





Universitat Autònoma de Barcelona

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <http://es.creativecommons.org/blog/licencias/>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>



Universitat Autònoma de Barcelona

Departament de Bioquímica i Biologia Molecular

Institut de Biotecnologia i Biomedicina

**Bioinformatic analysis on the
determinants of protein
aggregation and
conformational conversion**

Valentín Iglesias Mas

Barcelona, March 2021

Universitat Autònoma de Barcelona

Departament de Bioquímica i Biologia Molecular

Institut de Biotecnologia i Biomedicina

Bioinformatic analysis on the determinants of protein aggregation and conformational conversion

Doctoral thesis submitted by Valentín Iglesias Mas in candidacy for the degree of Ph.D. in Biochemistry, Molecular Biology and Biomedicine from the Universitat Autònoma de Barcelona.

The described work has been performed at the Departament de Bioquímica and Biologia Molecular and at the Institut de Biotecnologia i Biomedicina, under the supervision of Prof. Salvador Ventura Zamora.

Valentín Iglesias Mas

Prof. Salvador Ventura Zamora

Barcelona, March 2021

“Somewhere, something incredible is waiting to be known.”

Carl Sagan

Summary

Protein aggregation has moved from being an almost neglected biophysical curiosity to a central research field mostly due to aggregating proteins causing debilitating conditions in humans. The aggregation propensity of polypeptidic sequences is primarily dictated by their amino acid sequence, which delimits the possible interactions between amino acids. Different factors can modulate aggregation propensity. Achieving an energetic stable folded native state usually conceals aggregation prone-regions preventing aberrant self-oligomerization. Not all proteins fold into a defined three-dimensional structure; intrinsically disordered proteins are a group of polypeptides without a defined spatial architecture and therefore are significantly exposed to solvent; which increases the risk of forming aberrant contacts. A special case of disordered proteins or proteins with disordered regions are prions and prion-like proteins. These are characterized by low complexity regions with a cryptic aggregation propensity and able to self-template an aberrant conformation that self-assembles into aggregates.

Bioinformatics has assisted the study of these different kinds of proteins and protein structural levels by providing a toolbox of algorithms to model their behaviour in physiology and disease. These computational models were designed using methodology approximations that exploited the available knowledge at that time. Our understanding of the phenomena that govern processes such as protein aggregation is growing rapidly; therefore, the underlying principles behind these programs should be continuously revisited.

The present thesis provides a bioinformatics analysis of the phenomena behind protein compaction from multiple angles. By analysing protein aggregation in the native state, we propose improvements to both functionality and usability of a state-of-the-art globular prediction method. At the same time, the effect of pH (as a first approach integrating protein environment on calculations) on intrinsically disordered proteins aggregation and conditional folding was analysed. The obtained results will be used to build publicly accessible web servers as cost-effective tools for multiple research lines. The phenomenon behind prion and prion-like conversion will be studied to gain insight into the determinants that regulate this conversion and the functional role of proteins that undergo this transition; an aspect often overshadowed by their association with neurological diseases.

Overall, the work presented in this thesis attempts to understand fundamental inter- and intra-molecular determinants governing protein compaction in near-native and in changing environmental conditions, as a proxy to understand the role of this process in physiology and disease.

Resum

L'agregació de proteïnes ha passat de ser gairebé una curiositat biofísica sense major interès a un dels camps més actius de la recerca, especialment des que es va esbrinar que podia ser la causa de diverses malalties en humans. L'agregació en proteïnes ve determinada en un primer terme per la seva seqüència aminoacídica, que és qui delimita les possibles interaccions entre els seus aminoàcids. Diferents factors modulen aquesta propensió intrínseca a agregar. Sovint les proteïnes assoleixen un plegament natiu que és energèticament més estable i que usualment amaga regions propenses a agregar, i d'aquesta forma es prevé una oligomerització no funcional. No totes les proteïnes requereixen un plegament amb una estructura tridimensional definida; les proteïnes intrínsecament desordenades són un grup de polipèptids que manquen una arquitectura espacial definida, amb lo qual tenen una significativament major exposició al solvent; fet que incrementa el seu risc de formar contactes aberrants. Un cas especial de proteïnes desordenades o amb regions desestructurades són els prions i les proteïnes del tipus prió. Aquestes proteïnes es caracteritzen per tenir regions amb una baixa complexitat amb regions amb propensió críptica a agregar, que són capaces d'automodelar una conformació aberrant que s'acobla en forma d'agregats.

La bioinformàtica ha assistit en l'estudi d'aquests diferents grups de proteïnes i dels diferents nivells estructurals que adopten, dotant-nos d'un seguit d'eines en forma d'algoritmes per modelar els seus comportaments en processos fisiopatològics. Aquests models computacionals van ser dissenyats fent servir el coneixement del qual es disposava en el seu moment. Però el ràpid increment en l'enteniment dels fenòmens que dirigeixen els processos com l'agregació proteica fan imperatiu una contínua revisió i millora en el desenvolupament d'aquests programes.

La present tesi presenta una anàlisi bioinformàtica dels fenòmens darrere la compactació de proteïnes des de múltiples angles. Analitzant l'agregació de proteïnes des de l'estat natiu, proposem millores a la funcionalitat i la usabilitat d'un dels programes de predicció de referència. Tanmateix, s'analitzarà l'efecte del pH (com un primer intent d'integrar la situació on es troba la proteïna als càlculs) en els processos d'agregació i de plegament condicional en proteïnes intrínsecament desordenades. Els resultats obtinguts seran utilitzats per construir servidors web de caràcter obert, pensats com a solucions efectives a la vegada que econòmiques per a múltiples línies de recerca. El fenomen darrere la conversió priònica o de tipus prió serà analitzada per entendre els determinants que ho regulen i el rol funcional de les proteïnes que es sotmeten a aquesta transició; un aspecte sovint eclipsat per la seva associació amb malalties neurològiques.

En general, el treball presentat en aquesta tesi intenta comprendre els determinants inter i intramoleculars que regeixen la compactació de les proteïnes, tant en condicions natives com canviants, i d'aquesta manera d'entendre el paper d'aquest procés tant en condicions fisiològiques com quan esdevé malaltia.

List of publications

This thesis is composed of the following published works:

- I. Iglesias, V., de Groot, N. S. & Ventura, S. (2015) Computational analysis of candidate prion-like proteins in bacteria and their role, *Frontiers in microbiology*. **6**, 1123.
- II. Iglesias, V., Zambrano, R., Conchillo-Sole, O., Illa, R., Rousseau, F., Schymkowitz, J., Sabate, R., Daura, X. & Ventura, S. (2015) PrionW: a server to identify proteins containing glutamine/asparagine rich prion-like domains and their amyloid cores, *Nucleic Acids Research*, 1-7.
- III. Iglesias, V., Pallares, I., de Groot, N. S., Sant'Anna, R., Biosca, A., Fernandez-Busquets, X. & Ventura, S. (2018) Discovering Putative Prion-Like Proteins in Plasmodium falciparum: A Computational and Experimental Analysis, *Frontiers in microbiology*. **9**, 1737.
- IV. Iglesias, V., Conchillo-Sole, O., Batlle, C. & Ventura, S. (2019) AMYCO: evaluation of mutational impact on prion-like proteins aggregation propensity, *BMC bioinformatics*. **20**, 24.
- V. Iglesias, V., Kuriata, A., Pujols, J., Kurcinski, M., Kmiecik, S. & Ventura, S. (2019) Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility, *Nucleic Acids Res.* **47**, W300-W307.
- VI. Iglesias, V., Paladin, L., Juan-Blanco, T., Pallares, I., Aloy, P., Tosatto, S. C. E. & Ventura, S. (2019) In silico Characterization of Human Prion-Like Proteins: Beyond Neurological Diseases, *Frontiers in physiology*. **10**, 314.
- VII. Iglesias, V., Santos, J., Santos-Suárez, J., Mangiagalli, M., Brocca, S., Pallarès, I. & Ventura, S. (2020) pH-Dependent Aggregation in Intrinsically Disordered Proteins Is Determined by Charge and Lipophilicity, *Cells*. **9**, E145.
- VIII. Iglesias, V., Santos, J., Pintado, C., Santos-Suarez, J. & Ventura, S. (2020) DispHred: A Server to Predict pH-Dependent Order-Disorder Transitions in Intrinsically Disordered Proteins, *International journal of molecular sciences*. **21**.
- IX. Pintado, C., Santos, J., Iglesias, V* & Ventura, S.* (2020) SolupHred: A Server to Predict the pH-dependent Aggregation of Intrinsically Disordered Proteins, *Bioinformatics* * co-corresponding authorship

Other articles co-authored which are not part of this thesis:

- I. Pallares, I., Iglesias, V. & Ventura, S. (2015) The Rho Termination Factor of *Clostridium botulinum* Contains a Prion-Like Domain with a Highly Amyloidogenic Core, *Frontiers in microbiology*. **6**, 1516.
- II. Batlle, C., Fernandez, M. R., Iglesias, V. & Ventura, S. (2017) Perfecting prediction of mutational impact on the aggregation propensity of the ALS-associated hnRNPA2 prion-like protein, *FEBS letters*.
- III. Iglesias, V., Batlle, C., Navarro, S. & Ventura, S. (2017) Prion-like proteins and their computational identification in proteomes, *Expert review of proteomics*. **14**, 335-350.
- IV. Batlle, C., de Groot, N. S., Iglesias, V., Navarro, S. & Ventura, S. (2017) Characterization of Soft Amyloid Cores in Human Prion-Like Proteins, *Sci Rep*. **7**, 12134.
- V. Kuriata, A., Iglesias, V., Kurcinski, M., Ventura, S. & Kmieciak, S. (2019) Aggrescan3D standalone package for structure-based prediction of protein aggregation properties, *Bioinformatics*. **35**, 3834-3835.
- VI. Hatos, A., et al. (2020) DisProt: intrinsic protein disorder annotation in 2020, *Nucleic Acids Res*. **48**, D269-D276.
- VII. Carija, A., Pinheiro, F., Iglesias, V. & Ventura, S. (2019) Computational Assessment of Bacterial Protein Structures Indicates a Selection Against Aggregation, *Cells*. **8**.
- VIII. Fernandez, M. R., Pallares, I., Iglesias, V., Santos, J. & Ventura, S. (2019) Formation of Cross-Beta Supersecondary Structure by Soft-Amyloid Cores: Strategies for Their Prediction and Characterization, *Methods in molecular biology*. **1958**, 237-261.
- IX. Santos, J., Iglesias, V. & Ventura, S. (2020) Computational prediction and redesign of aberrant protein oligomerization, *Progress in molecular biology and translational science*. **169**, 43-83.
- X. Santos, J., Pujols, J., Pallarès, I., Iglesias, V. & Ventura, S. (2020) Computational prediction of protein aggregation: Advances in proteomics, conformation-specific algorithms and biotechnological applications, *Computational and structural biotechnology journal*, **18**, 1403-1413.
- XI. Biosca, A., Bouzon-Arnaiz, I., Spanos, L., Siden-Kiamos, I., Iglesias, V., Ventura, S. & Fernandez-Busquets, X. (2020) Detection of Protein Aggregation in Live Plasmodium Parasites, *Antimicrobial agents and chemotherapy*. **64**.

List of figures

- 1.1 Standard proteinogenic amino acids grouped by physicochemical properties conferred by their side chains.
- 1.2 Protein structure is organized in hierarchical levels.
- 1.3 Schematic energy landscape for protein folding and misfolding.
- 1.4 Aggregation propensity strategies for different levels of protein structure.
- 1.5 Compositional and soft-amyloid models of prion conversion.
- 3.1 Graphical Abstract: Aggrescan3D (A3D) 2.0 entails a major update to A3D web server.
- 3.2 The pipeline of A3D 2.0 server.
- 3.3 Aggregation propensity for different multimeric proteins, calculated in static or dynamic modes.
- 3.4 A3D 2.0 as a tool for the *in silico* redesign of more stable and soluble proteins.
- 3.5 Automated mutations for variable heavy (VH) segment of a human germline antibody.
- 3.6 A3D 2.0 redesigned main page.
- 4.1 Properties of PNT variants.
- 4.2 Lipophilicity-based prediction aggregation propensity at pH 7.4 against state-of-the-art aggregation predictor.
- 4.3 Modelling IDP pH-dependent solubility based on lipophilicity and net charge.
- 4.4 Prediction of experimental α -syn aggregation kinetic constants.
- 4.5 Linear correlation between IAPP fibrillation rate and predicted solubility at different pH.
- 4.6 Analysis of the effect of pH variations on A β -40 and tau K19 variant solubility.
- 4.7 Evaluation of the pH-dependent mechanism of fibrillation of functional amyloids.
- 4.8 SolupHred pipeline.
- 4.9 SolupHred web server interface.
- 4.10 Graphical Abstract: DispHred web server predicts pH dependant conditional disorder on IDPs.
- 4.11 Comparison of four different hydropathy scales at pH 7.0.
- 4.12 Charge–Hydropathy-based analysis of pH modulated order–disorder transitions.
- 4.13 DispHred web server interface.
- 5.1 Accuracy cut-off plot for PrionW.
- 5.2 PrionW predictions of prion-like domains and amyloid cores in the sequences of the genuine yeast prions NEW1, URE2 and RNQ1.
- 5.3. Screen shots of the PrionW web server.

- 5.4 Correlation between AMYCO and pRANK predictions and the aggregation propensity of human hnRNPA2 prion-like protein variants.
- 5.5 AMYCO web server main page.
- 5.6 Graphical representation of the AMYCO score.
- 6.1 Enrichment and clustering of PrLDs-containing proteins in bacteria accordingly to their biological process GO terms.
- 6.2 Enrichment and clustering of PrLDs-containing proteins in bacteria accordingly to their GO terms.
- 6.3 Number of different Pfam domains found in PrLDs -containing proteins.
- 6.4 PrLDs -containing proteins also contain multiple domains.
- 6.5 Structure of the domains located in the PrLD-containing proteins.
- 6.6 Clustering of GO terms and Pfam domains associated to PrLD-containing proteins in pathogen bacteria.
- 6.7 Computational analysis of the role of *P. falciparum* PrLD-containing proteins.
- 6.8 Soft-amyloid cores prediction in the three candidate proteins.
- 6.9 Predicted PrLD soft-amyloid cores secondary structure.
- 6.10 Binding of the predicted PrLD soft-amyloid cores to amyloid specific dyes.
- 6.11 Fibrillar structures formed by the predicted PrLD soft-amyloid cores.
- 6.12 Fluorescence microscopy analysis of the presence of protein aggregates in *P. falciparum*-infected RBCs.
- 6.13 Human prion-like proteins modularity.
- 6.14 PrLD distribution along the protein sequence.
- 6.15 Human prion-like proteins GO enrichment analysis.
- 6.16 Prion-like proteins expression in human tissues.
- 6.17 Human prion-like proteins disease association.
- 6.18 Human prion-like interactome.

List of tables

- 4.1 Fitting parameters resulting from the non-linear least squares parametrization.
- 4.2 Performance of pH-dependent and pH-independent hydrophobicity approaches in predicting pH-conditioned order–disorder transitions in a C–H analysis by applying Equation (4.2).
- 5.1 Performance of DIANA, LPSs and PrionW approaches in the prediction of experimental yeast prion-like proteins.

5.2 Performance of pRANK and AMYCO approaches in the prediction of mutation impact upon the aggregation of the human prion-like protein hnRNPA2.

5.3 AMYCO correctly predicts prion converting mutations on yeast proteins.

5.4 AMYCO predicts disease-causing mutations on human prion-like proteins

6.1 Predicted *Plasmodium falciparum* PrLD soft-amyloid cores.

6.2 Prion-like proteins are located nearer in the network than expected by chance.

Glossary

A3D – Aggrescan 3D

A β -40/ A β -42 – β -amyloid 40 and 42 residue peptides from amyloid-beta precursor protein

AD – Alzheimer’s disease

APR – Aggregation prone region

α -syn – α -synuclein

AUC – Area under curve

Bap – Biofilm associated proteins

BLAST – Basic Local Alignment Search Tool

BP – Biological Process

CC – Cellular Component

CD – Circular dichroism

CDR – Complementary-determining region

C-H plot – Charge-hydrophathy spatial distribution

CPEB – Cytoplasmic polyadenylation element binding protein

CR – Congo red

CSS – Cascading Style Sheets

Cter – Carboxi terminus

DAVID – Database for Annotation, Visualization and Integrated Discovery

DIANA – Defined Interval Amino acid Numerating Algorithm

DisGeNET – Database of gene-disease association

DNA – Deoxyribonucleic acid

DR – Disordered region

FN – False Negative

FP – False Positive

GFP – Green fluorescent protein

GRAVY – Grand average of hydrophathy

GO – Gene Ontology

HMM – Hidden Markov model

HTML – Hypertext markup language

IAPP – Islet amyloid polypeptide

IDP/IDR – Intrinsically disordered protein/region

JS – JavaScript

JSON – JavaScript Object Notation

LC – Low complexity

LCC – Largest connected component

MCC – Matthews correlation coefficient

MF – Molecular Function

MD – Molecular Dynamics

mRNA – Messenger RNA

MSA – Multiple system atrophy

MSD – Mean shortest distance

NCPR – Net charge per residue

NGP – Non-globular protein

Nter – Amino terminus

OMIM – Online Mendelian Inheritance in Man

PAPA – Prion Aggregation Prediction Algorithm (PAPA)

PD – Parkinson’s disease

PDB – Protein data bank

PFD – Prion forming domain

PLAAC – Prion-like amino acid composition

PNT – Nter moiety of the measles virus phosphoprotein

PPI – Protein-protein interaction

PQC – Protein quality control machinery

PrD – Prion Domain

PrLD – Prion-like domain

RBP – RNA-binding protein

RNA – Ribonucleic acid

ROC - Receiver operating characteristic

RRM – RNA recognition motifs

SEM – Standard error of the mean

SG – Stress granule

STAP – Structural Aggregation Propensity

SVM – Support vector machine

TEM – Transmission electron microscopy

Th-T – Thioflavin-T

TN – True Negative

TP – True Positive

TTR – Transthyretin

TSEs – Transmissible spongiform encephalopathies

UniprotKB – Universal protein resource

VH – Variable Heavy

ZIP – ZIP compression file format

Table of contents

List of publications

List of figures

List of tables

Glossary

1. Introduction

1.1 Protein folding

1.2 Protein misfolding and aggregation

1.2.1 Amyloids

1.2.1.1 Functional amyloids

1.2.2 Bioinformatic approaches to predict protein aggregation

1.3 Intrinsically disordered proteins

1.4 Prions and related phenomena

1.4.1 Bioinformatic approaches for prion-like detection

2. Objectives of the present thesis

3. Chapter I – Globular Protein Aggregation

3.1 Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility.

3.1.1 Abstract

3.1.3 Introduction

3.1.3 Methods

3.1.4 New features and updates

3.1.5 Description of the web server

3.1.6 References

4. Chapter II – Effect of pH in protein compaction

4.1 pH-dependent aggregation in intrinsically disordered proteins is determined by charge and lipophilicity.

4.1.1 Abstract

4.1.2 Introduction

4.1.3 Materials and methods

4.1.4 Results

4.1.5 Discussion

4.1.6 References

4.2 SolupHred: predicting pH-dependent aggregation of intrinsically disordered proteins.

4.2.1 Abstract

4.2.2 Introduction

4.2.3 Method

4.2.4 Implementation

- 4.2.5 Performance
- 4.2.6 Conclusions
- 4.2.7 References

4.3 DisPHred: a server to predict pH-dependent order-disorder transitions in intrinsically disordered proteins.

- 4.3.1 Abstract
- 4.3.2 Introduction
- 4.3.3 Materials and methods
- 4.3.4 Results
- 4.3.5 Discussion
- 4.3.6 References

5. Chapter III – Prediction of prion-like behaviour

5.1 PrionW: server for the prediction of glutamine/asparagine rich prion-like domains and their amyloid cores.

- 5.1.1 Abstract
- 5.1.2 Introduction
- 5.1.3 Methods
- 5.1.4 Performance
- 5.1.5 Server description
- 5.1.6 Conclusion
- 5.1.7 References

5.2 AMYCO: Evaluation of mutational impact on prion-like proteins aggregation propensity.

- 5.2.1 Abstract
- 5.2.2 Background
- 5.2.3 Implementation
- 5.2.4 Conclusions
- 5.2.5 Availability and requirements
- 5.2.6 References

6. Chapter IV – Characterization of prion-like proteins

6.1 Computational analysis of candidate prion-like proteins in Bacteria and their role.

- 6.1.1 Abstract
- 6.1.2 Introduction
- 6.1.3 Materials and methods
- 6.1.4 Results
- 6.1.5 Discussion
- 6.1.6 References

6.2 Discovering putative prion-like proteins in *Plasmodium falciparum*: A computational and experimental analysis.

- 6.2.1 Abstract
- 6.2.2 Introduction
- 6.2.3 Materials and methods
- 6.2.4 Results
- 6.2.5 Discussion

5.2.6 References

6.3 *In silico* characterization of human prion-like proteins: beyond neurological diseases.

6.3.1 Abstract

6.3.2 Introduction

6.3.3 Materials and methods

6.3.4 Results

6.3.5 Discussion

5.3.6 References

7. Concluding remarks

8. References

9. Appendices

9.1 Software used for this thesis

9.2 Databases used for this thesis

9.3 Operatives Systems used for this thesis

9.4 Web browsers used for this thesis

9.5 Supplementary Material

1. Introduction

1.1. PROTEIN FOLDING

Proteins are biological polymers mostly composed of a linear combination of 20 different amino acids. They share a basic skeleton but have different physicochemical properties conferred by their characteristic side chains (**Figure 1.1**). Proteins can have a wide variety of lengths, ranging from 30-40 amino acids to over 20,000. Usually, shorter versions are referred to as oligopeptides or polypeptides. Statistically, the potential different protein sequences for a given length can be calculated as 20^{length} , which for a standard 300 amino acid sequence provides $\sim 10^{400}$ possible unique protein sequences. There's a tight relationship between the amino acid sequence and the function the protein will develop, the space it will be located or the cellular moment it will be needed, and evolution has selected from this almost infinite pool of possible proteins the fittest for each case. All in all, proteins develop the majority of functions within the cells.

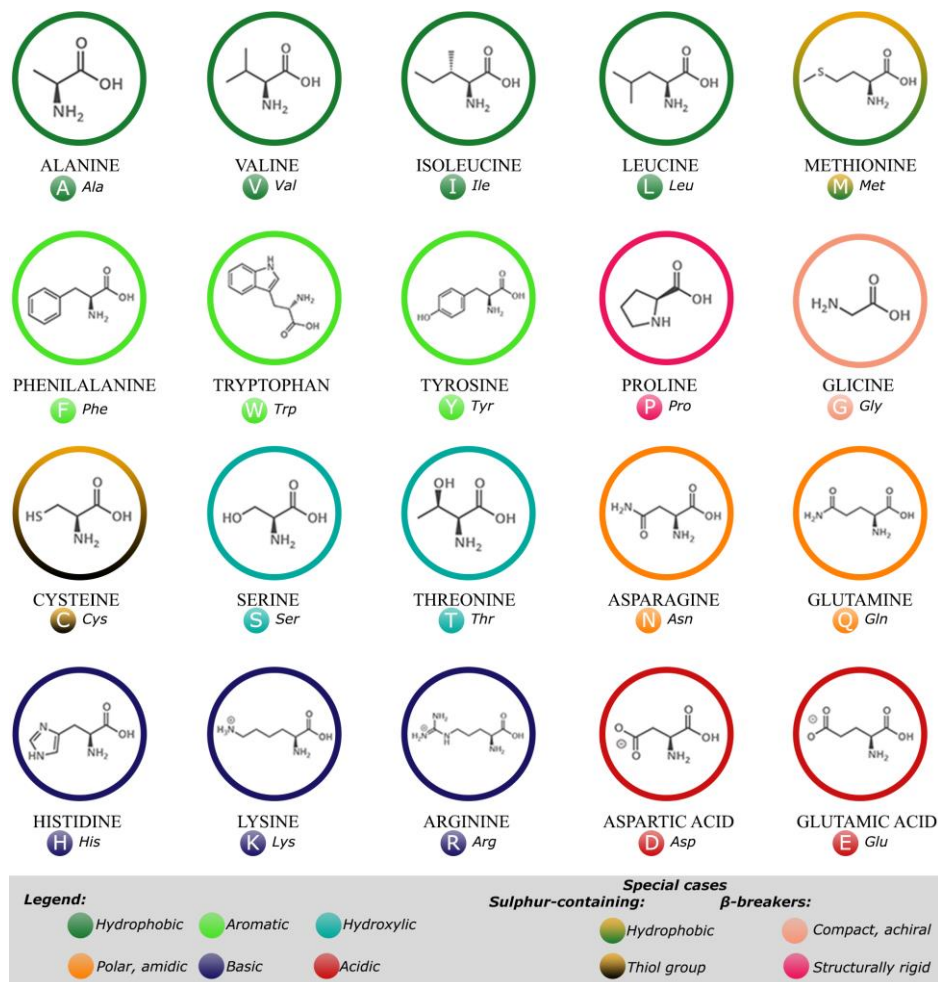


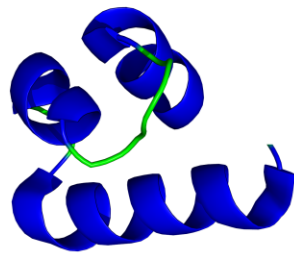
Figure 1.1 – Standard proteinogenic amino acids grouped by physicochemical properties conferred by their side chains. The 20 proteinogenic amino acids that are encoded by the standard genetic code have their groups joined by the α -carbon and are L-stereoisomers (except Glycine which does not possess a chiral centre). Under each amino acid name and their one- and three-letter code, represented by a circle and in italics respectively, is depicted.

The biosynthesis of proteins is carried out in the ribosomes, which sequentially interpret the codons in the mRNA, incorporate the respective amino acids and facilitate peptide bond formation between nascent polypeptide chain and newly incorporated residue.

PRIMARY STRUCTURE

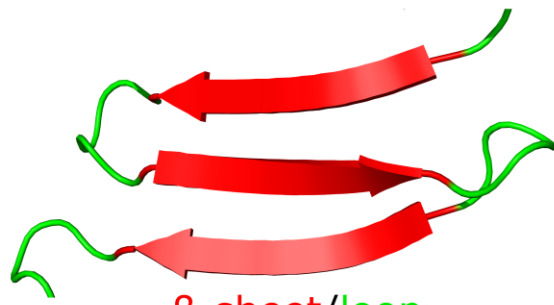
H_3N^+ -Met-Ala-Ser-Tyr-Cys-His-Trp-Gly-Gln-Glu-Ala- COO^-

SECONDARY STRUCTURE



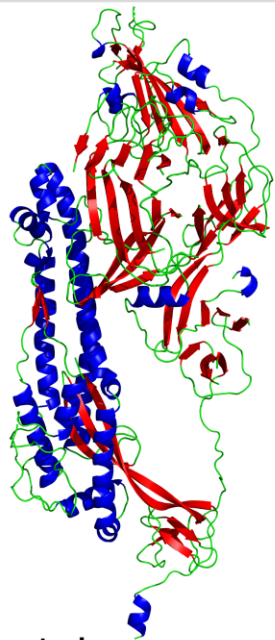
α -helix/loop

SECONDARY STRUCTURE



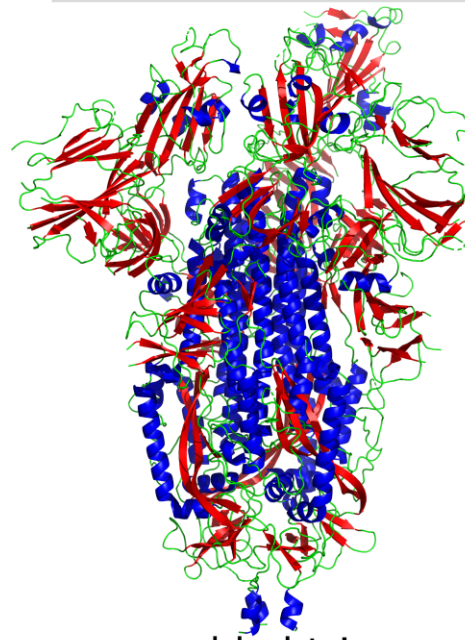
β -sheet/loop

TERTIARY STRUCTURE



protein monomer

QUATERNARY STRUCTURE



assembled trimer

Figure 1.2 – Protein structure is organized in hierarchical levels. The primary structure corresponds to the linear sequence of amino acids. Secondary structure is the first hierarchical three-dimensional arrangement. β -sheet and α -helix elements of secondary structure are depicted in red and blue respectively. Tertiary structure is composed of a folded protein monomer, with secondary structure elements and loops (coloured in green). Some proteins require a quaternary structure to be functional. In the case depicted above, three protein monomers cluster together forming a functional trimer as seen for monomeric and functional SARS-COV-2 Spike protein (PDB: 6VYB).

Most proteins must acquire a specific three dimensional structure, known as the native state (Anfinsen, 1973; Dobson, 2003) to be functional. The process by which polypeptides attain the native state from initially unstructured or partially structured conformations is referred to as protein folding. The study of the process was initiated by Christian B. Anfinsen in the 1960s' and 1970s'. By unfolding and refolding Ribonuclease A, he realised that the primary sequence of a polypeptide in its physiological condition (pH, temperature, and presence of partners or prosthetic groups) dictates its final spatial distribution and postulated the thermodynamic hypothesis; which stated that the native state of the protein constituted the Gibbs free energy minimum (Anfinsen, 1973). Nowadays, it is widely accepted that the protein native state can represent a local minimum, with intermolecular interactions forming supramolecular structures, such as those in protein aggregates (which will be further discussed in **Section 1.2**), can lead to lower energetic configurations. Nonetheless, Cyrus Levinthal noted that for a protein to stochastically explore all possible conformations would require higher times than the age of the universe itself; which contradicted the already known sub-second folding processes of proteins (Levinthal, 1969). This pointed to a cooperative folding scheme in which amino acids don't achieve its conformation independently. After decades of intense study, the actual consensus is that protein folding can be depicted as a folding energy landscape; in which achieving correctly folded stretches with have less potential energy, limit sterically and by establishing local interactions, non-folded stretches' possibilities in search for possible combinations (**Figure 1.3**).

