues or longer).

sample (heavy) RNase treated during lysis but prior to separation of low solubility proteins, and two more samples not treated with RNase. The pellet from the heavy and medium experiments, as well as the supernatant from the light experiment, were mixed in a 1:1:1 ratio by mass and analyzed by quantitative mass spectrometry (Figure 3.18a). We identified a group of proteins that do indeed appear more soluble following RNase treatment (Figure 3.18b and c). Many of these proteins were associated with RNA-associated GO terms (Table A.16Table A.17). However, there was no significant change to the trends observed for yeast, save for a slight reduction in beta sheet content for the LS bin proteins (Figure 3.18d). Therefore,while the solubility of RNA-binding proteins was affected by the presence or absence of RNase treatment, the solubility of most LS proteins was also not affected and most LS proteins remained in the lower solubility bin upon RNase treatment.

3.1.10 Coding sequences for LS proteins contain a lower GC content in yeast

After examining numerous traits of polypeptides that could influence their solubility, we next looked at nucleic acid-based traits to see if any of them correlated with solubility. Coding sequences for LS proteins tended to have a lower GC content than genes encoding NS and HS proteins (Figure 3.19). Higher GC content has

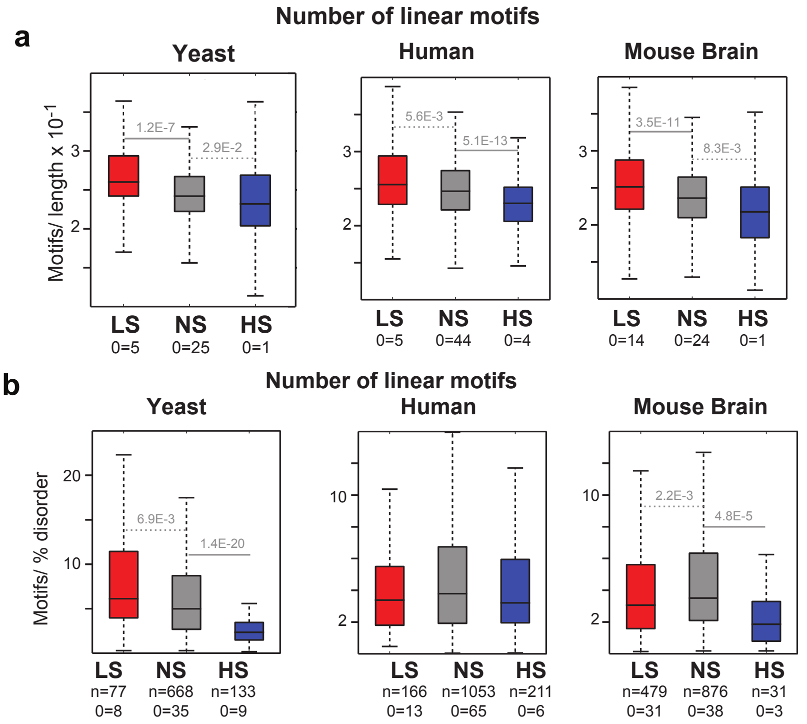

Figure 3.17: Number of eukarotic linear motifs (ELMs) within each protein, as predicted by the ANCHOR database. (a and b) The number of ELMs per protein, normalized to protein length (a) and percentage disorder (b). Only proteins predicted to be at least 10% disordered were considered. Figure reproduced from [4].

been associated with lower translation rates [100]. This appears to be consistent with findings that suggest lower translation rates provide more time for proteins to fold, and help to minimize misfolding [112] and thus aberrant amyloid or amorphous aggregation, potentially allowing for increased solubility.

3.1.11 LS proteins possess numerous traits that distinguish them from HS proteins

We have identified a number of traits that distinguish LS proteins from HS proteins (p-value summary in Table A.10, Table A.11, and Table A.12). In contrast with previous findings [29, 47], they are actually less abundant than HS proteins. They

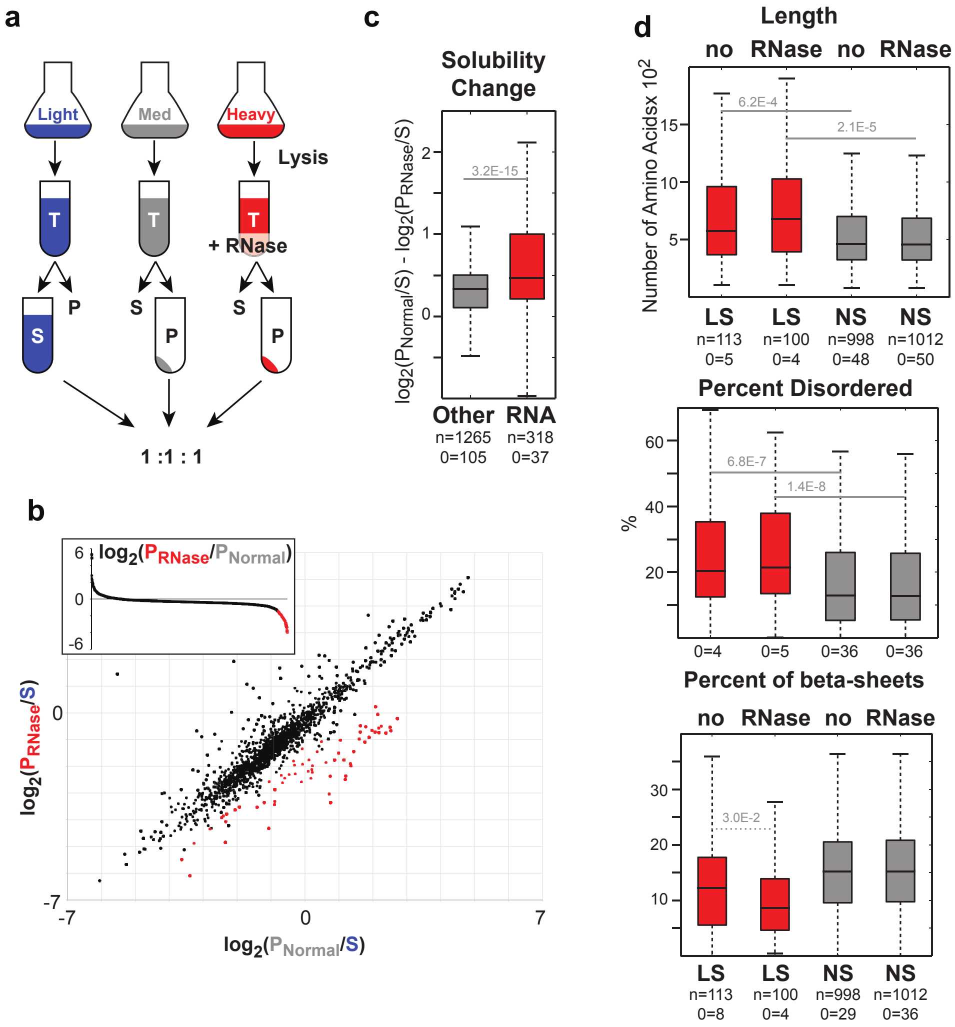

Figure 3.18: Analysis of RNase treatment on protein solubility. (a) Experimental setup used to determine RNA influence on protein solubility. (b) The $\log_2$ ratio distribution of low solubility proteins from yeast cell lysate treated with RNase (P_{RNase}) against those without the treatment (P_{Normal}) (top left). The scatter plot shows the comparison of $\log_2$ ratios between the two experiments; proteins which became more soluble after RNase treatment are highlighted in red. (c) Solubility change of proteins annotated with RNA associated GO terms and proteins without these GO terms after RNase treatment. (d) Boxplots showing the distribution in samples with and without RNase treatment of protein length (top), percent disorder (middle), and percent beta-sheets (bottom). Figure reproduced from [4].

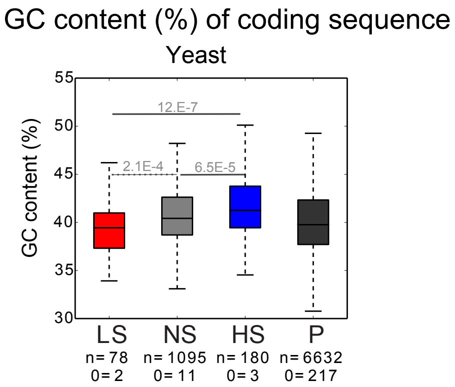

Figure 3.19: Percentage of coding sequences of each protein comprised of GC.

tend to be longer and more disordered than HS proteins. While they tend to be more aggregation prone (at least in yeast), their higher propensity for features that mediate protein-protein interactions such as ELMs and MoRFs, as well as features such as LCRs, suggests a relationship between low solubility and functional aggregation. While a small number of proteins associated with RNA-associated GO terms displayed a higher solubility upon treatment with RNase, this trend was not observed for majority of the proteins analyzed. In addition to traits of polypeptides affecting solubility, we also observed the LS proteins in yeast having a lower GC content than HS proteins, which could potentially allow HS proteins more time to fold during translation, contributing to their increased solubility.

3.2 Models to predict protein solubility

After analyzing various traits to check for relationships with solubility, we next aimed to use those traits to generate a model which would then aid in the prediction of solubility. With these models it is hoped that we can predict which proteins are of low solubility via *in silico* methods. *In silico* methods would allow for the prediction of a protein's solubility without any necessary *a priori* knowledge derived from experiments such as mass spectrometry and Western Blotting as previously detailed. Yeast was chosen as the organism to model as more comprehensive data is available on it compared to mouse and human, and it provides a simpler system

Table 3.1: mean squared error (MSE) and correlation coefficients from each of the five iterations of cross-validation, as well as the average values obtained.

Test set	MSE	Correlation coefficient
1	0.9246397	0.5158573
2	0.7163011	0.410455
3	0.8629595	0.4492167
4	0.6873905	0.5446077
5	0.7896966	0.4339884
Average	0.79619748	0.47082502

to build a model on as it possesses a simpler proteome without tissue specific compositional differences. The reduced time needed to generate strains for experiments would also speed up validation of the models.

3.2.1 Multiple regression model

One model we attempted to build in order to model protein solubility was a multiple linear regression model. A linear regression model has the general form

$$y_i = b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{ip} + c + \varepsilon_i$$

where y denotes one instance of the response variable (which is solubility in this case), x denotes the values of various traits such as length for that instance, b denotes the coefficient associated with each trait, c denotes a constant, and ε denotes the error for this instance.

The 85 protein traits analyzed (shown in Table A.10) were fitted in a stepwise fashion to a multiple regression model with elastic net regularization using the glmnet library for R ([45]). This was done by first finding the value at which the intercept alone would have the least deviance from the actual value, then at each subsequent iteration adding in the variable that explained the variation the most. The elastic net regularization was used to add a penalty for complexity to the model, minimizing overfitting, whereby the model would fit to and model noise within the data and lowering predictive performance. The available data was randomly split into 5 equal sized sets, with each set in turn used as a test set to assess

Table 3.2: Estimates of the coefficients from the regularized multiple regression model.

Trait	Estimate
(Intercept)	-8.17E-02
Membrane localization	2.18E+00
Vacuole localization	1.82E+00
Average charge per residue	1.43E+00
Number of disulphide bonds	4.00E-02
Percent abundance of Leu	2.07E-02
Number of close stop codons	8.64E-03
Net charge per protein	1.11E-03
Alpha helix propensity	2.32E-04
Net charge squared per protein	5.32E-05
Percent coiled coil	-2.53E-03
Percent disorder (IUPRED)	-2.80E-03
Percent negative residues	-5.50E-03
Percent abundance of Gln	-6.58E-03
Percent abundance of Glu	-1.42E-02
Percent abundance if Pro	-1.78E-02
Percent abundance of Ala	-3.29E-02
Number of MoRFs per unit length	-2.28E-01

the accuracy of the model, with the remaining 4 of them were used to as the training set to build the model. The sum of the squared differences between observed and predicted values for each data point was used to calculate the mean squared error (MSE), a measure of accuracy of the model.

The averaged MSE value over the 5 folds of cross-validation was 0.796 (Table 3.1). Table 3.2 shows the estimates of the coefficients for each trait that was included in the model. Given that the values of solubility ratios obtained from the mass spectrometry experiments range from roughly -5 to 5, the MSE obtained only affords a crude approximation. Given the performance of the multiple regression model, other models were explored with the aim of improved performance.

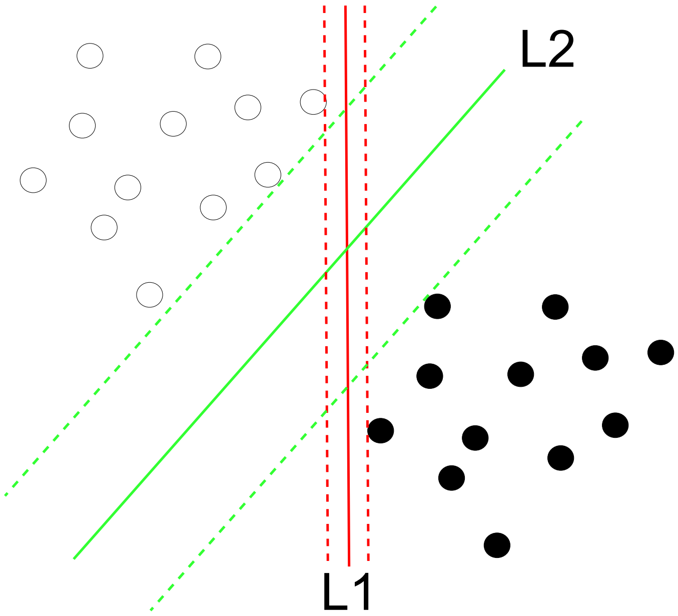

Figure 3.20: Illustration of SVM hyperplanes. While both the lines (hyperplanes) L1 and L2 can divide the two classes (black and white), L2 has a larger margin of separation and is chosen for higher performance. Two-dimensional space is used in this example.

3.2.2 Support vector machine

We next built a support vector machine to distinguish between LS and HS proteins, using the R library e1071, utilizing the the same traits that were considered for the regression model (Table A.10), save for the amyloid aggregation propensities as predicted by TANGO and AGGRESCAN, as the SVM is not compatible with categorical variables. Support vector machines project the data into a higher dimensional space using mapping functions, called kernel functions, and aim to find a multi-dimensional plane, referred to as a hyperplane, within this high dimensional space that can most effectively separate the two classes (in this case, LS and HS proteins) of data points. As shown in Figure 3.20, even though multiple hyperplanes might be able to divide the two classes, the one with the maximum separation distance is chosen. Usage of the two extreme bins from our dataset was done as an initial test to see if it was possible to distinguish the most dissimilar pro-

Table 3.3: The prediction performance of the SVM on distinguishing LS and HS proteins.

Test set	Sensitivity	Specificity	Precision	Accuracy	FDR	MCC	AUC
1	0.6667	0.9677	0.6667	0.8696	0.3333	0.6972	0.8831
2	0.8333	0.8571	0.8333	0.8478	0.1667	0.6844	0.8392
3	0.6875	0.9000	0.6875	0.8261	0.3125	0.6081	0.8147
4	0.8000	0.8889	0.8000	0.8696	0.2000	0.6471	0.8039
5	0.6842	0.8889	0.6842	0.8043	0.3158	0.5925	0.8063
Average	0.7343	0.9005	0.7343	0.8435	0.2657	0.6459	0.8294

teins based on the traits we already examined. Thus, LS and HS bins were pooled and then randomly split into five equal sample sizes. Each of the partitions in turn was used as a test set while the rest was used as the training set. The SVM was built with a linear kernel, and a cost of 10. The cost parameter provides a penalty to each possible hyperplane based on how many incorrect classifications that hyperplane makes. Higher cost values penalize incorrect classification more, but are prone to overfitting.

The performance of the SVM is shown in Table 3.3. Several measures were used to assess the performance of the SVM. Sensitivity, or true positive rate, is the ratio of true positives called by the model to the total number of positive data points. Specificity, or true negative rate, is the ratio of true negatives called by the model to the total number of negative data points. Precision, is the ratio of true positives called by the model to the total number of positives called by the model. Accuracy is the proportion of correctly called data points by the model. The false discovery rate (FDR) is the proportion of positives called by the model that are false positives. matthews correlation coefficient (MCC) is a measure of quality of the predictions by the model, ranging from -1 to 1, with 1 being perfect predictions, -1 being total disagreement with the observations, and 0 being no better than random chance. area under curve (AUC) represents the probability that the model will rank a chosen positive data point higher than a negative one, under the assumption that positive data points should rank higher. As shown in Table 3.3, the model is able to distinguish LS and HS proteins.

The SVM model obtained an FDR of 0.2657, which would mean that if using

Table 3.4: The prediction performance of the SVM on distinguishing LS and non-LS proteins.

Test set	Sensitivity	Specificity	Precision	Accuracy	FDR	MCC	AUC
1	0	0.9512	0	0.9123	1	-0.0458	0.4785
2	0.0909	0.9313	0.0909	0.8772	0.9091	0.0213	0.5102
3	0.5	0.9563	0.5	0.9244	0.5	0.4397	0.7119
4	0.375	0.9509	0.375	0.924	0.625	0.2805	0.6207
5	0.7143	0.9329	0.7143	0.924	0.2857	0.4403	0.6498
Average	0.336	0.9445	0.336	0.9124	0.664	0.2272	0.5942

the model as a preliminary *in silico* screen before biological validation, the error rate is low enough to still allow for narrowing down of candidates.

Next, the SVM was built using the LS, NS, and HS bins, rather than just the LS and HS bins, as that would provide a more useful tool for biologists. Performance of the model Table 3.4 is lower than that on just the LS and HS proteins, suggesting that it is indeed more challenging to distinguish between LS and NS proteins than it is to distinguish between LS and HS proteins. The high FDR of 0.664 will pose issues in that more often than not, an LS protein identified by the model will not actually be an LS protein. More improvements to the model will need to be made in order for it to be useful in analyzing proteins *in silico* prior to biological validation.

One possible avenue by which both models could be improved is the inclusion of more traits that correlate well with protein solubility. Future efforts to improve the model will include searching for these additional traits that can help improve the predictive power of models. It is hoped that with sufficient additions to both models, they can provide a tool to allow users to predict the solubility of proteins *in silico* without first having to run more time consuming and costly experiments such as mass spectrometry.

Chapter 4

Discussion

The section of the thesis will cover several areas

- Section 4.1: Ratios obtained from quantitative mass spectrometry are not directly indicative of absolute ratios
- Section 4.2: Feature analysis of LS proteins highlights differences between organisms and points to link between functional aggregation and low solubility
  - Section 4.2.1: Analysis of features of LS proteins highlights inter-organism differences
  - Section 4.2.2: LS proteins possess distinct features that differentiate them from other proteins
  - Section 4.2.3: LS proteins may be involved in functional aggregation
- Section 4.3: Generation of models to predict solubility of proteins
- Section 4.4: Future work

4.1 Ratios obtained from quantitative mass spectrometry are not directly indicative of absolute ratios

The aim of this project was to identify proteins that displayed lower solubility and characterize features that may contribute to a protein's solubility or lack thereof.

In order to gauge the proportion of a protein that is present in the low solubility fraction, as opposed to the amount of protein within the low solubility fraction, usage of the soluble fraction as a reference is necessary. In the absence of a reference such as the soluble fraction, it is not possible to distinguish between a highly abundant protein with high solubility and a low solubility protein with low abundance, as both proteins could very well have the same absolute abundance within the low solubility fraction. While the ratios obtained by the normalization method used are reflective of the partitioning of a protein between the low solubility and soluble fractions, they should not be taken directly as an absolute ratio of said partitioning. For example, in yeast the amount of protein recovered from the low solubility pellet was typically 2% of the amount recovered from the supernatant. Mixing the proteins obtained from each fraction in a $1{:}1$ ratio by mass would over-represent proteins in the pellet by a corresponding amount. The main advantage of the method used here is to allow for the accounting of protein abundance, and thus represent an improvement over previously used absolute quantification [29, 47].

4.2 Feature analysis of LS proteins highlights differences between organisms and points to association between functional aggregation and low solubility

After obtaining ratios representing the solubility of various proteins, we categorized proteins into lower solubility (LS), normal solubility (NS), and higher solubility (HS) based on their solubility. We then moved on to examine what traits distinguished LS proteins from other proteins. We examined several traits, and found some to aid in distinguishing LS proteins from other proteins, as well as to understand better what role LS proteins might have in relation to the proteome.

4.2.1 Analysis of features of LS proteins highlights inter-organism differences

Some of the features we examined, such as length, showed a consistent trend across all three model organisms. In many cases, however, there was agreement between mouse and human, but not with yeast. This could be due to greater evolutionary distance between yeast and the other two species.

One such case was when amyloid aggregation prone proteins were predicted to be more prevalent in the LS bin in yeast, but more prevalent in the HS bin in mouse and human. One possibility is that *S. cerevisiae* is more tolerant of amyloid aggregation-prone proteins as it is able to retain harmful aggregates within the mother cell during budding, allowing the daughter cell to be free of harmful aggregated species [115]. Another possibility is that *S. cerevisiae* is simply less able to disaggregate proteins as it lacks Hsp110 disaggregases [101], resulting in amyloid aggregation-prone proteins forming a larger portion of the LS fraction than metazoans. This is consistent with the observation that LS proteins are predicted to be more amyloid aggregation prone in yeast but not in our human and mouse samples Figure 3.6. As a unicellular fungal organism, this ensures the generation of offspring with greater fitness. Mouse and human cells do not possess these mechanisms for dealing with aggregated species. As multicellular organisms, the fitness of the whole organism would necessitate some mechanism of disposing of aggregated proteins, as both daughter cells after cell division are still part of the whole organism and contribute to its fitness. In view of this, it is likely that the significance of LS proteins in yeast might differ greatly from LS proteins in metazoans.

4.2.2 LS proteins possess distinct features that differentiate them from other proteins

As mentioned in Section 4.2.1, LS proteins tended to display more amyloid aggregation propensity in yeast, but not in mouse and human. The first possibility discussed was the evolutionary distance between yeast, a fungal organism, and the human and mouse samples, which are metazoan. A second explanation for this observation might be that since the algorithms used utilize hydrophobicity as well as beta-sheet propensity to predict amyloid aggregation [78], proteins that contain lower hydrophobicity scores would be predicted to be less amyloid aggregation prone. Consistent with this, mouse and human LS proteins did in fact obtain lower hydrophobicity scores relative to NS and HS proteins. LS proteins in the mouse and human datasets had a lower abundance for certain residues possessing hydrophobic side chains, such as isoleucine, leucine, and valine, which might result in these proteins being deemed less amyloid aggregation prone by the algorithms. Given that many proteins in the LS bin have previously not been characterized as amyloid aggregation prone, it is possible that they may possess novel features contributing to their low solubility that might not be taken into consideration by existing algorithms.

Several other features we examined highlighted trends that are consistent with work by other previously published studies. In human and mouse, proteins in the LS bin were found to be enriched in coils, relative to the NS and HS bins, consistent with the higher percentage of predicted disorder. Our finding that low solubility proteins tended to be more disordered is consistent with published work by Lai et al. that disordered proteins can participate in the assembly of functional aggregates. The finding by Ng et al. that less soluble, albeit ubiquitinated, proteins (albeit after heat stress) are more disordered, and that longer proteins were depleted in the soluble fraction supports this notion. Longer proteins potentially have more capacity to contain regions capable of participating in interactions with other proteins. Even after normalizing for length, proteins in the LS bin also contained more MoRFs, ELMs, and LCRs, features which are known to mediate intermolecular interactions between proteins. LCRs and MoRFs have been shown to bind multiple partners [90], consistent with the idea of disordered proteins forming interactions with multiple other partner proteins. Previous work by Kato et al. shows that LCRs being necessary and sufficient for the formation of hydrogels by proteins, underscoring the role of LCRs in functional assemblies.

Phosphorylation within unstructured regions of disordered proteins has also been shown to regulate formation of aggregates [51]. Mouse and human LS proteins were enriched in serine residues which are commonly utilized as phosphorylation targets. Given the association between aberrant hyperphosphorylation and glycosylation with pathologies such as alzheimer's disease (AD) [5, 13, 49], an enrichment for serine that would normally result in benign structures may be responsible for the assembly of harmful aggregates in the case of certain proteins.

Proteins associated with RNA were enriched in the LS bin, which is consistent with the functional aggregation hypothesis. RRM-1 domains, which are known to be involved in RNA binding, were indeed enriched in the LS bin in the human and mouse datasets. GO analysis also highlighted the enrichment of RNA processing and RNA binding of proteins within the LS bin of all three organisms. Consistent with this, the LS bin in mouse showed an enrichment for arginine, which is commonly involved in binding to nucleic acids, as well as a tendency to have a more positive net charge. LCRs have also been known to play a role in the assembly of RNA granules, a functional aggregate that stores mRNAs and allows an additional layer of regulation of gene expression [64, 102]. RNA packaging and transport to cellular extremities is essential to the complex architecture of neuronal cells [37]. While the solubility of RNA-related proteins was affected by the presence of RNase, these proteins were not exclusive to the low solubility fraction, and trends observed in the absence of RNase persisted after RNase addition. Macromolecular assemblies containing RNA are therefore unlikely to be the dominating feature of the low solubility fraction.

Many of the features that low solubility proteins possess suggest that their low solubility status might be due to biologically relevant interactions with other macromolecules within the cell, rather than merely aberrant interactions that need to be abrogated.

4.2.3 LS proteins may be involved in functional aggregation

LS proteins were found to possess several traits which are involved in protein-protein interactions, such as linear motifs, MoRFs, LCRs, and RNA binding regions. The tendency to contain more of these features raises the possibility that LS proteins are multivalent, allowing a single protein to form interactions with multiple partners simultaneously. Multivalent proteins can thus interact and form the building blocks for functional macromolecular complexes. This suggests that low solubility proteins may be involved in the formation of functional aggregates [24], which are distinct from toxic aggregates caused by aberrant folding or other insults. Functional aggregates are macromolecular assemblies in biological systems in a dynamic and reversible fashion [127]. Such functional assemblies can be formed via liquid-liquid demixing [75] and phase transitions *in vivo*, which result in such assemblies forming a phase distinct from the aqueous solution. This separation from the aqueous phase would be consistent with their presence in the low solubility fraction. LS proteins in yeast were also enriched in glutamine, which is known to form non-toxic aggregates [53], consistent with the idea that aggregates formed by LS proteins are functional rather than toxic.

The traits that distinguish LS proteins are associated with protein-protein interactions. Coupled with a tendency to be longer, LS proteins are thus able to assemble into functional macromulecular complexes. This highlights the notion that LS proteins may be of lower solubility because they are assembled into functional, rather than toxic, aggregates.

4.3 Generation of models to predict solubility of proteins

After looking at many traits and examining their relationship with solubility, we then utilized these traits to build models to predict the solubility of proteins *in silico*. A linear regression model was built in a stepwise fashion, adding in traits that could help improve the accuracy of the model, with elastic net regularization to help prevent overfitting. A support vector machine was also built to distinguish between LS and HS proteins. Both models utilized cross validation, splitting the data into equal sized portions and using each as a test set in turn, with the non-testing set being used for training the model.

While the regularized regression model is not able to provide precise estimates of the ratio of proteins, it does allow for an approximation of said ratios. Interestingly, not all of the traits that displayed an ability to distinguish LS proteins from NS and HS proteins were selected for inclusion into the model. Notably, length, linear motifs, and LCRs were not selected for inclusion into the model. This suggests that they could possibly be correlated with one or more of the traits that were included in the model. This would limit their contribution to the predictive power of the model and possibly prevent their inclusion into the model. The number of close stop codons (codons one base pair substitution away from a stop codon) [1] was included in the model, highlighting how features not in the actual protein sequence may be correlated to solubility and thus useful in predicting it. The ability to predict a numerical ratio without further categorization, coupled with the current margins of error, make it difficult to use the model as a preliminary step prior to biological validation. Incorporating a cutoff to identify proteins as LS would be an approach to explore, and could improve usability of the model.

The support vector machine is able to effectively distinguish between LS andHS proteins with a low FDR and high accuracy. However, the model is unable to reli-

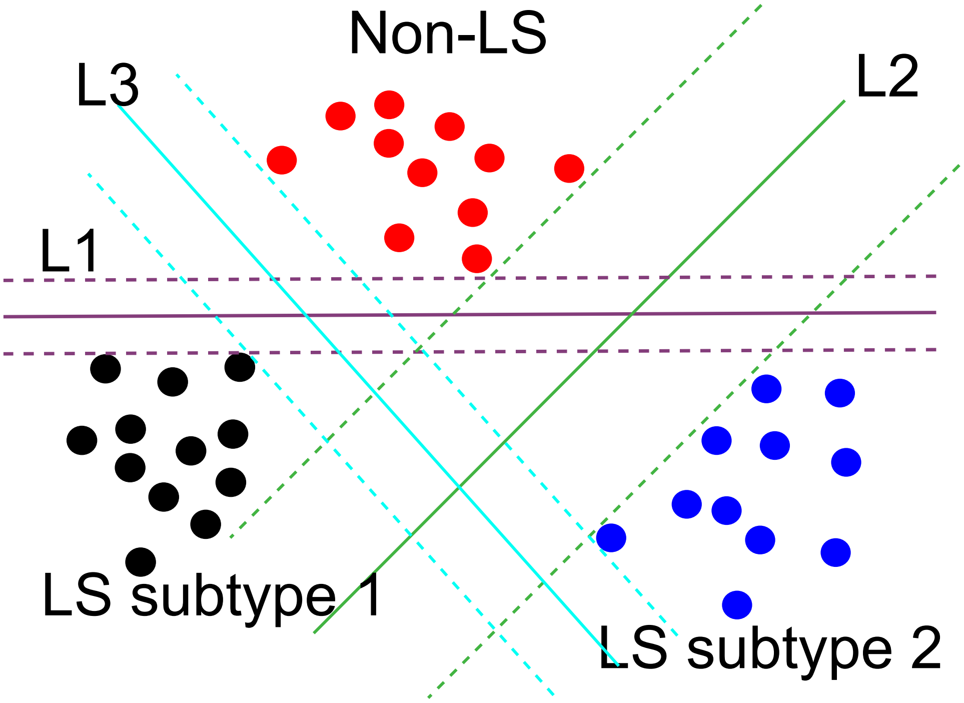

Figure 4.1: SVM and distinguishing multiple subtypes of a class. Potential subtypes of LS proteins might reduce separation margins if subtypes are grouped together (L1). Separating subtypes for classification may improve separation margin (L2 and L3).

ably detect LS proteins and distinguish them from NS proteins. This is likely due to the fact that LS and NS proteins are likely to be much more similar than LS and HS proteins are, and thus more challenging to distinguish. Identifying more traits that differentiate LS proteins from other proteins will be crucial in improving the predictive power of the models. With sufficiently high accuracy and low FDR, the model will be able to provide a useful preliminary *in silico* step to identify proteins of a lower solubility which a user can then attempt to validate using *in vivo* or *in vitro* methods.

Currently, both models work on the assumption that LS proteins are a specific subset of proteins that can be defined by a common set of traits. Given that proteins could conceivably be detected as low solubility by being within toxic aggregates as well as part of large functional assemblies, it is quite possible that LS proteins might be comprised of two or more distinct classes of proteins. If this is indeed so, the models might have difficulty attempting to distinguish the LS proteins and non-LS proteins. For instance, if LS proteins were composed of subtype 1, with high trait A, B, C and low D, E, F, and subtype 2 with low A, B, C, and high D, E, F, the linear regression model would be hard pressed to find a single set of coefficient values whereby both subtypes 1 and 2 would be scored highly. Contributions to the prediction from A, B, and C would tend to oppose D, E, and F, making it difficult to distinguish LS proteins from proteins that scored high in all of the traits as well as proteins that scored low in all of the traits. In order to address this, each subtype might be modeled separately, being assigned their own set of coefficients. Likewise, for the SVM if the two subtypes of LS proteins are different enough, it is possible that a separating margin such as L1 (Figure 4.1) might only be able to distinguish both LS protein subtypes (in blue and black) from non-LS proteins (red) with only a small separation margin, resulting in reduced performance of the model. Attempting to distinguish just one subtype from non-LS proteins such as with separating margin L2 might allow for a wider separation margin. This wider separation margin would allow for more confident and accurate separation of proteins. In addition to identification of additional traits, future work will also involve clustering proteins to determine if there are indeed various subtypes of LS proteins. Adjusting the models to distinguish one specific subtype at a time is a potential avenue to improve the performance of the SVM and the regression model.

4.4 Future work

The traits examined here have provided a plausible explanation for why low solubility proteins are indeed low solubility. Further work could follow up on modifying certain traits correlated with solubility (such as length), while keeping other traits as close to unchanged as possible, and monitoring any change in solubility. This work, while technically challenging, will aid in establishing whether the relationship seen in this study is causal, or merely correlated.

Examining additional traits may also shed more light on factors that contribute to the solubility of a protein. The amino acid index (AAindex) [65] is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids. Given the presence of certain localized features, such as MoRFs, it would be interesting to assess the scores of sliding windows of various sizes across proteins using the amino acid index. This approach aims to identify certain stretches within proteins that may be correlated with solubility. There are also large scale datasets characterizing half-lives of yeast proteins [11] as well as localization upon stress [86] that could be examined for their correlation to solubility. Identifying additional factors such as these and including them in the models will also likely contribute to the quality of the models developed.

In addition to examining various protein traits, future work will also involve investigating whether LS proteins are composed of multiple subtypes of proteins. Clustering algorithms such as Markov clustering [38] could be used to cluster the LS proteins and determine if there are indeed different subtypes of LS proteins. If LS proteins are comprised of different subtypes, modelling individual subtypes would be useful in improving the performance of the regression and SVM models in predicting LS proteins.

Currently, the yeast data which was used to build the models only covers approximately one-third of the proteome. A deeper mass spectrometry run which covers much more of the proteome would provide more data points with which to train the model on, increasing predictive power, while ironically reducing the number of proteins not identified in the mass spectrometry run that a user might actually need to predict via the model. Certain proteins might be of too low abundance to be detected and quantified by mass spectrometry, or possibly removed in the pre-clearance step, and be unobtainable by our current methods, which could potentially limit how many more data points we can acquire with a deeper mass spectrometry experiment.

Both of the models in Section 3.2 have various improvements that can be made to them. The linear model could be converted into a binary classifier to hopefully overcome its large error margin, as well as increase usability with predictions of ”Low solubility” or ”Not low solubility” being more intuitive and easier to work with than a continuous variable ratio. The SVM, while it has performed reasonably in distinguishing LS proteins from their HS counterparts, could be potentially more useful and powerful if it trained on the full dataset. This would allow it to distinguish between NS and LS proteins, which would require more power and precision than the current model that distinguishes LS and HS proteins.

By gaining a better understanding of what traits proteins possess that contribute to their solubility, it can allow us to better understand the mechanisms of protein solubility. With tools and models that allow for the prediction of protein solubility *in silico*, it will be possible to design experiments, keeping protein solubility in mind, without having to actually assess the solubility of the entire proteome empirically.

Chapter 5

Conclusion

Protein solubility is an integral component of protein homeostasis. Disruption of homeostasis can result in toxic aggregates that are detrimental to cell fitness. Neurodegenerative diseases are crippling diseases that have been associated with protein aggregation in cells. Understanding more about protein solubility and aggregation will be crucial to gleaning insight into the pathologies and designing treatments.

In order to examine traits associated with low solubility, we utilized quatitative mass spectrometry and an internal standard to account for protein abundance to allow us to obtain the solubility of proteins. After classifying proteins as low, normal, or high solubility, we examined several features of proteins and analyzed them for correlation with solubility. We have thus identified a number of features that distinguish low solubility proteins under unstressed conditions.

Several of the features we examined exhibited trends that were consistent across all three model organisms studied, in spite of the vast evolutionary distances between some organisms. In many cases, human and mouse samples showed a similarity that was not shared by mouse proteins, highlighting the evolutionary disparity between the fungal and metozoan systems, and suggesting that factors underlying solubility might differ greatly in these systems. Proteins found to be of low solubility were found to be longer, less abundant, and more disordered. Said proteins also contained more coiled regions, LCRs, ELMs, and MoRFs, suggesting a relationship between solubility and number of potential interaction partners. This points to a possible connection between low solubility proteins and functional aggregates. LS protein encoding genes also had a lower GC content, highlighting a relationship between coding sequence and the solubility of the encoded protein.

We also generated two models with which to estimate protein solubility, a linear regression model as well as a support vector machine. Both models provide usable estimates for solubility, but improving their accuracy will require uncovering more traits that correlate with protein solubility. Accurate algorithms to predict protein solubility will aid greatly in the experimental biology that will be crucial in understanding this complex aspect of protein homeostasis.

The work presented here highlights several traits that characterize low solubility proteins, as well as highlighting the possibility of low solubility proteins being of low solubility due to a role in functional aggregation. The models generated are starting steps towards providing a high throughput *in silico* platform for predicting protein solubility.