Proteins can consist of a unique or multiple compact structures, called domains. Protein domains are tertiary structure elements that are stable, fold, evolve and usually function autonomously (Janin and Wodak, 1983). Domains act as evolutionary independent, modular structures that contribute to the overall protein functions. Those characteristics make protein domains a building block for protein evolution, with an almost infinite number of ways to combine domains to accommodate function (Russell, 1994). Accordingly, multidomain proteins are thought to arise from single-domain proteins via domain insertion in different genes; which could be accomplished by means of exon shuffling or as a side effect of transposable elements (Russell, 1994). In this way, the SH3 domain is a small, ~60 amino acids, globular domain which is involved in protein binding, and is found in around 300 unrelated human proteins (Saksela and Permi, 2012). Similarly, multi-domain fusions are widely used in synthetic protein engineering, as it allows the rational combination of domains leading to a chimeric protein with predictable function and structure. Green fluorescent protein (GFP) is a single-domain globular protein that emits green fluorescence when exposed to blue light. It plays a main role in the bioluminescence of the jellyfish. Shimomura, Chalfie and Tsien were awarded the 2008 Nobel prize in Chemistry for the discovery (in the jellyfish *Aequorea Victoria*) and development of methodologies to work with GFP. Lately, the GFP domain has been widely used in fusion proteins as a reporter, as it folds into a fluorescent-emitting β -barrel, allowing researchers a simple localization of the chimeric protein even in crowded environments such as cytoplasm or nucleus (Ormo, et al., 1996).

Proteins reside in a very crowded, densely packed medium, in which they interact with partners, the aqueous medium, salts, nucleotides; which ultimately imply fluctuations of their native structure. This

structural dynamism may transiently expose hydrophobic regions, which are originally concealed in the native state, exposing them to non-native intermolecular contacts. This transient exposure might kick-start the misfolding and ultimately the aggregation of unfolded, folding intermediates or even well-folded proteins, being this mechanism concentration dependent (Chartier-Harlin, et al., 2004; Khurana and Lindquist, 2010; Singleton, et al., 2003).

1.2. PROTEIN MISFOLDING AND AGGREGATION

Despite the energetic investment the cell dedicates to ensure a correct protein folding, proteins do not always succeed to fold into their native states. Incorrectly folded or misfolded proteins not only imply a deficient function but may accumulate in a process known as protein aggregation. There is a significant interest in understanding protein misfolding and aggregation mainly driven because this mechanism is responsible for a large number of human disorders, which range from neurodegenerative diseases such as Alzheimer's (AD) and Parkinson's disease, Amyotrophic lateral sclerosis (ALS) to certain types of cancer or type II diabetes (Chiti and Dobson, 2006; Chiti and Dobson, 2017; Graña-Montes, et al., 2017).

Proteins aggregate through a variety of conformers, from nascent, unfolded, partially folded or even completely folded structures (**Figure 1.3**). Initial aggregates are usually clusters of monomers which retain certain structural features of their pervious state (Chiti and Dobson, 2017). Bigger oligomeric aggregates can grow into amorphous or native like-assemblies or matureate into more compact stable species. This usually requires internal reorganization into β -rich oligomers and eventually the formation of insoluble fibrils characterized by cross- β diffraction patterns known as amyloids (which will be further explored **Section 1.2.1**) (Chiti and Dobson, 2017) .

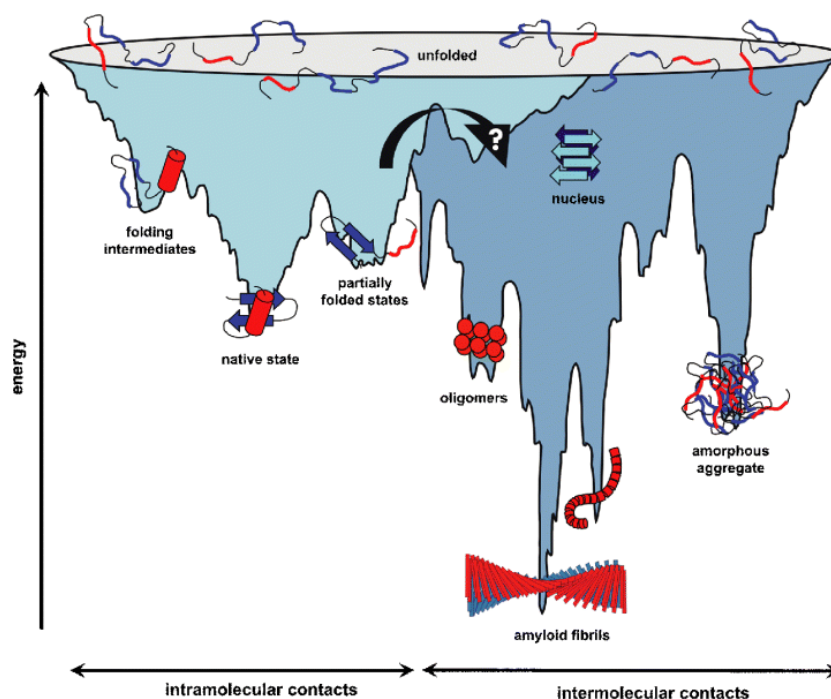


Figure 1.3 – Schematic energy landscape for protein folding and misfolding. The surface shows how attaining favourable intramolecular contacts funnel the energetic minimization towards the native state, decreasing its conformational freedom. On- or off-pathway folding intermediates occupy energetic wells (folding intermediates and partially folded states). Intermolecular interactions allow the formation of aberrant aggregates (amorphous aggregates, β -sheet-rich oligomers, and amyloid fibrils) which represent lower energetic conformations for the system. Figure reproduced with permission from (Jahn and Radford, 2005).

Protein folding and aggregation are regarded as competing processes as they are both driven by similar physicochemical principles such as the hydrophobic effect and hydrogen bonding (Cheon, et al., 2007; Jahn and Radford, 2008; Kauzmann, 1959; Monsellier and Chiti, 2007); although the stabilization of native states comes from specific intramolecular interactions, whereas protein aggregates are mainly stabilized by backbone-to-backbone intermolecular contacts (Auer, et al., 2008). The establishment of an energetically stable native structure minimizes the contacts needed for aggregation, kinetically impairing this reaction (Auer, et al., 2008; Monsellier and Chiti, 2007). However, cells need to synthesize and degrade proteins regularly as a response to external and internal stimuli; therefore, the interplay between protein stability in the native state and an assumable energetic expense to degrade them makes proteins marginally stable (Bartlett and Radford, 2009; Dobson, et al., 1998). In this scenario, changing environmental conditions or certain mutations which destabilize the native state can leave hydrophobic regions exposed to solvent igniting misfolding and aggregation (Bartlett and Radford, 2009; Dobson, et al., 1998). Moreover, protein interaction surfaces frequently require hydrophobic patches (Castillo and Ventura, 2009), but quaternary complexes keep those regions inaccessible to form aberrant contacts, which is key to prevent their aggregation (Santos, et al., 2020b; Yee, et al., 2019). However, disease-causing mutations often destabilize the formation of these complexes, leaving high-aggregation prone monomers a free-way to access aggregation pathways (Santos, et al., 2020b; Yee, et al., 2019).

Aggregation propensity is primarily dictated by its primary sequence (Graña-Montes, et al., 2017; Lopez de la Paz and Serrano, 2004). The amino acidic composition and sequential position dictates the possible interactions between residues, which determine the overall aggregation propensity, rate or even the possibility of amyloid formation (Monsellier and Chiti, 2007). Previous studies have shown that not all protein sequence is equally important for aggregation, but instead that short sequence fragments can promote the full protein to aggregate. These stretches, known as aggregation-prone regions (APRs) or *hot spots* of aggregation, are enriched in hydrophobic amino acids; aromatic (F, W, Y) and aliphatic (V, L, I) (Rousseau, et al., 2006; Ventura, et al., 2004). Therefore, APRs arise from combinations of residues with complementary physicochemical determinants that promote aggregation, namely hydrophobicity, tendency to preferentially adopt a given secondary structure and a low net charge per residue (NCPR) in this specific region.

Hydrophobicity has been identified as a major player in protein compaction; the burial from solvent of the hydrophobic-core, or hydrophobic collapse, being a key driver of globular protein folding (Lindorff-Larsen, et al., 2005) and inter and intramolecular hydrophobic contacts being essential for many quaternary structures or protein-protein interfaces (Castillo and Ventura, 2009). Hydrophobicity is also considered a major driving force of non-native oligomerization. Previous studies showed that mutations of polar to non-polar residues increase the aggregation rate; while mutating a non-polar for a polar amino

acid usually decreases or even abrogates it (Jahn and Radford, 2008). Amino acids specific stereochemistry shape their tendency to adopt different secondary structures and this affects their propensity to facilitate aggregation. Aberrant deposits frequently show β -sheet rich structures, in agreement with the observation that an enrichment in residues with a higher propensities to form β -sheets increased aggregation rates (Chiti, et al., 2002) and the pre-existence of β -strands in the native state intensified protein aggregation, requiring less rearrangement to form amyloid-like aggregates (Pallares, et al., 2004). On the other hand, amino acids with low tendency to form β -sheets (also known as “ β -breakers”) such as P and G, tend to disfavour aggregation (Monsellier and Chiti, 2007; Parrini, et al., 2005; Wood, et al., 1995). Charged residues have an important influence on protein deposition, both by the repulsion effect exerted by equal charges and by an entropic penalty for oligomerization, which negatively impacts most short-ranged intermolecular interactions required for protein aggregation (Reumers, et al., 2009). Chiti and co-workers showed that mutations which did not affect secondary structure, but involved the substitution of uncharged for charged residues decreased aggregation rates, while charged to uncharged changes increased it (Chiti, et al., 2002). Both β -breakers and charged residues are often found surrounding highly hydrophobic regions (Reumers, et al., 2009; Rousseau, et al., 2006) and are thought to be an evolutionary mechanism to discourage non-native contacts between aggregation-prone sequence stretches (Rousseau, et al., 2006), thus acting like aggregation gatekeepers (Rousseau, et al., 2006).

These discussed factors are intrinsic to the protein sequence and can be modulated by environmental conditions which impact kinetically, thermodynamically and structurally the deposition process. Protein concentration, local pH, temperature and the ionic strength are the extrinsic determinants with a greatest effect on aggregation (DuBay, et al., 2004). Protein concentration impact the thermodynamic and kinetic aspects of the aggregation reaction; as being a high-order reaction it is highly dependent on the polypeptide molarity (Tartaglia and Vendruscolo, 2009). Cells keep a tight control of protein expression; with protein levels found to be anti-correlated with their aggregation propensity (Tartaglia and Vendruscolo, 2009). This led Vendruscolo, Tartaglia, Dobson and Pechmann to postulate their *life on the edge* hypothesis: proteins have co-evolved their function and aggregation propensity, but they are present at their solubility limit (Tartaglia, et al., 2007). Temperature, on the other hand, influences the proteins conformational energy landscape varying the Gibbs free energy of each species as well as the activation energy to access them (Graña-Montes, et al., 2017; Lehninger, et al., 2005). As a general trend, working at higher protein concentrations and at higher temperatures accelerates the rate of protein aggregation. pH influences the protonation state of the charged residues (acidic D, E and basic R, K and H), thus modifying the local charge as well as the net charge of the complete protein. This affects the attraction and repulsion effects, thus the formation of electrostatic interactions (Lehninger, et al., 2005) and these residues’ hydrophobicity (Zamora, et al., 2019). In the present thesis we will explore the influence of pH on conditional folding and on protein aggregation. Finally, low salt concentrations stabilize proteins increasing their solubility, while high salt concentrations shields charges non-specifically, reducing the protein effective net charge and thus the repulsion effects between polypeptides.

Cells have evolved intricate strategies to control and minimize deleterious non-functional intermolecular contacts which could lead to aggregation and amyloid formation, conforming what is called the proteostasis network. The most remarkable is the protein quality control machinery (PQC) which comprises chaperones and chaperonins, proteases, ubiquitin ligases, proteasome and autophagy (Chiti and Dobson, 2017). Different molecular chaperones assist protein folding as early as when the protein is being translated, others can prompt unfolding and refolding of non-native conformations and subsequent refolding (Kim, et al., 2013; Patzelt, et al., 2001). Notably, chaperones can recognize the exposure of gatekeeper residues flanking hydrophobic regions (Kim, et al., 2013; Patzelt, et al., 2001). Misfolded proteins that elude this network can be recognized and degraded by the ubiquitin-proteasome system (Kaufman, et al., 2002; Kim, et al., 2013).

1.2.1. AMYLOIDS

Amyloids are supramolecular insoluble assemblies in which connected β -strands form β -sheets which stack consecutive protein molecules perpendicular to the fibre axis. Amyloid fibres share several common features such as binding to specific dyes such as Thioflavin-T (Th-T) and Congo Red (CR), detergent and proteolytic resistance, an enrichment in β -sheet secondary structure which shows specific signals in circular dichroism (CD) and the presence of cross- β signals on X-ray diffraction patterns. Polypeptides that assemble into this morphologically and structurally similar architecture are neither related in sequence, nor in native conformation, still, amyloid fibrils have been identified for a large number of diverse proteins from all kingdoms of life (Chiti and Dobson, 2017; Otzen and Riek, 2019). Eventually, under certain conditions (pH, temperature, salt concentrations, presence or absence of binding partners), it is possible to force virtually any protein to form amyloid assemblies (Chiti and Dobson, 2017; Knowles, et al., 2014). Altogether the access to amyloid structures seems to be a generic property of protein chains, rather than being specifically encoded in the sequence of amino acids (Chiti and Dobson, 2006).

Amyloids have received substantial interest mostly because amyloid depositions have been found for at least 37 peptides or proteins linked with human pathologies (Chiti and Dobson, 2017). As stated above they do not share sequential, structural or functional similarities and their aggregation occurs in a variety of different tissues (Chiti and Dobson, 2006; Chiti and Dobson, 2017; Uversky and Fink, 2004). Proteins forming amyloids in the central nervous system give rise to neurodegenerative conditions; they include β -amyloid peptides ($A\beta$ -40 and $A\beta$ -42) in AD, tau in AD, pick disease and frontotemporal dementia, α -synuclein (α -syn) in PD and multiple system atrophy (MSA) or Huntingtin in Huntington disease (Chiti and Dobson, 2006; Chiti and Dobson, 2017; Uversky and Fink, 2004). Non-neuropathic proteins can form amyloid aggregates in a specific tissue such as IAPP in type II diabetes, in which deposits form in the pancreas, or being systemic such as fragments of immunoglobulin light chains in light-chain amyloidosis, to complicate more the scenario, for some proteins, depending on the mutation, they aggregate in a specific tissue or are systemic (Chiti and Dobson, 2006; Chiti and Dobson, 2017). This latter is the case of

Transthyretin (TTR); a tetrameric protein that functions as a thyroxine transporter and aggregates systemically causing senile systemic amyloidosis disease (Chiti and Dobson, 2006; Chiti and Dobson, 2017; Pinheiro, et al., 2020; Sant'Anna, et al., 2016). However, different mutations destabilize the quaternary complex provoking TTR to form amyloid fibrils in the brain, which causes leptomeningeal amyloidosis or in the myocardium causing familial amyloid cardiomyopathy (Chiti and Dobson, 2006; Chiti and Dobson, 2017; Pinheiro, et al., 2020; Sant'Anna, et al., 2016). AD and PD are the most prevalent neurodegenerative conditions and affect an estimate of 50 million and 7 million people, respectively, especially those aged 65 and above (Brookmeyer, et al., 2007; Prince, 2015). As elder global population increases, the number of affected individuals is expected to duplicate by 2050, generating a huge social and economic burden (Brookmeyer, et al., 2007; Collaborators, 2018; Prince, 2015). Until recently the only available treatment for amyloidosis were palliative cares, which roughly slowed the progression of the diseases. However, work in TTR identified small molecules that significantly reduced the disease progression by stabilizing the protein quaternary structure (Bulawa, et al., 2012; Sant'Anna, et al., 2016). As of the time of elaborating this thesis, one of those drugs, *Tafamidis* is being used in the clinic in Europe and Japan, while *Tolcapone*, an already FDA-approved molecule is ongoing phase IIa clinical trials for different TTR-amyloidosis (Gamez, et al., 2019; Reig, et al., 2015). Similar endeavours are taking place for other amyloidosis with several drugs ongoing different stages of clinical trials (Nuvolone and Merlini, 2017; Pujols, et al., 2018; Pujols, et al., 2020).

1.2.1.1. FUNCTIONAL AMYLOIDS

Amyloid deposits have been traditionally regarded as undesirable pathogenic agents. However, amyloid fibres unique physicochemical and mechanical properties make them ideal to fulfil several specific biological functions that cannot be exerted by individual protein subunits. Indeed, organisms belonging to all kingdoms of life have evolved amyloid conformations for specific physiological tasks (Camara-Almiron, et al., 2018; Loquet, et al., 2018; McGlinchey and Lee, 2018; Otzen and Riek, 2019; Pallares, et al., 2015; Santos and Ventura, 2020), and witty strategies to avoid cytotoxic effects, when in their hosts, such as membrane-bounded compartmentation or modulation of assembly by pH, post-translational modifications, protease processing or shifting the direction of the reaction by modifying reactant concentrations (Jackson and Hewitt, 2017; Otzen, 2010; Otzen and Riek, 2019). A well-characterized functional amyloid application is biofilm formation in different bacteria. Biofilms are a self-produced extracellular matrices composed of polysaccharides, proteins, lipids and nucleic acids which protect bacteria from antimicrobials, chemical stresses, shear forces or the immune system, allowing communities formed of diverse groups of bacteria and fungi to thrive (Flemming and Wingender, 2010). Bacteria from the genus *Escherichia* (Curli fibres from CsgA and CsgB proteins), *Salmonella* (Curli fibres from CsgA and CsgB proteins) and *Pseudomonas* (Fap proteins) secrete into the biofilm proteins that assemble into amyloid aggregates, conferring high mechanical firmness (Chapman, et al., 2002; Zeng, et al., 2015). *Staphylococcus* secrete Biofilm associated proteins (Bap) that assemble into an amyloid in the

biofilm in response to environmental conditions (pH and Ca²⁺ levels), acting as an amyloid-switch-like mechanism (Taglialegna, et al., 2016).

Amyloid architecture is also functionally exploited in eukaryotes (Chiti and Dobson, 2017). Studies in *Antheraea Polyphemus*, silk moths, have shown that the main proteic component of eggshells forms amyloid fibrils (Iconomidou, et al., 2000). These amyloids would confer the oocyte and embryo mechanical and environmental protection, while allowing the biologically required gas exchange. During seed maturation in *Pisum sativum*, garden pea, Viciclin accumulates as amyloid fibrils conferring a source of amino acids for seed germination, growth, and possibly being a pathogen defence mechanism (Antonets, et al., 2020; Santos and Ventura, 2020). Further research will show if these vegetal functional amyloids are also present in other taxonomic groups. In humans and other mammals, pigment cell-specific protein Pmel17 amyloid formation is responsible for the deposition of melanin, thus playing a crucial role in the maturation of melanosomes (Watt, et al., 2013). Pmel17 deposition is tightly regulated, being transported through several endosomal compartments as a proprotein before being proteolytically processed to its aggregational form, and requiring the acidic (pH~5) environment of melanosomes to aggregate (Otzen and Riek, 2019; Watt, et al., 2013). It is expected that further research will find more cases of functional amyloid in living organisms, as it occurred with pathogenic amyloids, which were initially thought to be anecdotic.

The amyloid structure has been recently exploited to generate building blocks for functionalized self-assembled nanostructures such as nanotubes, nanocomposites, scaffolds for cell growth and biocatalysis, adhesives, hydrogels, biosensors or for energy conversion (Diaz-Caballero, et al., 2018; Knowles and Mezzenga, 2016; Li, et al., 2012; Wang and Ventura, 2020). This approach is extremely promising; however, certain technical limitations must be still overcome, specially the loss of the globular structure suffered during the rapid transition to β -sheet rich pre-amyloid conformations, which hamper or inactivate the protein function (Wang and Ventura, 2020).

1.2.2. BIOINFORMATIC APPROACHES TO PREDICT PROTEIN AGGREGATION

Growing knowledge on the physicochemical, sequential, and structural determinants of protein aggregation have propelled the development of mathematical models to predict the propensity to aggregate. The analysis of protein aggregation requires to consider the conditions in which this reaction occurs: the conformation, interacting partners and protein environment. Different conformational levels impose different constraints to aggregation, therefore dedicated computational algorithms are needed for each particular case (Santos, et al., 2020a; Santos, et al., 2020b). The first generation of aggregation predictors were designed to search for linear APRs, therefore they only required the primary sequence as an input. These algorithms can be divided in different categories according to the nature of the determinants of protein aggregation they evaluate (Grana-Montes, et al., 2012; Santos, et al., 2020a). Phenomenological predictors are characterized by applying experimentally derived scoring systems. This category includes algorithms such as Zyggregator, TANGO or AGGRESCAN (Conchillo-Sole, et al., 2007; de Groot, et al., 2012; Fernandez-Escamilla, et al., 2004; Tartaglia and Vendruscolo, 2008). Zyggregator applies an equation that accounts for hydrophobicity, secondary structure propensity and net charge,

built upon the changes in aggregation rate promoted by point mutations, while also pondering the solubilizing effect of gatekeeper residues. TANGO, on the other hand, evaluates the population of secondary structure from empirically and statistically derived amino acidic preferences. It is commonly accepted that regions with a tendency for β -sheet >5% over >5 consecutive amino acids reflect an APR. Noteworthy, TANGO allows tuning of extrinsic parameters such as ionic strength, temperature and pH; which modify bonding energies and thus secondary structure propensity (Lacroix, et al., 1998). Finally, AGGRESCAN evaluates the input sequence on an aggregation propensity scale obtained *in vivo*. Briefly, Ventura and co-workers mutated the central domain of A β -42 fused to GFP to all other 19 possible residues, and measured the emitted fluorescence (a reporter of the protein fusion solubility) (de Groot, et al., 2006). A second kind of algorithms to predict aggregation from the primary sequence corresponds to those that, in a way or another, are structure-based. They evaluate the conformational compatibility of the sequence with an amyloid fold. PASTA, FoldAmyloid, and Waltz are representatives of this class (Garbuzynskiy, et al., 2010; Maurer-Stroh, et al., 2010; Walsh, et al., 2014). The first two use scoring systems derived from protein structures; PASTA applies an energetic function which evaluates the possible parallel and anti-parallel β -pairing by considering the interaction potential and hydrogen-bonding for non-consecutive residues, while FoldAmyloid evaluates the hydrogen-bonding propensity and the packing density, under the premise that it is higher in hydrophobic stretches. Waltz uses a position-dependent matrix, which was trained upon evaluating the ability to form amyloids of over 200 hexapeptides' by electron microscopy, circular dichroism, Fourier-transform infrared spectroscopy and X-ray diffraction. A third group of programs combines the output of several predictors weighting their predictions and generating a consensus. In this way they try to minimize the possible bias any algorithm may have, thus increasing robustness (Graña-Montes, et al., 2017; Santos, et al., 2020a). AMYLPRED 2 or MetAmyl algorithms apply this rationale (Emily, et al., 2013; Tsolis, et al., 2013). The first generates consensus over 11 different algorithms but allows the user to customise the final output by deselecting some predictors, which is advised for redundant methods. MetAmyl applies instead four methodologies which showed lower redundancy and scores according to a linear combination of them. Lately, new aggregation predictors that exploit machine-learning strategies have arisen. APPNN and NETCSSP which use neural networks or FISH Amyloid, which applies a non-classical machine learning strategy, have applied these different methodologies to rank physicochemical and biochemical signatures in amyloids (such as β -propensity, hydrophobicity or by identifying specific patterns) to predict aggregation propensities (Familia, et al., 2015; Gasior and Kotulska, 2014; Kim, et al., 2009).

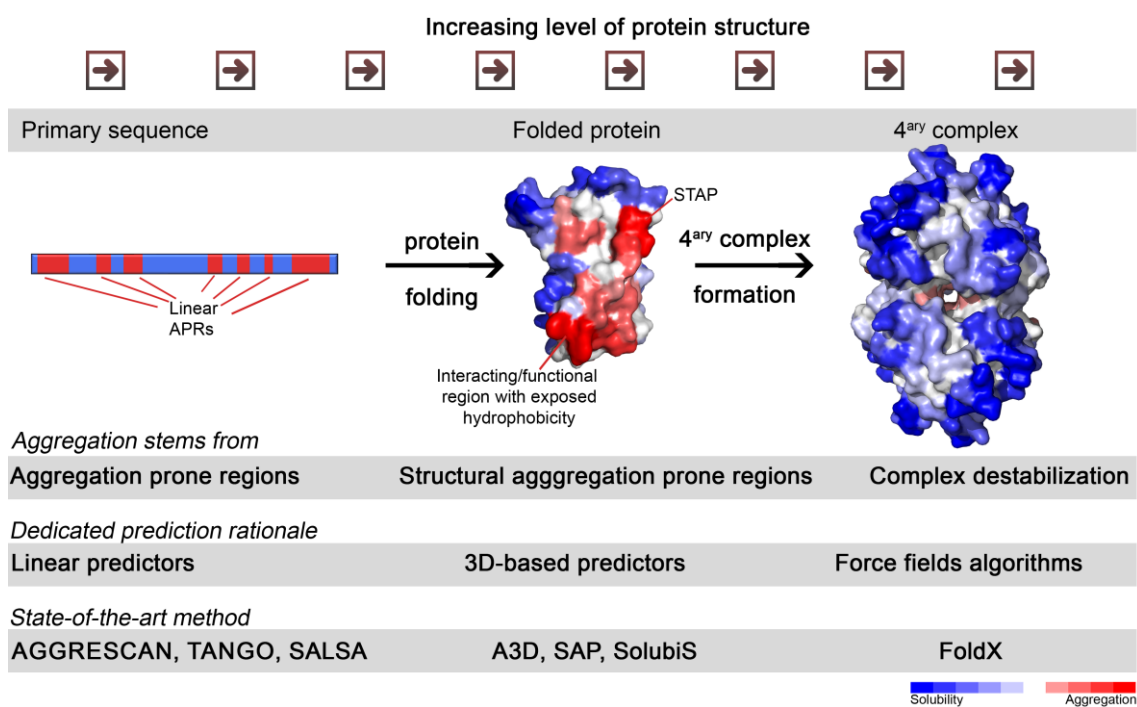


Figure 1.4 – Aggregation propensity strategies for different levels of protein structure. Linear predictors such as TANGO, AGGRESCAN or SALSA identify APRs, for which are most suitable for IDPs and non-folded polypeptides. Folded proteins expose STAPs for which 3D based predictors such as A3D, Camsol or SAP are recommended. For quaternary structures in which aggregation prone free monomers expose STAPs, the complex stability is the main source of aggregation. Therefore, force fields such as FoldX are vital to model protein deposition. Depicted PDB structures (in A3D colour-code) correspond to monomeric and tetrameric TTR (PDB: 1F41).

All in all, the aforementioned predictions methods have shown to be useful to disentangle the sequential determinants behind several disease-related proteins (Belli, et al., 2011). These approaches are especially suitable specially for IDPs or nascent proteins which have not acquired yet the native fold. Nonetheless their applicability to already folded structures is limited. In globular folds the three-dimensional disposition of the amino acids modifies their intrinsic aggregation propensities; with the establishment of contacts between non-contiguous amino acids and the hiding of certain residues inside a compact hydrophobic core. As a general trend, linear predictors overestimate the real aggregation propensity of folded domains. To overcome this limitation, Trout and co-workers applied molecular dynamics (MD) simulations and calculated the resulting solvent accessible areas to weight the hydrophobicity of each individual residue in a near-native environment. They named the resulting parameter as Spatial Aggregation Propensity (SAP) (Chennamsetty, et al., 2009) This first structure-based methodology was intended to be a cost-effective approach to generate more soluble protein-based biotherapeutics, especially antibodies. Structure-based algorithms use the three-dimensional protein coordinates as an input, instead of the sequence. These algorithms evaluate solvent-exposed hydrophobic patches, known as Structural APRs (STAP); which often overlap with interfaces or functional surfaces (Santos, et al., 2020b; van der Kant, et al., 2017). Examples of them are: SAP, SolubiS, CamSol and Aggrescan3D (A3D) (Sormanni, et al., 2015; Van Durme, et al., 2016; Zambrano, et al., 2015). SolubiS identifies linear APRs using TANGO and correct their propensity according to the local stability of the folded structure using the FoldX force field (Schymkowitz, et al., 2005; Van Durme, et al., 2016). This makes it able to analyse the structural

context of APRs, but because the primary prediction uses a linear predictor, this comes at the cost of being blind to STAPs. Camsol is the structural evolution of the linear predictor Zyggregator. It uses a linear combination of physicochemical properties derived from the primary sequence: hydrophobicity, charge, α -helix and β -sheet propensities and scans the sequence with a 7 amino acid window. Next it applies structural corrections to those calculations (Sormanni, et al., 2015). Remarkably, Camsol applies a semi-automated redesign strategy that identifies poorly soluble stretches and performs amino acids substitutions or insertions to improve solubility. Users can select the number of stretches to be engineered and select the functional residues to remain unchanged. As a proof of principle, they redesigned the gammabody A β (33-42), an anti-A β antibody-based molecule, showing an increase in the solubility of the engineered variants, while maintaining A β 42 binding capacity.

A3D is our group implementation of a structure-based aggregation predictor. It applies the *in vivo* aggregation propensity scale of AGGRESCAN corrected by its solvent-exposure by applying a spherical solvent exposure boundary (similar to SAP). A3D incorporates FoldX force field to minimize energetic clashes in the input structure and a *dynamic mode* in which the CABS-flex protocol, an efficient alternative to classical all-atom MD, is used to model protein flexibility in its native state, thus uncovering transiently populated conformations (Jamroz, et al., 2013; Kuriata, et al., 2018). To test the algorithm, they focused on β 2-microglobulin, a protein that forms amyloids in patients on long-term haemodialysis, ultimately causing haemodialysis-associated amyloidosis (Floege and Ketteler, 2001). Several mutations that accelerate amyloid formation have been described. A3D dynamic mode was able to rank the mutations' effect on experimentally observed amyloid propensity. I7A, one of the worst prognosis mutants, truncates an aliphatic group, thus being considered as more soluble by linear predictors. Instead, A3Ds' dynamic mode is able to model the transient exposure of hydrophobic residues hidden in the *wild type* β 2-microglobulin, thus explaining the experimentally observed increase in amyloidogenicity. A3D was also used to redesign a fast-folding, aggregation-resistant GFP variant, as well as redesigns of human antibodies (Gil-Garcia, et al., 2018).