Bibliography

[1] F. Agostini, M. Vendruscolo, and G. G. Tartaglia. Sequence-based prediction of protein solubility. *Journal of Molecular Biology*, 421(2-3): 237–241, Aug. 2012. ISSN 1089-8638. doi:10.1016/j.jmb.2011.12.005. → pages 12, 50

[2] H. Aguilaniu. Asymmetric Inheritance of Oxidatively Damaged Proteins During Cytokinesis. *Science*, 299(5613):1751–1753, Mar. 2003. ISSN 00368075, 10959203. doi:10.1126/science.1080418. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1080418. → pages 2, 19

[3] E. M. Ahmed. Hydrogel: Preparation, characterization, and applications. *Journal of Advanced Research*, July 2013. ISSN 20901232. doi:10.1016/j.jare.2013.07.006. URL http://linkinghub.elsevier.com/retrieve/pii/S2090123213000969. → pages 4

[4] R. F. Albu, G. T. Chan, M. Zhu, E. T. C. Wong, F. Taghizadeh, X. Hu, A. E. Mehran, J. D. Johnson, J. Gsponer, and T. Mayor. A feature analysis of lower solubility proteins in three eukaryotic systems. *Journal of Proteomics*, Oct. 2014. ISSN 1876-7737. doi:10.1016/j.jprot.2014.10.011. → pages iii, 6, 9, 14, 16, 18, 20, 22, 23, 25, 26, 27, 29, 31, 33, 35, 37, 38

[5] A. d. C. Alonso, T. Zaidi, M. Novak, I. Grundke-Iqbal, and K. Iqbal. Hyperphosphorylation induces self-assembly of into tangles of paired helical filaments/straight filaments. *Proceedings of the National Academy of Sciences*, 98(12):6923–6928, June 2001. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.121119298. URL http://www.pnas.org/cgi/doi/10.1073/pnas.121119298. → pages 48

[6] A. Alves-Rodrigues, L. Gregori, and M. E. Figueiredo-Pereira. Ubiquitin, cellular inclusions and their role in neurodegeneration. *Trends in Neurosciences*, 21(12):516–520, Dec. 1998. ISSN 01662236.

doi:10.1016/S0166-2236(98)01276-4. URL http://linkinghub.elsevier.com/retrieve/pii/S0166223698012764. → pages 3

[7] E. Angot, J. A. Steiner, C. M. Lema Tom, P. Ekstrm, B. Mattsson, A. Bjrklund, and P. Brundin. Alpha-Synuclein Cell-to-Cell Transfer and Seeding in Grafted Dopaminergic Neurons In Vivo. *PLoS ONE*, 7(6): e39465, June 2012. ISSN 1932-6203. doi:10.1371/journal.pone.0039465. URL http://dx.plos.org/10.1371/journal.pone.0039465. → pages 3

[8] S. Auli, T. T. Le, F. Moda, S. Abounit, S. Corvaglia, L. Casalis, S. Gustincich, C. Zurzolo, F. Tagliavini, and G. Legname. Defined -synuclein prion-like molecular assemblies spreading in cell culture. *BMC Neuroscience*, 15(1):69, 2014. ISSN 1471-2202. doi:10.1186/1471-2202-15-69. URL http://www.biomedcentral.com/1471-2202/15/69. → pages 3

[9] M. Babu, J. Vlasblom, S. Pu, X. Guo, C. Graham, B. D. M. Bean, H. E. Burston, F. J. Vizeacoumar, J. Snider, S. Phanse, V. Fong, Y. Y. C. Tam, M. Davey, O. Hnatshak, N. Bajaj, S. Chandran, T. Punna, C. Christopolous, V. Wong, A. Yu, G. Zhong, J. Li, I. Stagljar, E. Conibear, S. J. Wodak, A. Emili, and J. F. Greenblatt. Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. *Nature*, 489(7417):585–589, Sept. 2012. ISSN 0028-0836, 1476-4687. doi:10.1038/nature11354. URL http://www.nature.com/doifinder/10.1038/nature11354. → pages 19

[10] F. Bardag-Gorce, J. Vu, L. Nan, N. Riley, J. Li, and S. W. French. Proteasome inhibition induces cytokeratin accumulation in vivo. *Experimental and Molecular Pathology*, 76(2):83–89, Apr. 2004. ISSN 00144800. doi:10.1016/j.yexmp.2003.11.004. URL http://linkinghub.elsevier.com/retrieve/pii/S0014480003001382. → pages 2

[11] A. Belle, A. Tanay, L. Bitincka, R. Shamir, and E. K. O'Shea. Quantification of protein half-lives in the budding yeast proteome. *Proceedings of the National Academy of Sciences of the United States of America*, 103(35):13004–13009, Aug. 2006. ISSN 0027-8424. doi:10.1073/pnas.0605420103. → pages 53

[12] J. L. Biedler, S. Roffler-Tarlov, M. Schachner, and L. S. Freedman. Multiple neurotransmitter synthesis by human neuroblastoma cell lines and clones. *Cancer Research*, 38(11 Pt 1):3751–3757, Nov. 1978. ISSN 0008-5472. → pages 17

[13] M. Broncel, J. Falenski, S. Wagner, C. Hackenberger, and B. Koksch. How Post-Translational Modifications Influence Amyloid Formation: A Systematic Study of Phosphorylation and Glycosylation in Model Peptides. *Chemistry - A European Journal*, 16(26):7881–7888, May 2010. ISSN 09476539. doi:10.1002/chem.200902452. URL http://doi.wiley.com/10.1002/chem.200902452. → pages 48

[14] M. Bucciantini, E. Giannoni, F. Chiti, F. Baroni, L. Formigli, J. Zurdo, N. Taddei, G. Ramponi, C. M. Dobson, and M. Stefani. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, 416(6880):507–511, Apr. 2002. ISSN 0028-0836. doi:10.1038/416507a. → pages 1, 2

[15] M. R. Bufalino, B. DeVeale, and D. van der Kooy. The asymmetric segregation of damaged proteins is stem cell-type dependent. *The Journal of Cell Biology*, 201(4):523–530, May 2013. ISSN 0021-9525, 1540-8140. doi:10.1083/jcb.201207052. URL http://www.jcb.org/cgi/doi/10.1083/jcb.201207052. → pages 2, 19

[16] J. Cheng, H. Saigo, and P. Baldi. Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. *Proteins: Structure, Function, and Bioinformatics*, 62(3):617–629, Nov. 2005. ISSN 08873585. doi:10.1002/prot.20787. URL http://doi.wiley.com/10.1002/prot.20787. → pages 12

[17] J. M. Cherry, E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. T. Chan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, K. Karra, C. J. Krieger, S. R. Miyasato, R. S. Nash, J. Park, M. S. Skrzypek, M. Simison, S. Weng, and E. D. Wong. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(Database issue):D700–705, Jan. 2012. ISSN 1362-4962. doi:10.1093/nar/gkr1029. → pages 9, 12

[18] F. Chiti. Mutational analysis of the propensity for amyloid formation by a globular protein. *The EMBO Journal*, 19(7):1441–1449, Apr. 2000. ISSN 14602075. doi:10.1093/emboj/19.7.1441. URL http://emboj.embopress.org/cgi/doi/10.1093/emboj/19.7.1441. → pages 1

[19] F. Chiti and C. M. Dobson. Protein misfolding, functional amyloid, and human disease. *Annual Review of Biochemistry*, 75:333–366, 2006. ISSN 0066-4154. doi:10.1146/annurev.biochem.75.101304.123901. → pages 3

[20] F. Chiti and C. M. Dobson. Amyloid formation by globular proteins under native conditions. *Nature Chemical Biology*, 5(1):15–22, Jan. 2009. ISSN 1552-4450. doi:10.1038/nchembio.131. URL http://www.nature.com/doifinder/10.1038/nchembio.131. → pages 3

[21] F. Chiti, M. Stefani, N. Taddei, G. Ramponi, and C. M. Dobson. Rationalization of the effects of mutations on peptide andprotein aggregation rates. *Nature*, 424(6950):805–808, Aug. 2003. ISSN 0028-0836, 1476-4679. doi:10.1038/nature01891. URL http://www.nature.com/doifinder/10.1038/nature01891. → pages 3

[22] S. A. Comyn, G. T. Chan, and T. Mayor. False start: cotranslational protein ubiquitination and cytosolic protein quality control. *Journal of Proteomics*, 100:92–101, Apr. 2014. ISSN 1876-7737. doi:10.1016/j.jprot.2013.08.005. → pages 2

[23] O. Conchillo-Sol, N. S. de Groot, F. X. Avils, J. Vendrell, X. Daura, and S. Ventura. AGGRESCAN: a server for the prediction and evaluation of ”hot spots” of aggregation in polypeptides. *BMC Bioinformatics*, 8(1):65, 2007. ISSN 14712105. doi:10.1186/1471-2105-8-65. URL http://www.biomedcentral.com/1471-2105/8/65. → pages 4, 12

[24] A. Cumberworth, G. Lamour, M. Babu, and J. Gsponer. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochemical Journal*, 454(3):361–369, Sept. 2013. ISSN 0264-6021, 1470-8728. doi:10.1042/BJ20130545. URL http://www.biochemj.org/bj/454/bj4540361.htm. → pages 49

[25] N. E. Davey, K. Van Roey, R. J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel, and T. J. Gibson. Attributes of short linear motifs. *Molecular BioSystems*, 8(1):268, 2012. ISSN 1742-206X, 1742-2051. doi:10.1039/c1mb05231d. URL http://xlink.rsc.org/?DOI=c1mb05231d. → pages 34

[26] D. C. David, N. Ollikainen, J. C. Trinidad, M. P. Cary, A. L. Burlingame, and C. Kenyon. Widespread Protein Aggregation as an Inherent Part of Aging in C. elegans. *PLoS Biology*, 8(8):e1000450, Aug. 2010. ISSN 1545-7885. doi:10.1371/journal.pbio.1000450. URL http://dx.plos.org/10.1371/journal.pbio.1000450. → pages 21

[27] M. P. C. David, G. P. Concepcion, and E. A. Padlan. Using simple artificial intelligence methods for predicting amyloidogenesis in antibodies. *BMC*

bioinformatics, 11:79, 2010. ISSN 1471-2105. doi:10.1186/1471-2105-11-79. → pages 4

[28] S. J. Davis, E. A. Davies, M. G. Tucknott, E. Y. Jones, and P. A. van der Merwe. The role of charged residues mediating low affinity protein-protein recognition at the cell surface by CD2. *Proceedings of the National Academy of Sciences of the United States of America*, 95(10):5490–5494, May 1998. ISSN 0027-8424. → pages 28

[29] L. M. F. de Godoy, J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Frhlich, T. C. Walther, and M. Mann. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, 455(7217):1251–1254, Oct. 2008. ISSN 0028-0836, 1476-4687. doi:10.1038/nature07341. URL http://www.nature.com/doifinder/10.1038/nature07341. → pages 11, 21, 22, 37, 46

[30] N. S. de Groot, F. X. Aviles, J. Vendrell, and S. Ventura. Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. *The FEBS journal*, 273 (3):658–668, Feb. 2006. ISSN 1742-464X. doi:10.1111/j.1742-4658.2005.05102.x. → pages 5

[31] T. F. DeLuca, J. Cui, J.-Y. Jung, K. C. St. Gabriel, and D. P. Wall. Roundup 2.0: enabling comparative genomics for over 1800 genomes. *Bioinformatics*, 28(5):715–716, Mar. 2012. ISSN 1367-4803, 1460-2059. doi:10.1093/bioinformatics/bts006. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bts006. → pages 11

[32] G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, 4(5):P3, 2003. ISSN 1465-6914. → pages 11

[33] E. W. Dent and P. W. Baas. Microtubules in neurons as information carriers. *Journal of Neurochemistry*, 129(2):235–239, Apr. 2014. ISSN 00223042. doi:10.1111/jnc.12621. URL http://doi.wiley.com/10.1111/jnc.12621. → pages 21

[34] C. M. Dobson. Protein folding and misfolding. *Nature*, 426(6968): 884–890, Dec. 2003. ISSN 1476-4687. doi:10.1038/nature02261. → pages 2

[35] Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, Aug. 2005. ISSN 1367-4803, 1460-2059. doi:10.1093/bioinformatics/bti541. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bti541. → pages 12

[36] Z. Dosztanyi, B. Meszaros, and I. Simon. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, 25(20):2745–2746, Oct. 2009. ISSN 1367-4803, 1460-2059. doi:10.1093/bioinformatics/btp518. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp518. → pages 12

[37] E. Doxakis. RNA binding proteins: a common denominator of neuronal function and dysfunction. *Neuroscience Bulletin*, 30(4):610–626, Aug. 2014. ISSN 1673-7067, 1995-8218. doi:10.1007/s12264-014-1443-7. URL http://link.springer.com/10.1007/s12264-014-1443-7. → pages 49

[38] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7): 1575–1584, Apr. 2002. ISSN 1362-4962. → pages 53

[39] S. Escusa-Toret, W. I. M. Vonk, and J. Frydman. Spatial sequestration of misfolded proteins by a dynamic chaperone pathway enhances cellular fitness during stress. *Nature Cell Biology*, 15(10):1231–1243, Sept. 2013. ISSN 1465-7392, 1476-4679. doi:10.1038/ncb2838. URL http://www.nature.com/doifinder/10.1038/ncb2838. → pages 2

[40] A.-M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, and L. Serrano. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology*, 22(10): 1302–1306, Oct. 2004. ISSN 1087-0156. doi:10.1038/nbt1012. URL http://www.nature.com/doifinder/10.1038/nbt1012. → pages 4, 12

[41] A. D. Ferrao-Gonzales, S. O. Souto, J. L. Silva, and D. Foguel. The preaggregated state of an amyloidogenic protein: Hydrostatic pressure converts native transthyretin into the amyloidogenic state. *Proceedings of the National Academy of Sciences*, 97(12):6445–6450, June 2000. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.97.12.6445. URL http://www.pnas.org/cgi/doi/10.1073/pnas.97.12.6445. → pages 1

[42] R. D. Finn, J. Clements, and S. R. Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(suppl): W29–W37, July 2011. ISSN 0305-1048, 1362-4962. doi:10.1093/nar/gkr367. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkr367. → pages 11

[43] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta. Pfam: the protein families database. *Nucleic Acids Research*, 42(D1):D222–D230, Jan. 2014. ISSN 0305-1048, 1362-4962. doi:10.1093/nar/gkt1223. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1223. → pages 11

[44] D. M. Fowler, A. V. Koulov, W. E. Balch, and J. W. Kelly. Functional amyloid–from bacteria to humans. *Trends in Biochemical Sciences*, 32(5): 217–224, May 2007. ISSN 0968-0004. doi:10.1016/j.tibs.2007.03.003. → pages 3

[45] J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010. ISSN 1548-7660. → pages 13, 40

[46] B. Frost, R. L. Jacks, and M. I. Diamond. Propagation of Tau Misfolding from the Outside to the Inside of a Cell. *Journal of Biological Chemistry*, 284(19):12845–12852, May 2009. ISSN 0021-9258, 1083-351X. doi:10.1074/jbc.M808759200. URL http://www.jbc.org/cgi/doi/10.1074/jbc.M808759200. → pages 3

[47] S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman. Global analysis of protein expression in yeast. *Nature*, 425(6959):737–741, Oct. 2003. ISSN 0028-0836, 1476-4679. doi:10.1038/nature02046. URL http://www.nature.com/doifinder/10.1038/nature02046. → pages 11, 21, 22, 37, 46

[48] N. Gilks. Stress Granule Assembly Is Mediated by Prion-like Aggregation of TIA-1. *Molecular Biology of the Cell*, 15(12):5383–5398, Sept. 2004. ISSN 1059-1524. doi:10.1091/mbc.E04-08-0715. URL http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E04-08-0715. → pages 3, 4

[49] C.-X. Gong, F. Liu, I. Grundke-Iqbal, and K. Iqbal. Post-translational modifications of tau protein in Alzheimers disease. *Journal of Neural*

Transmission, 112(6):813–838, June 2005. ISSN 0300-9564, 1435-1463. doi:10.1007/s00702-004-0221-0. URL http://link.springer.com/10.1007/s00702-004-0221-0. → pages 48

[50] J. Gsponer and M. Babu. Cellular Strategies for Regulating Functional and Nonfunctional Protein Aggregation. *Cell Reports*, 2(5):1425–1437, Nov. 2012. ISSN 22111247. doi:10.1016/j.celrep.2012.09.036. URL http://linkinghub.elsevier.com/retrieve/pii/S2211124712003671. → pages 12, 21

[51] J. Gsponer, M. E. Futschik, S. A. Teichmann, and M. M. Babu. Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation. *Science*, 322(5906):1365–1368, Nov. 2008. ISSN 0036-8075, 1095-9203. doi:10.1126/science.1163581. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1163581. → pages 48

[52] J. L. Guo and V. M.-Y. Lee. Seeding of Normal Tau by Pathological Tau Conformers Drives Pathogenesis of Alzheimer-like Tangles. *Journal of Biological Chemistry*, 286(17):15317–15331, Apr. 2011. ISSN 0021-9258, 1083-351X. doi:10.1074/jbc.M110.209296. URL http://www.jbc.org/cgi/doi/10.1074/jbc.M110.209296. → pages 3

[53] R. Halfmann, S. Alberti, R. Krishnan, N. Lyle, C. O'Donnell, O. King, B. Berger, R. Pappu, and S. Lindquist. Opposing Effects of Glutamine and Asparagine Govern Prion Formation by Intrinsically Disordered Proteins. *Molecular Cell*, 43(1):72–84, July 2011. ISSN 10972765. doi:10.1016/j.molcel.2011.05.013. URL http://linkinghub.elsevier.com/retrieve/pii/S1097276511003807. → pages 28, 49

[54] S. L. Hands and A. Wyttenbach. Neurotoxic protein oligomerisation associated with polyglutamine diseases. *Acta Neuropathologica*, 120(4): 419–437, Oct. 2010. ISSN 1432-0533. doi:10.1007/s00401-010-0703-0. → pages 25

[55] P. M. Harrison and M. Gerstein. A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes. *Genome Biology*, 4(6):R40, 2003. ISSN 1465-6914. doi:10.1186/gb-2003-4-6-r40. → pages 12, 13

[56] B. B. Holmes and M. I. Diamond. Prion-like Properties of Tau Protein: The Importance of Extracellular Tau as a Therapeutic Target. *Journal of*

Biological Chemistry, 289(29):19855–19861, July 2014. ISSN 0021-9258, 1083-351X. doi:10.1074/jbc.R114.549295. URL http://www.jbc.org/cgi/doi/10.1074/jbc.R114.549295. → pages 3

[57] P. V. Hornbeck, I. Chabra, J. M. Kornhauser, E. Skrzypek, and B. Zhang. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *PROTEOMICS*, 4(6):1551–1561, June 2004. ISSN 1615-9853, 1615-9861. doi:10.1002/pmic.200300772. URL http://doi.wiley.com/10.1002/pmic.200300772. → pages 12

[58] J. D. Hunter. Matplotlib: A 2d Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. ISSN 1521-9615. doi:10.1109/MCSE.2007.55. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4160265. → pages 11

[59] D.-H. Hyun, M. Lee, B. Halliwell, and P. Jenner. Proteasomal inhibition causes the formation of protein aggregates containing a wide range of proteins, including nitrated proteins. *Journal of Neurochemistry*, 86(2): 363–373, July 2003. ISSN 0022-3042. → pages 2

[60] S. Idicula-Thomas and P. V. Balaji. Understanding the relationship between the primary structure of proteins and their amyloidogenic propensity: clues from inclusion body formation. *Protein engineering, design & selection: PEDS*, 18(4):175–180, Apr. 2005. ISSN 1741-0126. doi:10.1093/protein/gzi022. → pages 4

[61] A. Iwata. HDAC6 and Microtubules Are Required for Autophagic Degradation of Aggregated Huntingtin. *Journal of Biological Chemistry*, 280(48):40282–40292, Sept. 2005. ISSN 0021-9258, 1083-351X. doi:10.1074/jbc.M508786200. URL http://www.jbc.org/cgi/doi/10.1074/jbc.M508786200. → pages 2

[62] J. A. Johnston. Aggresomes: A Cellular Response to Misfolded Proteins. *The Journal of Cell Biology*, 143(7):1883–1898, Dec. 1998. ISSN 00219525. doi:10.1083/jcb.143.7.1883. URL http://www.jcb.org/cgi/doi/10.1083/jcb.143.7.1883. → pages 2

[63] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2): 195–202, Sept. 1999. ISSN 0022-2836. doi:10.1006/jmbi.1999.3091. → pages 12

[64] M. Kato, T. Han, S. Xie, K. Shi, X. Du, L. Wu, H. Mirzaei, E. Goldsmith, J. Longgood, J. Pei, N. Grishin, D. Frantz, J. Schneider, S. Chen, L. Li, M. Sawaya, D. Eisenberg, R. Tycko, and S. McKnight. Cell-free Formation of RNA Granules: Low Complexity Sequence Domains Form Dynamic Fibers within Hydrogels. *Cell*, 149(4):753–767, May 2012. ISSN 00928674. doi:10.1016/j.cell.2012.04.017. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867412005144. → pages 4, 48, 49

[65] S. Kawashima, H. Ogata, and M. Kanehisa. AAindex: Amino Acid Index Database. *Nucleic Acids Research*, 27(1):368–369, Jan. 1999. ISSN 0305-1048. → pages 52

[66] Y. E. Kim, M. S. Hipp, A. Bracher, M. Hayer-Hartl, and F. U. Hartl. Molecular chaperone functions in protein folding and proteostasis. *Annual Review of Biochemistry*, 82:323–355, 2013. ISSN 1545-4509. doi:10.1146/annurev-biochem-060208-092442. → pages 2

[67] G. Kleiger and T. Mayor. Perilous journey: a tour of the ubiquitinproteasome system. *Trends in Cell Biology*, 24(6):352–359, June 2014. ISSN 09628924. doi:10.1016/j.tcb.2013.12.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0962892413002274. → pages 2

[68] T. P. J. Knowles, M. Vendruscolo, and C. M. Dobson. The amyloid state and its association with protein misfolding diseases. *Nature Reviews Molecular Cell Biology*, 15(6):384–396, May 2014. ISSN 1471-0072, 1471-0080. doi:10.1038/nrm3810. URL http://www.nature.com/doifinder/10.1038/nrm3810. → pages 3

[69] R. R. Kopito. Aggresomes, inclusion bodies and protein aggregation. *Trends in Cell Biology*, 10(12):524–530, Dec. 2000. ISSN 09628924. doi:10.1016/S0962-8924(00)01852-3. URL http://linkinghub.elsevier.com/retrieve/pii/S0962892400018523. → pages 2

[70] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3):567–580, Jan. 2001. ISSN 0022-2836. doi:10.1006/jmbi.2000.4315. → pages 12

[71] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, May 1982. ISSN 0022-2836. → pages 12, 29

[72] J. Lai, C. H. Koh, M. Tjota, L. Pieuchot, V. Raman, K. B. Chandrababu, D. Yang, L. Wong, and G. Jedd. Intrinsically disordered proteins aggregate at fungal cell-to-cell channels and regulate intercellular connectivity. *Proceedings of the National Academy of Sciences*, 109(39):15781–15786, Sept. 2012. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1207467109. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1207467109. → pages 4, 48

[73] S. H. Lecker. Protein Degradation by the Ubiquitin-Proteasome Pathway in Normal and Disease States. *Journal of the American Society of Nephrology*, 17(7):1807–1819, June 2006. ISSN 1046-6673, 1533-3450. doi:10.1681/ASN.2006010083. URL http://www.jasn.org/cgi/doi/10.1681/ASN.2006010083. → pages 2

[74] B. Levine, N. Mizushima, and H. W. Virgin. Autophagy in immunity and inflammation. *Nature*, 469(7330):323–335, Jan. 2011. ISSN 1476-4687. doi:10.1038/nature09782. → pages 2

[75] P. Li, S. Banjade, H.-C. Cheng, S. Kim, B. Chen, L. Guo, M. Llaguno, J. V. Hollingsworth, D. S. King, S. F. Banani, P. S. Russo, Q.-X. Jiang, B. T. Nixon, and M. K. Rosen. Phase transitions in the assembly of multivalent signalling proteins. *Nature*, 483(7389):336–340, Mar. 2012. ISSN 1476-4687. doi:10.1038/nature10879. → pages 49

[76] S. Li, T. Izumi, J. Hu, H. H. Jin, A.-A. A. Siddiqui, S. G. Jacobson, D. Bok, and M. Jin. Rescue of enzymatic function for disease-associated RPE65 proteins containing various missense mutations in non-active sites. *The Journal of Biological Chemistry*, 289(27):18943–18956, July 2014. ISSN 1083-351X. doi:10.1074/jbc.M114.552117. → pages 1

[77] T.-h. Lin, R. F. Murphy, and Z. Bar-Joseph. Discriminative motif finding for predicting protein subcellular localization. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 8(2):441–451, Apr. 2011. ISSN 1557-9964. doi:10.1109/TCBB.2009.82. → pages 12

[78] R. Linding, J. Schymkowitz, F. Rousseau, F. Diella, and L. Serrano. A Comparative Study of the Relationship Between Protein Structure and -Aggregation in Globular and Intrinsically Disordered Proteins. *Journal of Molecular Biology*, 342(1):345–353, Sept. 2004. ISSN 00222836. doi:10.1016/j.jmb.2004.06.088. URL http://linkinghub.elsevier.com/retrieve/pii/S0022283604007715. → pages 47

[79] J. Lowe, A. Blanchard, K. Morrell, G. Lennox, L. Reynolds, M. Billett, M. Landon, and R. J. Mayer. Ubiquitin is a common factor in intermediate filament inclusion bodies of diverse type in man, including those of Parkinson's disease, Pick's disease, and Alzheimer's disease, as well as Rosenthal fibres in cerebellar astrocytomas, cytoplasmic bodies in muscle, and mallory bodies in alcoholic liver disease. *The Journal of Pathology*, 155(1):9–15, May 1988. ISSN 0022-3417, 1096-9896. doi:10.1002/path.1711550105. URL http://doi.wiley.com/10.1002/path.1711550105. → pages 3

[80] K. Lundmark, G. T. Westermark, S. Nystrom, C. L. Murphy, A. Solomon, and P. Westermark. Transmissibility of systemic amyloidosis by a prion-like mechanism. *Proceedings of the National Academy of Sciences*, 99(10):6979–6984, May 2002. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.092205999. URL http://www.pnas.org/cgi/doi/10.1073/pnas.092205999. → pages 3

[81] S. K. Maji, M. H. Perrin, M. R. Sawaya, S. Jessberger, K. Vadodaria, R. A. Rissman, P. S. Singru, K. P. R. Nilsson, R. Simon, D. Schubert, D. Eisenberg, J. Rivier, P. Sawchenko, W. Vale, and R. Riek. Functional Amyloids As Natural Storage of Peptide Hormones in Pituitary Secretory Granules. *Science*, 325(5938):328–332, July 2009. ISSN 0036-8075, 1095-9203. doi:10.1126/science.1173155. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1173155. → pages 3

[82] L. Malinovska, S. Kroschwald, M. C. Munder, D. Richter, and S. Alberti. Molecular chaperones and stress-inducible protein-sorting factors coordinate the spatiotemporal distribution of protein aggregates. *Molecular Biology of the Cell*, 23(16):3041–3056, Aug. 2012. ISSN 1059-1524. doi:10.1091/mbc.E12-03-0194. URL http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E12-03-0194. → pages 2

[83] S. Maurer-Stroh, M. Debulpaep, N. Kuemmerer, M. Lopez de la Paz, I. C. Martins, J. Reumers, K. L. Morris, A. Copland, L. Serpell, L. Serrano, J. W. H. Schymkowitz, and F. Rousseau. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature Methods*, 7(3):237–242, Mar. 2010. ISSN 1548-7105. doi:10.1038/nmeth.1432. → pages 4

[84] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. Chang, and C.-C. Lin. *Misc Functions of the Department of Statistics (e1071)*. Sept. 2014. → pages 13

[85] E. Mezey, A. Dehejia, G. Harta, M. Papp, M. Polymeropoulos, and M. Brownstein. Alpha synuclein in neurodegenerative disorders: Murderer or accomplice? *Nature Medicine*, 4(7):755–757, July 1998. ISSN 1078-8956. doi:10.1038/nm0798-755. URL http://www.nature.com/doifinder/10.1038/nm0798-755. → pages 3

[86] R. Narayanaswamy, M. Levy, M. Tsechansky, G. M. Stovall, J. D. O'Connell, J. Mirrielees, A. D. Ellington, and E. M. Marcotte. Widespread reorganization of metabolic enzymes into reversible assemblies upon nutrient starvation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(25):10147–10152, June 2009. ISSN 1091-6490. doi:10.1073/pnas.0812771106. → pages 53

[87] A. H. M. Ng, N. N. Fang, S. A. Comyn, J. Gsponer, and T. Mayor. System-wide Analysis Reveals Intrinsically Disordered Proteins Are Prone to Ubiquitylation after Misfolding Stress. *Molecular & Cellular Proteomics*, 12(9):2456–2467, Sept. 2013. ISSN 1535-9476, 1535-9484. doi:10.1074/mcp.M112.023416. URL http://www.mcponline.org/cgi/doi/10.1074/mcp.M112.023416. → pages iii, 5, 11, 22, 32, 48

[88] J. D. O'Connell, M. Tsechansky, A. Royall, D. R. Boutz, A. D. Ellington, and E. M. Marcotte. A proteomic survey of widespread protein aggregation in yeast. *Molecular bioSystems*, 10(4):851–861, Apr. 2014. ISSN 1742-2051. doi:10.1039/c3mb70508k. → pages 5

[89] C. W. O'Donnell, J. Waldisphl, M. Lis, R. Halfmann, S. Devadas, S. Lindquist, and B. Berger. A method for probing the mutational landscape of amyloid structure. *Bioinformatics (Oxford, England)*, 27(13): i34–42, July 2011. ISSN 1367-4811. doi:10.1093/bioinformatics/btr238. → pages 4

[90] C. J. Oldfield, J. Meng, J. Y. Yang, M. Q. Yang, V. N. Uversky, and A. K. Dunker. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics*, 9(Suppl 1):S1, 2008. ISSN 1471-2164. doi:10.1186/1471-2164-9-S1-S1. URL http://www.biomedcentral.com/1471-2164/9/S1/S1. → pages 48

[91] T. E. Oliphant. Python for Scientific Computing. *Computing in Science & Engineering*, 9(3):10–20, 2007. ISSN 1521-9615. doi:10.1109/MCSE.2007.58. URL

http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4160250. → pages 11

[92] H. Olzscha, S. M. Schermann, A. C. Woerner, S. Pinkert, M. H. Hecht, G. G. Tartaglia, M. Vendruscolo, M. Hayer-Hartl, F. U. Hartl, and R. M. Vabulas. Amyloid-like Aggregates Sequester Numerous Metastable Proteins with Essential Cellular Functions. *Cell*, 144(1):67–78, Jan. 2011. ISSN 00928674. doi:10.1016/j.cell.2010.11.050. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867410013723. → pages 1, 21

[93] M. T. Pastor, A. Esteras-Chopo, and L. Serrano. Hacking the code of amyloid formation: the amyloid stretch hypothesis. *Prion*, 1(1):9–14, Mar. 2007. ISSN 1933-690X. → pages 4

[94] A. Pastore and P. A. Temussi. The two faces of Janus: functional interactions and protein aggregation. *Current Opinion in Structural Biology*, 22(1):30–37, Feb. 2012. ISSN 0959440X. doi:10.1016/j.sbi.2011.11.007. URL http://linkinghub.elsevier.com/retrieve/pii/S0959440X11002016. → pages 3, 4

[95] N. P. Pavletich, K. A. Chambers, and C. O. Pabo. The DNA-binding domain of p53 contains the four conserved regions and the major mutation hot spots. *Genes & Development*, 7(12B):2556–2564, Dec. 1993. ISSN 0890-9369. → pages 3

[96] S. Pechmann, E. D. Levy, G. G. Tartaglia, and M. Vendruscolo. Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 106(25): 10159–10164, June 2009. ISSN 1091-6490. doi:10.1073/pnas.0812414106. → pages 4

[97] S. Pu, J. Wong, B. Turner, E. Cho, and S. J. Wodak. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, 37(3):825–831, Feb. 2009. ISSN 0305-1048, 1362-4962. doi:10.1093/nar/gkn1005. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkn1005. → pages 10

[98] P. Puigb, I. G. Bravo, and S. Garcia-Vallve. CAIcal: a combined set of tools to assess codon usage adaptation. *Biology Direct*, 3:38, 2008. ISSN 1745-6150. doi:10.1186/1745-6150-3-38. → pages 12

[99] P. Puntervoll. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Research*, 31(13): 3625–3630, July 2003. ISSN 1362-4962. doi:10.1093/nar/gkg545. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkg545. → pages 12

[100] X. Qu, J.-D. Wen, L. Lancaster, H. F. Noller, C. Bustamante, and I. Tinoco. The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature*, 475(7354):118–121, July 2011. ISSN 1476-4687. doi:10.1038/nature10126. → pages 37

[101] H. Rampelt, J. Kirstein-Miles, N. B. Nillegoda, K. Chi, S. R. Scholz, R. I. Morimoto, and B. Bukau. Metazoan Hsp70 machines use Hsp110 to power protein disaggregation. *The EMBO Journal*, 31(21):4221–4235, Nov. 2012. doi:10.1038/emboj.2012.264. → pages 47

[102] M. A. M. Reijns, R. D. Alexander, M. P. Spiller, and J. D. Beggs. A role for Q/N-rich aggregation-prone regions in P-body localization. *Journal of Cell Science*, 121(15):2463–2472, Aug. 2008. ISSN 0021-9533, 1477-9137. doi:10.1242/jcs.024976. URL http://jcs.biologists.org/cgi/doi/10.1242/jcs.024976. → pages 49

[103] P. Reis-Rodrigues, G. Czerwieniec, T. W. Peters, U. S. Evani, S. Alavez, E. A. Gaman, M. Vantipalli, S. D. Mooney, B. W. Gibson, G. J. Lithgow, and R. E. Hughes. Proteomic analysis of age-dependent changes in protein solubility identifies genes that modulate lifespan: Aging, protein solubility and lifespan in *C. elegans*. *Aging Cell*, 11(1):120–127, Feb. 2012. ISSN 14749718. doi:10.1111/j.1474-9726.2011.00765.x. URL http://doi.wiley.com/10.1111/j.1474-9726.2011.00765.x. → pages 21

[104] P.-H. Ren, J. E. Lauckner, I. Kachirskaia, J. E. Heuser, R. Melki, and R. R. Kopito. Cytoplasmic penetration and persistent infection of mammalian cells by polyglutamine aggregates. *Nature Cell Biology*, 11(2):219–225, Feb. 2009. ISSN 1465-7392, 1476-4679. doi:10.1038/ncb1830. URL http://www.nature.com/doifinder/10.1038/ncb1830. → pages 3

[105] J. Reumers, S. Maurer-Stroh, J. Schymkowitz, and F. Rousseau. Protein sequences encode safeguards against aggregation. *Human Mutation*, 30(3): 431–437, Mar. 2009. ISSN 1098-1004. doi:10.1002/humu.20905. → pages 22

[106] C. A. Ross and M. A. Poirier. Protein aggregation and neurodegenerative disease. *Nature Medicine*, 10(7):S10–S17, July 2004. ISSN 1078-8956.

doi:10.1038/nm1066. URL
http://www.nature.com/doifinder/10.1038/nm1066. → pages 3

[107] F. Rousseau, L. Serrano, and J. W. H. Schymkowitz. How evolutionary pressure against protein aggregation shaped chaperone specificity. *Journal of Molecular Biology*, 355(5):1037–1047, Feb. 2006. ISSN 0022-2836. doi:10.1016/j.jmb.2005.11.035. → pages 22

[108] A. Ruepp, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, M. Stransky, B. Waegele, T. Schmidt, O. N. Doudieu, V. Stumpflen, and H. W. Mewes. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Research*, 36(Database):D646–D650, Dec. 2007. ISSN 0305-1048, 1362-4962. doi:10.1093/nar/gkm936. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkm936. → pages 10

[109] I. Sadowski, B.-J. Breitkreutz, C. Stark, T.-C. Su, M. Dahabieh, S. Raithatha, W. Bernhard, R. Oughtred, K. Dolinski, K. Barreto, and M. Tyers. The PhosphoGRID Saccharomyces cerevisiae protein phosphorylation site database: version 2.0 update. *Database*, 2013(0): bat026–bat026, May 2013. ISSN 1758-0463. doi:10.1093/database/bat026. URL http://database.oxfordjournals.org/cgi/doi/10.1093/database/bat026. → pages 12

[110] A. M. Salazar, E. J. Silverman, K. P. Menon, and K. Zinn. Regulation of synaptic Pumilio function by an aggregation-prone domain. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 30(2): 515–522, Jan. 2010. ISSN 1529-2401. doi:10.1523/JNEUROSCI.2523-09.2010. → pages 4

[111] T. R. Serio and S. L. Lindquist. Protein-only inheritance in yeast: something to get [PSI+]-ched about. *Trends in Cell Biology*, 10(3):98–105, Mar. 2000. ISSN 0962-8924. → pages 4

[112] M. Y. Sherman and S.-B. Qian. Less is more: improving proteostasis by translation slow down. *Trends in Biochemical Sciences*, 38(12):585–591, Dec. 2013. ISSN 0968-0004. doi:10.1016/j.tibs.2013.09.003. → pages 37

[113] A. Shevchenko, M. Wilm, O. Vorm, O. N. Jensen, A. V. Podtelejnikov, G. Neubauer, A. Shevchenko, P. Mortensen, and M. Mann. A strategy for identifying gel-separated proteins in sequence databases by MS alone. *Biochemical Society Transactions*, 24(3):893–896, Aug. 1996. ISSN 0300-5127. → pages 8

[114] E. M. Sontag, W. I. Vonk, and J. Frydman. Sorting out the trash: the spatial nature of eukaryotic protein quality control. *Current Opinion in Cell Biology*, 26:139–146, Feb. 2014. ISSN 09550674. doi:10.1016/j.ceb.2013.12.006. URL http://linkinghub.elsevier.com/retrieve/pii/S0955067413001932. → pages 2