Evolutionary pressure on oligomeric proteins has acted at several levels. Interacting regions often overlap with hydrophobic stretches (Castillo and Ventura, 2009). This implies that monomeric subunits have solvent exposed STAPs that even though masked once the complex is formed, still remain exposed until quaternary structure formation is complete. For several pathological-related proteins, bad prognosis mutations have been identified to negatively impact the complex stability, thus favouring the dissociation of aggregation-prone monomeric units. This is the case of TTR and SOD-1, in which quaternary dissociation becomes a rate limiting step in pathological aggregation (Nordlund and Oliveberg, 2008; Quintas, et al., 2001; Sant'Anna, et al., 2016; Santos, et al., 2020b). For these proteins, evaluating the impact of mutations or redesigns on STAPs accompanied by a structural stability evaluation, such as those performed using the FoldX force field have been found profoundly useful (Gil-Garcia, et al., 2018; Schymkowitz, et al., 2005). FoldX calculates the free energy of unfolding (ΔG) by summing up the stabilisation/destabilisation effect of Van der Waals, Hydrogen bonds, water bridges, molecule solvation and electrostatic

contributions, each multiplied by a weight obtained by fitting empirically data for 339 datapoints for 9 different proteins (Guerois, et al., 2002). All in all, the aforementioned strategies constitute cost-effective tools in understanding mutational impact on aberrant disease-linked aggregation, as well as to optimize protein solubility for biotechnological and pharmaceutical applications.

1.3. INTRINSICALLY DISORDERED PROTEINS

Despite the broadly accepted paradigm stating that a protein needs to acquire a unique and relatively rigid 3D structure to develop a function resulted useful in anticipating function for structural proteins, different kinds of receptors or enzymes, increasing knowledge on the nature of protein coding sequences made scientists reappraise it (Romero, et al., 1998; Wright and Dyson, 1999). As more sequences were being added to Swissprot database, it became clearer that a significant number of them contained long regions predicted to be disordered (Romero, et al., 1998) referred to as intrinsically disordered regions (IDRs) and that this feature would have been counter-selected by evolution in case they would be devoid of any function (Wright and Dyson, 1999). Full-length proteins which lack a defined three-dimensional structure are referred to as intrinsically disordered proteins (IDPs). IDPs encompass a spectrum of unstructured conformations states from fully unstructured to partially structured and include random coils or (pre-)molten globules, and their flexibility is tightly connected to the variety of functions they develop (Tompa, 2002). IDPs can be classified according to their functions: entropic chains, behaving as linkers or spacers, or regarding the nature of their binding: which can be transient such as to display sites for post-translational modifications or chaperones that identify misfolded proteins or RNA, or more prolonged-binding as effectors modulating partner activity, as assemblers or as scavengers that store or hide ligands (van der Lee, et al., 2014). IDPs are sequentially characterized by having fewer APRs, a higher net charge, an enrichment in P and depletion of hydrophobic residues (Monsellier and Chiti, 2007; Tompa, 2002; Walsh, et al., 2012; Xue, et al., 2010). These strategies help to maintain their solubility despite their constant solvent exposure, acting as an evolutionary strategy to minimize aggregation (Monsellier and Chiti, 2007). Several prediction methods have been developed to identify IDRs/IDPs based on their compositional bias, their physicochemical signature, or their absence in three-dimensional protein structures, thus providing a valuable toolbox to study protein disorder (Linding, et al., 2003; Meszaros, et al., 2018; Prilusky, et al., 2005; Uversky, et al., 2000; Walsh, et al., 2012). Moreover, the creation of a manually curated database of experimentally characterized proteins with IDRs and fully unstructured IDPs have helped to study and characterize this large and widespread group of proteins (Hatos, et al., 2020).

1.4. PRIONS AND RELATED PHENOMENA

Prusiner called *prion* the infective proteinaceous particles capable of inducing different mammalian neurodegenerative diseases; known as transmissible spongiform encephalopathies (TSEs) (Prusiner, 1982). The causative agent was determined to be an endogenous cellular protein, prion protein (PrP) able to post-translationally convert from the soluble, native state into an infectious, self-templating

and self-propagating toxic conformation without an evident need for nucleic acids to be transmitted, even between individuals (Kraus, et al., 2013; Prusiner, 1982). These TSEs comprise scrapie in sheep and goats, chronic wasting disease in cervids or bovine spongiform encephalopathy in cattle and bovine spongiform encephalopathy or mad cow disease. This latter is to date the only prion disease proven to be zoonotically transmitted to humans (Davenport, et al., 2015). In *Homo sapiens*, TSEs include Creutzfeldt-Jacob disease, kuru, Gerstmann–Straüssler–Scheinker syndrome and fatal familial insomnia (Chiti and Dobson, 2017; Sikorska and Liberski, 2012).

Wickner reasoned that Ure2 and Sup35 proteins from *Saccharomyces cerevisiae*, baker's yeast, which behaved as non-Mendelian genetic elements were also self-propagating and transmissible protein isoforms, similar to PrP, thus expanding the classification of prions beyond mammals and disease (Wickner, 1994). Since then, the identification of *bona fide* yeast prions has significantly increased, specially thanks to a large-scale analysis by Lindquist and co-workers in which 28 proteins showed self-templating and self-propagating abilities (Alberti, et al., 2009). Yeast prions share some traits which are similar to those found in PrP: i) they self-template and self-propagate the prion conformation, converting the soluble protein into prionic species (Wickner and Kelly, 2016), ii) are inherited in a non-Mendelian way, inducing all the progeny to bear the prion-state (Brown and Lindquist, 2009; Cox, 1965; Uptain and Lindquist, 2002; Wickner, 1994), iii) can spontaneously epimutate between the prion and soluble state with a conversion ratio of $\sim 10^{-6}$ per generation (Brown and Lindquist, 2009; Lancaster, et al., 2010; Tank, et al., 2007), iv) different conformational prion strains render different biological phenotypes (Aguzzi, et al., 2007; Tank, et al., 2007; Wickner and Kelly, 2016), v) the prion state forms insoluble amyloids (Serio and Lindquist, 2001; Tank, et al., 2007) (with the exceptions of $[\beta]$, $[\text{GAR}^+]$ and $[\text{SMAUG}^+]$ yeast prions (Brown and Lindquist, 2009; Chakravarty, et al., 2020; Itakura, et al., 2020; Roberts and Wickner, 2003)), vi) the dependence of chaperones to maintain and propagate the prion state; which are assumed to act by severing prion fibres (Hosoda, et al., 2003; Newby and Lindquist, 2013; Serio and Lindquist, 2001), thus increasing the number of accessible fibril ends and facilitating the transmission of smaller aggregates (Cascarina and Ross, 2014; Halfmann, et al., 2011) and vii) their difficulty to overcome the species barrier (Chen, et al., 2007; Shida, et al., 2020).

Conventionally, yeast prions are written in capital letters to represent its phenotypical dominancy and between brackets to indicate its cytoplasmic inheritance. Accordingly, $[\text{PSI}^+]$, $[\text{NU}^+]$, $[\text{URE3}]$, $[\text{PIN}^+]$, $[\text{SWI}^+]$, $[\text{ISP}^+]$, $[\text{MOT}^+]$, $[\text{OCT}^+]$, $[\text{MOD}^+]$, $[\text{PUB1}]$, $[\text{RNQ}^+]$, $[\text{NUP100}^+]$, $[\text{SMAUG}^+]$ and $[\text{ESI}^+]$ indicate the prion state for Sup35, New1, Ure2, Rnq1, Swi1, Sfp1, Mot3, Cyc8, Mod5, Pub1, Rnq1, Nup100, Vts1 and Snt1 proteins (Chakravarty, et al., 2020; Halfmann, et al., 2012; Halfmann, et al., 2012; Harvey, et al., 2020; Itakura, et al., 2020; Liebman and Chernoff, 2012; Serio and Lindquist, 2001; Wickner, et al., 2015).

It is commonly accepted that yeast prions are beneficial, a bet-hedging mechanism that that would let isogenic colonies thrive under selective environmental conditions (Halfmann, et al., 2012; Harvey, et al., 2020; Newby and Lindquist, 2013; Serio and Lindquist, 2001). For instance Sup35, a ribosomal translation termination factor (Serio and Lindquist, 2001; Ter-Avanesyan, et al., 1993; True and Lindquist, 2000). In its prion aggregate form, $[\text{PSI}^+]$ its functionality is compromised, allowing read through stop-codons (Cox,

1965; Serio and Lindquist, 2001; True and Lindquist, 2000). This reveals a previously hidden genetic load, such as the genes needed to overcome adenine auxotrophy, which becomes beneficial in adenine deficient environments, thus giving a selective advantage to cells bearing the prion variant (Brown and Lindquist, 2009). Yeast cells in glucose-containing media repress alternative carbon sources, even when glucose is present in small amounts (Brown and Lindquist, 2009). [GAR+] prion arises spontaneously in presence of glucosamine, a nonmetabolizable mimetic of glucose, allowing the use of multiple carbon sources, ultimately making the colony to grow in those conditions (Brown and Lindquist, 2009). However, there is not an absolute consensus on the role of yeast prions and several authors argue that the low prevalence of them in wild yeast populations and the prion-infected individuals' slower growth should be regarded as unequivocal signs of their detrimental nature (McGlinchey, et al., 2011; Nakayashiki, et al., 2005).

Sequentially, yeast prions present a low complexity (LC, (i.e., regions that are enriched in a small subset of amino-acid residue types) IDR enriched in Q and N and depleted in hydrophobic and charged residues known as prion domain (PrD) (Ross, et al., 2005; Uptain and Lindquist, 2002). PrDs are necessary and sufficient to carry out prion self-templating and self-propagating activities (Alberti, et al., 2009; Halfmann, et al., 2012; Masison, et al., 1997). Moreover, PrD are of modular nature (Ter-Avanesyan, et al., 1993), they are present in proteins displaying also globular domains (Liu, et al., 2002; Masison, et al., 1997; Ross, et al., 2005), and can be fused to unrelated globular proteins while retaining their prion function (Alberti, et al., 2009; Li and Lindquist, 2000; Ross, et al., 2005; Toombs, et al., 2012). This modular architecture is exploited to test the prion-forming capacity of different natural or synthetic domains; by replacing with them the N-terminal PrD in Sup35 and (by overexpression of this chimeric Sup35 fusion), testing its ability to convert to the prion state (Toombs, et al., 2012). Furthermore, this modularity has allowed the development of artificial functionalized nanomaterials in which a PrD is fused to one or more globular domains, that generate amyloid fibrils composed of the PrD with folded functional domains hanging from them (Knowles and Mezzenga, 2016; Wang and Ventura, 2020). Functionalized PrDs display slower and tuneable aggregation kinetics (compared to functionalized 'classical' amyloids) which generally allow higher preservation of the globular structure, and thus of their enzymatic activities (Wang and Ventura, 2020).

The mechanism by which yeast prions switch conformation towards the amyloid state was at first proposed to be driven by a large number of weak interactions along the PrD, in what is referred to as the compositional model of prion formation (Ross, et al., 2005; Toombs, et al., 2012). Previous studies from our group have suggested that specific soft-amyloidogenic stretches (with milder aggregation potential than classical amyloids, and distributed among more amino acids) inside the IDRs of a PrD could play a crucial role in structural conversion, proposing the *soft-amyloid stretch* model of prion formation (Sabate, et al., 2015). These short amyloid cores, which are present in yeast prions, can form amyloid fibres and promote full protein prion conversion by their own (Sant'Anna, et al., 2016). Recent studies have shown these soft-amyloids are indispensable for prion propagation in mammalian cells (Duernberger, et al.,

2018). The interplay between compositional bias and the presence of short-amyloid stretches will be further explored in the present thesis.

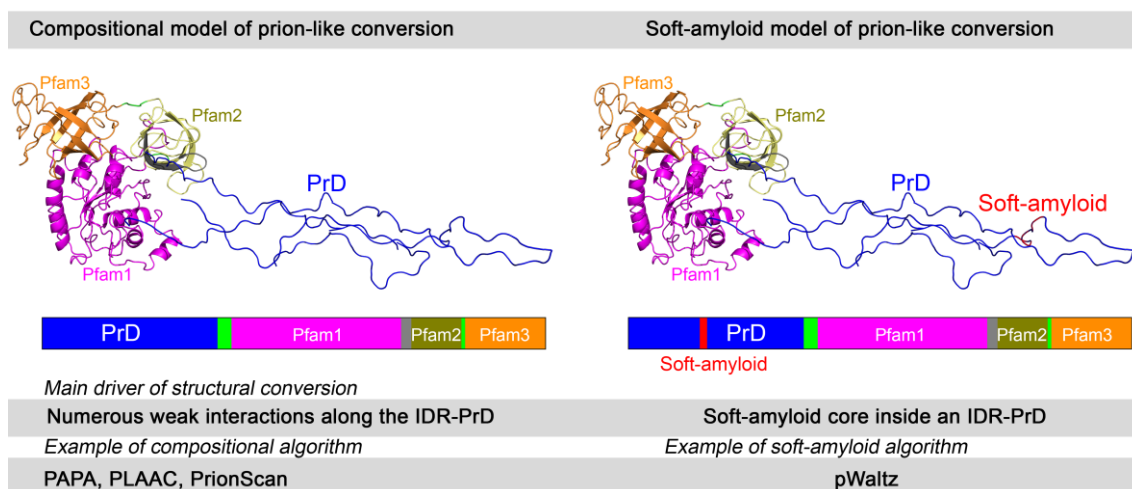


Figure 1.5 – Compositional and soft-amyloid models of prion conversion. Yeast protein Sup35 is above depicted structurally (PDB: 4CRN) and sequentially following the same colour pattern: blue for its PrD, red for the soft-amyloid core and pink, orange and green for the globular Pfam domains.

There exist proteins that display prion-like behaviour, with structural conversion being regulated by physiological signals and where the prion-like conformation displays a novel function, but do not fulfil all the prion conduct, specially the inter-individual transmission (Batlle, et al., 2017c). These are not necessarily sequentially related to yeast prions. For instance, long-term memory consolidation has been found to be dependent on the formation of an amyloid aggregate in CPEB (cytoplasmic polyadenylation element binding protein) family proteins for *Aplysia californica*, sea slug or sea hares, *Drosophyla melanogaster*, fruit fly (CPEB Orb2 protein); and mouse (Fioriti, et al., 2015; Majumdar, et al., 2012; Si, et al., 2003). CPEB can exist as a monomer and a self-sustaining amyloid; the interconversion of which is tightly regulated by the cell (in contrast to pathogenic amyloids) (Si, 2015). Once CPEB is in its stable prion-aggregate form, it could regulate synaptic mRNAs altering the protein composition of the synapse, and thus the neuronal output (Si, 2015). However, the most widespread connotation of the prion-like term refers to proteins with sequential or other signatures similar to those in yeast prions. Their LC domains, analogous to yeast PrD, are referred to as prion-like domains (PrLD) (Si, 2015). Similar to yeast prions, prion-like proteins are largely involved in transcription and translation (by modifying affinity of complexes that bind DNA or DNA compaction, RNA processing...) thus regulating the flow of genetic information in the cell (King, et al., 2012; Malinovska, et al., 2013).

Alongside, the term prionoid has been suggested for proteins involved in misfolded diseases that can self-propagate a misfolded conformation to healthy cells but not (at least spontaneously) between individuals (Batlle, et al., 2017c). Examples of prionoids can be proteins involved in neurodegenerative diseases as A β and tau in AD; α -syn in PD and multiple system atrophy; SOD-1 in ALS and frontotemporal dementia; but also p53 in several cancer types (Batlle, et al., 2017c; Costa, et al., 2016). These, contrary to prion-like proteins, do not share evolutionary, structural, or sequential relationship between them or with yeast prions (Batlle, et al., 2017c).

Unravelling the mechanisms of prion-like proteins in disease and physiology in different organisms require the identification and characterization of novel prions across species. However, prion intrinsic sequential bias makes the aggregation and amyloid prediction methods described in **Section 1.2.2** ineffective at identifying prions and prion-like proteins (Fernandez, et al., 2017; Linding, et al., 2004; Toombs, et al., 2012). Therefore, there is a need to develop approaches that specifically identify prions and prion-like proteins in a fast and accurate way; a topic that will be addressed in the present thesis.

1.4.1. BIOINFORMATIC APPROACHES FOR PRION-LIKE DETECTION

Computational efforts in detecting prion sequences were pioneered by Michelitsch and Weissman under the consideration that if both known prions Sup35 and Ure2 and several neurodegenerative human diseases linked proteins were Q/N rich, screening against this compositional bias could reveal similar proteins in different organisms, therefore increasing the understanding of the aforementioned phenomena (Michelitsch and Weissman, 2000). They defined a sliding window of 80 residues (based on the size of Sup35 PrD) and searched for a minimum content of 30 Q+N per window (based on *Komagataella pastoris*, budding yeast, homologue of Sup35) and named the method Defined Interval Amino acid Numerating Algorithm (DIANA). DIANA would then return the highest Q+N scoring window per sequence. Public complete proteomes from bacteria, archaea and three model eukaryotes (*C. elegans*, *D. melanogaster* and yeast) along with the available sequences from human, mice and *Arabidopsis thaliana* were scanned. Of interest, they found their Q/N-rich regions were essentially absent from thermophilic bacteria and archaea and far more frequent in eukaryotic proteomes, which was attributed to a possible role in protein-protein interaction (PPI) mediated by these stretches. Most notably, this approach allowed the identification of two new yeast prions: New1 and Rnq1.

Toombs and co-workers used a scrambled version of Sup35 that forms prions without overexpression and determined the most important stretches for prion conversion. Next, they performed random mutagenesis in the main 8-amino acid segment and sequenced those variants which could form the [PSI⁺] phenotype. From this dataset, they generated a scoring system of over and under-represented amino acids in prion domains (Toombs, et al., 2010). They found a bias towards hydrophobic residues, against charged and Ps, but most surprisingly their dataset showed no bias towards Qs and Ns despite Q/N are highly overrepresented in yeast PrD. They defined the prion propensity as the log-odds ratio of the frequency of occurrence of each amino acid among the prion-forming clones, relative to the starting library. Finally, they realised they could achieve higher predictive performance by averaging scores using 41 residues scoring windows on top of FoldIndex, a protein disorder predictor (Prilusky, et al., 2005), but not by incorporating classical amyloid prediction methods. This method was implemented in a computational algorithm named Prion Aggregation Prediction Algorithm (PAPA) (Toombs, et al., 2012). PAPA was further used to design synthetic prions performing a computational controlled shuffling generating variants in which the Sup35 Q+N content remained unaltered. Two of the computer-designed sequences which scored positive and three negatives were tested for [PSI⁺] phenotype when substituting the cellular Sup35 PrD, attaining a perfect correlation between predictions and phenotype. PAPA was

further used to identify the GAFA factor, a transcription factor from *Drosophila* that was able to induce [PIN+] phenotype when replacing the Sup35 PrD (Tariq, et al., 2013). It has been successfully applied to identify and delimit PrLD as a first step for the further identification of their soft-amyloid cores; which will be further described in upcoming paragraphs. Recently, Cascarina and Ross developed a modified version of PAPA, essentially lowering its threshold, to explore protein sequence variation at genetic, post-transcriptional, and post-translational levels, with the intention to identify possible prion-like conversions arising from mutations, thus increasing the number of potential human prion-like candidates (Cascarina and Ross, 2020).

Lindquist' lab ventured in a massive effort to identify prions and prion-like determinants in yeast (Alberti, et al., 2009). They generated a hidden Markov Model (HMM), trained on the PrD of known prions Sup35, Rnq1, Ure2 and New1 and used it to rank the whole yeast genome. The top scoring 100 candidates were then tested for their aggregation potential, stability of these aggregates, amyloid formation and prion-potential (by showing [PIN+] phenotype when substituting the Sup35 PrD). Sequentially, the positive candidates showed an underrepresentation of charged residues, Ps and Qs; but enriched in Ns; which was unexpected as both Qs and Ns had been regarded as exerting a similar role in prion formation (Michelitsch and Weissman, 2000). By applying a witty strategy, they could confirm transcription factor Mot3 as a *bona fide* yeast prion. This dataset generated by means of a computational and experimental collaboration allowed the development or testing of third-party bioinformatics approaches (Batlle, et al., 2017c; Espinosa Angarica, et al., 2013; Sabate, et al., 2015; Toombs, et al., 2012). Finally, Lindquist and co-workers updated the amino acid frequencies for the prion state of the HMM from 4 yeast prions to 28 of the candidates which showed higher experimental prion propensity and added the possibility to adjust background frequencies for different species. They deployed the algorithm to a web server and standalone application and named it the prion-like amino acid composition (PLAAC) prediction algorithm (Couthouis, et al., 2011; Lancaster, et al., 2014). PLAAC incorporates PAPA and Foldindex calculations and retrieves them in the program's output. Since published, PLAAC has been widely accepted by the community and used for bioinformatic screenings leading to the discoveries such as the one of *A. thaliana* transcriptional factor Luminidependens that regulates flowering time, the first plant protein able to switch to [PIN+] phenotype when its PrLD replaced that of Sup35 PrD (Chakrabortee, et al., 2016), or most recently in a multi-species screening rendering the first proteins in Archaea able to functionally replace Sup35 PrD with their PrLD (Zajkowski, et al., 2021). As with PAPA, PLAAC has been successfully applied to identify and delimit PrLD for the subsequent identification of their soft-amyloid cores; this strategy will be further explored in upcoming paragraphs.

A collaboration between Sancho and Ventura's labs explored the compositional determinants of yeast PrD. By selecting the 29 PrD which showed amyloid formation and switching behaviour in Lindquist's approach, they calculated their amino acid frequencies and adjusted a threshold with 18 Q/N-rich sequences without prion capacity. As expected, a positive bias towards Q and especially N residues was found, but also towards S and Y; while charged residues, C and W were underrepresented. This algorithm, which was named PrionScan, was made public through a web server and Perl code (Espinosa Angarica, et

al., 2014; Espinosa Angarica, et al., 2013). PrionScan was used to scan the annotated proteins in UniprotKB database rendering notable differences in different taxa. Virus and archaea held less than 10 prions per proteome, while in bacteria, fungi, plants and animals that number ranged from tens to hundreds. Remarkably, the proteomes of *Dictyostelium discoideum*, slime mold and *Plasmodium falciparum*, the most prevalent parasite causing human malaria, showed prion predictions for 20% and 10% of their proteins respectively, which could be due to its high number of proteins carrying LC, N/Q-rich regions. PrionScan has a built-in database with precomputed predictions for all sequences in UniprotKB database. It regularly scans UniprotKB database releases, currently holding > 28.000 predictions.

Ventura and co-workers approached the prion conversion from a different angle. They realised the PAPA-derived scoring system was highly dependent on a short, 8-residue stretch, in which hydrophobic residues were enriched while charged amino acids and P were penalised. These observations are in accordance with that of classical amyloids and, by evaluating the PAPA predicted prion-promoting sequences with the amyloid predictor Waltz, they found a trend to be slightly amyloidogenic. These hints made them propose the soft-amyloid stretch hypothesis, reasoning that a larger stretch with lower and more spread amyloid propensity than in classical amyloids, when embedded in a Q/N-rich IDR could play a role in prion conversion. They fixed a window length of 21 residues which corresponded to the minimum transmissible β -fold as seen in HET-s prion from the fungus *Podospora anserina*, accommodated the Waltz position-specific amyloid matrix to this length and named the approach pWALTZ (Sabate, et al., 2015). Thus, they proposed that it was not only the composition of a PrD which would determine its prion potential but also its capacity to physically accommodate a cryptic amyloid sequence. pWALTZ was tested on top of Lindquist' dataset showing higher discrimination potency than PAPA. Moreover, it was able to identify soft-amyloid cores in the PAPA-shuffled prion-forming sequences and in disease-related mutations of human prion-like proteins. pWALTZ is not designed for detecting IDR or prion or prion-like domains. To avoid false positive soft-amyloid cores predictions it is advised to input the sequences known to correspond to IDRs and suspected to behave as PrLD. Since its development, pWALTZ has been extensively used, having a great success in complementing PrLD predicting software such as PAPA or PLAAC, in the localization of soft-amyloid cores in yeast prions, the identification of human prion-like protein amyloid cores and in the discovery of Rho transcription terminator factor from *Clostridium botulinum*, the first prion identified in bacteria (Batlle, et al., 2017a; Pallares, et al., 2015; Pallares and Ventura, 2017; Sant'Anna, et al., 2016; Yuan and Hochschild, 2017).

All in all, the development of fast and more accurate *in silico* tools coupled to the exponential growth in protein sequences is expected to allow better understanding of the physiological purpose of these proteins. Regardless of the different assumptions behind these predictions, judging by their successes, it is likely that to a certain extent both composition and specific sequences would play an active role in prion self-assembly. Hence a complementary approach would probably ensure a higher success rate as candidates would have to satisfy both compositional and amyloidogenic requirements, similar to those found on *bona fide* yeast prions. It is worth to mention that all the aforementioned prion-like prediction methods are based on the compositional and sequential features of a relatively limited number of yeast

prions. This entails a possible bias towards similar proteins, possibly leaving aside prions that deviate from these premises. For instance [β], [GAR+] and [SMAUG+] yeast prions do not form amyloid aggregates (Brown and Lindquist, 2009; Chakravarty, et al., 2020; Itakura, et al., 2020; Roberts and Wickner, 2003) and mammalian PrP or fungal HET-s are not Q/N-enriched (Balguerie, et al., 2003; Batlle, et al., 2017c; Shorter and Lindquist, 2005). Only by identifying new prions in non-related organisms will we be able to ascribe sequential and compositional requirements for prion-like mechanisms across species; and these innovations will likely require important program adjustments.

2. Objectives

The works we describe here have the common objective of increasing our understanding of the determinants behind the process of protein aggregation. To pursue this aim, experimental data will be analysed, rationalised and the underlying processes computationally modelled. When pertinent, the gained knowledge will be implemented into user-friendly algorithms, freely accessible to the scientific community. The specific objectives of the present thesis can be summarised in the following points:

- Study the determinants behind globular protein aggregation in near-native environments, evaluating improvements to current algorithms, including the benefits of considering fluctuations and stability when dealing with protein quaternary structure predictions.
- Identify the effect of protein environment, namely the solution pH, in modulating protein structural transitions and aggregation.
- Study the determinants underlying yeast prion conversion. Apply this knowledge to improve the prediction of proteins able to experiment prion-like structural conversions.
- Identify prion-like proteins across different species. Explore if this kind of structural conversion might be an evolutionary conserved mechanism using state-of-the-art functional characterization resources.

3. Chapter I – Globular Protein Aggregation

3.1 Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility

Valentin Iglesias^{1†}, Aleksander Kuriata^{2†}, Jordi Pujols¹, Mateusz Kurcinski², Sebastian Kmiecik^{2*} and Salvador Ventura^{1*}

¹ Insitut de Biotecnologia i Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra (Barcelona) 08193, Spain.

² Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw, 02-089 Warsaw, Poland.

†These authors contributed equally to this work. *To whom correspondence should be addressed.

Author Contributions: Software, validation, data curation, writing—original draft preparation.

3.1.1 ABSTRACT

Protein aggregation is a hallmark of a growing number of human disorders and constitutes a major bottleneck in the manufacturing of therapeutic proteins. Therefore, there is a strong need of computational methods that can anticipate the aggregative properties of protein variants linked to disease and assist the engineering of soluble protein-based drugs. A few years ago, our groups developed a method for structure-based prediction of aggregation properties that considers the dynamic fluctuations of proteins. The method has been made available as the Aggrescan3D (A3D) web server and applied in numerous studies of protein structure-aggregation relationship. Here, we present a major update of the A3D web server to the version 2.0. The new features include: extension of dynamic calculations to significantly larger and multimeric proteins, simultaneous prediction of changes in protein solubility and stability upon mutation, rapid screening for functional protein variants with improved solubility, a REST-ful service to incorporate A3D calculations in automatic pipelines, and a newly designed, enhanced web server interface.

Availability and Implementation: : A3D 2.0 does not require previous registration and is freely available at: <http://biocomp.chem.uw.edu.pl/A3D2/>.

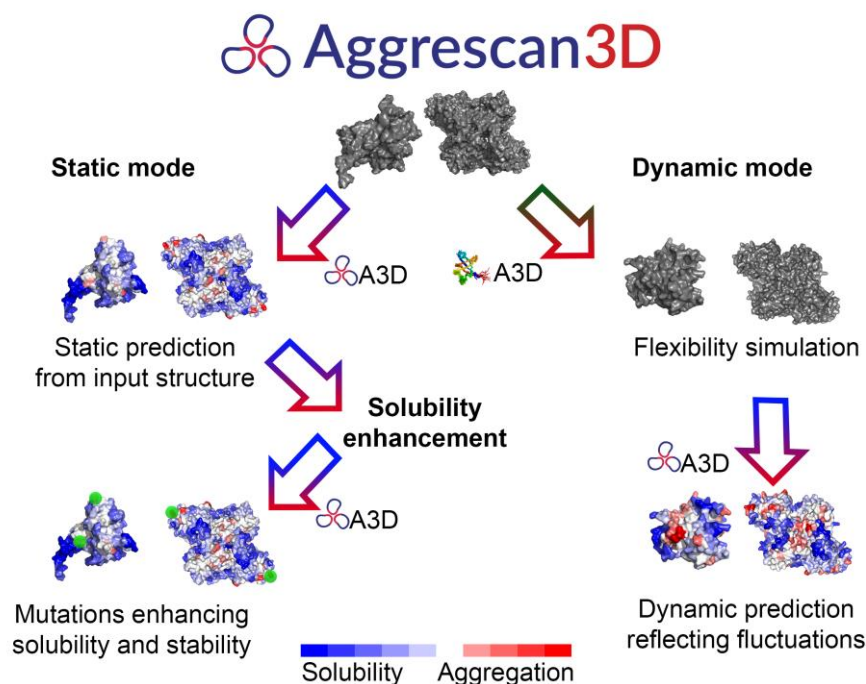


Figure 3.1 – Graphical abstract: Aggrescan3D (A3D) 2.0 entails a major update to A3D web server. It includes: extended dynamic calculations, prediction of changes in stability upon mutation or automatic screening for improved solubility protein variants.

3.1.2 INTRODUCTION

Protein aggregation lies behind more than 40 human diseases, ranging from neurodegenerative disorders to some types of cancers or diabetes type II (Chiti and Dobson, 2017; de Oliveira, et al., 2020; Invernizzi, et al., 2012). In addition, aggregation is a major limitation in the production, storage and administration of life-saving protein pharmaceuticals, like antibodies and replacement enzymes, since it both reduces the percentage of therapeutically active molecules and increases immunogenic responses (Hamrang, et al., 2013).