[115] R. Spokoini, O. Moldavski, Y. Nahmias, J. England, M. Schuldiner, and D. Kaganovich. Confinement to Organelle-Associated Inclusion Structures Mediates Asymmetric Inheritance of Aggregated Protein in Budding Yeast. *Cell Reports*, 2(4):738–747, Oct. 2012. ISSN 22111247. doi:10.1016/j.celrep.2012.08.024. URL http://linkinghub.elsevier.com/retrieve/pii/S2211124712002641. → pages 2, 47

[116] F. Sun, V. Anantharam, D. Zhang, C. Latchoumycandane, A. Kanthasamy, and A. G. Kanthasamy. Proteasome inhibitor MG-132 induces dopaminergic degeneration in cell culture and animal models. *NeuroToxicology*, 27(5):807–815, Sept. 2006. ISSN 0161813X. doi:10.1016/j.neuro.2006.06.006. URL http://linkinghub.elsevier.com/retrieve/pii/S0161813X06001689. → pages 2

[117] The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 42(D1):D191–D198, Jan. 2014. ISSN 0305-1048, 1362-4962. doi:10.1093/nar/gkt1140. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1140. → pages 9, 12

[118] J. Tian, N. Wu, J. Guo, and Y. Fan. Prediction of amyloid fibril-forming segments based on a support vector machine. *BMC bioinformatics*, 10 Suppl 1:S45, 2009. ISSN 1471-2105. doi:10.1186/1471-2105-10-S1-S45. → pages 4

[119] A. C. Tsolis, N. C. Papandreou, V. A. Iconomidou, and S. J. Hamodrakas. A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins. *PloS One*, 8(1):e54175, 2013. ISSN 1932-6203. doi:10.1371/journal.pone.0054175. → pages 4

[120] N. D. Udeshi, P. Mertins, T. Svinkina, and S. A. Carr. Large-scale identification of ubiquitination sites by mass spectrometry. *Nature Protocols*, 8(10):1950–1960, Sept. 2013. ISSN 1754-2189, 1750-2799. doi:10.1038/nprot.2013.120. URL http://www.nature.com/doifinder/10.1038/nprot.2013.120. → pages 8

[121] R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, and M. M. Babu. Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews*, 114(13):6589–6631, July 2014. ISSN 0009-2665, 1520-6890. doi:10.1021/cr400525m. URL http://pubs.acs.org/doi/abs/10.1021/cr400525m. → pages 32

[122] S. van der Walt, S. C. Colbert, and G. Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, 13(2):22–30, Mar. 2011. ISSN 1521-9615. doi:10.1109/MCSE.2011.37. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5725236. → pages 11

[123] S. Ventura, J. Zurdo, S. Narayanan, M. Parreno, R. Mangues, B. Reif, F. Chiti, E. Giannoni, C. M. Dobson, F. X. Aviles, and L. Serrano. Short amino acid stretches can mediate amyloid formation in globular proteins: The Src homology 3 (SH3) case. *Proceedings of the National Academy of Sciences*, 101(19):7258–7263, May 2004. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.0308249101. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0308249101. → pages 3

[124] D. Vilchez, I. Saez, and A. Dillin. The role of protein clearance mechanisms in organismal ageing and age-related diseases. *Nature Communications*, 5:5659, 2014. ISSN 2041-1723. doi:10.1038/ncomms6659. → pages 2

[125] B. S. Wang, R. A. Grant, and C. O. Pabo. Selected peptide extension contacts hydrophobic patch on neighboring zinc finger and mediates dimerization on DNA. *Nature Structural Biology*, 8(7):589–593, July 2001. ISSN 1072-8368. doi:10.1038/89617. → pages 29

[126] J. Ward, J. Sodhi, L. McGuffin, B. Buxton, and D. Jones. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *Journal of Molecular Biology*, 337(3):635–645, Mar. 2004. ISSN 00222836. doi:10.1016/j.jmb.2004.02.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0022283604001482. → pages 12

[127] S. Weber and C. Brangwynne. Getting RNA and Protein in Phase. *Cell*, 149(6):1188–1191, June 2012. ISSN 00928674.

doi:10.1016/j.cell.2012.05.022. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867412006344. → pages 49

[128] I. B. Wilde, M. Brack, J. M. Winget, and T. Mayor. Proteomic Characterization of Aggregating Proteins after the Inhibition of the Ubiquitin Proteasome System. *Journal of Proteome Research*, 10(3): 1062–1072, Mar. 2011. ISSN 1535-3893, 1535-3907. doi:10.1021/pr1008543. URL http://pubs.acs.org/doi/abs/10.1021/pr1008543. → pages 5, 21

[129] R. L. Woltjer. Proteomic determination of widespread detergent insolubility, including A but not tau, early in the pathogenesis of Alzheimer's disease. *The FASEB Journal*, Sept. 2005. ISSN 0892-6638, 1530-6860. doi:10.1096/fj.05-4263fje. URL http://www.fasebj.org/cgi/doi/10.1096/fj.05-4263fje. → pages 3

[130] C. W. Wong, V. Quaranta, and G. G. Glenner. Neuritic plaques and cerebrovascular amyloid in Alzheimer disease are antigenically related. *Proceedings of the National Academy of Sciences of the United States of America*, 82(24):8729–8732, Dec. 1985. ISSN 0027-8424. → pages 3

[131] G. Xu, S. M. Stevens, F. Kobiessy, H. Brown, S. McClung, M. S. Gold, and D. R. Borchelt. Identification of Proteins Sensitive to Thermal Stress in Human Neuroblastoma and Glioma Cell Lines. *PLoS ONE*, 7(11):e49021, Nov. 2012. ISSN 1932-6203. doi:10.1371/journal.pone.0049021. URL http://dx.plos.org/10.1371/journal.pone.0049021. → pages 5

[132] F. Yang, Y. Shen, D. G. Camp, and R. D. Smith. High-pH reversed-phase chromatography with fraction concatenation for 2d proteomic analysis. *Expert Review of Proteomics*, 9(2):129–134, Apr. 2012. ISSN 1478-9450, 1744-8387. doi:10.1586/epr.12.15. URL http://informahealthcare.com/doi/abs/10.1586/epr.12.15. → pages 8

[133] Y. Yoshimura, Y. Lin, H. Yagi, Y.-H. Lee, H. Kitayama, K. Sakurai, M. So, H. Ogi, H. Naiki, and Y. Goto. Distinguishing crystal-like amyloid fibrils and glass-like amorphous aggregates from their kinetics of formation. *Proceedings of the National Academy of Sciences*, 109(36):14446–14451, Sept. 2012. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1208228109. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1208228109. → pages 2

[134] C. Zhou, B. Slaughter, J. Unruh, A. Eldakak, B. Rubinstein, and R. Li. Motility and Segregation of Hsp104-Associated Protein Aggregates in

Budding Yeast. *Cell*, 147(5):1186–1196, Nov. 2011. ISSN 00928674. doi:10.1016/j.cell.2011.11.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867411012918. → pages 2

[135] S. Zibaee, O. S. Makin, M. Goedert, and L. C. Serpell. A simple algorithm locates beta-strands in the amyloid fibril core of alpha-synuclein, Abeta, and tau using the amino acid sequence alone. *Protein Science: A Publication of the Protein Society*, 16(5):906–918, May 2007. ISSN 0961-8368. doi:10.1110/ps.062624507. → pages 4

[136] H. Y. Zoghbi and H. T. Orr. Glutamine repeats and neurodegeneration. *Annual Review of Neuroscience*, 23:217–247, 2000. ISSN 0147-006X. doi:10.1146/annurev.neuro.23.1.217. → pages 25

Appendix A

Supporting Materials

Table A.1: GO analysis (biological processes) for yeast LS proteins

Go annotation	GO term	No. in LS	No. in category	p-value
GO:0016072	rRNA metabolic process	35	248	1.40E-19
GO:0006364	rRNA processing	34	239	5.70E-19
GO:0042254	Ribosome biogenesis	39	351	1.20E-18
GO:0022613	Ribonucleoprotein complex biogenesis	39	397	1.00E-16
GO:0034470	ncRNA processing	35	335	2.70E-15
GO:0034660	ncRNA metabolic process	36	393	6.00E-14
GO:0000462	Maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	19	83	1.10E-12
GO:0030490	Maturation of SSU-rRNA	19	85	1.70E-12
GO:0006396	RNA processing	36	515	2.50E-10

Table A.1: GO analysis (biological processes) for yeast LS proteins

Go annotation	GO term	No. in LS	No. in category	p-value
GO:0000480	Endonucleolytic cleavage in 5'-ETS of tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	9	26	5.30E-06
GO:0000447	Endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	10	38	7.80E-06
GO:0000472	Endonucleolytic cleavage to generate mature 5'-end of SSU-rRNA from (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	9	28	1.00E-05
GO:0000479	Endonucleolytic cleavage of tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	10	40	1.30E-05
GO:0000478	Endonucleolytic cleavages during rRNA processing	10	40	1.30E-05
GO:0000967	rRNA 5'-end processing	9	29	1.40E-05
GO:0034471	ncRNA 5'-end processing	9	29	1.40E-05
GO:0000460	Maturation of 5.8S rRNA	12	69	1.50E-05

Table A.1: GO analysis (biological processes) for yeast LS proteins

Go annotation	GO term	No. in LS	No. in category	p-value
GO:0000466	Maturation of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	12	69	1.50E-05
GO:0000966	RNA 5'-end processing	9	30	1.90E-05
GO:0000469	Cleavages during rRNA processing	10	59	4.40E-04
GO:0045943	Positive regulation of transcription from RNA polymerase I promoter	6	12	6.70E-04
GO:0006356	Regulation of transcription from RNA polymerase I promoter	6	18	6.60E-03

Table A.2: GO analysis (molecular function) for yeast LS proteins

Go annotation	GO term	No. in LS	No. in category	p-value
GO:0030515	snoRNA binding	10	19	1.10E-09
GO:0003723	RNA binding	25	504	3.00E-04
GO:0004386	Helicase activity	11	106	1.80E-03
GO:0003724	RNA helicase activity	7	42	1.10E-02
GO:0070035	Purine NTP-dependent helicase activity	9	84	1.30E-02
GO:0008026	ATP-dependent helicase activity	9	84	1.30E-02

Table A.3: GO analysis (biological processes) for human LS proteins

Go annotation	GO term	No. in LS	No. in category	p-value
GO:0008380	RNA splicing	20	284	4.40E-09
GO:0016071	mRNA metabolic process	22	370	7.00E-09
GO:0006397	mRNA processing	20	321	3.70E-08
GO:0000398	Nuclear mRNA splicing; via spliceosome	15	153	5.10E-08
GO:0000377	RNA splicing; via transesterification reactions with bulged adenosine as nucleophile	15	153	5.10E-08
GO:0000375	RNA splicing; via transesterification reactions	15	153	5.10E-08
GO:0006333	Chromatin assembly or disassembly	14	127	6.20E-08
GO:0006325	Chromatin organization	21	378	8.10E-08
GO:0051276	Chromosome organization	22	485	1.00E-06
GO:0006396	RNA processing	23	547	1.60E-06
GO:0034621	Cellular macromolecular complex subunit organization	17	357	7.10E-05
GO:0045449	Regulation of transcription	47	2601	5.60E-04
GO:0034622	Cellular macromolecular complex assembly	15	318	5.80E-04
GO:0007049	Cell cycle	23	776	7.60E-04
GO:0034728	Nucleosome organization	9	93	1.10E-03
GO:0006350	Transcription	40	2101	1.80E-03
GO:0006974	Response to DNA damage stimulus	15	373	3.80E-03

Table A.3: GO analysis (biological processes) for human LS proteins

Go annotation	GO term	No. in LS	No. in category	p-value
GO:0043933	Macromolecular complex subunit organization	20	710	1.00E-02
GO:0007017	Microtubule-based process	12	253	1.00E-02
GO:0065003	Macromolecular complex assembly	19	665	1.50E-02
GO:0033554	Cellular response to stress	17	566	2.70E-02
GO:0006259	DNA metabolic process	16	506	2.80E-02
GO:0006281	DNA repair	12	284	3.00E-02

Table A.4: GO analysis (molecular function) for human LS proteins

Go annotation	GO term	No. in LS	No. in category	p-value
GO:0003677	DNA binding	54	2331	1.20E-08
GO:0003723	RNA binding	29	718	1.60E-08
GO:0003682	Chromatin binding	11	150	2.50E-04

Table A.5: GO analysis (biological processes) for mouse LS proteins

Go annotation	GO term	No. in LS	No. in category	p-value
GO:0006412	Translation	64	319	2.40E-36
GO:0008380	RNA splicing	51	201	2.40E-33
GO:0016071	mRNA metabolic process	58	302	2.00E-31
GO:0006397	mRNA processing	53	262	1.60E-29

Table A.5: GO analysis (biological processes) for mouse LS proteins

Go annotation	GO term	No. in LS	No. in category	p-value
GO:0006396	RNA processing	58	437	1.00E-22
GO:0022900	Electron transport chain	26	112	2.50E-14
GO:0006091	Generation of precursor metabolites and energy	33	261	6.70E-11
GO:0007010	Cytoskeleton organization	31	326	6.30E-07
GO:0000377	RNA splicing; via transesterification reactions with bulged adenosine as nucleophile	12	37	1.00E-06
GO:0000375	RNA splicing; via transesterification reactions	12	37	1.00E-06
GO:0000398	Nuclear mRNA splicing; via spliceosome	12	37	1.00E-06
GO:0030029	Actin filament-based process	18	176	2.00E-03
GO:0006403	RNA localization	11	67	6.50E-03
GO:0043244	Regulation of protein complex disassembly	9	43	1.10E-02
GO:0043242	Negative regulation of protein complex disassembly	8	35	2.40E-02
GO:0051236	Establishment of RNA localization	10	66	4.00E-02
GO:0050657	Nucleic acid transport	10	66	4.00E-02
GO:0050658	RNA transport	10	66	4.00E-02
GO:0022613	Ribonucleoprotein complex biogenesis	14	137	4.10E-02

Table A.6: GO analysis (molecular function) for mouse LS proteins

Go annotation	GO term	No. in LS	No. in category	p-value
GO:0003735	Structural constituent of ribosome	59	151	1.30E-50
GO:0005198	Structural molecule activity	82	450	2.80E-43
GO:0003723	RNA binding	97	672	5.70E-43
GO:0003779	Actin binding	37	288	1.40E-12
GO:0008092	Cytoskeletal protein binding	42	414	4.80E-11
GO:0008137	NADH dehydrogenase (ubiquinone) activity	11	24	5.60E-08
GO:0003954	NADH dehydrogenase activity	11	24	5.60E-08
GO:0050136	NADH dehydrogenase (quinone) activity	11	24	5.60E-08
GO:0016655	Oxidoreductase activity; acting on NADH or NADPH; quinone or similar compound as acceptor	11	27	2.20E-07
GO:0003729	mRNA binding	14	54	2.70E-07
GO:0016651	Oxidoreductase activity; acting on NADH or NADPH	12	51	1.90E-05
GO:0000166	Nucleotide binding	92	2183	5.50E-04
GO:0019843	rRNA binding	8	24	5.60E-04
GO:0003697	Single-stranded DNA binding	8	33	5.60E-03
GO:0005516	Calmodulin binding	13	114	1.30E-02

Table A.6: GO analysis (molecular function) for mouse LS proteins

Go annotation	GO term	No. in LS	No. in category	p-value
GO:0015078	Hydrogen ion transmembrane transporter activity	11	82	1.60E-02
GO:0015077	Monovalent inorganic cation transmembrane transporter activity	11	87	2.60E-02

Table A.7: Analysis for enrichment of Pfam domains for yeast LS proteins

Domain	In LS	Total LS	in NS	No. Total NS	p-value	significant
WD40	11	96	23	1095	3.19E-05	No
Histone	4	96	1	1095	1.87E-04	No
AAA_12	3	96	0	1095	5.09E-04	No
AAA_11	3	96	0	1095	5.09E-04	No
Glyco_hydro_72	3	96	0	1095	5.09E-04	No
Utp12	3	96	0	1095	5.09E-04	No
DEAD	6	96	16	1095	6.06E-03	No
NOP5NT	2	96	0	1095	6.43E-03	No
NOSIC	2	96	0	1095	6.43E-03	No
Nop	2	96	0	1095	6.43E-03	No
XPG_I	2	96	0	1095	6.43E-03	No
XPG_N	2	96	0	1095	6.43E-03	No
Myosin_TH1	2	96	0	1095	6.43E-03	No
Tubulin	2	96	0	1095	6.43E-03	No
Tubulin_C	2	96	0	1095	6.43E-03	No
PH	2	96	1	1095	1.83E-02	No

Table A.7: Analysis for enrichment of Pfam domains for yeast LS proteins

Domain	In LS	Total LS	in NS	No. Total NS	p-value	significant
HA2	2	96	1	1095	1.83E-02	No
OB_NTP_bind	2	96	1	1095	1.83E-02	No
Helicase_C	6	96	24	1095	2.84E-02	No
Myosin_head	2	96	2	1095	3.47E-02	No
PI3_PI4_kinase	2	96	2	1095	3.47E-02	No

Table A.8: Analysis for enrichment of Pfam domains for human LS proteins

Domain	In LS	Total LS	in NS	No. Total NS	p-value	significant
RRM_1	18	167	20	1190	6.52E-08	Yes
Septin	7	170	0	1201	4.04E-07	Yes
Filament	6	170	0	1201	3.36E-06	Yes
Bromodomain	7	170	2	1201	1.17E-05	No
Histone	5	169	0	1201	2.71E-05	No
HMG_box	5	168	2	1201	4.50E-04	No
PHD	5	168	2	1200	4.51E-04	No
HMG_box_2	4	170	1	1201	1.03E-03	No
AAA_33	3	169	0	1199	1.86E-03	No
WHIM3	3	170	0	1201	1.88E-03	No
Homeobox	3	169	1	1201	6.72E-03	No
SAP	3	167	2	1199	1.49E-02	No
LTD	2	169	0	1201	1.51E-02	No
WHIM1	2	169	0	1201	1.51E-02	No
Filament_head	2	170	0	1201	1.53E-02	No

Table A.8: Analysis for enrichment of Pfam domains for human LS proteins

Domain	**In LS**	**Total LS**	**in NS**	**No. Total NS**	**p-value**	**significant**
CUT	2	170	0	1201	1.53E-02	No
BAR	2	170	0	1201	1.53E-02	No
EFhand_Ca_insen	2	170	0	1201	1.53E-02	No
Macro	2	170	0	1201	1.53E-02	No
EF1_GNE	2	170	0	1201	1.53E-02	No
EF-1_beta_acid	2	170	0	1201	1.53E-02	No
GATA	2	170	0	1201	1.53E-02	No
DLIC	2	170	0	1201	1.53E-02	No
zf-PARP	2	170	0	1201	1.53E-02	No
Rtt106	2	170	0	1201	1.53E-02	No
NOPS	2	170	0	1201	1.53E-02	No
RRM_6	5	170	9	1201	2.17E-02	No
Chromo	2	169	1	1201	4.17E-02	No
SWIRM	2	170	1	1201	4.21E-02	No
ANTH	2	170	1	1201	4.21E-02	No
2-oxoacid_dh	2	170	1	1201	4.21E-02	No
I_LWEQ	2	170	1	1201	4.21E-02	No
zf-C2H2_4	2	170	1	1200	4.22E-02	No

Table A.9: Analysis for enrichment of Pfam domains for mouse LS proteins

Domain	**In LS**	**Total LS**	**in NS**	**No. Total NS**	**p-value**	**significant**
RRM_1	50	525	5	1255	4.00E-22	Yes
Spectrin	12	530	0	1255	4.30E-07	Yes

Table A.9: Analysis for enrichment of Pfam domains for mouse LS proteins

Domain	In LS	Total LS	in NS	No. Total NS	p-value	significant
Filament	10	530	0	1255	5.01E-06	Yes
Myosin_head	8	530	0	1255	5.82E-05	No
PDZ	17	521	10	1251	3.51E-04	No
RRM_6	8	527	1	1255	3.75E-04	No
Guanylate_kin	9	528	2	1253	5.03E-04	No
EFhand_Ca_insen	6	530	0	1255	6.72E-04	No
Filament_head	6	530	0	1255	6.72E-04	No
Ras	1	530	25	1254	1.87E-03	No
RRM_5	5	527	0	1255	2.23E-03	No
SH3_2	10	528	5	1253	3.19E-03	No
EF-hand_6	6	530	1	1254	3.53E-03	No
C2	12	530	7	1254	3.63E-03	No
SAM_1	7	529	2	1255	3.90E-03	No
I-set	7	530	2	1255	3.93E-03	No
Band_7	4	529	0	1255	7.67E-03	No
Histone	4	529	0	1255	7.67E-03	No
Myosin_tail_1	4	529	0	1255	7.67E-03	No
Sorb	4	530	0	1255	7.71E-03	No
Linker_histone	4	530	0	1255	7.71E-03	No
DUF1899	4	530	0	1255	7.71E-03	No
NAC	4	530	0	1255	7.71E-03	No
VGCC_beta4Aa_N	4	530	0	1255	7.71E-03	No
SAP	5	529	1	1255	1.02E-02	No
TPR_11	0	530	13	1254	1.37E-02	No
Proteasome	0	530	14	1255	1.45E-02	No
GKAP	3	529	0	1255	2.60E-02	No
dsrm	3	529	0	1255	2.60E-02	No

Table A.9: Analysis for enrichment of Pfam domains for mouse LS proteins

Domain	In LS	Total LS	in NS	No. Total NS	p-value	significant
CaMKII_AD	3	529	0	1255	2.60E-02	No
Ribosomal_L7Ae	3	529	0	1255	2.60E-02	No
Sec7	3	529	0	1255	2.60E-02	No
Collagen	3	530	0	1255	2.61E-02	No
CNH	3	530	0	1255	2.61E-02	No
C1q	3	530	0	1255	2.61E-02	No
BAG	3	530	0	1255	2.61E-02	No
DUF2051	3	530	0	1255	2.61E-02	No
LTD	3	530	0	1255	2.61E-02	No
Myosin_N	3	530	0	1255	2.61E-02	No
Fox-1_C	3	530	0	1255	2.61E-02	No
Cast	3	530	0	1255	2.61E-02	No
PurA	3	530	0	1255	2.61E-02	No
Agenet	3	530	0	1255	2.61E-02	No
Tropomyosin	3	530	0	1255	2.61E-02	No
PH_9	4	528	1	1255	2.92E-02	No
IQ	4	530	1	1255	2.94E-02	No
DUF1900	4	530	1	1255	2.94E-02	No
SH3_1	10	529	8	1251	3.38E-02	No
TPR_1	0	530	10	1252	3.89E-02	No
Aldedh	0	530	10	1255	3.90E-02	No

Table A.10: Table of p-values for feature analysis of yeast proteins

Analysis	LS vs NS	HS vs NS	LS vs HS
Percent G	2.44E-04	5.38E-02	1.49E-01

Table A.10: Table of p-values for feature analysis of yeast proteins

Analysis	LS vs NS	HS vs NS	LS vs HS
Percent A	7.27E-04	2.35E-02	3.20E-05
Percent V	2.12E-03	2.29E-04	9.70E-01
Percent L	4.89E-01	3.26E-06	3.21E-03
Percent I	5.10E-01	5.30E-10	3.23E-03
Percent P	3.20E-02	1.90E-02	3.09E-03
Percent F	3.92E-01	8.62E-03	3.54E-01
Percent Y	9.15E-01	8.93E-05	1.82E-02
Percent W	8.45E-01	4.78E-03	1.34E-01
Percent H	4.09E-01	3.21E-03	2.46E-01
Percent M	2.15E-01	7.91E-01	3.08E-01
Percent C	1.88E-01	3.57E-06	1.17E-04
Percent S	7.42E-10	1.71E-01	3.54E-08
Percent T	3.04E-01	1.67E-03	1.03E-02
Percent K	6.61E-01	6.82E-05	1.86E-02
Percent D	5.97E-01	2.09E-04	1.59E-02
Percent E	1.11E-01	7.46E-09	8.35E-06
Percent N	1.09E-04	7.17E-01	4.68E-03
Percent Q	9.07E-02	4.02E-02	1.21E-02
Percent R	4.52E-01	2.38E-03	4.36E-02
ER likelihood	2.12E-01	3.03E-08	9.79E-05
Golgi likelihood	4.27E-02	6.83E-12	8.52E-08
Vacuole likelihood	1.07E-05	6.12E-06	1.82E-11
Membrane likelihood	3.41E-03	1.20E-10	1.67E-07
Secretory likelihood	8.41E-03	1.14E-01	1.41E-03
Cytosol likelihood	6.67E-04	4.00E-03	2.76E-01
Peroxisome likelihood	1.94E-02	1.12E-04	7.75E-01
Mitochondria likelihood	8.00E-04	6.88E-02	1.08E-01
Nucleus likelihood	5.00E-04	1.54E-03	4.05E-01
Length	7.37E-06	8.58E-23	5.99E-17

Table A.10: Table of p-values for feature analysis of yeast proteins

Analysis	LS vs NS	HS vs NS	LS vs HS
Disorder prediction (DISOPRED)	2.16E-04	4.11E-07	7.39E-01
Disorder prediction (IUPRED)	2.40E-03	3.10E-08	3.11E-01
Number of LCRs	4.30E-12	4.40E-01	1.31E-10
Number of MoRFs (ANCHOR)	4.24E-04	1.58E-02	9.16E-02
Number of ELMs	1.36E-08	3.58E-21	2.94E-18
Number of LCRs per unit length	2.66E-07	2.09E-02	4.69E-02
Number of MoRFs (ANCHOR) per unit length	1.28E-02	2.27E-08	7.56E-02
Number of ELMs per unit length	1.03E-07	2.88E-02	9.70E-07
Percent helix	7.61E-02	3.91E-01	3.29E-01
Percent sheet	9.57E-02	1.29E-02	5.65E-01
Percent coil	1.50E-03	1.66E-01	1.37E-01
Percent polar	2.03E-06	1.63E-02	1.31E-06
Percent hydrophobic	4.08E-06	1.50E-03	1.49E-01
Percent positive	4.12E-01	2.72E-01	8.75E-01
Percent negative	1.06E-01	2.46E-10	3.78E-06
Abundance	3.01E-10	4.49E-04	5.41E-14
Number of disulfide bonds	1.40E-04	2.31E-18	3.96E-15
Number of phosphorylation sites	1.02E-01	6.64E-01	1.09E-01
Number of disordered regions per unit length	2.03E-02	5.43E-08	7.78E-02
Number of disordered regions	4.47E-07	2.22E-02	6.90E-09
Number of coiled coil regions	6.63E-01	1.26E-01	5.74E-01
Percent coiled coiled	7.55E-01	8.03E-02	4.57E-01
Number of MoRFs (ANCHOR) per percent disorder	2.86E-02	4.80E-13	6.70E-03
Number of ELMs per percent disorder	7.66E-01	5.23E-23	7.35E-11
Number of LCRs per Percent disorder	5.33E-09	7.11E-04	3.89E-14

Table A.10: Table of p-values for feature analysis of yeast proteins

Analysis	LS vs NS	HS vs NS	LS vs HS
Number of MoRFs per Percent disorder	1.68E-02	9.83E-01	2.37E-02
Number of MoRFs per disordered patch	8.37E-01	1.01E-01	4.29E-01
Codon Adaptation Index (CAI)	6.59E-01	1.30E-02	2.83E-01
Number of transmembrane helices	2.68E-01	6.51E-01	3.54E-01
Hydrophobicity (GRAVY index)	7.55E-02	4.96E-12	8.15E-03
Number of codons	7.76E-06	6.52E-16	1.89E-14
Percent GC content	2.09E-04	6.49E-05	1.15E-07
Number of close stop codons	9.82E-01	1.72E-07	1.40E-03
Percent rate amino acids (C W H M)	6.94E-01	1.66E-04	5.97E-02
Number of MoRF residues (ANCHOR)	3.35E-04	2.32E-02	6.14E-02
Number of MoRF residues (ANCHOR) per unit length	4.54E-03	2.00E-06	4.59E-01
Number of MoRFs (ANCHOR) per disordered patch	5.60E-03	2.37E-04	9.20E-01
Percent aromatic residues (F Y W)	8.22E-01	8.88E-06	1.49E-02
Number of aromatic patches	3.01E-04	2.44E-06	2.77E-09
Number of hydrophobic patches	1.21E-03	9.13E-09	1.50E-08
Number of negatively charged patches	3.55E-03	2.61E-04	3.35E-06
Number of positively charged patches	5.69E-01	2.74E-08	9.87E-05
Number of polar patches (Q N)	2.76E-03	2.22E-03	2.35E-05
Number of polar patches (S T)	4.88E-02	6.06E-03	7.83E-04
Number of aromatic patches per unit length	8.09E-01	1.91E-01	6.44E-01
Number of hydrophobic patches per unit length	4.97E-01	7.56E-02	2.01E-01

Table A.10: Table of p-values for feature analysis of yeast proteins

Analysis	LS vs NS	HS vs NS	LS vs HS
Number of negatively charged patches per unit length	7.60E-01	8.62E-05	1.71E-02
Number of positively charged patches per unit length	3.38E-02	6.44E-01	1.63E-01
Number of polar patches (Q N) per unit length	7.80E-01	6.18E-03	1.33E-01
Number of polar patches (S T) per unit length	4.25E-02	1.94E-05	1.06E-04
Net charge per protein	3.95E-04	6.31E-08	2.62E-07
Net charge squared per protein	2.02E-09	5.26E-05	1.63E-12
Net charge per residue	2.21E-02	1.10E-05	3.32E-05
Net charge per residue squared	9.49E-04	7.26E-02	7.29E-02

Table A.11: Table of p-values for feature analysis of human proteins

Analysis	LS vs NS	HS vs NS	LS vs HS
Percent G	1.64E-01	3.83E-04	6.70E-01
Percent A	8.17E-01	9.18E-05	5.22E-02
Percent V	1.36E-03	7.57E-08	8.90E-09
Percent L	1.27E-03	3.67E-03	2.75E-06
Percent I	8.96E-09	5.01E-03	3.59E-12
Percent P	2.59E-01	6.41E-06	3.10E-04
Percent F	1.08E-04	3.35E-12	1.34E-12
Percent Y	5.81E-01	3.03E-05	6.69E-03
Percent W	1.23E-03	4.33E-08	1.08E-09
Percent H	3.56E-02	9.66E-04	8.09E-01
Percent M	1.53E-01	1.32E-02	4.78E-03
Percent C	4.83E-04	5.74E-06	3.01E-09

Table A.11: Table of p-values for feature analysis of human proteins

Analysis	LS vs NS	HS vs NS	LS vs HS
Percent S	1.19E-07	4.31E-09	6.81E-17
Percent T	5.51E-02	8.91E-01	1.16E-01
Percent K	6.15E-01	1.77E-06	1.72E-02
Percent D	5.80E-01	7.93E-08	3.55E-04
Percent E	1.18E-01	1.11E-01	6.36E-01
Percent N	9.56E-02	7.67E-01	1.07E-01
Percent Q	8.70E-01	1.35E-05	1.21E-02
Percent R	1.86E-01	8.43E-20	1.05E-10
Length	5.92E-02	1.21E-10	7.14E-09
Disorder prediction (DISOPRED)	7.67E-13	1.79E-39	1.91E-37
Disorder prediction (IUPRED)	8.01E-11	3.91E-26	7.09E-31
Number of LCRs	1.11E-08	1.70E-24	1.06E-28
Number of MoRFs (ANCHOR)	6.55E-09	1.46E-29	5.43E-32
Number of ELMs	1.85E-02	4.03E-14	4.67E-12
Number of LCRs per unit length	1.04E-08	7.87E-20	6.17E-25
Number of MoRFs (ANCHOR) per unit length	2.24E-09	1.56E-23	1.82E-27
Number of ELMs per unit length	1.95E-02	5.08E-13	7.06E-10
Percent helix	1.11E-01	3.59E-03	1.21E-03
Percent sheet	4.97E-03	1.03E-08	1.91E-09
Percent coil	8.68E-05	3.57E-11	2.89E-13
Percent polar	9.50E-02	2.57E-04	2.26E-04
Percent hydrophobic	5.51E-02	5.13E-18	8.48E-11
Percent positive	5.42E-02	1.35E-31	3.58E-18
Percent negative	1.09E-01	5.57E-05	4.01E-01
Abundance	5.92E-01	4.11E-24	1.87E-09
Number of disulfide bonds	1.53E-01	1.02E-03	5.92E-01
Number of phosphorylation sites	4.51E-06	3.27E-08	2.08E-13

Table A.11: Table of p-values for feature analysis of human proteins

Analysis	**LS vs NS**	**HS vs NS**	**LS vs HS**
Number of disordered regions per unit length	4.52E-04	1.14E-15	4.91E-15
Number of disordered regions	1.53E-04	1.11E-31	9.83E-23
Number of coiled coil regions	5.01E-04	3.66E-06	9.75E-10
Percent coiled coiled	6.23E-04	1.25E-05	4.46E-09
Number of MoRFs (ANCHOR) per percent disorder	5.87E-08	1.93E-17	8.23E-21
Number of ELMs per percent disorder	8.30E-05	1.62E-10	6.66E-13
Number of LCRs per Percent disorder	1.43E-03	2.22E-11	2.36E-12
Number of MoRFs per Percent disorder	2.80E-03	6.42E-11	2.22E-11
Number of MoRFs per disordered patch	1.48E-01	1.17E-12	5.79E-09
Codon Adaptation Index (CAI)	5.42E-01	3.36E-01	9.78E-01
Percent rate amino acids (C W H M)	4.20E-05	2.13E-04	3.01E-09
Number of MoRF residues (ANCHOR)	2.79E-08	2.14E-30	1.29E-31
Number of MoRF residues (ANCHOR) per unit length	1.55E-08	3.09E-26	3.15E-28
Number of MoRFs (ANCHOR) per disordered patch	3.63E-07	7.32E-23	6.13E-28
Percent aromatic residues (F Y W)	3.04E-03	7.43E-15	1.00E-10
Net charge per protein	9.19E-01	5.79E-43	6.17E-17
Net charge squared per protein	2.02E-01	1.06E-24	2.02E-13
Net charge per residue	6.94E-01	1.23E-35	2.00E-13
Net charge per residue squared	3.69E-01	3.41E-09	6.54E-03

Table A.12: Table of p-values for feature analysis of mouse proteins

Analysis	LS vs NS	HS vs NS	LS vs HS
Percent G	6.47E-01	4.51E-01	5.66E-01
Percent A	1.12E-01	9.13E-01	8.15E-01
Percent V	1.44E-19	2.36E-03	9.69E-01
Percent L	1.51E-12	3.38E-02	3.45E-05
Percent I	7.14E-13	2.00E-01	1.78E-01
Percent P	8.77E-05	1.67E-01	1.13E-02
Percent F	7.95E-14	7.48E-01	9.76E-03
Percent Y	4.86E-01	1.12E-01	2.23E-01
Percent W	2.54E-04	2.10E-01	9.18E-01
Percent H	9.03E-01	9.16E-01	9.69E-01
Percent M	5.76E-06	6.70E-01	4.35E-02
Percent C	6.09E-18	4.17E-01	3.75E-02
Percent S	3.49E-08	8.72E-01	2.40E-02
Percent T	1.23E-03	8.43E-02	5.46E-01
Percent K	8.22E-01	2.64E-04	2.14E-03
Percent D	3.32E-06	7.00E-02	5.84E-04
Percent E	6.95E-01	1.17E-01	1.46E-01
Percent N	5.26E-01	4.44E-01	5.81E-01
Percent Q	1.70E-03	4.68E-01	9.83E-02
Percent R	4.43E-37	2.02E-04	7.61E-12
Length	5.06E-10	3.90E-11	3.15E-13
Disorder prediction (DISOPRED)	5.49E-42	1.81E-01	7.06E-08
Disorder prediction (IUPRED)	3.31E-36	1.12E-01	6.73E-07
Number of LCRs	4.74E-38	5.82E-02	1.30E-10
Number of MoRFs (ANCHOR)	6.46E-38	2.17E-03	8.00E-13
Number of ELMs	1.90E-13	3.39E-11	2.73E-14
Number of LCRs per unit length	3.18E-30	4.84E-01	2.12E-05
Number of MoRFs (ANCHOR) per unit length	3.32E-33	2.45E-01	1.50E-05