The growing concern about protein aggregation has fuelled the development of over twenty predictive algorithms (Meric, et al., 2017; Pallares and Ventura, 2017; Santos, et al., 2020a). A majority of methods identify and score protein aggregation prone regions (APRs) relying only on protein sequence. Those programs find difficulties predicting APRs of folded globular proteins, failing to detect APRs when residues are not contiguous in sequence or mistaking APRs for the buried hydrophobic core. These problems motivated the development of a second generation of algorithms that use structure-based approaches for their predictions (Graña-Montes, et al., 2017; Santos, et al., 2020a). In 2015, our group in collaboration with S. Kmieciks' lab, developed the Aggrescan3D (A3D) web server for prediction of aggregation properties of protein structures (Zambrano, et al., 2015). The A3D method was shown to outperform sequence- and composition-based algorithms when dealing with proteins in their native-like states (Pujols, et al., 2018; Zambrano, et al., 2015).

A3D works by integrating the 3D information of protein structures and evaluating the contribution of solvent-exposed APRs. The method projects experimental aggregation propensities onto a protein structure. Aggregation propensity is calculated for spherical regions centred on every residue α -carbon using the intrinsic amino acid aggregation scale from the AGGRESCAN method (Conchillo-Sole, et al., 2007; de Groot, et al., 2012), the first sequence-based algorithm to exploit empirical *in vivo* data. This provides a structurally corrected aggregation value (A3D score) for each particular amino acid, depending on its specific conformational context, discarding the negligible contribution of hydrophobic residues buried in the core of folded proteins and focusing on protein surfaces. The dynamic structural fluctuations of proteins in solution influences the degree of exposure of APRs and for this reason, A3D incorporates the CABS-flex approach, an efficient alternative to classical all-atom molecular dynamics (Jamroz, et al., 2013; Kuriata, et al., 2018) for fast simulations of protein flexibility in its dynamic mode. Moreover, A3D allows the introduction of user-defined mutations to rationally design more soluble protein variants or to test the impact of disease-linked mutations on the aggregation propensity.

Among other applications, A3D has been exploited to understand the binding of chaperones to their targets (Pulido, et al., 2016), to study the binding of antimicrobial proteins to membranes (Pulido, et al., 2016), to rationalize the yield of engineered nanobodies (Soler, et al., 2016), to study the aggregation properties of pathogenic (Bhandare and Ramaswamy, 2018; Zerovnik, 2017) and non-pathogenic (Katina, et al., 2017) globular proteins or to assist the design of biotechnologically relevant proteins (Gil-Garcia, et al., 2018; Xia, et al., 2016).

In this work, we present a major update of the original A3D, which significantly extends its capabilities. A3D 2.0 incorporates three major feature upgrades.

- protein flexibility simulations using new CABS-flex standalone package (Kurcinski, et al., 2018), which extends the dynamic mode analysis range to proteins up to 4,000 residues long and consisting of up to 10 chains.
- protein stability calculations using the FoldX force field (Schymkowitz, et al., 2005), allowing to account for the impact of amino acid substitutions on the overall structure stability.
- an “automated mutations” tool that identifies high scoring residues in structural APRs and suggests protein variants with optimized solubility.

These features were implemented to address the major A3D drawbacks according to users’ feedback. (i) protein size limitations in the dynamic mode, which were restricted only to single-chain proteins shorter than 400 amino acids; (ii) user-introduced mutations could negatively impact protein stability, resulting in unfolding and increased aggregation; (iii) the design of improved solubility variants required significant knowledge about the structural and aggregational determinants of proteins and, thus, was not accessible to many potential users.

Additionally, A3D 2.0 incorporates an updated REST-full service that allows the user to incorporate its calculations in automatic pipelines and a newly designed interface that facilitates extended *in situ* interactive result analysis and data interpretation.

3.1.3 METHODS

A3D prediction protocol

The original A3D server was described in detail previously (Pujols, et al., 2018; Zambrano, et al., 2015). A3D server can be run in Static Mode (default) or Dynamic Mode. The static mode was validated by predicting the solubility of a large set of protein mutational variants, whereas the dynamic mode allowed to display disease relevant APRs not identified by alternative approaches (Pujols, et al., 2018; Zambrano, et al., 2015). The present update retains the main principles of the original web server and here we only detail major methodological modifications. The overview of the method pipeline is presented in **Figure 3.2**.

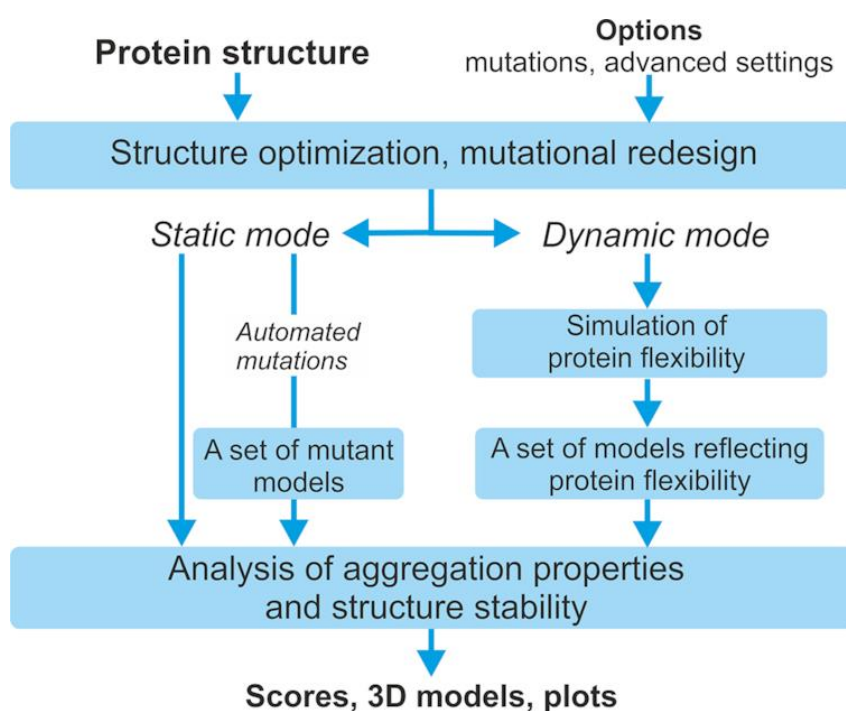


Figure 3.2 – The pipeline of A3D 2.0 server.

Calculation of the impact of introduced mutations on protein thermodynamic stability

Users can introduce individual or multiple mutations before or after running A3D 2.0. The selected mutations are modelled on top of the protein structure using FoldX (Schymkowitz, et al., 2005) and the predicted change in stability, relative to the reference molecule, is calculated. Positive and negative values indicate decreased and increased stabilities upon mutation, respectively.

Automated mutation workflow

The input structure is optimized using FoldX (Schymkowitz, et al., 2005) and the most aggregation-prone residues identified according to their A3D score. These residues are individually mutated to solubilizing charged amino acids (arginine, aspartic acid, glutamic acid and lysine), excluding those positions specified by the user. The changes in aggregation propensity and stability are calculated for each potential point mutant and short-listed according to these values, up to a maximum of 12 suggested changes. Only the two most solubilizing mutations for each particular position are shown, in order to maximize the number of positions that can be potentially engineered (up to 6).

3.1.4 NEW FEATURES AND UPDATES

Analysis of the impact of protein flexibility in the aggregation properties of large and multimeric proteins.

In its dynamic mode, A3D was able to capture the influence of structural flexibility on protein aggregation by incorporating the CABS-flex protocol, an efficient alternative to classical all-atom molecular dynamics (Jamroz, et al., 2013; Jamroz, et al., 2014; Jamroz, et al., 2013; Kurcinski, et al., 2018; Kuriata, et al., 2018). A set of protein models (in an all-atom resolution) reflecting the most dominant structural fluctuations in the near-native ensemble are generated with CABS-flex for each input structure. Then, the highest A3D scoring model is selected as a proxy of the most aggregation-prone conformer in solution. Although this feature uncovered structural APRs not accessible to other structure-based predictors (Zambrano, et al., 2015), its use was restricted to relatively small, single chain proteins, which impeded the analysis of many biomedical and biotechnologically important proteins. With A3D 2.0 we extended the dynamic mode to larger and multimeric proteins by dedicating significantly larger computational resources to web server jobs and rewriting the CABS-flex code (Kurcinski, et al., 2018).

We used A3D 2.0 to analyse the influence of protein dynamics on the aggregation properties of multimeric proteins, using a data set of 163 proteins (69 homodimers, 54 heterodimers and 60 antibodies). In the dynamic mode, A3D 2.0 rendered 12 models for each input structure and calculated their individual A3D scores. These values were then compared with the ones obtained for the same proteins ran in static mode. We found the input static structures to be the least aggregation-prone in a large majority of cases, both for the complete set and when the three protein categories were analysed separately (**Figure 3.3A**). We averaged the A3D scores of the 12 models for each individual protein as a proxy for the aggregation propensity of its native-like ensemble. The resulting average value was higher than that of the static structure in 80 % of the cases. These observations have important implications, since most alternative structure-based aggregation predictors work directly on PDB structures and, therefore, they might underscore the aggregation of multimeric proteins by ignoring the contribution of transiently exposed APRs. The effect is illustrated in **Figure 3.3B** for bevacizumab, a humanized monoclonal antibody prescribed for the treatment of different types of cancers (Gridelli, et al., 2018). The Fab domain of bevacizumab is very aggregation-prone, and, accordingly, the antibody must be formulated at low concentrations (Courtois, et al., 2016; Oliva, et al., 2014). The comparative static and dynamic analysis of bevacizumab Fab fragment (2 chains, 4 domains) suggests that structural fluctuations result in an increased aggregation-prone area, with newly exposed APRs ready to establish intermolecular interactions. The same effect was observed for other therapeutic antibodies, replacement enzymes such as α -galactosidase or important pharmaceutical targets such as insulin and androgen receptors.

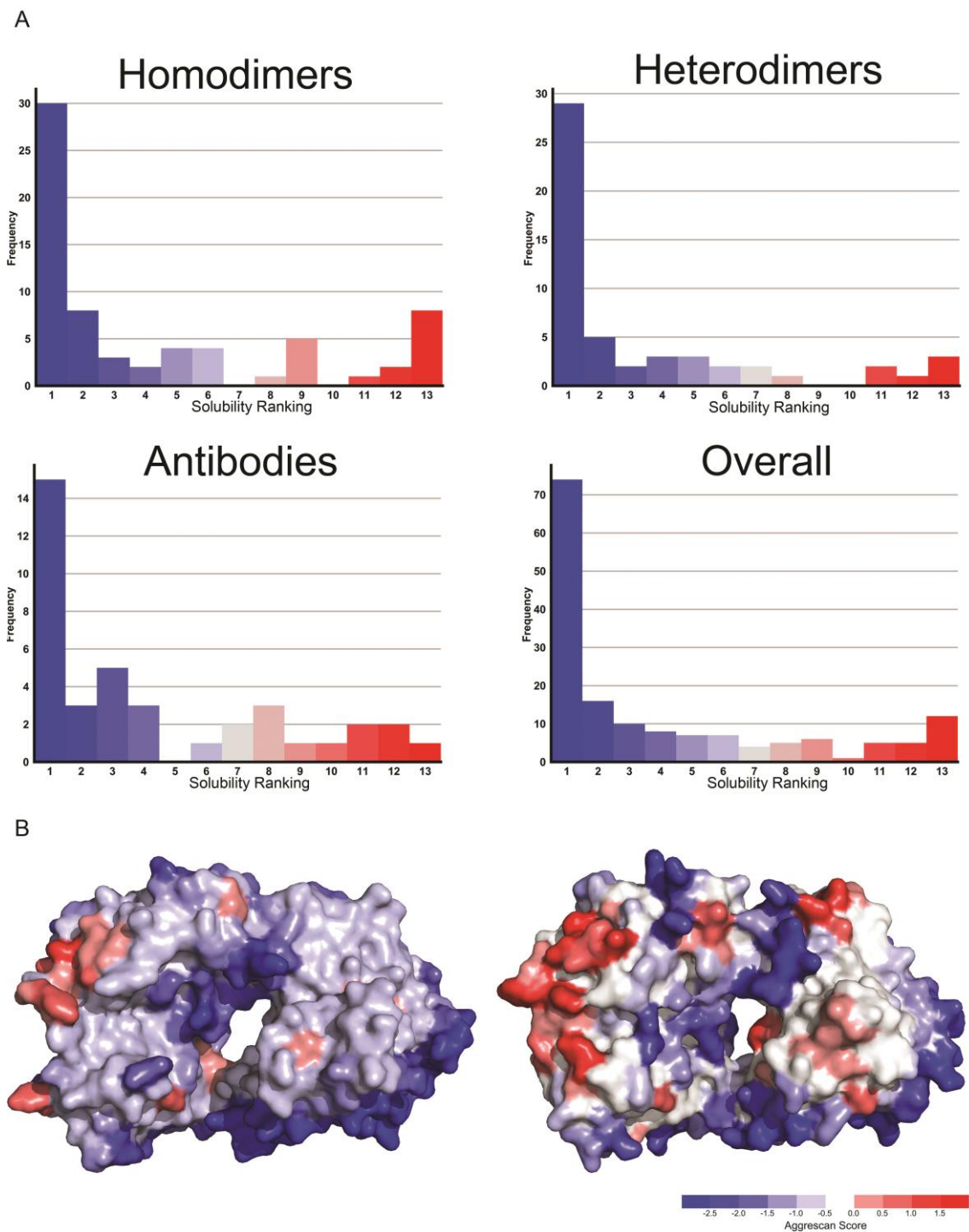


Figure 3.3 – Aggregation propensity for different multimeric proteins, calculated in static or dynamic modes. A) The aggregation propensity of the static input structure relative to that of the 12 dynamic models is represented for homodimers, heterodimers, antibodies, or the complete set. In the colour scale, dark blue indicates the static structure being the most soluble (ranking 1) and dark red the static structure being the most aggregation-prone (ranking 13). **B)** Monoclonal antibody bevacizumab Fab fragment (PDB: 1BJ1) ran on static (left) or dynamic (right) modes.

Simultaneous analysis of the impact of user-selected mutations in protein solubility and stability

The A3D server allowed users to mutate one or more selected residues in the structure, pre- or post-analysis, in order to evaluate the impact in protein aggregation. However, these mutations might also affect the protein thermodynamic stability, an effect that was not contemplated at that time. Indeed, previous work from our group has shown that there is a strict correlation between the destabilizing impact of a given mutation and the increase it promotes in protein aggregation (Castillo, et al., 2010; Espargaro, et al., 2008). Thus, the solubilizing gain of a residue substitution can be completely cancelled if it negatively impacts the protein stability.

Mutations at the protein surface are generally better tolerated than residue changes in the protein interior (Franzosa and Xia, 2009). However, when we used A3D to identify the top solubilizing point mutations for a set of 75 globular proteins, it turned out that 10 % of these superficial changes (32/324) destabilized the protein over 1 kcal/mol according to FoldX. This motivated us to introduce a simultaneous prediction of protein solubility and stability changes upon mutation in A3D 2.0; to identify mutations that decrease globular proteins aggregation propensities without compromising their stability and function. This approach was exploited to design of a fast-folding, aggregation-resistant GFP variant (Gil-Garcia, et al., 2018) (**Figure 3.4**). The analysis of the original GFP structure with A3D 2.0, indicated the existence of three exposed hydrophobic residues at the protein surface, whose mutation to either K or D would be equally solubilizing. However, the energetic analysis indicated that mutations to K would be neutral, whereas mutations to D would destabilize the protein. Two GFP variants in which the three hydrophobic residues were changed either to K or D were recombinantly expressed. As predicted, the triple K GFP mutant (GFP/KKK) was highly soluble, preserved the native structure and was fully functional, whereas the triple D variant (GFP/DDD) was inactive and could not be purified. Importantly, the behaviour of GFP/KKK and GFP/DDD designs, could not be anticipated by any other alternative sequence- or structure-based algorithm. A3D 2.0 advises now against the experimental characterization of destabilized re-designs ($\Delta\Delta G > 1$ kcal/mol), irrespective of their A3D scores. $\Delta\Delta G$ values are provided in the “Project details” tab.

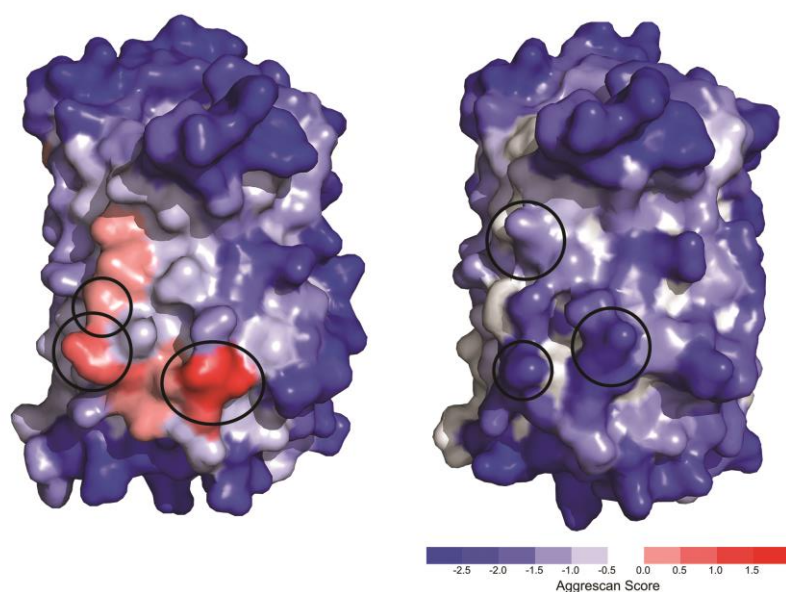


Figure 3.4 – A3D 2.0 as a tool for the *in silico* redesign of more stable and soluble proteins. The original GFP (left) (PDB: 2B3Q:A) and engineered GFP/KKK mutant (right) (PDB: 6FWW) coloured according to their A3D score. Mutations which lowered aggregation propensity, while maintaining protein stability are encircled. The mutated variant was experimentally shown to be 2-fold more resistant against aggregation (Gil-Garcia, et al., 2018).

Automated design of solubility improved protein variants

The search for soluble functional variants of therapeutic proteins is a challenging task, usually addressed using combinatorial experimental approaches, such as phage display (Sidhu, 2000). A goal for any aggregation prediction algorithm is to provide a routine that can substitute these experiments, saving time and costs. Ideally, this routine should be simple enough to be accessible to non-expert users. With these two objectives in mind, we implemented the “automated mutations” tool in A3D 2.0, accessible at the server front page through the “Enhance protein solubility” option.

The “automated mutations” tool identifies the most aggregation-prone patches at the protein surface and virtually mutates their residues by charged amino acids, under the assumption that they will act as “gatekeepers”, counteracting protein self-association. Then it provides a ranked list of point mutations, where both the solubilizing and energetic effects are taken into account, in such a way that the user can discard potentially solubilizing, but destabilizing mutations.

The optimization of the solubility of antibodies is especially challenging, because, in these molecules, the tight binding to their targets depends on the presence of exposed APRs at their complementarity-determining regions (CDRs). This is the reason why computer-(Sormanni, et al., 2015) or experiment-(Perchiacca, et al., 2014) based designs usually target residues within or close to the CDRs; however, these changes might significantly compromise the antibody affinity. A3D 2.0 addresses this problem by allowing users to exclude from the virtual screening functionally relevant residues, i.e. CDRs in antibodies or active sites in enzymes.

The “automated mutation” tool has been used for the redesign of an aggregation-prone Variable Heavy (VH) segment of the human antibody germline (Teplyakov, et al., 2016). Soluble variants of this antibody were previously evolved by phage display, but all the introduced mutations clustered at one of the CDRs (Dudgeon, et al., 2012). A3D 2.0 was ran pre-excluding residues at the CDRs. Mutations at three different residues outside these domains were automatically suggested (**Figure 3.5A**). A designed VH variant containing the 3 top ranked mutations was recombinantly expressed and characterized (**Figure 3.5B**), turning to be significantly more resistant against aggregation than the original germline antibody (Gil-Garcia, et al., 2018).

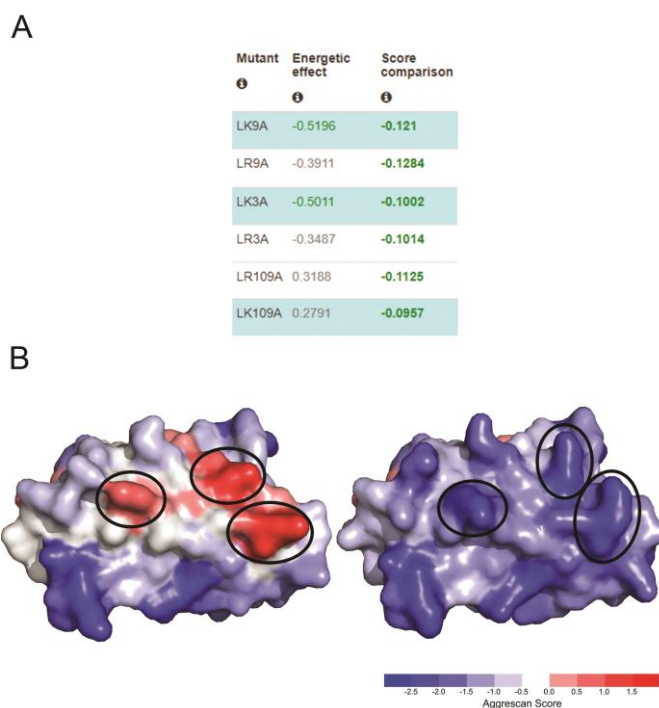


Figure 3.5 – Automated mutations for variable heavy (VH) segment of a human germline antibody. A) A3D 2.0 automated mutations output when the residues at the three antibody CDRs were excluded from the screening. **B)** The blue highlighted mutations in panel A were combined to render triple mutant engineered antibody. Structures of wild type (PDB: 5I19) and the mutant, as predicted by A3D 2.0. Solubilizing mutations are encircled. The engineered antibody variant was experimentally shown to be 3-fold more resistant against aggregation (Gil-Garcia, et al., 2018).

3.1.5 DESCRIPTION OF THE WEB SERVER

Input interface and requirements

The only required input is a protein structure in PDB format (**Figure 3.6**). Users can submit as a PDB code or upload a local structure in the ‘Input structure’ panel. Optionally, users can select desired chain(s) identifier(s) (only provided chains will be used in the analysis). In the ‘Options’ panel, several additional options can be chosen:

- Project name - the name under which the project will be displayed (and which can be used to find it via the project name search on the top of the page)
- Email address - the server will notify the user when the job has started and ended on the provided an email address
- Stability calculations - if selected ('Yes' by default), the submitted structure will be energetically minimized before the A3D analysis using FoldX and stability calculated in case mutations are defined.
- Dynamic mode - in this mode, the input structure's flexibility will be simulated using the CABS-flex software. A set of predicted models reflecting the flexibility of the input structure will be analysed and scored for aggregation propensity. Note: this option is incompatible with the "Enhance protein solubility" option.
- Mutate residues - Selecting this option will prompt a new window, which allows introducing the desired mutation(s), which will be carried out using FoldX. Note: this option cannot be used with the 'Enhance protein solubility' option.
- Distance of aggregation analysis - in the A3D method, the intrinsic aggregation propensity of each particular amino acid in the structure is modulated by its specific structural context. Aggregation propensity is calculated for spherical regions centred on every residue α -carbon. This option allows changing the size of said region, allowing for more and less granular approaches.
- Enhance protein solubility – Please see Methods for details on the automated mutations workflow behind this option. Selecting it will prompt a new window to open upon submitting, where the user can prevent chosen residues from being mutated. Note: this option is incompatible with 'Mutate residues' or 'Dynamic mode' options.
- Do not show my job in the results page - if the box is ticked the job will not be visible to other A3D 2.0 users.

Figure 3.6 – A3D 2.0 redesigned main page. On the upper right box the user is prompted to input the PDB formatted protein structure. Immediately under it, A3D 2.0 allows different options for a which will influence the final prediction. A3D 2.0 allows users to retrieve their previous jobs by searching on the uppermost search text box the specified project name or jobid. Alternatively, these can be retrieved by a manual search under the “Queue” link on the top-left side.

Output interface

For each submitted job, the output interface is organized under the following tabs: ‘Project details’, ‘Aggrescan3D plot’, ‘Aggrescan3D score’, ‘Structure’, ‘Automated mutations’ (available if the job was submitted with the option ‘Enhance protein solubility’), ‘Dynamic mode details’ (available if the job was submitted in the ‘Dynamic mode’) and ‘Gallery’. The content of these tabs is presented and described in the online documentation. Here, we present only short descriptions:

- ‘Project details’ tab – contains information about the specified options used to run the job and links to download the job data. It also provides stability calculations, when it applies.
- ‘Aggrescan3D plot’ tab – presents A3D analysis results in the form of an interactive online plot for a selected protein chain.
- ‘Aggrescan3D score’ tab – presents A3D analysis results in the form of an interactive table together with "mutate" buttons in the right side of the table, which will resubmit the job with the chosen mutations.
- ‘Structure’ tab – allows viewing an analysed structure in an interactive way. The residues are coloured in shades from dark blue (high soluble residues), through white (no predicted influence on

aggregation properties), to dark red (aggregation prone residues). A set of visualization options such as tagging specific residues, rotating the molecule or showing it as video are available.

- ‘Dynamic mode details’ tab – presents A3D analysis results for a set of models reflecting flexibility of the input structure. The results are organized in the table and interactive plots presenting scores for particular models.
- ‘Automated mutations’ tab – presents A3D analysis results for a set of mutant models generated using option ‘Enhance protein solubility’. The results are organized in the table and in interactive plots which represent the scores for each particular protein variant.
- ‘Gallery’ tab - contains all screenshots taken by users (using ‘Take snapshot’ buttons) in previous tabs.

Online documentation

The server provides useful documentation which can be found under the “Tutorial” tab (available from the main menu). Additionally, the web interface provides short help notes that are available close to the presented content. The online documentation is updated on a regular basis according to users’ needs or the server improvement.

Command-line availability

The A3D 2.0 server can be also operated from the command line using RESTful web services. The instructions for using the RESTful service are available from the online tutorial (accessible from the main menu).

Server architecture

The Aggrescan3D 2.0 server is a HTML based service dynamically generated using the Flask framework and the jinja2 templating engine. The user data is stored using a MySQL database upon submission and unique id and a status are assigned for each job. The server notifies the user of its progress by a job status, which is ‘pending’ when the server is waiting for a computational cluster response, ‘in queue’ when there are no resources available yet, ‘running’ and then finally either ‘done’ or ‘error’. The simulation is carried out using the Aggrescan3D standalone software (Kuriata, et al., 2019) (that is available at: <http://bitbucket.org/lcbio/aggrescan3d>) and other previously described programs (with the RSA calculations done by FreeSASA software (Meszaros, et al., 2019)). The structures are presented in an interactive way using the 3Dmol library (HTML5/Javascript). The A3D score is plotted using the D3.js library (HTML5/Javascript) and the model and mutant comparison plots are generated using the Bokeh library (Python/Javascript). The PDB structures are obtained using RESTful services. The A3D 2.0 website handles user’s requests using an Apache2 server. The A3D 2.0 server is free, open to all users and there is no login requirement.

3.1.6 REFERENCES

- Bhandare, V.V. and Ramaswamy, A. (2018) The proteinopathy of D169G and K263E mutants at the RNA Recognition Motif (RRM) domain of tar DNA-binding protein (tdp43) causing neurological disorders: A computational study, *Journal of biomolecular structure & dynamics*, **36**, 1075-1093.
- Castillo, V., *et al.* (2010) Deciphering the role of the thermodynamic and kinetic stabilities of SH3 domains on their aggregation inside bacteria, *Proteomics*, **10**, 4172-4185.
- Conchillo-Sole, O., *et al.* (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides, *BMC bioinformatics*, **8**, 65.
- Courtois, F., *et al.* (2016) Rational design of therapeutic mAbs against aggregation through protein engineering and incorporation of glycosylation motifs applied to bevacizumab, *mAbs*, **8**, 99-112.
- Chiti, F. and Dobson, C.M. (2017) Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade, *Annu Rev Biochem*, **86**, 27-68.
- de Groot, N.S., *et al.* (2012) AGGRESCAN: method, application, and perspectives for drug design, *Methods in molecular biology*, **819**, 199-220.
- Dudgeon, K., *et al.* (2012) General strategy for the generation of human antibody variable domains with increased aggregation resistance, *Proc Natl Acad Sci U S A*, **109**, 10879-10884.
- Espargaro, A., *et al.* (2008) The in vivo and in vitro aggregation properties of globular proteins correlate with their conformational stability: the SH3 case, *Journal of molecular biology*, **378**, 1116-1131.
- Franzosa, E.A. and Xia, Y. (2009) Structural determinants of protein evolution are context-sensitive at the residue level, *Molecular biology and evolution*, **26**, 2387-2395.
- Gil-Garcia, M., *et al.* (2018) Combining structural aggregation propensity and stability predictions to re-design protein solubility, *Molecular pharmaceutics*.
- Gridelli, C., *et al.* (2018) Safety and Efficacy of Bevacizumab Plus Standard-of-Care Treatment Beyond Disease Progression in Patients With Advanced Non-Small Cell Lung Cancer: The AvaALL Randomized Clinical Trial, *JAMA oncology*, **4**, e183486.
- Hamrang, Z., Rattray, N.J. and Pluen, A. (2013) Proteins behaving badly: emerging technologies in profiling biopharmaceutical aggregation, *Trends in biotechnology*, **31**, 448-458.
- Invernizzi, G., *et al.* (2012) Protein aggregation: mechanisms and functional consequences, *The international journal of biochemistry & cell biology*, **44**, 1541-1554.
- Jamroz, M., Kolinski, A. and Kmiecik, S. (2013) CABS-flex: Server for fast simulation of protein structure fluctuations, *Nucleic Acids Res*, **41**, W427-431.
- Jamroz, M., Kolinski, A. and Kmiecik, S. (2014) CABS-flex predictions of protein flexibility compared with NMR ensembles, *Bioinformatics*, **30**, 2150-2154.
- Jamroz, M., *et al.* (2013) Consistent View of Protein Fluctuations from All-Atom Molecular Dynamics and Coarse-Grained Dynamics with Knowledge-Based Force-Field, *Journal of chemical theory and computation*, **9**, 119-125.
- Katina, N.S., *et al.* (2017) sw ApoMb Amyloid Aggregation under Nondenaturing Conditions: The Role of Native Structure Stability, *Biophysical journal*, **113**, 991-1001.
- Kurcinski, M., *et al.* (2018) CABS-flex standalone: a simulation environment for fast modeling of protein flexibility, *Bioinformatics*.
- Kuriata, A., *et al.* (2018) CABS-flex 2.0: a web server for fast simulations of flexibility of protein structures, *Nucleic Acids Res*, **46**, W338-W343.
- Meric, G., Robinson, A.S. and Roberts, C.J. (2017) Driving Forces for Nonnative Protein Aggregation and Approaches to Predict Aggregation-Prone Regions, *Annual review of chemical and biomolecular engineering*, **8**, 139-159.
- Meszaros, B., *et al.* (2019) PhaSePro: the database of proteins driving liquid-liquid phase separation, *Nucleic Acids Res*.
- Oliva, A., Llabres, M. and Farina, J.B. (2014) Capability measurement of size-exclusion chromatography with a light-scattering detection method in a stability study of bevacizumab using the process capability indices, *Journal of chromatography. A*, **1353**, 89-98.
- Pallares, I. and Ventura, S. (2017) Advances in the prediction of protein aggregation propensity, *Current medicinal chemistry*.
- Perchiacca, J.M., Lee, C.C. and Tessier, P.M. (2014) Optimal charged mutations in the complementarity-determining regions that prevent domain antibody aggregation are dependent on the antibody scaffold, *Protein engineering, design & selection : PEDS*, **27**, 29-39.
- Pujols, J., Pena-Diaz, S. and Ventura, S. (2018) AGGRESCAN3D: Toward the Prediction of the Aggregation Propensities of Protein Structures, *Methods in molecular biology*, **1762**, 427-443.
- Pulido, D., *et al.* (2016) Insights into the Antimicrobial Mechanism of Action of Human RNase6: Structural Determinants for Bacterial Cell Agglutination and Membrane Permeation, *International journal of molecular sciences*, **17**, 552.
- Pulido, P., *et al.* (2016) Specific Hsp100 Chaperones Determine the Fate of the First Enzyme of the Plastidial Isoprenoid Pathway for Either Refolding or Degradation by the Stromal Clp Protease in Arabidopsis, *PLoS genetics*, **12**, e1005824.
- Ricardo Graña-Montes, J.P.-P., Carlota Gómez-Picanyol and Ventura, a.S. (2017) Prediction of Protein Aggregation and Amyloid Formation. In Rigden, D.J. (ed), *From Protein Structure to Function with Bioinformatics*. Springer, pp. 205-263.
- Schymkowitz, J., *et al.* (2005) The FoldX web server: an online force field, *Nucleic Acids Res*, **33**, W382-388.