Table A.12: Table of p-values for feature analysis of mouse proteins

Analysis	LS vs NS	HS vs NS	LS vs HS
Number of ELMs per unit length	7.19E-12	7.72E-03	4.77E-06
Percent helix	1.30E-08	4.21E-03	6.33E-05
Percent sheet	2.44E-12	3.95E-03	3.47E-01
Percent coil	5.30E-19	9.31E-03	2.01E-06
Percent polar	1.74E-03	1.90E-01	2.62E-02
Percent hydrophobic	7.73E-21	9.72E-01	6.94E-04
Percent positive	4.53E-22	2.55E-01	1.18E-02
Percent negative	5.55E-02	3.59E-02	1.14E-02
Abundance	5.79E-05	3.57E-04	1.69E-06
Number of disulfide bonds	5.89E-01	3.66E-07	1.29E-04
Number of phosphorylation sites	1.47E-06	3.62E-01	5.95E-03
Number of disordered regions per unit length	1.98E-12	5.20E-01	6.29E-03
Number of disordered regions	5.79E-25	3.86E-06	5.99E-14
Number of coiled coil regions	6.38E-11	4.76E-02	2.45E-04
Percent coiled coiled	2.82E-09	6.14E-02	6.47E-04
Number of MoRFs (ANCHOR) per percent disorder	7.24E-24	7.09E-02	4.15E-01
Number of ELMs per percent disorder	3.25E-11	5.21E-02	8.91E-01
Number of LCRs per Percent disorder	4.44E-19	1.22E-01	2.00E-06
Number of MoRFs per Percent disorder	5.33E-18	3.10E-04	6.53E-11
Number of MoRFs per disordered patch	2.93E-06	9.85E-01	1.72E-01
Codon Adaptation Index (CAI)	8.61E-03	5.06E-02	2.91E-01
Hydrophobicity (GRAVY index)	1.20E-41	6.50E-01	2.16E-05
Percent rate amino acids (C W H M)	8.10E-11	7.29E-01	1.88E-02
Number of MoRF residues (ANCHOR)	2.33E-37	4.46E-03	2.11E-12

Table A.12: Table of p-values for feature analysis of mouse proteins

Analysis	LS vs NS	HS vs NS	LS vs HS
Number of MoRF residues (ANCHOR) per unit length	4.95E-34	1.63E-01	8.41E-07
Number of MoRFs (ANCHOR) per disordered patch	9.38E-30	5.25E-01	3.64E-05
Percent aromatic residues (F Y W)	4.99E-07	4.46E-01	3.66E-01
Net charge per protein	5.31E-18	6.98E-03	3.69E-07
Net charge squared per protein	9.52E-32	2.38E-04	2.10E-13
Net charge per residue	4.04E-16	8.91E-02	2.84E-05
Net charge per residue squared	9.21E-14	1.02E-01	1.47E-01

Table A.13: Enrichment analysis for low complexity regions in yeast LS proteins relative to NS proteins

Residue	LS with LCR	LS without LCR	NS with LCR	NS without LCR	p-value	Trend direction
A	0	207	21	1220	6.00E-02	depleted
C	4	203	14	1227	3.10E-01	enriched
D	15	192	111	1130	5.10E-01	depleted
E	24	183	199	1042	1.20E-01	depleted
F	5	202	19	1222	3.70E-01	enriched
G	3	204	33	1208	4.70E-01	depleted
H	3	204	15	1226	7.30E-01	enriched
I	5	202	22	1219	5.80E-01	enriched
L	47	160	294	947	7.90E-01	depleted
K	10	197	23	1218	1.90E-02	enriched
M	0	207	6	1235	6.00E-01	depleted
N	31	176	172	1069	6.70E-01	enriched

Table A.13: Enrichment analysis for low complexity regions in yeast LS proteins relative to NS proteins

Residue	LS with LCR	LS without LCR	NS with LCR	NS without LCR	p-value	Trend direction
P	3	204	39	1202	2.60E-01	depleted
Q	6	201	60	1181	2.80E-01	depleted
R	5	202	12	1229	8.30E-02	enriched
S	41	166	146	1095	2.40E-03	enriched
T	2	205	23	1218	5.60E-01	depleted
V	0	207	9	1232	3.70E-01	depleted
W	0	207	7	1234	6.00E-01	depleted
Y	3	204	15	1226	7.30E-01	enriched

Table A.14: Enrichment analysis for low complexity regions in human LS proteins relative to NS proteins

Residue	LS with LCR	LS without LCR	NS with LCR	NS without LCR	p-value	Trend direction
A	16	521	91	2143	2.60E-01	depleted
C	15	522	84	2149	3.00E-01	depleted
D	14	523	59	2175	1.00E+00	depleted
E	107	430	375	1859	8.70E-02	enriched
F	1	536	13	2221	4.90E-01	depleted
G	35	502	124	2109	4.10E-01	enriched
H	7	530	25	2209	6.60E-01	enriched
I	0	537	3	2231	1.00E+00	depleted
L	80	457	349	1885	7.40E-01	depleted
K	5	531	47	2187	7.70E-02	depleted

Table A.14: Enrichment analysis for low complexity regions in human LS proteins relative to NS proteins

Residue	LS with LCR	LS without LCR	NS with LCR	NS without LCR	p-value	Trend direction
M	4	532	13	2221	7.60E-01	enriched
N	3	534	10	2223	7.30E-01	enriched
P	70	466	331	1903	3.40E-01	depleted
Q	49	488	203	2031	1.00E+00	enriched
R	40	496	145	2089	4.40E-01	enriched
S	66	471	293	1941	6.70E-01	depleted
T	7	529	28	2206	8.30E-01	enriched
V	3	534	9	2225	7.10E-01	enriched
W	0	537	5	2229	5.90E-01	depleted
Y	9	527	19	2215	9.30E-02	enriched

Table A.15: Enrichment analysis for low complexity regions in mouse LS proteins relative to NS proteins

Residue	LS with LCR	LS without LCR	NS with LCR	NS without LCR	p-value	Trend direction
A	53	1235	65	1186	2.20E-01	depleted
C	37	1251	46	1204	2.70E-01	depleted
D	15	1273	39	1212	8.00E-04	depleted
E	213	1075	241	1010	7.80E-02	depleted
F	5	1283	8	1243	4.20E-01	depleted
G	77	1211	59	1191	1.90E-01	enriched
H	22	1266	10	1240	4.90E-02	enriched
I	2	1286	4	1247	4.50E-01	depleted

Table A.15: Enrichment analysis for low complexity regions in mouse LS proteins relative to NS proteins

Residue	LS with LCR	LS without LCR	NS with LCR	NS without LCR	p-value	Trend direction
L	132	1156	174	1076	5.00E-03	depleted
K	3	1285	21	1229	1.20E-04	depleted
M	8	1280	11	1239	5.00E-01	depleted
N	9	1279	9	1241	1.00E+00	depleted
P	227	1061	204	1046	4.00E-01	enriched
Q	120	1167	111	1140	7.30E-01	enriched
R	108	1180	35	1215	5.10E-10	enriched
S	199	1089	160	1090	6.00E-02	enriched
T	15	1273	23	1227	1.90E-01	depleted
V	4	1283	14	1236	1.70E-02	depleted
W	2	1286	2	1249	1.00E+00	depleted
Y	30	1258	7	1243	1.80E-04	enriched

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0006364	rRNA processing	YLR197W; YOL144W; YGL078C; YJL069C; YGR090W; YEL026W; YLR129W; YDR449C; YPL126W; YOL077C; YPR137W; YMR229C; YHR196W; YJL109C; YGR128C; YOR078W; YJR041C; YKL014C; YPL043W; YOR004W; YDL208W; YER082C; YOR310C; YKL172W; YCR057C; YPL266W; YLL011W; YGL120C; YDR324C; YLR196W; YOR119C; YBL004W; YCR031C; YHR148W; YGL171W; YLR222C; YJR002W; YMR093W; YDL213C; YCL059C; YPL157W; YDL014W; YDR398W; YLR175W	3.84E-28

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0016072	rRNA metabolic process	YLR197W; YOL144W; YGL078C; YJL069C; YGR090W; YEL026W; YLR129W; YDR449C; YPL126W; YOL077C; YPR137W; YMR229C; YHR196W; YJL109C; YGR128C; YOR078W; YJR041C; YKL014C; YPL043W; YOR004W; YDL208W; YER082C; YOR310C; YKL172W; YCR057C; YPL266W; YLL011W; YGL120C; YDR324C; YLR196W; YOR119C; YBL004W; YCR031C; YHR148W; YGL171W; YLR222C; YJR002W; YMR093W; YDL213C; YCL059C; YPL157W; YDL014W; YDR398W; YLR175W	1.96E-27

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0042254	ribosome biogenesis	YDR060W; YLR197W; YOL144W; YGL078C; YJL069C; YGR090W; YEL026W; YLR129W; YDR449C; YPL126W; YOL077C; YPR137W; YMR229C; YHR196W; YJL109C; YGR128C; YNR053C; YOR078W; YJR041C; YKL014C; YPL043W; YOR004W; YDL208W; YER082C; YOR310C; YKL172W; YCR057C; YPL266W; YLL011W; YGL120C; YDR324C; YLR196W; YOR119C; YBL004W; YCR031C; YHR148W; YGL171W; YLR222C; YJR002W; YMR093W; YDL213C; YCL059C; YPL157W; YDL014W; YDR398W; YJR066W; YLR175W; YLR003C	2.80E-25

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0022613	ribonucleoprotein complex biogenesis	YDR060W; YLR197W; YOL144W; YGL078C; YJL069C; YGR090W; YEL026W; YLR129W; YDR449C; YPL126W; YOL077C; YPR137W; YMR229C; YHR196W; YJL109C; YGR128C; YNR053C; YOR078W; YJR041C; YKL014C; YPL043W; YOR004W; YDL208W; YER082C; YOR310C; YKL172W; YCR057C; YPL266W; YLL011W; YGL120C; YDR324C; YLR196W; YOR119C; YBL004W; YCR031C; YHR148W; YGL171W; YLR222C; YJR002W; YMR093W; YDL213C; YCL059C; YPL157W; YDL014W; YDR398W; YJR066W; YLR175W; YLR003C	7.73E-23
GO:0000462	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	YJL069C; YGR090W; YLR129W; YEL026W; YDR449C; YPL126W; YMR229C; YJL109C; YHR196W; YGR128C; YOR078W; YOR004W; YER082C; YOR310C; YCR057C; YLL011W; YGL120C; YDR324C; YOR119C; YBL004W; YCR031C; YLR222C; YJR002W; YMR093W; YCL059C; YDL014W; YDR398W	3.16E-22

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0030490	maturation of SSU-rRNA	YJL069C; YGR090W; YLR129W; YEL026W; YDR449C; YPL126W; YMR229C; YJL109C; YHR196W; YGR128C; YOR078W; YOR004W; YER082C; YOR310C; YCR057C; YLL011W; YGL120C; YDR324C; YOR119C; YBL004W; YCR031C; YLR222C; YJR002W; YMR093W; YCL059C; YDL014W; YDR398W	6.39E-22
GO:0034470	ncRNA processing	YLR197W; YOL144W; YGL078C; YJL069C; YGR090W; YEL026W; YLR129W; YDR449C; YPL126W; YOL077C; YPR137W; YMR229C; YHR196W; YJL109C; YGR128C; YOR078W; YJR041C; YKL014C; YPL043W; YOR004W; YDL208W; YER082C; YOR310C; YKL172W; YCR057C; YPL266W; YLL011W; YGL120C; YDR324C; YLR196W; YOR119C; YBL004W; YCR031C; YHR148W; YGL171W; YLR222C; YJR002W; YMR093W; YDL213C; YCL059C; YPL157W; YDL014W; YDR398W; YLR175W	7.64E-22

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0034660	ncRNA metabolic process	YLR197W; YOL144W; YGL078C; YJL069C; YGR090W; YEL026W; YLR129W; YDR449C; YPL126W; YOL077C; YPR137W; YMR229C; YHR196W; YJL109C; YGR128C; YOR078W; YJR041C; YKL014C; YPL043W; YOR004W; YDL208W; YER082C; YOR310C; YKL172W; YCR057C; YPL266W; YLL011W; YGL120C; YDR324C; YLR196W; YOR119C; YBL004W; YCR031C; YHR148W; YGL171W; YLR222C; YJR002W; YMR093W; YDL213C; YCL059C; YPL157W; YDL014W; YDR398W; YLR175W	5.33E-19

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0006396	RNA processing	YLR197W; YOL144W; YGL078C; YJL069C; YGR090W; YEL026W; YLR129W; YDR449C; YPL126W; YOL077C; YPR137W; YMR229C; YHR196W; YJL109C; YGR128C; YOR078W; YJR041C; YKL014C; YPL043W; YOR004W; YDL208W; YER082C; YOR310C; YKL172W; YCR057C; YPL266W; YLL011W; YGL120C; YDR324C; YLR196W; YOR119C; YBL004W; YCR031C; YHR148W; YGL171W; YLR222C; YJR002W; YMR093W; YDL213C; YCL059C; YPL157W; YDL014W; YDR398W; YLR175W	2.13E-14
GO:0000479	endonucleolytic cleavage of tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	YCR057C; YJL069C; YLR129W; YDR449C; YBL004W; YMR229C; YJL109C; YLR222C; YJR002W; YOR078W; YCL059C; YOR004W; YER082C; YDL208W; YOR310C	9.60E-12
GO:0000478	endonucleolytic cleavages during rRNA processing	YCR057C; YJL069C; YLR129W; YDR449C; YBL004W; YMR229C; YJL109C; YLR222C; YJR002W; YOR078W; YCL059C; YOR004W; YER082C; YDL208W; YOR310C	9.60E-12

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0000472	endonucleolytic cleavage to generate mature 5'-end of SSU-rRNA from (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	YCR057C; YJL069C; YLR129W; YDR449C; YBL004W; YMR229C; YJL109C; YLR222C; YJR002W; YOR078W; YOR004W; YER082C; YOR310C	4.46E-11
GO:0000967	rRNA 5'-end processing	YCR057C; YJL069C; YLR129W; YDR449C; YBL004W; YMR229C; YJL109C; YLR222C; YJR002W; YOR078W; YOR004W; YER082C; YOR310C	7.48E-11
GO:0034471	ncRNA 5'-end processing	YCR057C; YJL069C; YLR129W; YDR449C; YBL004W; YMR229C; YJL109C; YLR222C; YJR002W; YOR078W; YOR004W; YER082C; YOR310C	7.48E-11
GO:0000966	RNA 5'-end processing	YCR057C; YJL069C; YLR129W; YDR449C; YBL004W; YMR229C; YJL109C; YLR222C; YJR002W; YOR078W; YOR004W; YER082C; YOR310C	1.22E-10

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0000447	endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	YCR057C; YJL069C; YLR129W; YDR449C; YBL004W; YMR229C; YJL109C; YLR222C; YJR002W; YOR078W; YCL059C; YOR004W; YER082C; YOR310C	1.25E-10
GO:0000480	endonucleolytic cleavage in 5’-ETS of tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	YMR229C; YLR222C; YCR057C; YJL109C; YJR002W; YJL069C; YLR129W; YDR449C; YOR004W; YER082C; YBL004W; YOR310C	6.21E-10
GO:0000460	maturation of 5.8S rRNA	YCR057C; YJL069C; YLR129W; YGL120C; YDR449C; YBL004W; YMR229C; YLR222C; YJL109C; YJR002W; YOR078W; YCL059C; YKL014C; YOR004W; YER082C; YOR310C	2.62E-09

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0000466	maturation of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	YCR057C; YJL069C; YLR129W; YGL120C; YDR449C; YBL004W; YMR229C; YLR222C; YJL109C; YJR002W; YOR078W; YCL059C; YKL014C; YOR004W; YER082C; YOR310C	2.62E-09
GO:0000469	cleavages during rRNA processing	YCR057C; YJL069C; YLR129W; YDR449C; YBL004W; YMR229C; YJL109C; YLR222C; YJR002W; YOR078W; YCL059C; YOR004W; YER082C; YDL208W; YOR310C	3.97E-09
GO:0045943	positive regulation of transcription from RNA polymerase I promoter	YJL109C; YHR196W; YGR128C; YMR093W; YDR324C; YPL126W; YDR398W	3.39E-05
GO:0006356	regulation of transcription from RNA polymerase I promoter	YJL109C; YHR196W; YGR128C; YMR093W; YDR324C; YPL126W; YDR398W	6.16E-04
GO:0042274	ribosomal small subunit biogenesis	YJR002W; YPL266W; YDL213C; YLR129W; YCL059C; YER082C; YLR003C; YHR148W; YCR031C	6.27E-03

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0000154	rRNA modification	YLR197W; YPL266W; YDL014W; YDL208W; YLR175W; YPR137W	7.47E-03
GO:0045941	positive regulation of transcription	YDR224C; YDR324C; YGL150C; YPL126W; YJL109C; YHR196W; YGR128C; YGR270W; YMR093W; YGL133W; YDR169C; YDR398W; YBR009C	1.10E-01
GO:0010628	positive regulation of gene expression	YDR224C; YDR324C; YGL150C; YPL126W; YJL109C; YHR196W; YGR128C; YGR270W; YMR093W; YGL133W; YDR169C; YDR398W; YBR009C	1.16E-01
GO:0051173	positive regulation of nitrogen compound metabolic process	YDR224C; YDR324C; YGL150C; YPL126W; YJL109C; YHR196W; YGR128C; YGR270W; YMR093W; YGL133W; YDR169C; YDR398W; YBR009C	1.80E-01
GO:0045935	positive regulation of nucleobase; nucleoside; nucleotide and nucleic acid metabolic process	YDR224C; YDR324C; YGL150C; YPL126W; YJL109C; YHR196W; YGR128C; YGR270W; YMR093W; YGL133W; YDR169C; YDR398W; YBR009C	1.80E-01

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0010557	positive regulation of macromolecule biosynthetic process	YDR224C; YDR324C; YGL150C; YPL126W; YJL109C; YHR196W; YGR128C; YGR270W; YMR093W; YGL133W; YDR169C; YDR398W; YBR009C	2.99E-01
GO:0031328	positive regulation of cellular biosynthetic process	YDR224C; YDR324C; YGL150C; YPL126W; YJL109C; YHR196W; YGR128C; YGR270W; YMR093W; YGL133W; YDR169C; YDR398W; YBR009C	3.47E-01
GO:0009891	positive regulation of biosynthetic process	YDR224C; YDR324C; YGL150C; YPL126W; YJL109C; YHR196W; YGR128C; YGR270W; YMR093W; YGL133W; YDR169C; YDR398W; YBR009C	3.47E-01
GO:0010604	positive regulation of macromolecule metabolic process	YDR224C; YDR324C; YGL150C; YPL126W; YJL109C; YHR196W; YGR128C; YGR270W; YMR093W; YGL133W; YDR169C; YDR398W; YBR009C	5.41E-01
GO:0045893	positive regulation of transcription; DNA-dependent	YJL109C; YHR196W; YGR128C; YGR270W; YMR093W; YDR324C; YGL150C; YPL126W; YDR169C; YDR398W	6.52E-01

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0051254	positive regulation of RNA metabolic process	YJL109C; YHR196W; YGR128C; YGR270W; YMR093W; YDR324C; YGL150C; YPL126W; YDR169C; YDR398W	7.43E-01
GO:0006355	regulation of transcription; DNA-dependent	YGR122W; YDR224C; YDR324C; YGL150C; YPL126W; YDR310C; YJL109C; YHR196W; YGR270W; YGR128C; YMR093W; YBL052C; YBL054W; YMR247C; YER088C; YMR080C; YGL133W; YDR169C; YMR307W; YDR398W	7.77E-01
GO:0009451	RNA modification	YLR197W; YPL266W; YGL078C; YPL157W; YDL014W; YOR004W; YDL208W; YLR175W; YPR137W	8.36E-01
GO:0051252	regulation of RNA metabolic process	YGR122W; YDR224C; YDR324C; YGL150C; YPL126W; YDR310C; YJL109C; YHR196W; YGR270W; YGR128C; YMR093W; YBL052C; YBL054W; YMR247C; YER088C; YMR080C; YGL133W; YDR169C; YMR307W; YDR398W	8.46E-01
GO:0045814	negative regulation of gene expression; epigenetic	YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YMR080C; YGL133W; YMR307W	9.28E-01

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0006342	chromatin silencing	YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YMR080C; YGL133W; YMR307W	9.28E-01
GO:0051172	negative regulation of nitrogen compound metabolic process	YOL081W; YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YMR080C; YGL133W; YMR307W; YBR009C	9.61E-01
GO:0045934	negative regulation of nucleobase; nucleoside; nucleotide and nucleic acid metabolic process	YOL081W; YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YMR080C; YGL133W; YMR307W; YBR009C	9.61E-01
GO:0040029	regulation of gene expression; epigenetic	YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YMR080C; YGL133W; YMR307W	9.78E-01
GO:0016458	gene silencing	YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YMR080C; YGL133W; YMR307W	9.82E-01
GO:0016481	negative regulation of transcription	YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YMR080C; YGL133W; YMR307W; YBR009C	9.84E-01

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0010629	negative regulation of gene expression	YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YMR080C; YGL133W; YMR307W; YBR009C	9.90E-01
GO:0031327	negative regulation of cellular biosynthetic process	YOL081W; YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YMR080C; YGL133W; YMR307W; YBR009C	9.91E-01
GO:0006348	chromatin silencing at telomere	YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YGL133W	9.93E-01
GO:0009890	negative regulation of biosynthetic process	YOL081W; YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YMR080C; YGL133W; YMR307W; YBR009C	9.93E-01
GO:0045892	negative regulation of transcription; DNA-dependent	YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YMR080C; YGL133W; YMR307W	9.97E-01
GO:0051253	negative regulation of RNA metabolic process	YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YMR080C; YGL133W; YMR307W	9.97E-01

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0010558	negative regulation of macromolecule biosynthetic process	YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YMR080C; YGL133W; YMR307W; YBR009C	1.00E+00
GO:0045449	regulation of transcription	YGR122W; YDR224C; YGL150C; YDR324C; YPL126W; YDR310C; YJL109C; YHR196W; YGR270W; YGR128C; YMR093W; YBL052C; YGR040W; YBL054W; YMR247C; YER088C; YMR080C; YGL133W; YDR169C; YMR307W; YDR398W; YBR009C	1.00E+00
GO:0000463	maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	YMR229C; YGL120C; YKL014C	1.00E+00
GO:0000470	maturation of LSU-rRNA	YMR229C; YGL120C; YKL014C	1.00E+00
GO:0010605	negative regulation of macromolecule metabolic process	YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YER088C; YDR310C; YMR080C; YGL133W; YMR307W; YBR009C	1.00E+00

Table A.16: GO analysis (biological processes) for yeast LS proteins without RNase treatment

GO annotation	GO term	Genes	p-value
GO:0042255	ribosome assembly	YDR060W; YGL078C; YKL014C; YOL077C; YCR031C	1.00E+00
GO:0042273	ribosomal large subunit biogenesis	YDR060W; YOL144W; YGL078C; YGL120C; YOL077C	1.00E+00
GO:0031118	rRNA pseudouridine synthesis	YDL208W; YLR175W	1.00E+00

Table A.17: GO analysis (biological processes) for yeast LS proteins with RNase treatment

GO annotation	GO term	Genes	p-value
GO:0006364	rRNA processing	YLR197W; YOL144W; YGL078C; YPL266W; YEL026W; YGL120C; YBL004W; YJL109C; YJR041C; YPL157W; YKL014C; YDL014W; YDL208W; YLR175W; YOR310C	4.40E-02
GO:0016072	rRNA metabolic process	YLR197W; YOL144W; YGL078C; YPL266W; YEL026W; YGL120C; YBL004W; YJL109C; YJR041C; YPL157W; YKL014C; YDL014W; YDL208W; YLR175W; YOR310C	6.48E-02
GO:0000154	rRNA modification	YLR197W; YPL266W; YDL014W; YDL208W; YLR175W	8.29E-02

Table A.17: GO analysis (biological processes) for yeast LS proteins with RNase treatment

GO annotation	GO term	Genes	p-value
GO:0006338	chromatin remodeling	YFL013C; YOR141C; YNL088W; YGL150C; YGL133W; YDL002C; YBR245C; YLR095C	1.24E-01
GO:0042254	ribosome biogenesis	YLR197W; YOL144W; YPL266W; YGL078C; YEL026W; YGL120C; YBL004W; YJL109C; YNR053C; YJR041C; YPL157W; YKL014C; YDL014W; YJR066W; YDL208W; YLR175W; YOR310C	2.24E-01
GO:0016568	chromatin modification	YFL013C; YBL052C; YPL116W; YOR141C; YNL088W; YMR247C; YGL150C; YGL133W; YDL002C; YBR245C; YLR095C; YBR009C	2.38E-01
GO:0051172	negative regulation of nitrogen compound metabolic process	YGR122W; YDR224C; YGL150C; YDR310C; YLL004W; YBR245C; YOL081W; YBL052C; YMR247C; YGL133W; YNL167C; YMR307W; YBR009C	2.55E-01
GO:0045934	negative regulation of nucleobase; nucleoside; nucleotide and nucleic acid metabolic process	YGR122W; YDR224C; YGL150C; YDR310C; YLL004W; YBR245C; YOL081W; YBL052C; YMR247C; YGL133W; YNL167C; YMR307W; YBR009C	2.55E-01

Table A.17: GO analysis (biological processes) for yeast LS proteins with RNase treatment

GO annotation	GO term	Genes	p-value
GO:0006325	chromatin organization	YFL013C; YPL116W; YDR224C; YOR141C; YGL150C; YDL002C; YBR245C; YBL052C; YNL088W; YMR247C; YGL133W; YBR009C; YLR095C	2.65E-01
GO:0016481	negative regulation of transcription	YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YDR310C; YGL133W; YNL167C; YLL004W; YMR307W; YBR245C; YBR009C	3.32E-01
GO:0031327	negative regulation of cellular biosynthetic process	YGR122W; YDR224C; YGL150C; YDR310C; YLL004W; YBR245C; YOL081W; YBL052C; YMR247C; YGL133W; YNL167C; YMR307W; YBR009C	3.67E-01
GO:0010629	negative regulation of gene expression	YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YDR310C; YGL133W; YNL167C; YLL004W; YMR307W; YBR245C; YBR009C	3.68E-01
GO:0009890	negative regulation of biosynthetic process	YGR122W; YDR224C; YGL150C; YDR310C; YLL004W; YBR245C; YOL081W; YBL052C; YMR247C; YGL133W; YNL167C; YMR307W; YBR009C	3.89E-01

Table A.17: GO analysis (biological processes) for yeast LS proteins with RNase treatment

GO annotation	GO term	Genes	p-value
GO:0022613	ribonucleoprotein complex biogenesis	YLR197W; YOL144W; YPL266W; YGL078C; YEL026W; YGL120C; YBL004W; YJL109C; YNR053C; YJR041C; YPL157W; YKL014C; YDL014W; YJR066W; YDL208W; YLR175W; YOR310C	6.19E-01
GO:0010558	negative regulation of macromolecule biosynthetic process	YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YDR310C; YGL133W; YNL167C; YLL004W; YMR307W; YBR245C; YBR009C	6.47E-01
GO:0034470	ncRNA processing	YLR197W; YOL144W; YGL078C; YPL266W; YEL026W; YGL120C; YBL004W; YJL109C; YJR041C; YPL157W; YKL014C; YDL014W; YDL208W; YLR175W; YOR310C	7.49E-01
GO:0051276	chromosome organization	YFL013C; YFL037W; YDR224C; YPL116W; YOR141C; YGL150C; YDL002C; YBR245C; YBL052C; YNL088W; YMR247C; YGL133W; YPL157W; YML085C; YLR095C; YBR009C	8.64E-01
GO:0045892	negative regulation of transcription; DNA-dependent	YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YDR310C; YGL133W; YNL167C; YLL004W; YMR307W	8.89E-01

Table A.17: GO analysis (biological processes) for yeast LS proteins with RNase treatment

GO annotation	GO term	Genes	p-value
GO:0051253	negative regulation of RNA metabolic process	YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YDR310C; YGL133W; YNL167C; YLL004W; YMR307W	8.98E-01
GO:0045449	regulation of transcription	YFL013C; YGR122W; YDR224C; YPL116W; YOR141C; YIL038C; YGL150C; YDR310C; YLL004W; YDL002C; YBR245C; YJL109C; YGR270W; YBL052C; YGR040W; YBL054W; YMR247C; YGL133W; YNL167C; YMR307W; YLR095C; YBR009C	9.56E-01
GO:0000742	karyogamy during conjugation with cellular fusion	YFL037W; YDR356W; YML085C; YHR073W	9.66E-01
GO:0010605	negative regulation of macromolecule metabolic process	YGR122W; YDR224C; YBL052C; YMR247C; YGL150C; YDR310C; YGL133W; YNL167C; YLL004W; YMR307W; YBR245C; YBR009C	9.84E-01
GO:0006355	regulation of transcription; DNA-dependent	YGR122W; YDR224C; YPL116W; YIL038C; YGL150C; YDR310C; YLL004W; YDL002C; YBR245C; YJL109C; YGR270W; YBL052C; YBL054W; YMR247C; YGL133W; YNL167C; YMR307W	9.86E-01

Table A.17: GO analysis (biological processes) for yeast LS proteins with RNase treatment

GO annotation	GO term	Genes	p-value
GO:0000741	karyogamy	YFL037W; YDR356W; YML085C; YHR073W	9.93E-01
GO:0045814	negative regulation of gene expression; epigenetic	YBL052C; YMR247C; YGL150C; YDR310C; YGL133W; YLL004W; YMR307W	9.94E-01
GO:0006342	chromatin silencing	YBL052C; YMR247C; YGL150C; YDR310C; YGL133W; YLL004W; YMR307W	9.94E-01
GO:0051252	regulation of RNA metabolic process	YGR122W; YDR224C; YPL116W; YIL038C; YGL150C; YDR310C; YLL004W; YDL002C; YBR245C; YJL109C; YGR270W; YBL052C; YBL054W; YMR247C; YGL133W; YNL167C; YMR307W	9.94E-01
GO:0006997	nucleus organization	YFL037W; YDR356W; YPL157W; YFR028C; YML085C; YHR073W	9.95E-01
GO:0034660	ncRNA metabolic process	YLR197W; YOL144W; YGL078C; YPL266W; YEL026W; YGL120C; YBL004W; YJL109C; YJR041C; YPL157W; YKL014C; YDL014W; YDL208W; YLR175W; YOR310C	9.97E-01

Table A.17: GO analysis (biological processes) for yeast LS proteins with RNase treatment

GO annotation	GO term	Genes	p-value
GO:0010551	regulation of specific transcription from RNA polymerase II promoter	YGR270W; YPL116W; YGL150C; YGL133W	9.99E-01
GO:0040029	regulation of gene expression; epigenetic	YBL052C; YMR247C; YGL150C; YDR310C; YGL133W; YLL004W; YMR307W	9.99E-01
GO:0016458	gene silencing	YBL052C; YMR247C; YGL150C; YDR310C; YGL133W; YLL004W; YMR307W	9.99E-01
GO:0000462	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	YJL109C; YGL120C; YEL026W; YDL014W; YBL004W; YOR310C	1.00E+00
GO:0030490	maturation of SSU-rRNA	YJL109C; YGL120C; YEL026W; YDL014W; YBL004W; YOR310C	1.00E+00
GO:0006357	regulation of transcription from RNA polymerase II promoter	YGR122W; YGR270W; YDR224C; YPL116W; YIL038C; YBL054W; YGL150C; YGL133W; YNL167C; YDL002C; YBR245C	1.00E+00

Table A.17: GO analysis (biological processes) for yeast LS proteins with RNase treatment

GO annotation	GO term	Genes	p-value
GO:0000747	conjugation with cellular fusion	YFL037W; YGR040W; YDR356W; YPR122W; YML085C; YHR073W; YKR031C	1.00E+00
GO:0009451	RNA modification	YLR197W; YPL266W; YGL078C; YPL157W; YDL014W; YDL208W; YLR175W	1.00E+00
GO:0006396	RNA processing	YLR197W; YOL144W; YDR194C; YPL266W; YGL078C; YEL026W; YIL038C; YGL120C; YBL004W; YJL109C; YJR041C; YPL157W; YKL014C; YDL014W; YDL208W; YLR175W; YOR310C	1.00E+00
GO:0032583	regulation of gene-specific transcription	YGR270W; YPL116W; YGL150C; YGL133W	1.00E+00
GO:0006348	chromatin silencing at telomere	YBL052C; YMR247C; YGL150C; YDR310C; YGL133W	1.00E+00
GO:0000746	conjugation	YFL037W; YGR040W; YDR356W; YPR122W; YML085C; YHR073W; YKR031C	1.00E+00
GO:0000478	endonucleolytic cleavages during rRNA processing	YJL109C; YDL208W; YBL004W; YOR310C	1.00E+00

Table A.17: GO analysis (biological processes) for yeast LS proteins with RNase treatment

GO annotation	GO term	Genes	p-value
GO:0000479	endonucleolytic cleavage of tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	YJL109C; YDL208W; YBL004W; YOR310C	1.00E+00
GO:0045941	positive regulation of transcription	YJL109C; YGR270W; YDR224C; YPL116W; YGL150C; YGL133W; YBR245C; YBR009C	1.00E+00
GO:0000460	maturation of 5.8S rRNA	YJL109C; YGL120C; YKL014C; YBL004W; YOR310C	1.00E+00
GO:0000466	maturation of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	YJL109C; YGL120C; YKL014C; YBL004W; YOR310C	1.00E+00
GO:0010628	positive regulation of gene expression	YJL109C; YGR270W; YDR224C; YPL116W; YGL150C; YGL133W; YBR245C; YBR009C	1.00E+00

Table A.17: GO analysis (biological processes) for yeast LS proteins with RNase treatment

GO annotation	GO term	Genes	p-value
GO:0010552	positive regulation of specific transcription from RNA polymerase II promoter	YGR270W; YPL116W; YGL150C	1.00E+00
GO:0051173	positive regulation of nitrogen compound metabolic process	YJL109C; YGR270W; YDR224C; YPL116W; YGL150C; YGL133W; YBR245C; YBR009C	1.00E+00
GO:0045935	positive regulation of nucleobase; nucleoside; nucleotide and nucleic acid metabolic process	YJL109C; YGR270W; YDR224C; YPL116W; YGL150C; YGL133W; YBR245C; YBR009C	1.00E+00
GO:0007126	meiosis	YOR373W; YLR219W; YFL037W; YNL088W; YFL009W; YJR066W; YML085C; YKR031C	1.00E+00
GO:0051327	M phase of meiotic cell cycle	YOR373W; YLR219W; YFL037W; YNL088W; YFL009W; YJR066W; YML085C; YKR031C	1.00E+00

Table A.17: GO analysis (biological processes) for yeast LS proteins with RNase treatment

GO annotation	GO term	Genes	p-value
GO:0006260	DNA replication	YCR028C-A; YGR109W-B; YNL088W; YDR310C; YPL157W; YLL004W; YHR031C	1.00E+00
GO:0051321	meiotic cell cycle	YOR373W; YLR219W; YFL037W; YNL088W; YFL009W; YJR066W; YML085C; YKR031C	1.00E+00
GO:0010557	positive regulation of macromolecule biosynthetic process	YJL109C; YGR270W; YDR224C; YPL116W; YGL150C; YGL133W; YBR245C; YBR009C	1.00E+00
GO:0043193	positive regulation of gene-specific transcription	YGR270W; YPL116W; YGL150C	1.00E+00
GO:0031328	positive regulation of cellular biosynthetic process	YJL109C; YGR270W; YDR224C; YPL116W; YGL150C; YGL133W; YBR245C; YBR009C	1.00E+00
GO:0009891	positive regulation of biosynthetic process	YJL109C; YGR270W; YDR224C; YPL116W; YGL150C; YGL133W; YBR245C; YBR009C	1.00E+00

Table A.17: GO analysis (biological processes) for yeast LS proteins with RNase treatment

GO annotation	GO term	Genes	p-value
GO:0006259	DNA metabolic process	YLR219W; YDR224C; YOR141C; YCR028C-A; YGL150C; YDR310C; YLL004W; YJR144W; YHR031C; YIL128W; YNL088W; YGR109W-B; YPL157W; YBR009C	1.00E+00
GO:0051181	cofactor transport	YNL078W; YPR122W; YBL037W; YPL249C	1.00E+00
GO:0007000	nucleolus organization	YPL157W; YFR028C	1.00E+00
GO:0031118	rRNA pseudouridine synthesis	YDL208W; YLR175W	1.00E+00
GO:0006350	transcription	YFL013C; YPL116W; YOR141C; YIL038C; YGL150C; YDR310C; YDL002C; YGR097W; YBR245C; YJL109C; YBL052C; YIL128W; YGL133W; YNL167C; YJR066W; YLR095C	1.00E+00
GO:0000279	M phase	YOR373W; YLR219W; YFL037W; YNL088W; YDR356W; YFL009W; YJR066W; YML085C; YLR175W; YPL124W; YKR031C	1.00E+00

Table A.17: GO analysis (biological processes) for yeast LS proteins with RNase treatment

GO annotation	GO term	Genes	p-value
GO:0000480	endonucleolytic cleavage in 5'-ETS of tricistronic rRNA transcript (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	YJL109C; YBL004W; YOR310C	1.00E+00
GO:0045132	meiotic chromosome segregation	YOR373W; YFL037W; YML085C	1.00E+00
GO:0010604	positive regulation of macromolecule metabolic process	YJL109C; YGR270W; YDR224C; YPL116W; YGL150C; YGL133W; YBR245C; YBR009C	1.00E+00
GO:0000469	cleavages during rRNA processing	YJL109C; YDL208W; YBL004W; YOR310C	1.00E+00
GO:0000472	endonucleolytic cleavage to generate mature 5'-end of SSU-rRNA from (SSU-rRNA; 5.8S rRNA; LSU-rRNA)	YJL109C; YBL004W; YOR310C	1.00E+00

Table A.17: GO analysis (biological processes) for yeast LS proteins with RNase treatment

GO annotation	GO term	Genes	p-value
GO:0034728	nucleosome organization	YDR224C; YGL150C; YBR245C; YBR009C	1.00E+00
GO:0000967	rRNA 5'-end processing	YJL109C; YBL004W; YOR310C	1.00E+00
GO:0034471	ncRNA 5'-end processing	YJL109C; YBL004W; YOR310C	1.00E+00