Sidhu, S.S. (2000) Phage display in pharmaceutical biotechnology, *Current opinion in biotechnology*, **11**, 610-616.

Soler, M.A., de Marco, A. and Fortuna, S. (2016) Molecular dynamics simulations and docking enable to explore the biophysical factors controlling the yields of engineered nanobodies, *Sci Rep*, **6**, 34869.

Sormanni, P., Aprile, F.A. and Vendruscolo, M. (2015) The CamSol method of rational design of protein mutants with enhanced solubility, *Journal of molecular biology*, **427**, 478-490.

Tepliyakov, A., *et al.* (2016) Structural diversity in a human antibody germline library, *mAbs*, **8**, 1045-1063.

Xia, X., *et al.* (2016) Engineering a Cysteine-Free Form of Human Fibroblast Growth Factor-1 for "Second Generation" Therapeutic Application, *Journal of pharmaceutical sciences*, **105**, 1444-1453.

Zambrano, R., *et al.* (2015) AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures, *Nucleic Acids Res*, **43**, W306-313.

Zerovnik, E. (2017) Putative alternative functions of human stefin B (cystatin B): binding to amyloid-beta, membranes, and copper, *Journal of molecular recognition : JMR*, **30**.

4 Chapter II – Effect of pH in protein compaction

4.1 pH-dependent aggregation in intrinsically disordered proteins is determined by charge and lipophilicity

Valentín Iglesias^{1†}, Jaime Santos^{1,†}, Juan Santos-Suárez², Marco Mangiagalli³, Stefania Brocca³, Irantzu Pallarès¹ and Salvador Ventura^{1,*}

¹ Insitut de Biotecnologia i Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra (Barcelona) 08193, Spain.

² Galicia Supercomputing Center (CESGA), Santiago de Compostela, Spain.

³ Department of Biotechnology and Biosciences, University of Milano-Bicocca, Italy.

†These authors contributed equally to this work. *To whom correspondence should be addressed.

Author Contributions: Methodology, formal analysis and validation, writing—review and editing.

4.1.1 ABSTRACT

Protein aggregation is associated with an increasing number of human disorders and premature aging. Moreover, it is a central concern in the manufacturing of recombinant proteins for biotechnological and therapeutic applications. Nevertheless, the unique architecture of protein aggregates is also exploited for functional purposes, from bacteria to humans. The relevance of this process in physiopathology has attracted interest in understanding and controlling aggregation, with the concomitant development of a toolbox of algorithms aimed to predict aggregation propensities. However, most of these programs are blind to the protein environment and, in particular, to the influence of the pH. Here, we developed an empirical equation to model the pH-dependent aggregation of intrinsically disordered proteins (IDPs) based on the assumption that both the global protein charge and lipophilicity depend on the solution pH. Upon its parametrization with a model IDP, this simple phenomenological approach showed unprecedented accuracy in predicting the dependence of the aggregation of both pathogenic and functional amyloidogenic IDPs on the pH. The algorithm might find utility for diverse applications, from large-scale analysis of IDPs aggregation properties to the design of novel reversible nanofibrillar materials.

4.1.2 INTRODUCTION

Protein aggregation is an inherent feature of polypeptides that lies behind the onset of a wide range of human pathologies, including Alzheimer's and Parkinson's diseases, type II diabetes or certain cancers (Chiti and Dobson, 2006; Chiti and Dobson, 2017; de Oliveira, et al., 2020; Invernizzi, et al., 2012). Moreover, aggregation often occurs during protein recombinant production and downstream processing,

becoming a major bottleneck for the marketing of protein-based drugs (Cromwell, et al., 2006; Lin, et al., 2000). Indeed, polypeptides are susceptible of suffering aggregation at every step during protein production, from recombinant expression and purification to formulation and storage (Cromwell, et al., 2006). This implies a constant monitorization and optimization of production conditions and processes, which is costly and time-consuming. However, protein aggregation is not always deleterious, and organisms exploit the particular properties of amyloid protein assemblies for beneficial purposes (Camara-Almiron, et al., 2018; Loquet, et al., 2018; McGlinchey and Lee, 2018). This evidence has inspired the use of aggregation-prone proteins and peptides to build up functionalized nanofibrils with applications in tissue engineering, drug delivery or as nanowires and nanosensors (Diaz-Caballero, et al., 2018; Diaz-Caballero, et al., 2018; Knowles and Mezzenga, 2016; Wei, et al., 2017).

The development of *in silico* tools able to predict protein aggregation propensities has provided scientists with a versatile toolbox to assist and guide basic research and protein engineering processes (Pallares and Ventura, 2017; Santos, et al., 2020a). These algorithms exploit the evidence that protein aggregation is driven by short and specific stretches, known as aggregation-prone regions (APRs), displaying particular physicochemical features: low net charge, high hydrophobicity and, frequently, a preference for β -sheet secondary structure (Graña-Montes, et al., 2017). AGGRESKAN, Amylpred, Amyloid Mutants, FoldAmyloid, MetAmyl, PASTA, Tango, Waltz or Zyggregator (Conchillo-Sole, et al., 2007; Fernandez-Escamilla, et al., 2004; Garbuzynskiy, et al., 2010; Maurer-Stroh, et al., 2010; O'Donnell, et al., 2011; Rousseau, et al., 2006; Sanchez de Groot, et al., 2005; Tartaglia, et al., 2008; Tsolis, et al., 2013; Walsh, et al., 2014) are some examples of this kind of software. However, most of these prediction methods are blind to the protein environment, despite it is well known that factors like temperature, ionic force or pH dramatically impact protein aggregation. Regarding pH, many protein products are purified, stored or formulated at pHs different from 7.0, the default pH in these algorithms. In particular, over 65% of antibodies, Fc fusion products and Fab conjugates are formulated at pH < 6.5 (Roberts, 2014; Wang, et al., 2007). Therefore, it is surprising that, despite the vast experimental data supporting the modulation of intrinsic protein aggregative properties by the solution pH, such effect has been essentially disregarded in computational approaches (Jha, et al., 2010).

Among the different intrinsic protein properties that can contribute to protein aggregation, hydrophobicity plays a major role. Indeed, APRs usually comprise highly hydrophobic sequence stretches (Riek and Eisenberg, 2016; Ventura, et al., 2004) and mutations of polar residues to nonpolar ones exacerbate aggregation, whereas changes in the opposite direction promote solubility (Jahn and Radford, 2008). It is therefore not surprising that hydrophobicity is given a major weight, directly or indirectly, in the different equations implemented in sequence-based aggregation predictors (Castillo, et al., 2011; Graña-Montes, et al., 2017). Notably, all the aforementioned algorithms assume that the lipophilicity of the sequence to be independent of the pH. However, it is well-known that the partition coefficients of the neutral and charged species of ionizable amino acids, therefore their hydrophobicity, depend on the pH of the solution (MacCallum and Tieleman, 2011; Simm, et al., 2016). Moreover, the electrostatic

properties of proteins -i.e. their net charge in a given solution- are also connected to the solution pH, being important determinants of protein solubility (Shaw, et al., 2001; Tedeschi, et al., 2017).

To the best of our knowledge, we present here the first approach to predict how the relative aggregation propensity of a given protein changes with the solution pH. Towards this objective, we exploited a recently developed, pH-dependent, lipophilicity scale of amino acids (Zamora, et al., 2019) and implemented a simple phenomenological equation that considers the effect of the pH on both the net charge and the lipophilicity of a protein sequence. We assayed the approach on top of intrinsically disordered proteins (IDPs), which lack defined secondary structure elements, to exclude any interference on calculation coming from structural constrains. With our approach we accurately predict the impact of the pH on the aggregation properties of well-known human disease-linked proteins like α -synuclein (α -syn), A β -40, the islet amyloid polypeptide (IAPP) and the tau K19 variant, as well as in biologically relevant functional amyloids such as the melanosomal protein Pmel17, the B domain of the Bap protein and the corticotropin-releasing hormone, which indicates that it might find general application in the prediction of the pH-dependent aggregation properties of IDPs.

4.1.3 MATERIALS AND METHODS

Generation of lipophilicity profiles.

The pH-dependent lipophilicity scale developed by Zamora and co-workers (Zamora, et al., 2019) using continuum solvation calculations was employed to infer the lipophilicity of each individual amino acid at the analysed pH. Our algorithm employs a sliding window system – as previously described for AGGRESCAN linear predictor (Conchillo-Sole, et al., 2007)- to generate the lipophilicity profile of any given protein. Briefly, the program calculates the average lipophilicity of a sliding window and assigns this value to the amino acid in the center of the window. The size of the window is defined in relation with the protein length: 5 residues for proteins shorter than 75 amino acids, 7 for longer than 75 but shorter than 175, 9 for longer than 175 but shorter than 300 and 11 for longer than 300. The resulting values can be employed to build/draw a lipophilicity profile along the protein sequence or to calculate a mean value of global protein lipophilicity.

Solubility modelling.

The experimental data was obtained from Tedeschi et al. (Tedeschi, et al., 2017). The pH-dependent experimental solubility of a model IDP was used as training set to parameterize a function that describes protein solubility as a function of pH. We selected two variables to model protein solubility: pH-dependent lipophilicity and net charge. pH-dependent lipophilicity was calculated as the average of the lipophilicity profile. Protein net charge was determined using the protein calculator v3.4 server (Putnam) run at the selected pH. These theoretical values were parameterized against the solubility experimental data using

Equation 4.1:

$$\text{Solubility} = \alpha * \text{Lipophilicity} + \beta * |\text{Net Charge}|^2 + \gamma * |\text{Net Charge}| + \delta, \quad (4.1)$$

For the parameterization we employed the non-linear least squares approach of Scipy Python module, being able to define the α , β , γ and δ parameters in equation (4.1).

Data analysis and fitting.

Kinetic constants for pH-dependent α -syn aggregation were obtained from Finke and Morris and Uverski and co-workers (Morris, et al., 2009; Uverski, et al., 2001). Fibrillation rates of IAPP aggregation were previously reported by Alexandrescu and colleagues (Jha, et al., 2014). Tau K19 amyloid formation data was extracted from Jeganathan and co-workers (Jeganathan, et al., 2008). Data on A β 40 solubility at different pHs was obtained from Fändrich and co-workers (Hortschansky, et al., 2005). Data on the effect of pH on functional amyloids was extracted from references (Maji, et al., 2009; Pfefferkorn, et al., 2010; Taglialegna, et al., 2016). Linear regression analysis was performed using Graphpad Prism 6. Trendy line and 95% confidence interval were plotted, and regression r-square was added to the graph. For linear regressions, the two-tailed p-value was calculated (Soper, 2018).

4.1.4 RESULTS

Rational analysis of the molecular determinants behind pH associated aggregation.

In order to develop a new theoretical model that can forecast the effect of pH on protein aggregation, we exploited a previous work on the N-terminus moiety of the measles virus phosphoprotein (PNT), an IDP model whose aggregation propensity was deeply analysed in relation with pH and its net charge (Tedeschi, et al., 2017). In collaboration with Brocca's lab, we engineered three PNT variants displaying different net charges and isoelectric points (pI) by reversing the sign of charged residues already present in the wild-type sequence, without mutating any other PNT residue (**Figure 4.1A-D** and Supplementary Material S4.1). In detail, the acidic PNT has a pI of 3.37 and includes 62 negatively charged residues while basic PNT has a pI of 9.61 and includes 37 positively charged and 23 negatively charged residues. Attempts to produce more basic PNTs, with further unbalanced composition, were unsuccessful. The solubility of each of these protein variants was assessed experimentally in a wide range of pH, thus generating an ideal dataset to parametrize a function intended to predict the pH-dependent aggregation propensity of protein sequences (Tedeschi, et al., 2017).

Due to the lack of a well-defined 3D structure, one can hypothesize that the physicochemical determinants of pH-dependent aggregation of IDPs are directly encoded in their amino acid sequence. We propose lipophilicity (hydrophobicity) and net charge as the main properties accounting for the differential aggregation propensity of any given protein at different pHs. One can argue that this is a rather simplistic approach, but existing methodologies only consider the net charge contribution, while they overlook the role of lipophilicity.

Although unmodified at their apolar residues, our PNT variants, exhibit different lipophilicity at neutral pH, since they differ in the identity of the charged amino acids (**Figure 4.1E**). In addition, because

the hydrophobicity of ionizable amino acids is dependent on the pH, the global protein lipophilicity (average lipophilicity score) in acidic or basic conditions might differ significantly from that calculated at pH 7.0 and this parameter should be taken into account together with the net charge of the polypeptide when forecasting protein solubility.

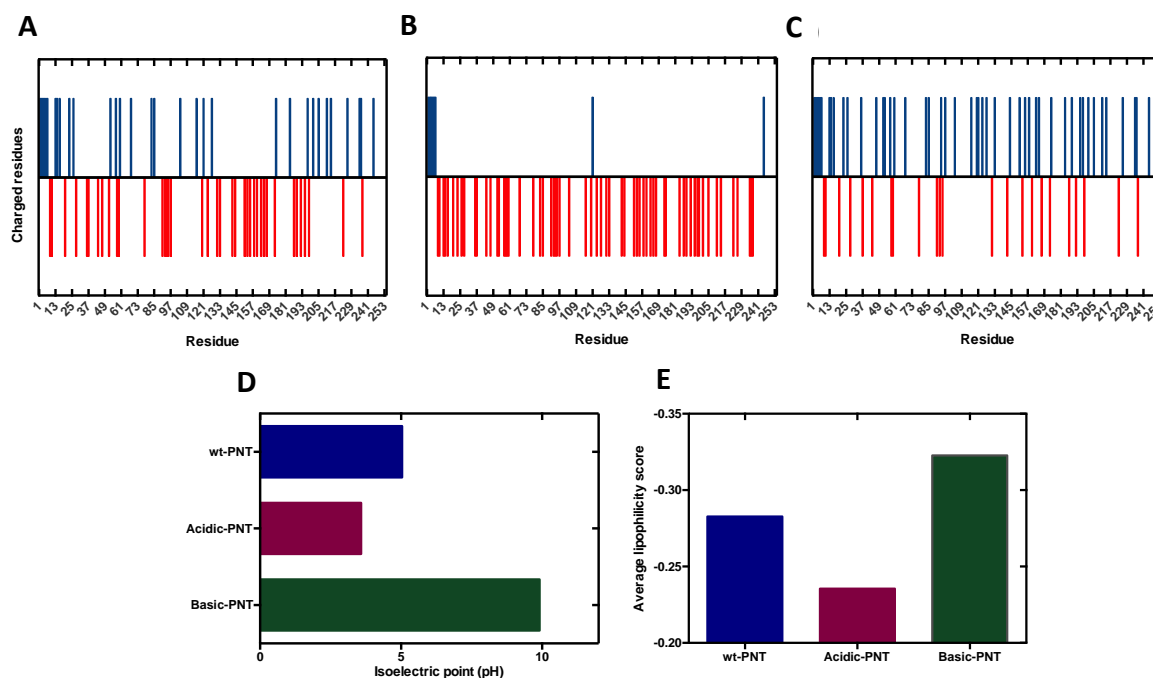


Figure 4.1 – Properties of PNT variants. A-C) Scheme of charge distribution in wild type-PNT, acidic-PNT and basic-PNT. Positive and negative residues are represented in red and blue, respectively. D) Isoelectric points of PNT variants. E) Average lipophilicity of PNT variants at pH 7.

Analysis and validation of the lipophilicity scale as a proxy for aggregation prediction.

To explore the relationship between amino acid lipophilicity, pH and protein aggregation, we exploited a pH-dependent amino acid lipophilicity scale recently derived by Zamora and co-workers (Zamora, et al., 2019). We compared the lipophilicity score of each amino acid at physiological pH (pH 7.4) with their *in vivo*-derived experimental aggregation coefficient (de Groot, et al., 2006; Sanchez de Groot, et al., 2005). We observed a highly significant correlation between aggregation and lipophilicity (p-value < 0.00001) (**Figure 4.2A**) as expected, since hydrophobic side chains are known to play determinant role in aberrant protein self-assembly (Fink, 1998).

Next, we compared the lipophilicity profile of three well-characterized disease-related proteins (*i.e.* A β 40, α -syn and IAPP) at physiological pH with their aggregation profile generated with AGGRESCAN. AGGRESCAN is an in-house developed algorithm, which implements the aforementioned *in vivo* derived aggregation propensity amino acid scale and stands as one of the most reliable algorithms to predict protein aggregation in close to *in vivo* conditions (Belli, et al., 2011). The lipophilicity and aggregation profiles of all the three proteins are in close agreement (**Figure 4.2B-D**), indicating that, at constant pH, the lipophilicity can be used as a proxy of aggregation propensity. Although other physicochemical determinants are certainly involved in protein aggregation, we assume here that they have less impact than lipophilicity or charge on pH modulated protein aggregation.

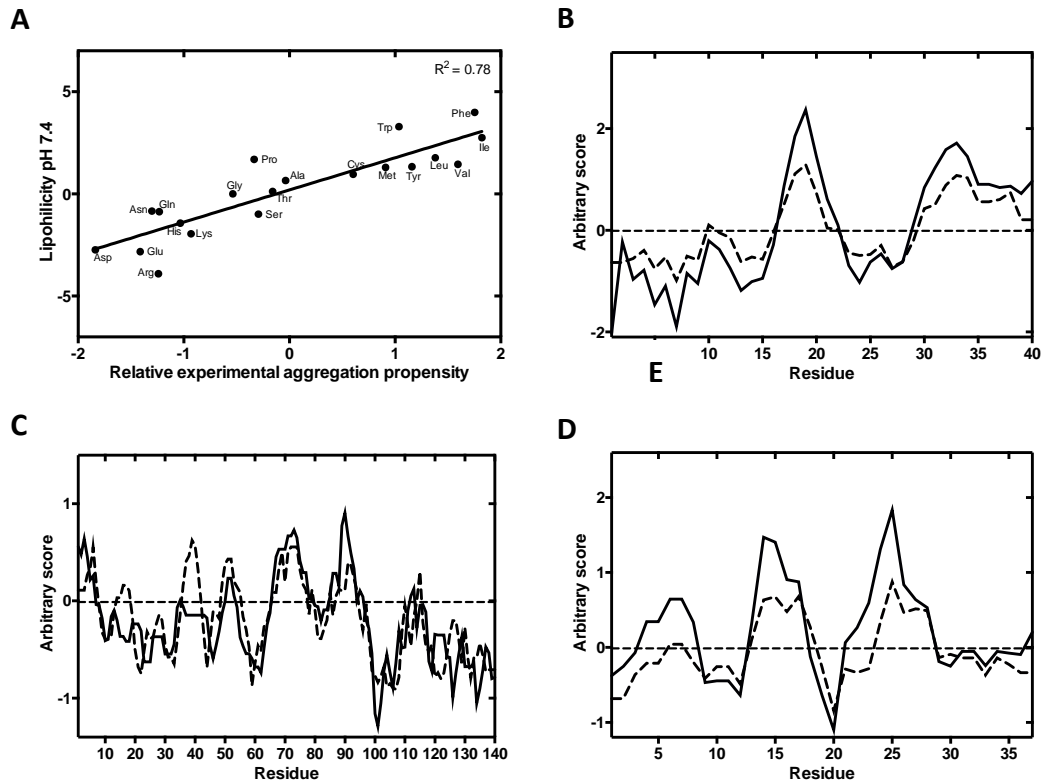


Figure 4.2 – Lipophilicity-based prediction aggregation propensity at pH 7.4 against state-of-the-art aggregation predictor. **A)** Linear correlation between amino acids *in vivo* aggregation propensity, as implemented in AGGRESCAN (de Groot, et al., 2006; Sanchez de Groot, et al., 2005), and their lipophilicity at pH 7.4. **B-D)** Overlap between AGGRESCAN-derived (dashed line) aggregation profiles and lipophilicity profiles (solid line) from Aβ40, α-syn and IAPP, respectively.

Modelling pH-dependent solubility using lipophilicity and net charge.

We next sought to build a model to determine the role of lipophilicity and net charge on pH-dependent protein aggregation. For each data point of our previous study with the PNTs, we calculated the protein net charge and the overall protein lipophilicity using a sliding window system analogous to that in AGGRESCAN (Conchillo-Sole, et al., 2007). Therefore, each data point is defined by its lipophilicity, net charge and experimental solubility, allowing their representation as a 3-axis scatter plot. The visual inspection of their spatial distribution of lipophilicity and net charge in relation with experimental solubility reveals a dispersion that resembles a quadratic 3D-surface. Thus, to model this relationship we defined an empirical formula (**Equation 4.1**) that describes a bivariate polynomial model with a quadratic component, suitable to address a 3D-surface regression in our dataset. Next, to parameterize this equation, we applied a non-linear least squares approach. As a result of the fitting, we calculated parameters α , β , γ and δ (**Table 4.1**). The resulting model, built using **Equation 4.1**, delineates a 3D-surface, where the solubility is defined as a function of net charge and lipophilicity (**Figure 4.3A**). Remarkably, the values derived from the equation shows a significant correlation with the observed solubility data (**Figure 4.3B**) (p -value < 0.00001). In contrast, a mere charge-dependent model, as the one implemented in competing approaches, fail to predict the experimental behavior of the dataset (p -value < 0.1) (Supplementary Material S4.2). Overall, these results reinforce the hypothesis that pH-induced

lipophilicity fluctuations should be taken into consideration for an accurate prediction of protein aggregation.

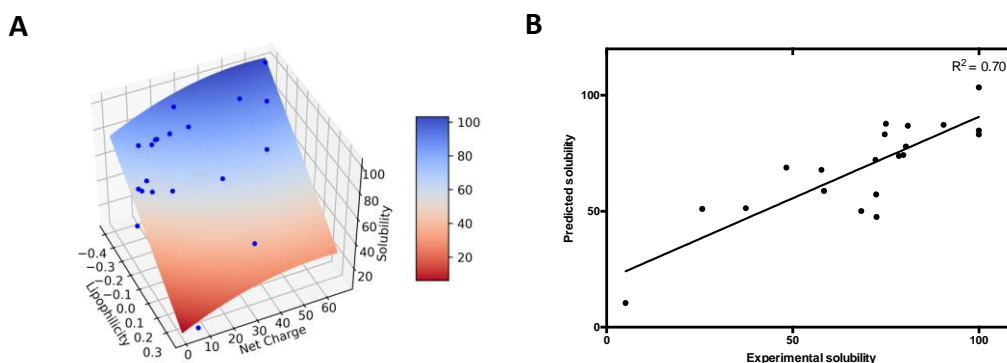


Figure 4.3 – Modeling IDP pH-dependent solubility based on lipophilicity and net charge. **A)** Experimental pH-dependent solubility modeled as a 3D surface plot. Experimental data from our previous work is represented as blue dots, and the 3D surface resultant from modeling is colored as a heat map, according to the corresponding predicted solubility as represented in the color bar. **B)** Correlation between the experimental and predicted solubility. Solid line corresponds to the fit of the data to a linear regression with a p-value < 0.00001.

Table 4.1 – Fitting parameters resulting from the non-linear least squares parametrization.

Parameter	α	β	γ	δ
Values	-97.82	-0.00747	0.8770	38.24

pH-dependent aggregation prediction in disease-linked proteins.

As a proof of principle of the predictive performance of the approach, we tested our charge and lipophilicity-dependent model in a set of well-characterized IDPs linked with conformational diseases. As discussed, IDPs represent an ideal test set for our model since they allow to consider almost exclusively the contribution of primary structure on aggregation, excluding folding and protein stability contributions. The obtained pH-dependent aggregation profile for each protein was compared with available experimental data in the literature by assessing the linear regression between experimental and predicted solubility values.

α -Synuclein (α -syn)

Parkinson’s disease (PD) is the second most prevalent neurodegenerative disorder. Brains from PD patients exhibit the recurrent presence of intracellular proteinaceous aggregates, mainly composed by α -syn. These deposits, known as Lewy Bodies, represent the main neuropathological hallmark of the disease and are responsible for eliciting cellular toxicity and causing neuronal death (Emamzadeh, 2016; Goedert, et al., 2013; Spillantini, et al., 1997). From a molecular perspective, α -syn is a 140-residues IDP, highly expressed in the synapses of dopaminergic neurons that has been shown to assemble *in vitro* into amyloid fibrils under different conditions (Lashuel, et al., 2013; Lassen, et al., 2016; Villar-Pique, et al., 2016). Owing to the connection between α -syn and PD, there is a great interest to define the determinants of α -syn aggregation. In that context, Uversky and co-workers described the effect of pH on α -syn solubility

(Uversky, et al., 2001); later on, Finke and Morris fitted the data into formal aggregation kinetic equations (Morris, et al., 2009). To assess whether the effect of pH on α -syn aggregation could be anticipated by our equation, we compared the predicted α -syn solubility with the experimental aggregation kinetic data parameters in a wide range of pHs. We found an excellent correlation between our predicted solubility and both the elongation constants and latency times of the reaction (**Figure 4.4**). In α -syn the majority of the charged residues are segregated in the C-terminal of the protein, while the hydrophobicity is clustered in its central NAC domain. It is remarkably that such dual distribution does not seem to compromise the performance of the approach.

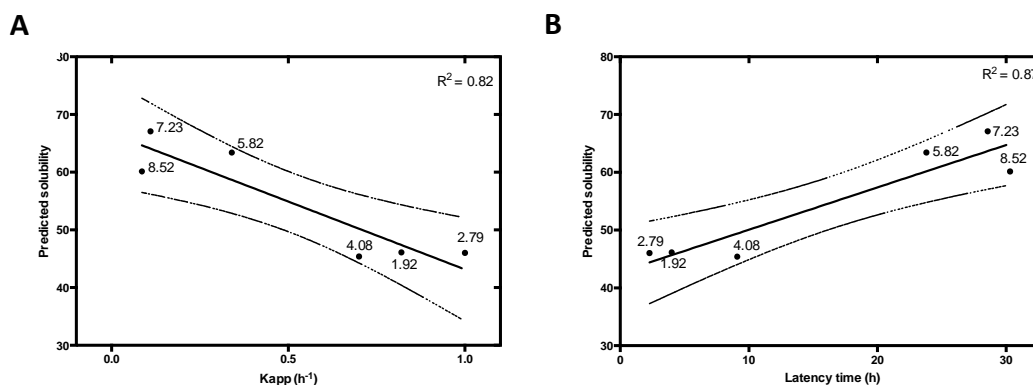


Figure 4.4 – Prediction of experimental α -syn aggregation kinetic constants. Correlation between **A**) the elongation constant K_{app} and **B**) latency time and the predicted protein solubility at different pH (1.92, 2.79, 4.08, 5.82, 7.23, 8.52). Experimental data was extracted from Morris and Finke’s work (Morris, et al., 2009). Each point represents an experimental data point labelled with its corresponding pH. A linear regression (solid line) and its 95% confidence interval (dashed line) were applied to fit the data with a p-value < 0.05 for K_{app} and < 0.01 for latency time.

Islet amyloid polypeptide (IAPP)

Aggregates of IAPP are present in the extracellular space of the islet of Langerhans in patients suffering from type II diabetes (Westermarck, et al., 2008). IAPP is an intrinsically disordered peptide hormone co-stored with insulin and involved in glycemic control (Denroche and Verchere, 2018). Under pathological conditions, IAPP forms extracellular amyloid deposits causing the degeneration of pancreatic β -cells (Mukherjee, et al., 2017). This behavior is thought to be dependent on the environmental pH (Akter, et al., 2016; Khemtémourian, et al., 2011), being slightly acidic pH of the secretory granules ($pH \approx 5.5$) able to protect IAPP from aggregation, while the extracellular environment pH ($pH \approx 7.4$) pro-aggregational. The pH-dependent fibrillation of IAPP was studied by Alexandrescu and co-workers, uncovering a strong pH dependency for this peptide (Jha, et al., 2014). This work provided us with a complete set of kinetic data over a wide pH range to further test our model. IAPP fibrillation rates are tightly connected to the solution pH, a trend that can be predicted with high confidence by applying our equation (**Figure 4.5**).

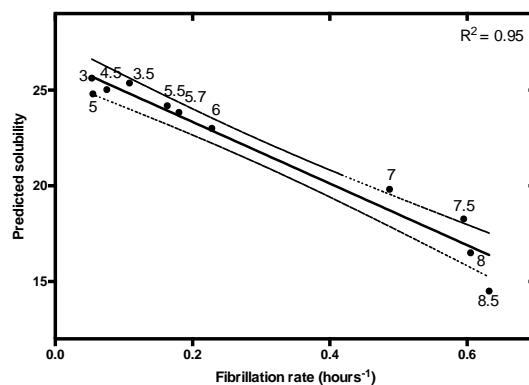


Figure 4.5 – Linear correlation between IAPP fibrillation rate and predicted solubility at different pH (3, 3.5, 4.5, 5, 5.5, 5.7, 6, 7, 7.5, 8.5, 9). Data on IAPP fibrillation were extracted from Alexandrescu and co-workers (Jha, et al., 2014). Data was fitted to linear regression (solid line) with a p-value <0.00001 and its 95% confidence interval was represented (dashed line).

Alzheimer’s disease related proteins: amyloid-beta peptides and tau protein

Alzheimer’s disease (AD) is the most prevalent neurodegenerative disorder and is characterized by a progressive cognitive impairment. The molecular pathology of AD is characterized by the combined presence of two distinct types of aberrant protein deposits in brain tissue: extracellular amyloid deposits -amyloid plaques- and intraneuronal neurofibrillary tangles (Lane, et al., 2018). The β -amyloid peptides A β -40 and A β -42 are intrinsically disordered proteolytic fragments of amyloid-beta precursor protein (Meng, et al., 2018) and their aggregates constitute the principal components of the amyloid plaques. Tau is an IDP (Eliezer, et al., 2005; Schweers, et al., 1994) whose main function is promoting microtubule assembly and stability. In AD, tau aggregation results in the assembly of abnormal neurofibrillary tangles. The aggregation reactions of these proteins have been extensively characterized due to their pivotal role in AD. Fändrich and co-workers addressed the effect of pH over A β -40 solubility, reporting a significant decrease in solubility below neutral pH (Hortschansky, et al., 2005). Our model is able to recapitulate this pH-dependence of A β -40 solubility with high accuracy (**Figure 4.6A**). Jeganathan and colleagues studied how pH affected tau K19 aggregation (Jeganathan, et al., 2008). Tau K19 is a truncated construct containing three microtubule binding repeats (R1, R3, and R4); whose aggregates show structural features that are reminiscent of those of the full-length tau protein (Dinkel, et al., 2011; Siddiqua and Margittai, 2010). Again, our algorithm successfully models the experimental behaviour of tau K19 aggregation at different pHs (**Figure 4.6B**).

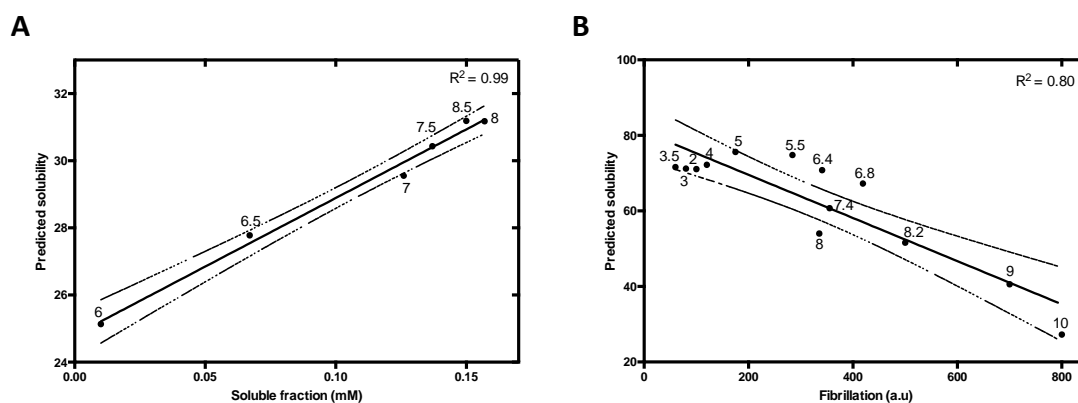


Figure 4.6 – Analysis of the effect of pH variations on A β -40 and tau K19 variant solubility. A) Correlation between A β -40 predicted and experimental solubility at different pH (6, 6.5, 7, 7.5, 8, 8.5). Experimental data was extracted from Fändrich and co-workers (Hortschansky, et al., 2005). **B)** Analysis of the experimental amyloid formation reported by Thioflavin S fluorescence emission, extracted from Jeganathan and co-workers at a range of pH from 3 to 10 (Jeganathan, et al., 2008). Data was fitted to linear regression (solid line) and its 95% confidence interval was represented (dashed line), with a p-value < 0.0001 in both cases.

Predicting the impact of pH on the aggregation of functional amyloids: context-dependent aggregation allows enclosure of functional self-assembly.

Amyloid fibrils have been traditionally considered pathogenic agents responsible for a set of devastating human disorders, such as the examples mentioned on the previous chapters. However, the last decade has seen a large body of evidence supporting that the amyloid architecture can be exploited to develop biological functions (Pham, et al., 2014). Functional amyloids work under physiologically conditions without any associated cytotoxicity (Jackson and Hewitt, 2017; Otzen, 2010), mainly because, in contrast to their toxic counterparts, coordinated cellular strategies have evolved to control their assembly. One of these strategies consists in confining aggregation inside a specific cellular compartment in a pH-dependent manner. This natural strategy provides an exceptional benchmark to validate our predictive model.

Pigment cell-specific melanosome protein.

The pigment cell-specific melanosome protein (Pmel17) is involved in the biogenesis and maturation of melanosomes, organelles specialized in melanin synthesis, present in melanocytes and epithelial cells in mammals. The specific role of Pmel17 is the formation of amyloid fibrils in the lumen of the melanosomes that optimize the sequestration and condensation of melanin (Jha, et al., 2014; Pham, et al., 2014). Pmel17 fibrillation occurs in the acidic environment of the early stage melanosome (pH=4-5). Lee and co-workers first reported the amyloidogenesis of the repeat domain (RPT) of Pmel17, describing a strong dependence on solution pH: a fast aggregation at pH 4, slower at pH 5 and 5.5 and no aggregation -and even fibril disaggregation- at pH 7 (Pfefferkorn, et al., 2010). Our algorithm successfully discriminates those three regimes of aggregation (**Figure 4.7A**).

Corticotropin-releasing hormone.

Maji and co-workers discovered in 2009 a novel activity of functional amyloids as storage of peptide hormones in secretory granules (Maji, et al., 2009). They described that peptide hormones fibrillate due to the low pH (≈ 5.5) of those granules and that, upon release to the extracellular environment ($\text{pH} \approx 7.4$), the fibrils gradually disassemble into the monomeric bioactive specie. In that work, they explore this effect *in vitro* on the corticotropin-releasing hormone (CRF) by inducing the formation of fibrils at acidic pH 5.5 and analyzing their disaggregation at higher pHs (pH 6 and 7.4). The experimental disaggregation was accelerated at pH 7.4. This behavior, with a gradual gain of solubility at increasing pHs and fast dissociation at pH 7.4, is fairly recapitulated by our model (**Figure 4.7B**).

B domain of the Bap protein.

Staphylococcus aureus Bap is an extracellular protein able to self-assemble at acidic pH (≈ 4.5), forming amyloid fibrils that scaffold the formation of a biofilm matrix (Taglialegna, et al., 2016). In the case of Bap, aggregation is confined in the extracellular environment where it functions as a pH sensor and -upon acidic conditions- orchestrates a multicellular response that elicits biofilm formation. Lasa, Valle and co-workers reported the aggregation of this protein, they identified an amyloidogenic domain (BapB) and characterized its pH-dependent aggregation (Taglialegna, et al., 2016). BapB forms amyloid fibrils at pH 4.5 that dissociate when the pH rises to attain the neutrality. Once more, our approach is able to predict such behavior (**Figure 4.7C**).

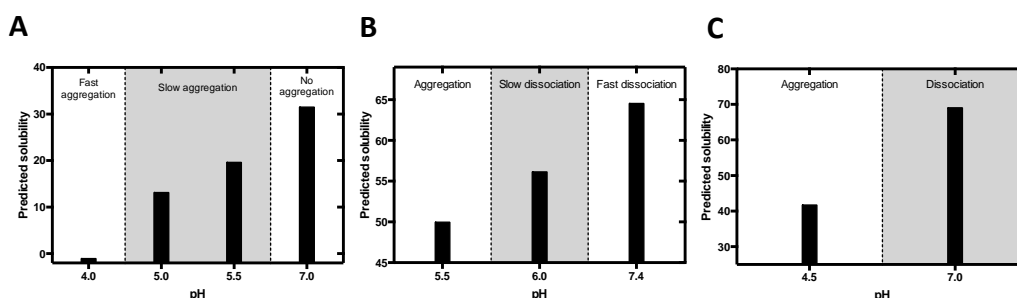


Figure 4.7 – Evaluation of the pH-dependent mechanism of fibrillation of functional amyloids. A) Pmel17, B) CRF, C) BapB predicted solubility against their physiological fibrillation and disaggregation tendencies. The different regions of aggregation are delimited by dotted lines.

4.1.3 DISCUSSION

In the last decades, the advances in the field of protein aggregation resulted in the development of over 40 different predictive methods to computationally assess protein deposition. Thus, we have at our disposal a wide variety of algorithms based on conceptually different molecular determinants to systematically predict protein aggregation. However, these approaches barely exploit the influence of the protein environment. This is important because solvent conditions impact solubility by modulating the hydrophobic effect, electrostatic interactions or the degree of protonation of the different ionizable groups. Here, we presented a novel phenomenological model whose aim is the evaluation of protein solubility as a function of solvent pH. Exploiting previous experimental data on the solubility of a charge-engineered model IDP, we were able to weight the contribution of lipophilicity and net charge to protein solubility and, subsequently, elaborate a phenomenological predictor with high accuracy in predicting pH-

dependent aggregation of IDPs. Our results indicate that in addition to the net charge, pH also modulates protein lipophilicity and that such control has a great impact on protein solubility.

Our algorithm demonstrates high accuracy in predicting pH modulation of aggregation propensity in a set of disease-associated IDPs, such as α -syn, IAPP, tau K19 fragment and A β -40. Moreover, we employed our approach to evaluate the aggregation propensity of three proteins reported to form functional amyloids *in vivo* upon pH shifts. Interestingly enough, in these proteins, evolution has exerted a positive selective pressure to attain a reversible fibrillation mechanism where pH controls the assembly and disassembly of the fibrils. Notably, we were able to predict such behavior by analyzing only protein primary structures, highlighting that this conformational transition is intrinsically imprinted in the polypeptide chain.

The main application of our prediction method would be the profiling of protein solubility along a continuous pH interval, since it demonstrates a remarkable accuracy in describing this behavior. The model is simple, and computation is fast, which should allow the analysis of large sequence datasets, including the complete complement of IDPs in a given proteome. It would be interesting to assess whether the IDPs residing in cellular compartments are optimized to display the maximum solubility at the specific compartment pH. The algorithm can also contribute to understand the role of changes in intracellular pH in protein phase separation reactions, since this phenomenon results from the coalescence of intrinsically disordered regions (Franzmann, et al., 2018). We also propose that our method may have an impact in the design of nanomaterials with pH-modulable assembling properties, which can transition between soluble and amyloid-like states simply by shifting the solution pH.

The method can also be used to assist the purification, formulation and storage of proteins of biotechnological and therapeutic interest, by predicting the range of pH in which they are more soluble, as long as they are intrinsically disordered, as in the case of peptidic hormones. For its use in the design of optimal solutions for globular proteins, like therapeutic antibodies, the concept should be first implemented in a structural predictor, were the intrinsic charge and lipophilic properties of amino acids would be modulated according to their conformational context. This step will be analogous to the evolution of AGGRESCAN (Conchillo-Sole, et al., 2007) into our structural A3D aggregation predictor (Kuriata, et al., 2019; Kuriata, et al., 2019; Zambrano, et al., 2015) and thus, perfectly attainable.

4.1.6 REFERENCES

- Akter, R., *et al.* (2016) Islet Amyloid Polypeptide: Structure, Function, and Pathophysiology, *Journal of diabetes research*, **2016**, 2798269.
- Belli, M., Ramazzotti, M. and Chiti, F. (2011) Prediction of amyloid aggregation *in vivo*, *EMBO reports*, **12**, 657-663.
- Camara-Almiron, J., *et al.* (2018) Beyond the expected: the structural and functional diversity of bacterial amyloids, *Critical reviews in microbiology*, **44**, 653-666.
- Castillo, V., *et al.* (2011) Prediction of the aggregation propensity of proteins from the primary sequence: aggregation properties of proteomes, *Biotechnology journal*, **6**, 674-685.
- Conchillo-Sole, O., *et al.* (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides, *BMC bioinformatics*, **8**, 65.
- Cromwell, M.E., Hilario, E. and Jacobson, F. (2006) Protein aggregation and bioprocessing, *The AAPS journal*, **8**, E572-579.

Chiti, F. and Dobson, C.M. (2006) Protein misfolding, functional amyloid, and human disease, *Annu Rev Biochem*, **75**, 333-366.

Chiti, F. and Dobson, C.M. (2017) Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade, *Annu Rev Biochem*, **86**, 27-68.

de Groot, N.S., *et al.* (2006) Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities, *FEBS J*, **273**, 658-668.

Denroche, H.C. and Verchere, C.B. (2018) IAPP and type 1 diabetes: implications for immunity, metabolism and islet transplants, *Journal of molecular endocrinology*, **60**, R57-R75.

Diaz-Caballero, M., *et al.* (2018) Prion-based nanomaterials and their emerging applications, *Prion*, **12**, 266-272.

Diaz-Caballero, M., *et al.* (2018) Minimalist Prion-Inspired Polar Self-Assembling Peptides, *ACS nano*, **12**, 5394-5407.

Dinkel, P.D., *et al.* (2011) Variations in filament conformation dictate seeding barrier between three- and four-repeat tau, *Biochemistry*, **50**, 4330-4336.

Eliezer, D., *et al.* (2005) Residual structure in the repeat domain of tau: echoes of microtubule binding and paired helical filament formation, *Biochemistry*, **44**, 1026-1036.

Emamzadeh, F.N. (2016) Alpha-synuclein structure, functions, and interactions, *Journal of research in medical sciences : the official journal of Isfahan University of Medical Sciences*, **21**, 29.

Fernandez-Escamilla, A.M., *et al.* (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins, *Nature biotechnology*, **22**, 1302-1306.

Fink, A.L. (1998) Protein aggregation: folding aggregates, inclusion bodies and amyloid, *Folding & design*, **3**, R9-23.

Franzmann, T.M., *et al.* (2018) Phase separation of a yeast prion protein promotes cellular fitness, *Science*, **359**.

Garbuzynskiy, S.O., Lobanov, M.Y. and Galzitskaya, O.V. (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence, *Bioinformatics*, **26**, 326-332.

Goedert, M., *et al.* (2013) 100 years of Lewy pathology, *Nature reviews. Neurology*, **9**, 13-24.

Hortschansky, P., *et al.* (2005) The aggregation kinetics of Alzheimer's beta-amyloid peptide is controlled by stochastic nucleation, *Protein science : a publication of the Protein Society*, **14**, 1753-1759.

Jackson, M.P. and Hewitt, E.W. (2017) Why are Functional Amyloids Non-Toxic in Humans?, *Biomolecules*, **7**.

Jahn, T.R. and Radford, S.E. (2008) Folding versus aggregation: polypeptide conformations on competing pathways, *Archives of biochemistry and biophysics*, **469**, 100-117.

Jeganathan, S., *et al.* (2008) The natively unfolded character of tau and its aggregation to Alzheimer-like paired helical filaments, *Biochemistry*, **47**, 10526-10539.

Jha, R.K., *et al.* (2010) Computational design of a PAK1 binding protein, *Journal of molecular biology*, **400**, 257-270.

Jha, S., *et al.* (2014) pH dependence of amylin fibrillization, *Biochemistry*, **53**, 300-310.

Khemtemourian, L., *et al.* (2011) Low pH acts as inhibitor of membrane damage induced by human islet amyloid polypeptide, *Journal of the American Chemical Society*, **133**, 15598-15604.

Knowles, T.P. and Mezzenga, R. (2016) Amyloid Fibrils as Building Blocks for Natural and Artificial Functional Materials, *Advanced materials*, **28**, 6546-6561.

Kuriata, A., *et al.* (2019) Aggrescan3D standalone package for structure-based prediction of protein aggregation properties, *Bioinformatics*, **35**, 3834-3835.

Kuriata, A., *et al.* (2019) Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility, *Nucleic Acids Res*, **47**, W300-W307.

Lane, C.A., Hardy, J. and Schott, J.M. (2018) Alzheimer's disease, *European journal of neurology*, **25**, 59-70.

Lashuel, H.A., *et al.* (2013) The many faces of alpha-synuclein: from structure and toxicity to therapeutic target, *Nature reviews. Neuroscience*, **14**, 38-48.

Lassen, L.B., *et al.* (2016) Protein Partners of alpha-Synuclein in Health and Disease, *Brain pathology*, **26**, 389-397.

Lin, J.J., *et al.* (2000) Stability of human serum albumin during bioprocessing: denaturation and aggregation during processing of albumin paste, *Pharmaceutical research*, **17**, 391-396.

Loquet, A., Saupe, S.J. and Romero, D. (2018) Functional Amyloids in Health and Disease, *Journal of molecular biology*, **430**, 3629-3630.

MacCallum, J.L. and Tieleman, D.P. (2011) Hydrophobicity scales: a thermodynamic looking glass into lipid-protein interactions, *Trends Biochem Sci*, **36**, 653-662.

Maji, S.K., *et al.* (2009) Functional amyloids as natural storage of peptide hormones in pituitary secretory granules, *Science*, **325**, 328-332.

Maurer-Stroh, S., *et al.* (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices, *Nature methods*, **7**, 237-242.

McGlinchey, R.P. and Lee, J.C. (2018) Why Study Functional Amyloids? Lessons from the Repeat Domain of Pmel17, *Journal of molecular biology*, **430**, 3696-3706.

Meng, F., *et al.* (2018) Highly Disordered Amyloid-beta Monomer Probed by Single-Molecule FRET and MD Simulation, *Biophysical journal*, **114**, 870-884.

Morris, A.M., Watzky, M.A. and Finke, R.G. (2009) Protein aggregation kinetics, mechanism, and curve-fitting: a review of the literature, *Biochim Biophys Acta*, **1794**, 375-397.

Mukherjee, A., *et al.* (2017) Induction of IAPP amyloid deposition and associated diabetic abnormalities by a prion-like mechanism, *The Journal of experimental medicine*, **214**, 2591-2610.

O'Donnell, C.W., *et al.* (2011) A method for probing the mutational landscape of amyloid structure, *Bioinformatics*, **27**, i34-42.

Otzen, D. (2010) Functional amyloid: turning swords into plowshares, *Prion*, **4**, 256-264.

Pallares, I. and Ventura, S. (2017) Advances in the prediction of protein aggregation propensity, *Current medicinal chemistry*.

Pfefferkorn, C.M., McGlinchey, R.P. and Lee, J.C. (2010) Effects of pH on aggregation kinetics of the repeat domain of a functional amyloid, Pmel17, *Proc Natl Acad Sci U S A*, **107**, 21447-21452.

Pham, C.L., Kwan, A.H. and Sunde, M. (2014) Functional amyloid: widespread in Nature, diverse in purpose, *Essays in biochemistry*, **56**, 207-219.

Putnam, C. Protein Calculator.

Ricardo Graña-Montes, J.P.-P., Carlota Gómez-Picanyol and Ventura, a.S. (2017) Prediction of Protein Aggregation and Amyloid Formation. In Rigden, D.J. (ed), *From Protein Structure to Function with Bioinformatics*. Springer, pp. 205-263.

Riek, R. and Eisenberg, D.S. (2016) The activities of amyloids from a structural perspective, *Nature*, **539**, 227-235.

Roberts, C.J. (2014) Therapeutic protein aggregation: mechanisms, design, and control, *Trends in biotechnology*, **32**, 372-380.

Rousseau, F., Schymkowitz, J. and Serrano, L. (2006) Protein aggregation and amyloidosis: confusion of the kinds?, *Current opinion in structural biology*, **16**, 118-126.

Sanchez de Groot, N., et al. (2005) Prediction of "hot spots" of aggregation in disease-linked polypeptides, *BMC structural biology*, **5**, 18.

Schweers, O., et al. (1994) Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure, *The Journal of biological chemistry*, **269**, 24290-24297.

Shaw, K.L., et al. (2001) The effect of net charge on the solubility, activity, and stability of ribonuclease Sa, *Protein science : a publication of the Protein Society*, **10**, 1206-1215.

Siddiqua, A. and Margittai, M. (2010) Three- and four-repeat Tau coassemble into heterogeneous filaments: an implication for Alzheimer disease, *The Journal of biological chemistry*, **285**, 37920-37926.

Simm, S., et al. (2016) 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification, *Biological research*, **49**, 31.

Soper, D.S. (2018) p-Value Calculator for Correlation Coefficients.

Spillantini, M.G., et al. (1997) Alpha-synuclein in Lewy bodies, *Nature*, **388**, 839-840.

Taglialegna, A., et al. (2016) Staphylococcal Bap Proteins Build Amyloid Scaffold Biofilm Matrices in Response to Environmental Signals, *PLoS pathogens*, **12**, e1005711.

Tartaglia, G.G., et al. (2008) Prediction of aggregation-prone regions in structured proteins, *Journal of molecular biology*, **380**, 425-436.

Tedeschi, G., et al. (2017) Aggregation properties of a disordered protein are tunable by pH and depend on its net charge per residue, *Biochimica et biophysica acta. General subjects*, **1861**, 2543-2550.

Tsolis, A.C., et al. (2013) A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins, *PLoS one*, **8**, e54175.

Uversky, V.N., Li, J. and Fink, A.L. (2001) Evidence for a partially folded intermediate in alpha-synuclein fibril formation, *The Journal of biological chemistry*, **276**, 10737-10744.

Ventura, S., et al. (2004) Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case, *Proc Natl Acad Sci U S A*, **101**, 7258-7263.

Villar-Pique, A., Lopes da Fonseca, T. and Outeiro, T.F. (2016) Structure, function and toxicity of alpha-synuclein: the Bermuda triangle in synucleinopathies, *Journal of neurochemistry*, **139 Suppl 1**, 240-255.

Walsh, I., et al. (2014) PASTA 2.0: an improved server for protein aggregation prediction, *Nucleic Acids Res*, **42**, W301-307.

Wang, W., et al. (2007) Antibody structure, instability, and formulation, *Journal of pharmaceutical sciences*, **96**, 1-26.

Wei, G., et al. (2017) Self-assembling peptide and protein amyloids: from structure to tailored function in nanotechnology, *Chemical Society reviews*, **46**, 4661-4708.

Westermarck, G.T., et al. (2008) Widespread amyloid deposition in transplanted human pancreatic islets, *The New England journal of medicine*, **359**, 977-979.

Zambrano, R., et al. (2015) AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures, *Nucleic Acids Res*, **43**, W306-313.

Zamora, W.J., Campanera, J.M. and Luque, F.J. (2019) Development of a Structure-Based, pH-Dependent Lipophilicity Scale of Amino Acids from Continuum Solvation Calculations, *The journal of physical chemistry letters*, **10**, 883-889.

4.2 SolupHred: predicting pH-dependent aggregation of intrinsically disordered proteins

Carlos Pintado^{1,†}, Jaime Santos^{1,†}, Valentín Iglesias^{1,*} and Salvador Ventura^{1,*}

¹ Insitut de Biotecnologia i Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra (Barcelona) 08193, Spain.

†These authors contributed equally to this work. *To whom correspondence should be addressed.

Author Contributions: Conceptualization, software, validation, data curation, writing—original draft preparation, supervision.

4.2.1 ABSTRACT

Summary: Polypeptides are exposed to changing environmental conditions that modulate their intrinsic aggregation propensities. Intrinsically disordered proteins (IDPs) constitutively expose their aggregation determinants to the solvent, thus being especially sensitive to its fluctuations. However, solvent conditions are often disregarded in computational aggregation predictors. We recently developed a phenomenological model to predict IDPs' solubility as a function of the solution pH, which is based on the assumption that both protein lipophilicity and charge depend on this parameter. The model anticipated solubility changes in different IDPs accurately. Here, we present SolupHred, a web-based interface that implements the aforementioned theoretical framework into a predictive tool able to compute IDPs aggregation propensities as a function of pH. SolupHred is the first dedicated software for the prediction of pH-dependent protein aggregation.

Availability and Implementation: The SolupHred web server is freely available for academic users at: <https://ppmclab.pythonanywhere.com/SolupHred>. It is platform-independent and does not require previous registration.

4.2.2 INTRODUCTION

Protein aggregation is a significant bottleneck in the production and storage of protein-based therapeutics, and it is associated with a wide range of human disorders. Accordingly, anticipating proteins' aggregation properties has attracted significant interest in biotechnology and biomedicine (Santos, et al., 2020a).

In intrinsically disordered proteins (IDPs), aggregation is not constrained by structural elements, and therefore it can be inferred directly from the primary sequence (Santos, et al., 2020b). More than 20 different algorithms have been built on this principle, achieving a remarkable overlap with experimental results. The lack of residues' protection by elements of secondary and tertiary structures, makes IDPs more sensitive to solvent conditions and environmental fluctuations than folded proteins (Uversky, 2009),

an effect which has been traditionally disregarded or barely parametrized in state-of-the-art aggregation predictors. Indeed, numerous data evidence that IDPs aggregation is strongly modulated by factors extrinsic to the sequence, such as ion concentration, ligands, or pH (Uversky, 2009).

However, the pH-dependent aggregation of IDPs is not always associated with a deleterious loss-of-function, and evolution exploits reversible fibrillation mechanisms, where pH modulates the mesoscopic assembly of functional amyloids to regulate their activities (Maji, et al., 2009). Thus, modelling the effect of pH on IDPs aggregation would offer an avenue to analyse context-dependent aggregation in physiological backgrounds and tightly regulate IDPs solubility in diverse biotechnological applications.

In a recent work addressed in **Section 4.1**, we elaborated a phenomenological model to predict IDPs aggregation as a function of pH, based on the assumption that protein lipophilicity and charge are both dependent on the solution pH (Santos, et al., 2020c). The model showed remarkable reliability in predicting the pH-dependent aggregation of disease-associated IDPs and in anticipating the pH-modulated assembly of functional amyloids. Here, this conceptual framework is implemented in SolupHred web server, the first computational tool dedicated to evaluating the effect of solution pH on IDPs aggregation. The SolupHred web server is free for academic users, allowing fast and reliable analysis of either individual IDPs or large sets of disordered sequences in the desired pH ranges.

4.2.3 METHODS

The SolupHred web server profiles the pH-dependent aggregation of the analysed disordered sequence(s) in a user-defined pH range. To do so, SolupHred computes the sequence lipophilicity and net charge at each pH and applies an empirical equation (**Equation 4.1**) to model the aggregation in each particular condition (Santos, et al., 2020c) (**Figure 4.8**):

Input interface: One or more disordered sequence(s) in FASTA format can be pasted or uploaded. Users can define the range of pHs -with the desired step size- in which solubility will be computed. Alternatively, solubility at a specific pH can be calculated (**Figure 4.9A**).

Computation of lipophilicity profile: The algorithm uses a size-dependent sliding widow to generate a lipophilicity profile for each sequence using a recently developed pH-dependent lipophilicity scale of amino acids (Zamora, et al., 2019). Mean lipophilicity is computed as the average of all individual residue scores in the profile.

Net charge calculation: Residue partial charge is calculated using the Henderson-Hasselbalch equation. Global net charge corresponds to the absolute value of the sum of all individual residues' partial charges.

Solubility calculation: Mean lipophilicity and global net charge are combined in the equation described by Santos and co-workers (**Equation 4.1**) to predict solubility in the selected pH range.

Output presentation: The results page (**Figure 4.9B**) displays two clickable links containing a JSON file with all stored information and a downloadable ZIP file with all generated results (CSV and JSON files and

figures). An interactive table appears below with the main results, showing pHs where solubility is maximum and minimum along with the solubility scores. Besides, the pH intervals in which proteins have 10% of their maximum or minimum (10% max/min) solubility are displayed. Clicking identifiers will open the correspondent graph showing solubility variations in the specified pH range with the 10% max/min solubility in blue and red, respectively.

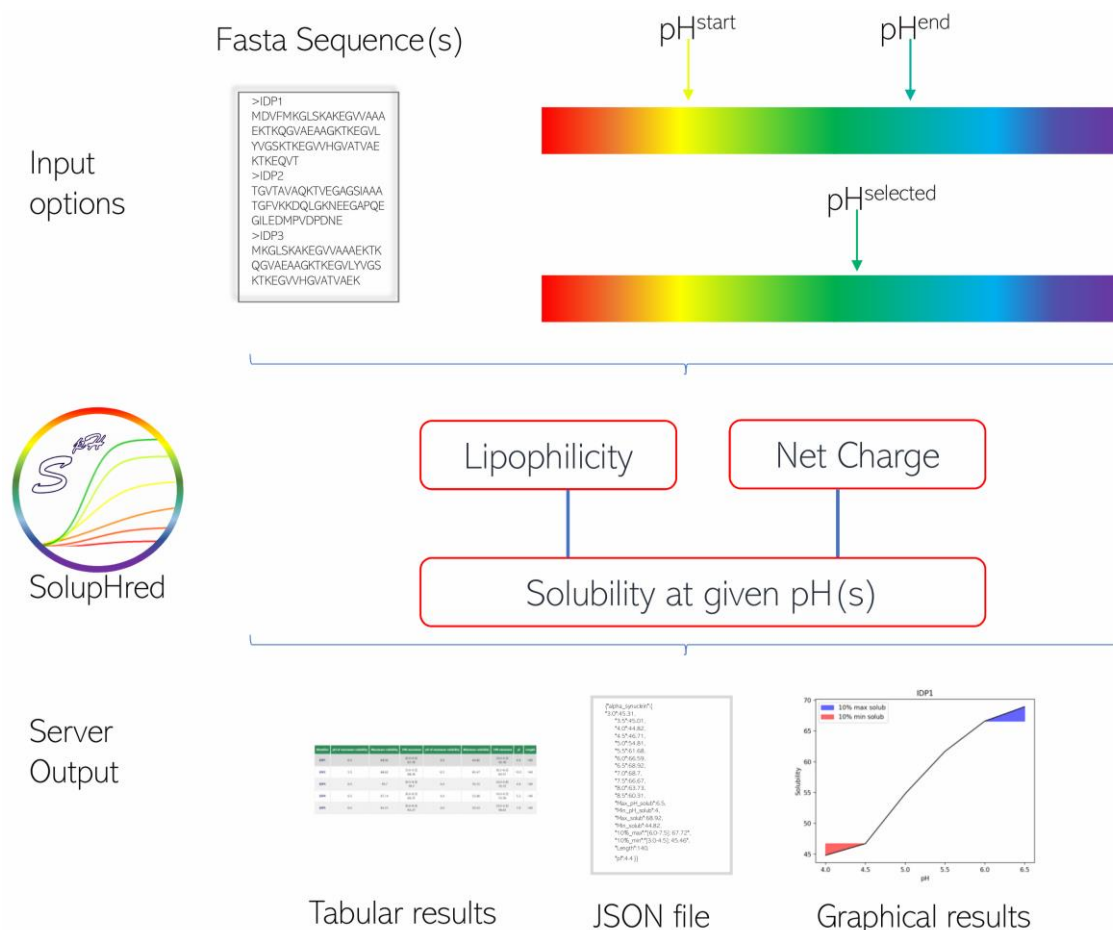


Figure 4.8 – SolupHred pipeline.

4.2.3 IMPLEMENTATION


SolupHred script is written in Python and uses Python3.7 as the interpreter. The web interface was built using HTML/CSS/JavaScript. Inputs and outputs are processed by Django CGI scripts written in Python.

4.2.4 PERFORMANCE



SolupHred implements a phenomenological equation to calculate pH-dependent solubility of IDPs, whose performance has been previously validated on a set of disease-associated IDPs (Section 4.1) and functional amyloids (Santos, et al., 2020c), with an excellent correlation between experiments and predictions (Supplementary Material S4.3). In the experimentally validated dataset analysed, SolupHred

predicts with high accuracy (0.91) whether the deviation from pH neutrality results in increased or decreased aggregation for each protein (Supplementary Material S4.4).

A



SolupHred
The pH-dependent solubility predictor of IDPs

SolupHred is the first phenomenological predictor that considers protein environment pH when calculating aggregation propensity of Intrinsically Disordered Proteins (IDPs).

Submission Help References Contact

Introduce your sequence(s) in [FASTA format](#).
All sequence(s) will be treated as disordered. Alternatively users can explore pH dependent disorder [here](#).

Browse... No file selected. [Example](#)

```
>alpha_synuclein
MDVFMKGLSKAKEGVVAAAEKTKQGVAAEAGKTKEGVLVYVGSKTKEGV
VHGVAIVAEKTKQVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVK
KDQLGKNEEGAPQEGILEDMPPDPNEAYEMPSEEGYQDYPEA
```

From pH: To pH: pH interval:

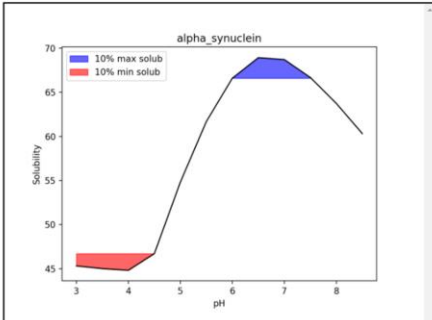
Predict solubility at specific pH:

B

[Visualize results in json format](#) [Download all results \(.zip\)](#) [Back to initial page](#)

Input summary:
1 sequence(s) introduced
From pH: 3
To pH: 8.5
pH interval: 0.5

Identifier	pH of maximum solubility	Maximum solubility	10% maximum	pH of minimum solubility	Minimum solubility	10% minimum	pI	Length
alpha_synuclein	6.5	68.92	[6.0-7.5] 67.72	4.0	44.82	[3.0-4.5] 45.46	4.4	140



pH	Solubility Score
3.0	45.31
3.5	45.01
4.0	44.82
4.5	46.71
5.0	54.81
5.5	61.68
6.0	66.59
6.5	68.92

Figure 4.9 – SolupHred web server interface. **A)** Web input page. The user can paste their FASTA-formatted sequences in the box or upload them as a file. By default SoluHred checks solubility in pH interval, but it allows users to test values at a specific pH. **B)** Output page for one protein in a range of pHs. Two clickable links appear on the upper left part of the screen with results retrievable in JSON format or a compressed ZIP file containing SolupHred calculations and generated figures. On the left part, a table depicts the pHs of maximum and minimum solubility along with the pH intervals in which proteins have 10% of their maximum or minimum (10% max/min) solubility. On the lower right side, a table with the solubility variations for the specified pH range is presented while at the bottom left a plot shows the aforementioned data, with the 10% max/min solubility coloured in blue and red, respectively.

Alternatively, for multiple proteins, this graph can be reachable by clicking the link in each protein identifier.

SolupHred is suitable for the analysis of large collections of proteins in a fast and comprehensive way, performing over 500 pH-datapoint calculations per second when benchmarked using the DisProt database (Hatos et al., 2020). The web server is limited to the 20 standard proteinogenic amino acids and assumes input proteins remain disordered in the user specified pH range (Santos, et al., 2020d) (this aspect will be discussed in **Section 4.3**).

4.2.4 CONCLUSIONS

SolupHred is a web application tool to predict IDPs' solubility as a function of the pH, which makes publicly accessible the predictive model we developed recently (Santos, et al., 2020c), (**Section 4.1**). It allows fast and accurate evaluations of the aggregation propensities of disordered sequences in a given range of pHs. SolupHred permits the large-scale analysis of disordered protein databases. The SolupHred output was designed to be easily incorporated into external computational pipelines dealing with IDPs properties.

We expect SolupHred to be adopted by the community as a fast, cost-effective way to decide the adequate conditions for performing aggregation experiments and the purification and storage of IDPs. We envision that, as SolupHred, next-generation programs will progressively incorporate extrinsic environmental factors in their predictions.

4.2.5 REFERENCES

- Maji, S.K., et al. (2009) Functional amyloids as natural storage of peptide hormones in pituitary secretory granules, *Science*, 325, 328-332.
- Santos, J., et al. (2020a) Computational prediction of protein aggregation: Advances in proteomics, conformation-specific algorithms and biotechnological applications, *Computational and structural biotechnology journal*, 18, 1403-1413.
- Santos, J., Iglesias, V. and Ventura, S. (2020b) Computational prediction and redesign of aberrant protein oligomerization, *Progress in molecular biology and translational science*, 169, 43-83.
- Santos, J., et al. (2020c) pH-Dependent Aggregation in Intrinsically Disordered Proteins Is Determined by Charge and Lipophilicity, *Cells*, 9, E145.
- Santos, J., et al. (2020d) DispHred: A Server to Predict pH-Dependent Order-Disorder Transitions in Intrinsically Disordered Proteins, *International journal of molecular sciences*, 21, 5814.
- Uversky, V.N. (2009) Intrinsically disordered proteins and their environment: effects of strong denaturants, temperature, pH, counter ions, membranes, binding partners, osmolytes, and macromolecular crowding, *The protein journal*, 28, 305-325.
- Hatos, A., et al. (2020) DisProt: intrinsic protein disorder annotation in 2020, *Nucleic Acids Res*, 48, D269-D276.
- Zamora, W.J., Campanera, J.M. and Luque, F.J. (2019) Development of a Structure-Based, pH-Dependent Lipophilicity Scale of Amino Acids from Continuum Solvation Calculations, *The journal of physical chemistry letters*, 10, 883-889.

4.3 DispHred: A server to predict pH-dependent order-disorder transitions in intrinsically disordered proteins.

Valentín Iglesias^{1†}, Jaime Santos^{1†}, Carlos Pintado¹, Juan Santos-Suárez¹ and Salvador Ventura^{1,*}

¹ Institut de Biotecnologia i Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain;

† These authors contributed equally to this work. *To whom correspondence should be addressed.

Author Contribution: Conceptualization, software, validation, data curation, writing—original draft preparation.

4.3.1 ABSTRACT

The natively unfolded nature of intrinsically disordered proteins (IDPs) relies on several physicochemical principles, of which the balance between a low sequence hydrophobicity and a high net charge appears to be critical. Under this premise, it is well-known that disordered proteins populate a defined region of the charge-hydrophobicity (C-H) space and that a linear boundary condition is sufficient to distinguish between folded and disordered proteins, an approach widely applied for the prediction of protein disorder. Nevertheless, it is evident that the C-H relation of a protein is not unalterable but can be modulated by factors extrinsic to its sequence. Here, we applied a C-H based analysis to develop a computational approach that evaluates sequence disorder as a function of pH, assuming that both protein net charge and hydrophobicity are dependent on pH solution. On that basis, we developed DispHred, the first pH-dependent predictor of protein disorder. Despite its simplicity, DispHred displays very high accuracy in identifying pH-induced order/disorder protein transitions. DispHred might be useful for diverse applications, from the analysis of conditionally disordered segments to the rational engineering of disordered proteins for diverse biotechnological applications. Importantly, since many disorder predictors use hydrophobicity as an input, the here developed framework can be implemented in other state-of-the-art algorithms.

Availability and Implementation: The DispHred web server is freely available for academic users at: <https://ppmclab.pythonanywhere.com/DispHred>. It is platform-independent and does not require previous registration.

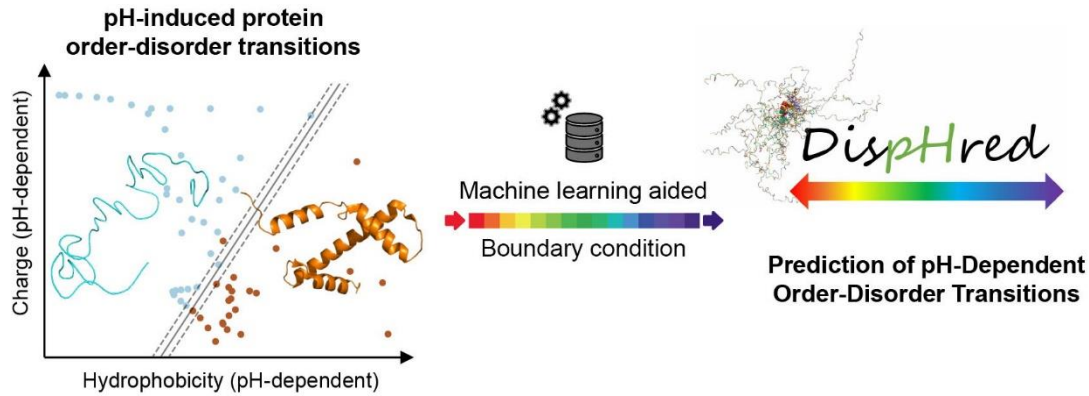


Figure 4.10 – Graphical Abstract: DispHred web server predicts pH dependant conditional disorder on IDPs. It is based on applying pH-dependence to the C-H phase diagram and utilizing a machine learning strategy to delimit the boundary condition.

4.3.2 INTRODUCTION

Intrinsically disordered proteins (IDPs) are a class of polypeptides that do not require a defined folded structure to execute their biological activities (Chen and Kriwacki, 2018; Dunker and Obradovic, 2001; Kulkarni and Kulkarni, 2019). The plasticity of these biomolecules allows them to interact with structurally diverse partners, and they are often involved in the wiring of protein networks, acting both as central hubs and as molecular switches (Wright and Dyson, 2015). The unfolded nature of IDPs is intrinsically encoded in their primary sequence, which is generally enriched in ionizable and polar residues and depleted of hydrophobic amino acids (Dyson, 2016). Thus, IDPs' extended conformation rely both on electrostatic repulsions between uncompensated charges and on a low hydrophobicity load, which prevents extensive protein compaction (Uversky, et al., 2000).

Based on the balance between attractive and repulsive forces in IDPs, Uversky and co-workers proposed that they populated a distinct region in the mean net charge-hydrophobicity (C-H) phase space diagram, and showed that by dividing this space with an empirically-obtained boundary line it was possible to discriminate between folded and disordered proteins (Uversky, et al., 2000). Under that premise, the disordered nature of a polypeptide sequence can be predicted by evaluating its C-H relationship in the aforementioned attraction-repulsion scheme. The C-H plot analysis has been applied for disorder prediction, it lies behind the popular FoldIndex algorithm (Prilusky, et al., 2005), and it is also computed by other multiparametric software (He, et al., 2009).

More than 50 prediction methods, based on different molecular principles, have been developed to assess protein disorder, thus providing a robust toolbox for identifying natively unfolded proteins or their regions (Dosztanyi, 2018; He, et al., 2009; Lieutaud, et al., 2016). Besides, new tools able to reverse-engineer the above-mentioned principles into a sequence allow now for the artificial design of disordered protein segments (Harmon, et al., 2016; Schramm, et al., 2017). Nevertheless, most of these methods are blind to the protein context, even if IDPs are extremely sensitive to environmental fluctuations (Jakob, et al., 2014; Uversky, 2009). Ligands, binding partners, or solvent conditions such as

ions concentration or pH, have been reported to induce conditional folding in IDPs (Fonin, et al., 2019; Smith and Jelokhani-Niaraki, 2012; Uversky, et al., 2000). Therefore, it is surprising to find out that those effects have been mostly disregarded in state-of-the-art computational approaches. Indeed, it becomes evident that the C-H relationship of a given protein is not constant since both protein net charge and hydrophobicity can be modulated by factors that are extrinsic to the protein sequence.

In a recent work addressed in **Section 4.1**, we showed that the solution's pH effect on IDPs solubility is not restricted to its effect on the charge of ionizable residues since the pH also modulates the sequence hydrophobicity, a traditionally neglected effect. Driven by this simple idea, we revisit here the C-H concept, on the evidence that both protein net charge and hydrophobicity are dependent on pH. By delineating a boundary condition similar to the one described by Uversky (Uversky, et al., 2000), we demonstrate that IDPs' pH-induced folding can be predicted just by evaluating the pH dependence of the C-H space diagram. This allowed us to develop DispHred, a first computational approach to predict protein disorder as a function of the pH. DispHred is freely available for academic users at <https://ppmclab.pythonanywhere.com/DispHred>. We envision the data presented here may prompt the development of a new generation of disorder predictors that include solvent conditions on their pipelines.

4.3.3 MATERIALS AND METHODS

Data collection.

The dataset of 111 experimentally verified fully disordered proteins was obtained from the Disprot database (DisProt 2020_06) (Hatos, et al., 2020) by selecting proteins with a 100% disorder coverage. The set of 150 fully folded sequences was randomly extracted from the Protein Data Bank (PDB) under the query single-chain structures larger than 100 residues and determined by X-ray crystallography.

Data regarding the effect of pH on protein disorder was extracted from the bibliography. Data regarding the pH-dependent folding of prothymosin was obtained from the characterization of Uversky and co-workers (Uversky, et al., 1999). Order-disorder pH-transition of the PEST region (201-268) from human c-Myc oncoprotein was analysed in Ansari and Swaminathan study (Ansari and Swaminathan, 2020). LL-37 pH-dependent helix formation was reported by Johansson and co-workers (Johansson, et al., 1998). Victor Muñoz and Luis Serrano reported the effect of solution pH on a model peptide Ac-AKAAKAKAAKAKAAKA-NH₂ (Munoz and Serrano, 1995). Data on the pH-modulated collapse of human histones were extracted from Munishkina and co-workers (Munishkina, et al., 2004). The analysis of the disordered A-domain of the Toc132 receptor disorder was performed by Lynn GL Richardson, Masoud Jelokhani-Niaraki, and Matthew D Smith (Richardson, et al., 2009). The conformational fluctuations of the 36-loop region of the influenza hemagglutinin were analysed by Chavela M. Carr and Peter S. Kim (Carr and Kim, 1993).

DispHred: Evaluation of hydrophobicity and charge as a function of pH.

To analyse the lipophilicity of protein sequences, we employed the pH-dependent lipophilicity scale developed by Zamora and co-workers (Zamora, et al., 2019). They used continuum solvation calculations, which allow us to calculate the hydrophobicity of a given residue at the desired pH. Then, DispHred uses a sliding window with a user-defined length to calculate the average hydrophobicity in the window and assigns it to the residue in the center. In the analysis performed in this article, we used a fixed window of 7 residues. The results are averaged to calculate the mean hydrophobicity of the sequence at the analysed pH.

Protein NCPR is calculated by applying the Henderson-Hasselbalch equation to derive the partial charge of each ionizable residue at the analysed pH. Then, global NCPR is calculated as the sum of all partial charges divided by the protein length. To calculate the Dis_{pH} score of a given window, the NCPR is calculated using the residues included in this particular window and its length.

Hydropathy scales performance analysis at neutral pH.

We delineated a C-H plot for each of the analysed hydropathy scales. Each scale was normalized from 0 to 1 according to the increased hydrophobicity of the protein residues; for the pH-dependent scale, we employed the values calculated at pH 7.0 (Zamora, et al., 2019). The performance of the different scales was evaluated using a ROC analysis, in which the true-positive rate is plotted against the false-positive rate. The ROC analysis was performed against a dataset of 111 fully disordered proteins and 150 single-chain folded proteins. The AUC was taken as an reporter of sensitivity and sensibility.

Support vector machine analysis

SVM was applied to define the optimal boundary line delimitating two classes of samples as folded or disordered. NCPR and pH-dependent hydrophobicity were calculated as previously stated for the 59 data points. Experimental data was labeled as ordered or disordered as described in the literature and employed for the machine learning process. To perform the analysis, we used the freely available machine learning library scikit-learn for Python (Pedregosa, et al., 2011). SVM kernel was set to “linear” to map the data on a two-dimensional space.

DispHred: Prediction of sequence disorder.

DispHred uses a C-H plot analysis to discriminate between folded and disordered sequences at the analysed pH by applying a defined boundary condition. For each pH, the mean hydrophobicity ($\langle H_{pH} \rangle$) and the absolute value of the NCPR are calculated. Then, the Dis_{pH} score is obtained by applying the SVM derived **Equation 4.2**. Positive and negative values are classified as folded or disordered, respectively. DispHred calculates the Dis_{pH} score at all the pHs in the desired range to profile sequence disorder as a function of pH. DispHred also analyses the Dis_{pH} score of the sliding windows to identify specific stretches whose disorder is affected by pH.

Performance analysis.

The sensitivity, specificity, precision, accuracy and false discovery rate when predicting order-disorder transitions was evaluated as follows: Sensitivity = $TP/(TP + FN)$; Specificity = $TN/(TN + FP)$; Precision = $TP/(TP + FP)$; Accuracy = $(TP + TN)/(TP + TN + FP + FN)$; and False Discovery Rate = $FP/(FP + TP)$. F1 Score and Matthews Correlation Coefficient were calculated as previously described in (Chicco and Jurman, 2020). TP, TN, FP and FN correspond to true positives, true negatives, false positives and false negatives, respectively.

DispHred web server

DispHred web server interface was built in HTML/CSS/JavaScript. It uses the Django 3.0 framework working with Python 3.7. The figures are generated using matplotlib library (Hunter, 2007). The server is platform-independent, free and open for academic users. It does not require previous login.

4.3.4 RESULTS

Validation of a pH-dependent hydrophathy scale for C-H plot-based predictions.

The original C-H analysis was developed using the Kyte-Doolittle hydrophathy scale to calculate the mean hydrophobicity of protein sequences (Kyte and Doolittle, 1982; Prilusky, et al., 2005; Uversky, et al., 2000). Here, we implement a novel amino acid pH-dependent hydrophathy scale developed by Zamora and co-workers (Zamora, et al., 2019), based on implicit solvation calculations, that allow us to evaluate the effect of the solution pH on sequence hydrophobicity. As a first step in developing our approach, we assessed the performance of this pH-dependent scale for C-H plot-based order-disorder predictions at neutral pH. Uversky and Dunker performed an extensive analysis of 19 diverse hydrophathy scales to compare their performance in C-H plot-based predictions (Huang, et al., 2014). They reported that the Guy hydrophathy scale (Guy, 1985) had the highest discriminative power, while Kyte-Doolittle performance was in the average of the 19 scales. Additionally, they developed a new scale that provided the best order-disorder discrimination (IDP-Hydrophathy) (Huang, et al., 2014).

We compared the pH-dependent hydrophathy (pH-dependent) scale with the Kyte-Doolittle, Guy, and IDP-Hydrophathy scales. First, we normalized the four scales between 0 and 1, assigning a value of 1 to the highest hydrophobicity. Then we calculated the values for the pH-dependent scale at pH 7.0. We found the highest correlation with the Guy scale ($R^2 = 0.72$), followed by the Kyte-Doolittle ($R^2 = 0.60$) and the IDP-hydrophathy ($R^2 = 0.51$) scales (**Figure 4.11A-C**). The correlation between Guy and Kyte-Doolittle scales is $R^2 = 0.78$. In contrast, as it happens for the pH-dependent scale, the correlation between the IDP-hydrophathy and the Guy or the Kyte-Doolittle scales is low, with $R^2 = 0.52$ and $R^2 = 0.33$, respectively. These low correlations stem mostly from the fact that, counter-intuitively, the IDP-hydrophathy scale considers P as the most hydrophilic residue, with a value of 0 in our normalized scale. Removing P from the correlation between the pH-dependent and IDP-hydrophathy scales increases R^2 to

0.70 and arbitrarily assigning this residue a value of 0 in the pH-dependent scale (pH-P-corrected scale) results in an $R^2 = 0.74$ (Supplementary Material S4.5A).

We next ensembled a dataset of 111 experimentally validated fully disordered proteins and 150 folded single-chain proteins with X-ray resolved structures to test the discriminatory power of the four scales in a C-H plot analysis. The ability to classify ordered and disordered sequences of each scale was assessed by applying a Receiver Operating Characteristic (ROC) method. The associated area under the curve (AUC) was used as a sensitivity-specificity reporter. The pH-dependent and the Kyte-Doolittle scales showed an identical discriminatory potential (AUC = 0.91), while the Guy and IDP-hydropathy scales demonstrated slightly higher performances (AUC = 0.94 and 0.98, respectively) (**Figure 4.11D**). The pH-P-corrected scale exhibited an AUC = 0.95 (Supplementary Material S4.5B), which suggests that the minimal value assigned to P in the IDP-hydropathy scale contributes to its higher discrimination.

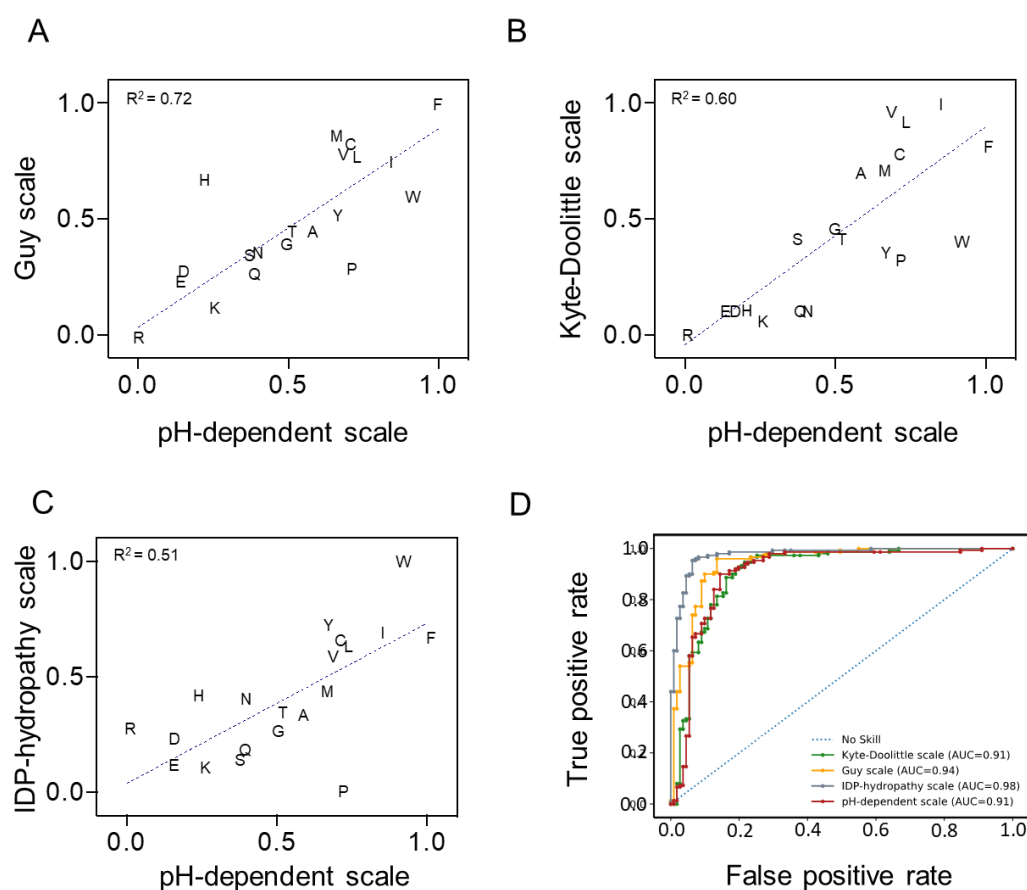


Figure 4.11 – Comparison of four different hydropathy scales. Correlation between pH-dependent scale and **A**) Guy, **B**) Kyte-Doolittle, and **C**) IDP-hydropathy scales. Amino acids are represented in their one-letter code. Hydropathy values are normalized between 0 and 1, corresponding to the minimum and maximum values for each scale. The R^2 value of the linear regression is shown in each graph. **D**) ROC curves showing the performance of the four scales in discriminating a dataset of fully disordered (n=111) and single-chain folded (n=150) proteins. Blue dotted represents no skill (AUC=0.50), green Kyte-Doolittle (AUC=0.91), yellow Guy (AUC=0.94), grey IDP-hydropathy (AUC=0.98) and brown pH-dependent (AUC=0.91) scales.

Overall, the analysis suggested that the pH-dependent scale compared well with the other analysed scales at pH 7.0, with a discriminatory power identical to the widely employed Kyte-Doolittle scale. Thus, this scale will allow us to extend the C-H predictive potential to the full pH scale without compromising the performance at neutral pH significantly. Despite its higher discrimination, we preferred not using the pH-P-corrected scale and keep the hydrophathy value obtained from implicit solvation calculations for P residues (Zamora, et al., 2019).

C-H space phase diagram and order-disorder boundary condition can anticipate pH-induced order-disorder transition of IDPs.

Next, we explored whether the C-H model would be a reliable tool to predict the pH-dependent order-disorder transition in IDPs. To that end, we performed a bibliographic search of structural data on IDPs that suffer a conditional folding at specific pHs. We collected 59 bibliographic pH datapoints for 7 disordered proteins and peptides (**Figure 4.12** and Supplementary Material S4.6). For each point, we calculated the protein net charge per residue (NCPR) and protein mean hydrophobicity at the given pH $\langle H_{pH} \rangle$. NCPR is calculated using the Henderson-Hasselbach equation, and $\langle H_{pH} \rangle$ is computed according to the pH-dependent scale developed by Zamora and co-workers (Zamora, et al., 2019). We plotted each datapoint in a 3-axis scatter plot according to its pH, $\langle H_{pH} \rangle$, and NCPR, employing a colour-code to indicating whether the protein was folded or disordered in this condition (**Figure 4.12**).

To develop a consistent C-H based order-disorder classification for the experimental data, we sought to seek the order-disorder boundary condition that allowed the maximal separation between the two states. Since the datasets for the different proteins diverged in size, nature, and source, we assumed that a classic iterative analysis might lead to overfitting and/or result in a biased boundary condition in case some data points were misclassified.

To minimize such limitations, we applied a support vector machine (SVM) learning strategy, a supervised feedforward network specifically designed to build a binary classifier and retrieve the boundary condition that maximizes the separation between observations (Vapnik, 1998; Vapnik, 2013). SVM-based analysis reduces overfitting and tolerates a certain degree of misclassified data points without forcing a bias, being robust classification strategies, and increasing their predictive potential when applied to new observations, especially near the boundary condition. Additionally, since SVM analysis takes into account a slight uncertainty and misclassification, it also provides a margin near the boundary line (Supplementary Material S4.7) that can be used as a confidence interval in a subsequent classification of new data points in predictive applications.

By using the above-described SVM-based analysis, we identified a linear boundary condition defined by **Equation 4.2**,

$$Dis_{pH} = 2.775 \langle H_{pH} \rangle - |NCPR| - 1.118 \quad (4.2)$$

that successfully discriminates between folded and disordered proteins with a Matthews Correlation Coefficient of 0.97 (Supplementary Material S4.7A, **Table 4.2**). Note that our boundary condition for order-disorder classifications is reasonably similar to that previously defined by Uversky and colleagues at neutral pH (**Equation 4.3**) (Uversky, et al., 2000):

$$I = 2.785 \langle H \rangle - |\langle R \rangle| - 1.151 \quad (4.3)$$

$\langle H \rangle$ and $\langle R \rangle$ corresponding the mean hydrophobicity and mean charge at neutral pH, respectively.

In contrast, applying the same SVM analysis but considering that hydrophobicity is independent of pH, we did not observe a consistent classification of the datapoints -Matthews Correlation Coefficient of 0.6- neither the boundary line satisfies the C-H relationship (Supplementary Material S4.7B, Supplementary Material S4.8).

As shown in **Figure 4.12A**, the boundary plane defined by **Equation 4.2**, satisfactorily delimited folding-unfolding transitions for the analysed IDPs, with only one datapoint wrongly predicted but still reasonably close to the boundary. This translates into 98 % accuracy in predicting the proteins' conformational states at any given pH (**Table 4.2**). On the contrary, by considering that hydrophobicity is independent of pH (and computing its value at pH 7.0 and under the same boundary condition **Equation 4.2**), we observed that the NCPR change alone could not discriminate between folded and disordered sequences (**Figure 4.12B** and **Table 4.2**). This observation evidences the importance of modeling the pH-dependent hydrophobicity when predicting protein disorder.

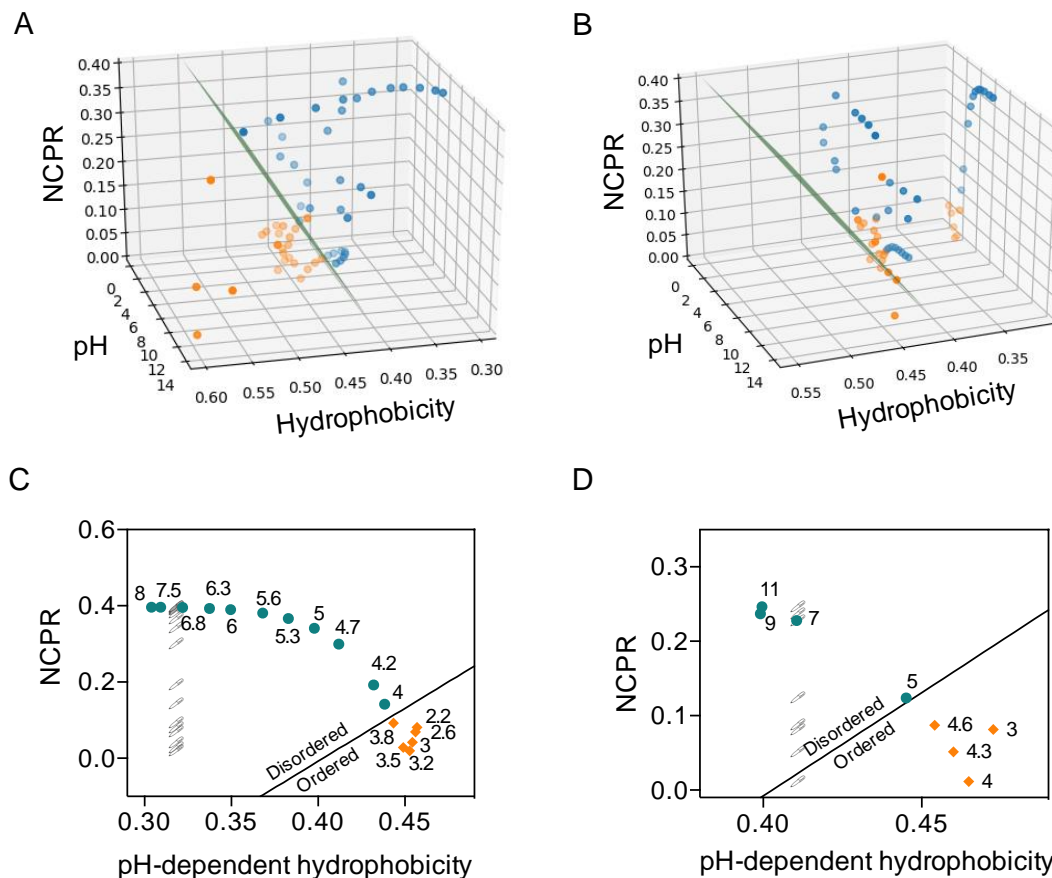


Figure 4.12 – C-H based analysis of pH modulated order-disorder transitions. 3-dimensional C-H plots containing 59 datapoints of 7 proteins at different pHs, computing pH influence over **A)** sequence NCPR and hydrophobicity, or **B)** assuming constant hydrophobicity values (as calculated at pH 7.0). Blue and orange points correspond to conditions in which protein/peptides are disordered and folded, respectively. The green surfaces delimit the boundary conditions between folded and disordered proteins as defined in Equation 4.2. **C-D)** Two-dimensional C-H plots of **C)** prothymosin and **D)** PEST-c-myc using the same color pattern than in panels A and B for folded-unfolded datapoints. A solid line represents the boundary condition. Open circles represent the same data points assuming constant hydrophobicity values (as calculated at pH 7.0).

Prothymosin is a classic example of an IDP at neutral pH which experiences a conditional folding at lower pHs, characterized by the gain of α -helical structure (Uversky, et al., 1999). The transition occurs between pH 3.5 and pH 5.0, with prothymosin being fully folded below pH 3.5 and fully unfolded above pH 5.0. In a two-dimensional projection of the data points for this protein, we can observe that all folded points fall below the boundary line, being thus accurately predicted (**Figure 4.12C**). We also observed that our pH-dependent C-H representation also succeeds in delineating the transition range (pH 3.5-5). Similarly, the disordered PEST region (201-268) from human c-Myc oncoprotein collapses into a folded conformation at pHs below 4.8 (Ansari and Swaminathan, 2020), a transition that is successfully identified by our pH-dependent C-H ratio (**Figure 4.12D**). Note that the same analysis considering a constant hydrophobicity is blind to these structural conversions (open circles in **Figure 4.12C** and **D**). The same trend can be observed in the two-dimensional C-H plots of the other 5 protein sets in **Figures 4.12A** and **4.12B** (Supplementary Material S4.6).

Table 4.2 – Performance of pH-dependent and pH-independent hydrophobicity approaches in predicting pH-conditioned order-disorder transitions in a C-H analysis by applying **Equation 4.2**. Unfolded sequences correctly predicted to be unfolded were classified as true positives. The highest values for each measure are indicated in bold.

	pH-dependent hydrophobicity	pH-independent hydrophobicity
Sensitivity	1.00	1.00
Specificity	0.96	0.21
Precision	0.97	0.65
False Discovery rate	0.03	0.35
Accuracy	0.98	0.68
F1 Score	0.99	0.79
Matthews Correlation Coefficient	0.97	0.37


The presented data demonstrates that the effect of pH on IDPs conditional folding can be successfully predicted by applying a pH-dependent C-H analysis. With these results in hand, we aimed to develop a computational tool for predicting protein disorder that considers implicitly the solution pH, which we named DispHred.

Rationale and implementation of DispHred, a pH-dependent predictor of sequence disorder.

DispHred uses the C-H space diagram analysis proposed by Uversky and co-workers and later implemented in FoldIndex (Prilusky, et al., 2005; Uversky, et al., 2000). Nevertheless, instead of considering constant net charges and hydrophobicity for each analysed sequence, DispHred assumes that the solution pH modulates both parameters. Thus, DispHred computes the protein NCPR and the mean hydrophobicity of a sequence as a function of pH. Then, DispHred applies the boundary condition defined by **Equation 4.2** to separate folded and disordered proteins. Dis_{pH} positive values correspond to sequences classified as folded and negative values to those classified as disordered at the analysed pH or pH range. The SVM approach provides a margin of ± 0.02 around the boundary line used as a confidence interval in the classification.

DispHred calculates the Dis_{pH} score for all the analysed pHs, profiling the pH-dependence disorder of a protein sequence, and thus including the pH dimension in the classical C-H phase diagram. DispHred runs a user-defined sliding window that enables the analysis of the folded/disordered regions in a protein sequence at every requested pH. Sequence stretches fall in three classes: i) regions that are predicted to be always folded in the analysed pH interval, ii) regions that are predicted to be always disordered in this pH interval, and iii) regions whose folded/disordered conformation is modulated by the pH.

A



UAB
Universitat Autònoma de Barcelona

DispHred is a protein disorder predictor over pH for IDPs/IDRs. Users can introduce or upload a fasta sequence for analysis or try the server provided example.

Submission Help References Contact

Insert Uniprot Accession number:

Or paste a [FASTA formatted](#) sequence.

Example

```

Human Prothymosin alpha
MSDAADVTSS EITTKDLKEK KEVVEEAENG RDAPANGNAE NEENGEQEA DNEVDEEEEG GEEEGGDEEAE SATGKRAAID D
EEDVDTKKQKTDED
  
```

Window size: 51 From pH: 1 To pH: 9 pH interval: 0.5

Predict disorder at specific pH:

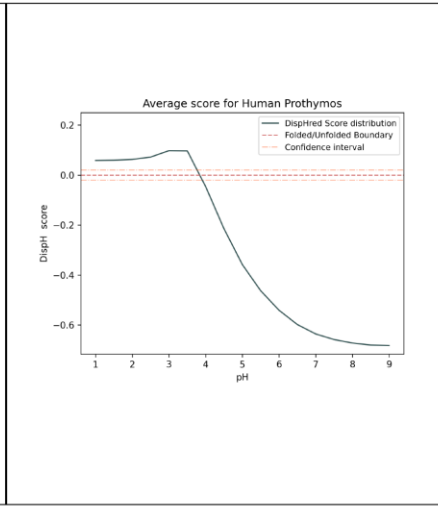
B

Prediction completed successfully

[Visualize results in json format](#)
[Download all results \(.zip\)](#)

Input summary:
Starting pH: 1
Final pH: 9
pH interval: 0.5
Window size: 51

pH	DispH Score	Hydrophobicity	NCPR
1.0	0.058	0.456	0.091
1.5	0.059	0.456	0.092
2.0	0.063	0.456	0.096
2.5	0.072	0.455	0.107
3.0	0.097	0.453	0.137
3.5	0.097	0.448	0.207
4.0	-0.046	0.437	0.319
4.5	-0.215	0.42	0.44
5.0	-0.358	0.397	0.521



pH-dependent disorder:

- Predicted as disordered in the selected pH range
- Predicted as folded in the selected pH range
- Predicted as pH-dependent disorder

```

1 MSDAAVD TSS EITTKDLKEK KEVVEEAENG RDAPANGNAE NEENGEQEA DNEVDEEEEG
61 GEEEEEEEG DGEEEDGDED EEAESATGKR AAEDDED DV DTKKQKTDED D
  
```

Figure 4.13 – DispHred web server interface. A) Web input page. The user can paste their FASTA-formatted sequence or insert a valid UniprotKB Accession number. DispHred works by default by checking disorder in a range of pHs but allows users to test values at a specific pH. By default, a 51-residue sliding window is populated, but users can personalize its length. **B)** Web results page for a selected range of pHs. Two clickable links appear on the upper left part of the screen with a JSON file or a ZIP file containing DispHred calculations and generated figures. On the central left part, a table shows the DispHred, hydrophobicity, and NCPR average scores for each pH. Clicking each pH will open a figure representing the Dis_{pH} score variation along the sequence for the selected pH. On the right, a figure representing the Dis_{pH} average score for each pH is shown. Scores above the red dashed line indicate predicted order. On

the bottom of the screen folded, disordered, and conditionally disordered regions for the pH interval are indicated in the sequence in green, red and blue respectively.

DispHred is free for academic users and does not require login. DispHred is available at <https://ppmclab.pythonanywhere.com/DispHred>. In the input page the user can (i) introduce a sequence in FASTA format or insert a valid UniprotKB Accession number, (ii) select the pH range and step size for the analysis or type a single specific pH and (iii) select the sliding window size (**Figure 4.13A**). After running the program, the user will be redirected to a results page containing the report of the analysis (**Figure 4.13B**): Dis_{pH} scores, mean hydrophobicity, and NCPR for each of the analysed pHs, a graph showing Dis_{pH} score as a function of pH, and clickable links that redirect to the sequence profile prediction at each desired pH. The protein regions exhibiting pH-dependent and pH-independent folded/disordered conformations are colored on top of the input sequence.

Users can retrieve all data in a JavaScript Object Notation (JSON) file or download all the generated data in a compressed ZIP file. A clickable example is provided in the input page to illustrate DispHred outputs.

4.3.5 DISCUSSION

Structural disorder is a fundamental trait of protein biology that complements the activities of structured proteins and domains by contributing flexibility and plasticity (Babu, et al., 2011; Oldfield and Dunker, 2014; Tompa, 2012). In contrast to folded proteins, IDPs exist as ensembles sampling a wide range of dynamic conformations in which the bulk of the primary sequence is highly exposed to the solvent. Accordingly, IDPs' properties display little dependence on structural elements and can be inferred from the primary sequence, which has allowed the design of computational tools for predicting, designing, and analyzing protein disorder (He, et al., 2009; Lieutaud, et al., 2016; Schramm, et al., 2017). At the same time, IDPs are extremely sensitive to environmental conditions; an effect often disregarded in predictive approaches. Among the different parameters that may affect IDPs properties, the solution pH has a significant impact, mainly due to the high prevalence of ionizable residues in these polypeptides (Payliss, et al., 2019; Santos, et al., 2020c; Smith and Jelokhani-Niaraki, 2012; Uversky, 2009). In this work, we demonstrated that the effect of pH on the disordered nature of a protein sequence can be easily predicted by evaluating the changes in protein charge and hydrophobicity as a function of this parameter. Even if the effect of pH over net charge is well-recognized, hydrophobicity is usually considered to be constant, disregarding its pH-dependence. However, we found that the evaluation of the pH-dependent hydrophobicity is fundamental for the accuracy of the order/disorder prediction in any given condition.

The analysis of the local or global hydrophobicity of protein sequences is a pivotal stage in many *in silico* pipelines aimed to predict protein disorder and its associated properties. A significant number of disorder predictors, such as FoldIndex or PONDR, rely on the direct or indirect analysis of hydrophobicity, a property that is also used to predict folding upon binding, RNA- DNA- interactions or post-translational modification sites in IDPs (Garner, et al., 1999; He, et al., 2009; Iakoucheva, et al., 2004; Lieutaud, et al.,

2016; Meng, et al., 2017; Prilusky, et al., 2005; Ward, et al., 2004; Xue, et al., 2010). Thus, the identification of hydrophathy scales suitable for such analyses attracted significant attention (Huang, et al., 2014). Our results indicate that by applying a recently developed pH-dependent hydrophathy scale, the contribution of this predictive physicochemical property to disorder prediction can be extended to the full pH scale. Thus, the implementation of pH-dependent hydrophathy scales, like the one used here, may increase applicability in currently available algorithms.

pH, ion concentrations, redox state, or post-translational modifications are known regulators of protein function by controlling the switch between the disordered and folded or partially folded states of polypeptides. Thus, although the conditional disorder's prediction is a challenging task, it is fundamental to elucidate the functionality of IDPs (Bardwell and Jakob, 2012; Jakob, et al., 2014). To advance in this direction, we developed DispHred, an online web server that exploits the C-H space analysis to predict protein disorder as a function of pH. Its main application is the profiling of protein disorder across a continuous pH interval, for which it demonstrates a high accuracy in classifying the pH-modulated order-disorder transitions for sequentially unrelated model proteins and peptides. Additionally, DispHred allows the assessment of the specific protein regions contributing the most to conditional disorder.

In essence, DispHred is the first disorder predictor dedicated to evaluating the effect of the solution pH and constitutes a proof-of-concept for the implementation of this kind of approach in future predictive endeavors. Intrinsically disorder tags are increasingly used to solubilize proteins and to engineer the pharmacological properties of protein and peptide pharmaceuticals (Minde, et al., 2013). We envision that DispHred can be of significant help in these and other biotechnological tasks.

4.3.6 REFERENCES

- Ansari, M.Z. and Swaminathan, R. (2020) Structure and dynamics at N- and C-terminal regions of intrinsically disordered human c-Myc PEST degron reveal a pH-induced transition, *Proteins*, **88**, 889-909.
- Babu, M.M., et al. (2011) Intrinsically disordered proteins: regulation and disease, *Curr Opin Struct Biol*, **21**, 432-440.
- Bardwell, J.C. and Jakob, U. (2012) Conditional disorder in chaperone action, *Trends Biochem Sci*, **37**, 517-525.
- Carr, C.M. and Kim, P.S. (1993) A spring-loaded mechanism for the conformational change of influenza hemagglutinin, *Cell*, **73**, 823-832.
- Chen, J. and Kriwacki, R.W. (2018) Intrinsically Disordered Proteins: Structure, Function and Therapeutics, *J Mol Biol*, **430**, 2275-2277.
- Chicco, D. and Jurman, G. (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, *BMC Genomics*, **21**, 6.
- Dosztanyi, Z. (2018) Prediction of protein disorder based on IUPred, *Protein Sci*, **27**, 331-340.
- Dunker, A.K. and Obradovic, Z. (2001) The protein trinity--linking function and disorder, *Nat Biotechnol*, **19**, 805-806.
- Dyson, H.J. (2016) Making Sense of Intrinsically Disordered Proteins, *Biophys J*, **110**, 1013-1016.
- Fonin, A.V., et al. (2019) Folding of poly-amino acids and intrinsically disordered proteins in overcrowded milieu induced by pH change, *Int J Biol Macromol*, **125**, 244-255.
- Garner, E., et al. (1999) Predicting Binding Regions within Disordered Proteins, *Genome Inform Ser Workshop Genome Inform*, **10**, 41-50.

- Guy, H.R. (1985) Amino acid side-chain partition energies and distribution of residues in soluble proteins, *Biophys J*, **47**, 61-70.
- Harmon, T.S., *et al.* (2016) GADIS: Algorithm for designing sequences to achieve target secondary structure profiles of intrinsically disordered proteins, *Protein Eng Des Sel*, **29**, 339-346.
- Hatos, A., *et al.* (2020) DisProt: intrinsic protein disorder annotation in 2020, *Nucleic Acids Res*, **48**, D269-D276.
- He, B., *et al.* (2009) Predicting intrinsic disorder in proteins: an overview, *Cell Res*, **19**, 929-949.
- Huang, F., *et al.* (2014) Improving protein order-disorder classification using charge-hydrophathy plots, *BMC Bioinformatics*, **15 Suppl 17**, S4.
- Hunter, J. (2007) Matplotlib: A 2D Graphics Environment, *Computing in Science & Engineering*, **9**, 90-95.
- Iakoucheva, L.M., *et al.* (2004) The importance of intrinsic disorder for protein phosphorylation, *Nucleic Acids Res*, **32**, 1037-1049.
- Jakob, U., Kriwacki, R. and Uversky, V.N. (2014) Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function, *Chem Rev*, **114**, 6779-6805.
- Johansson, J., *et al.* (1998) Conformation-dependent antibacterial activity of the naturally occurring human peptide LL-37, *J Biol Chem*, **273**, 3718-3724.
- Kulkarni, V. and Kulkarni, P. (2019) Intrinsically disordered proteins and phenotypic switching: Implications in cancer, *Prog Mol Biol Transl Sci*, **166**, 63-84.
- Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydrophatic character of a protein, *J Mol Biol*, **157**, 105-132.
- Lieutaud, P., *et al.* (2016) How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe, *Intrinsically Disord Proteins*, **4**, e1259708.
- Meng, F., Uversky, V.N. and Kurgan, L. (2017) Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions, *Cell Mol Life Sci*, **74**, 3069-3090.
- Minde, D.P., Halff, E.F. and Tans, S. (2013) Designing disorder: Tales of the unexpected tails, *Intrinsically disordered proteins*, **1**, e26790.
- Munishkina, L.A., Fink, A.L. and Uversky, V.N. (2004) Conformational prerequisites for formation of amyloid fibrils from histones, *J Mol Biol*, **342**, 1305-1324.
- Munoz, V. and Serrano, L. (1995) Elucidating the folding problem of helical peptides using empirical parameters. III. Temperature and pH dependence, *J Mol Biol*, **245**, 297-308.
- Oldfield, C.J. and Dunker, A.K. (2014) Intrinsically disordered proteins and intrinsically disordered protein regions, *Annu Rev Biochem*, **83**, 553-584.
- Payliss, B.J., Vogel, J. and Mittermaier, A.K. (2019) Side chain electrostatic interactions and pH-dependent expansion of the intrinsically disordered, highly acidic carboxyl-terminus of gamma-tubulin, *Protein Sci*, **28**, 1095-1105.
- Pedregosa, F., *et al.* (2011) Scikit-learn: Machine learning in Python, *the Journal of machine Learning research*, **12**, 2825-2830.
- Prilusky, J., *et al.* (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded, *Bioinformatics*, **21**, 3435-3438.
- Richardson, L.G., Jelokhani-Niaraki, M. and Smith, M.D. (2009) The acidic domains of the Toc159 chloroplast preprotein receptor family are intrinsically disordered protein domains, *BMC Biochem*, **10**, 35.
- Santos, J., *et al.* (2020) pH-Dependent Aggregation in Intrinsically Disordered Proteins Is Determined by Charge and Lipophilicity, *Cells*, **9**.
- Schramm, A., *et al.* (2017) InSiDDe: A Server for Designing Artificial Disordered Proteins, *Int J Mol Sci*, **19**.
- Smith, M.D. and Jelokhani-Niaraki, M. (2012) pH-induced changes in intrinsically disordered proteins, *Methods Mol Biol*, **896**, 223-231.
- Tomba, P. (2012) Intrinsically disordered proteins: a 10-year recap, *Trends Biochem Sci*, **37**, 509-516.

Uversky, V.N. (2009) Intrinsically disordered proteins and their environment: effects of strong denaturants, temperature, pH, counter ions, membranes, binding partners, osmolytes, and macromolecular crowding, *The protein journal*, **28**, 305-325.

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

Uversky, V.N., *et al.* (1999) Natively unfolded human prothymosin alpha adopts partially folded collapsed conformation at acidic pH, *Biochemistry*, **38**, 15009-15016.

Vapnik, V. (1998) *Statistical learning theory*. New York. Wiley.

Vapnik, V. (2013) *The nature of statistical learning theory*. Springer science & business media.

Ward, J.J., *et al.* (2004) The DISOPRED server for the prediction of protein disorder, *Bioinformatics*, **20**, 2138-2139.

Wright, P.E. and Dyson, H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation, *Nat Rev Mol Cell Biol*, **16**, 18-29.

Xue, B., *et al.* (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids, *Biochim Biophys Acta*, **1804**, 996-1010.

Zamora, W.J., Campanera, J.M. and Luque, F.J. (2019) Development of a Structure-Based, pH-Dependent Lipophilicity Scale of Amino Acids from Continuum Solvation Calculations, *The journal of physical chemistry letters*, **10**, 883-889.

5 Chapter III – Prediction of prion-like behaviour

5.1 PrionW: server for the prediction of glutamine/asparagine rich prion-like domains and their amyloid cores

Valentin Iglesias^{1†}, Rafael Zambrano^{1†}, Oscar Conchillo-Sole^{1†}, Ricard Illa¹, Frederic Rousseau², Joost Schymkowitz², Raimon Sabate³, Xavier Daura^{1,4} and Salvador Ventura^{1*}

¹ Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, 08193, Spain

² VIB Switch Laboratory and Department for Cellular and Molecular Medicine, KU Leuven, Leuven, Belgium

³ Institut de Nanociència i Nanotecnologia (IN²UB) and Departament de Físicoquímica, Facultat de Farmàcia, Universitat de Barcelona, Barcelona, Spain

⁴ Institut Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

†These authors contributed equally to this work. *To whom correspondence should be addressed.

Author Contribution: validation, data curation.

5.1.1 ABSTRACT

Prions are a particular type of amyloids with the ability to self-perpetuate and propagate *in vivo*. Prion-like conversion underlies important biological processes but is also connected to human disease. Yeast prions are the best understood transmissible amyloids. In these proteins, prion formation from an initially soluble state involves a structural conversion, driven, in many cases, by specific domains enriched in glutamine/asparagine (Q/N) residues. Importantly, domains sharing this compositional bias are also present in the proteomes of higher organisms, thus suggesting that prion-like conversion might be an evolutionary conserved mechanism. Previous work has shown that the identification and evaluation of the potency of amyloid nucleating sequences in putative prion domains allows discrimination of genuine prions. PrionW is a web application that exploits this principle to scan sequences in order to identify proteins containing Q/N enriched prion-like domains (PrLDs) in large datasets. When used to scan the complete yeast proteome, PrionW identifies previously experimentally validated prions with high accuracy. Users can analyse up to 10.000 sequences at a time. PrLD-containing proteins are identified and their putative PrLDs and soft-amyloid nucleating cores visualized and scored. The output files can be downloaded for further analysis.

Availability and Implementation: PrionW does not require previous registration and is freely available at: <http://bioinf.uab.cat/prionw/>.

5.1.2 INTRODUCTION

Prions are a class of proteins that can exist in at least two conformations, of which one is an amyloid state that is self-propagating and hence infectious as it can induce the conversion of identical protein sequences from the non-prion conformation to the amyloid state (Ashe and Aguzzi, 2012).

Although prions were discovered through the example of the mammalian pathogen PrP (Nystrom, et al., 2012), a host of functional prions have since been discovered, predominantly in fungi (Fowler, et al., 2006; Fowler, et al., 2007). Importantly, the distinction between prion proteins and other proteins capable of forming amyloids is blurring, notably in human diseases such as Alzheimer's or type-II diabetes, as it has been observed that amyloids of the proteins involved in these diseases are capable of cross-seeding amyloid formation of the soluble form of these proteins, both *in vitro* and *in vivo* lab conditions (Eisenberg and Jucker, 2012; Westermark and Westermark, 2010). Given that there is no epidemiological evidence that these amyloidogenic proteins are spreading in natural systems, the group has been called prion-like or 'prionoid' (Ashe and Aguzzi, 2012). This raises the question of what sequence determinants characterize a functional prion beyond mere amyloid propensity. A subset of prions, not including PrP, are multi domain proteins containing both globular domains and, usually, one Prion Domain (PrD) enriched in glutamine and asparagine (Q/N) residues that undergoes the structural rearrangement during prion conversion (Greenwald and Riek, 2010). Most known yeast prions, but not all, share this architecture. The sequence features of these PrDs overlap with those of intrinsically disordered regions (Malinowska, et al., 2013). It has been proposed that in contrast to the short stretches that are known to be sufficient to nucleate amyloid formation, Q/N based yeast prions have more diffuse nuclei, characterized by a large number of weak interactions between the side-chains of the PrD (Toombs, et al., 2010; Toombs, et al., 2012). However, we have demonstrated that the superimposition of an intrinsically disordered sequence region containing amyloid nucleating sequences in fact yields a more accurate classification of experimental prions from related Q/N-enriched sequences (Toombs, et al., 2010). In the current paper, we provide public access to our method by way of a web server.

5.1.3 METHOD

PrionW allows scanning individual protein sequences for the presence of Q/N rich PrLDs, as well as the scanning of large protein datasets (up to 10000 sequences) for proteomic analysis. The method behind PrionW assumes that in order to be a PrLD a protein sequence should fulfil the following requirements: a) contain a specific stretch with amyloid propensity, longer than classical amyloids, which we call soft-amyloids, able to selectively nucleate self-assembly into ordered, but brittle, amyloid structures, b) have a disordered structural context that readily permits self-assembly without requiring conformational unfolding and c) have an amino acid composition that allows the domain to be soluble at the physiological concentrations required for protein function yet display a basal amyloid propensity, to which N and Q residues would contribute significantly, promoting domain assembly in the presence of preformed amyloid seeds or when the concentration is increased.

PrionW analyses whether a given protein or protein fragment satisfies the above requirements in three sequential steps:

i) Identification of Disordered Regions (DRs) in protein sequences: PrionW analyses protein sequences to identify the presence of intrinsically disordered regions by implementing FoldIndex (Prilusky, et al., 2005)

with the default 51-amino acid window size. Only disordered segments of at least 60 contiguous residues are further evaluated, since this window size seems to suffice to attain a prion-like behaviour (Alberti, et al., 2009). When a protein contains two or more DRs, these regions are subsequently evaluated individually.

ii) Evaluation of Q/N enrichment: The proportion of Q+N residues in the detected DRs is calculated. The program moves through each individual sequence by single amino acid steps looking for the longer stretch of adjacent residues having a Q/N proportion equal or bigger than a given threshold. Again, these regions should be at least 60 residues long. The default is set at $\geq 25\%$ of Q/N residues, because the PrDs of most characterized yeast prions fulfil this requirement (Alberti, et al., 2009; Espinosa Angarica, et al., 2013). However, since Q/N enrichment for prion-like formation might change from organism to organism the user can select the minimum Q/N content. If the threshold is set to 0% the program will only search for disordered regions.

iii) Soft-amyloid core identification and scoring: The individual sequences fulfilling the requirements in steps 1 and 2 are further evaluated for the presence of a 21-residue long soft-amyloid core able to specifically nucleate its self-assembly according to the pWALTZ scoring function (Sabate, et al., 2015), an update of the scoring function in the well-established amyloid predictor WALTZ (Maurer-Stroh, et al., 2010). The default pWALTZ cut-off was set to 73.55, since this value provides the best accuracy for the discrimination of experimentally validated yeast PrDs from sequences displaying similar Q/N content but devoid of prionogenic potential (see Performance section and **Figure 5.1**). A lower cutoff can be useful to identify sequences in genomes with a basal prion propensity (Kim, et al., 2013). Accordingly, the user can select the pWALTZ cut-off in the 50-74 value range. pWALTZ values lower than 50.0 are not allowed as they do not permit discrimination of prion and non-prion sequences in the yeast dataset used for parameterization, since the accuracy of PrionW in this condition is below 0.5 (**Figure 5.1**). For a given protein sequence, the disordered Q/N rich region containing the highest-scoring soft-amyloid core is selected as the prion-like domain (PrLD) in this protein, as long as it passes the selected threshold.

5.1.4 PERFORMANCE

Yeast prions constitute ideal model systems to characterize prion-like behaviour. On the basis of compositional similarity to known prions, Lindquist's group used a hidden Markov model (HMM) to identify 100 prion candidates in the yeast genome (Alberti, et al., 2009). They scored 92 of them from 0 to 10 according to their performance in four different experimental assays for both amyloid and prion forming ability, higher scores indicating more prionogenic sequences. It turned out that in this, in principle, prion enriched set, only 13 % of the proteins scored ≥ 9 whereas 42 % scored ≤ 2 , demonstrating the extreme difficulty to discriminate real prions from non-prions when they all share a similar Q/N enriched compositional context. The predicted PrLDs of these proteins were used to build up a dataset in which we considered as non-prions (negatives) those sequences scoring ≤ 2 and being positive in one assay at maximum (39 sequences), because it means that they do not exhibit amyloid and prion forming

ability, and prions (positives) those domains being positive in all four assays and scoring ≥ 9 , with a total of 12 sequences, including the known prions New1, Rnq1, Swi1, Sup35 and Ure2 proteins (Supplementary Material S5.1). We speculated that the presence and the strength of short amyloid cores embedded in these PrLDs might account for their different prionogenic potential. This concept was implemented in the pWALTZ scoring function, allowing discrimination between positive and negative proteins in the above-mentioned 51-protein dataset with better accuracy than approaches based only on composition (Sabate, et al., 2015). However, despite its accuracy, a serious limitation of pWALTZ to analyse large protein datasets is that it needs to work on top of dissected putative PrLDs sequences, because the folded domains adjacent to these regions and, more generally, globular proteins usually contain one or more amyloid regions (Rousseau, et al., 2006), whose high aggregation potency blur any prediction. PrionW approaches this issue by considering the structural disorder and Q/N compositional bias characteristic of most yeast PrDs.

In our previous work, a 73.55 pWALTZ cut-off provided the best accuracy to discriminate prions from non-prions; however this value resulted from the analysis of the PrDs identified by the Lindquist's group HMM, which may or may not coincide with those sequences identified by PrionW on the basis of structural disorder and Q/N content for their further pWALTZ classification (see Methods). Thus, to parameterize PrionW, we analysed the 6719 proteins encoded in the *S. cerevisiae* S288c reference proteome for the presence of PrLDs using a fixed Q/N content $\geq 25\%$ and gradually increasing the pWALTZ cut-off from 35 to 90% in 0.1% steps. The accuracy of the method was calculated at each stringency level by evaluating the presence of positive and negative instances from the original 51-protein dataset in the returned proteome predictions (**Figure 5.1**). The best predictions were obtained with cut-offs ranging from 73.50 to 73.60, suggesting that the disordered Q/N rich domains identified by PrionW overlap significantly with the candidates identified using the HMM. 73.55 was selected therefore as the default pWALTZ value in PrionW. Using these default parameters PrionW returned a total of 61 predictions. They included 92% of the previously considered positives (11 sequences), only Puf4 being missing. In contrast, only 5% of the negative ones (2 sequences) were recovered. This corresponds to a sensitivity of 0.917, a specificity of 0.949, a precision of 0.846 an accuracy of 0.941 and a false discovery rate of only 0.154. These values (**Table 5.1**) indicate that our methodology produce fairly clean recovery sets with a rather low proportion of false positives. If we consider as positive sequences only the set of actual Q/N-rich prions: Cyc8, Mot3, New1, Rnq1, Sfp1, Swi1, Sup35 and Ure2, PrionW is able to recover the large majority of them from the yeast proteome with the default settings, missing only Cyc8.

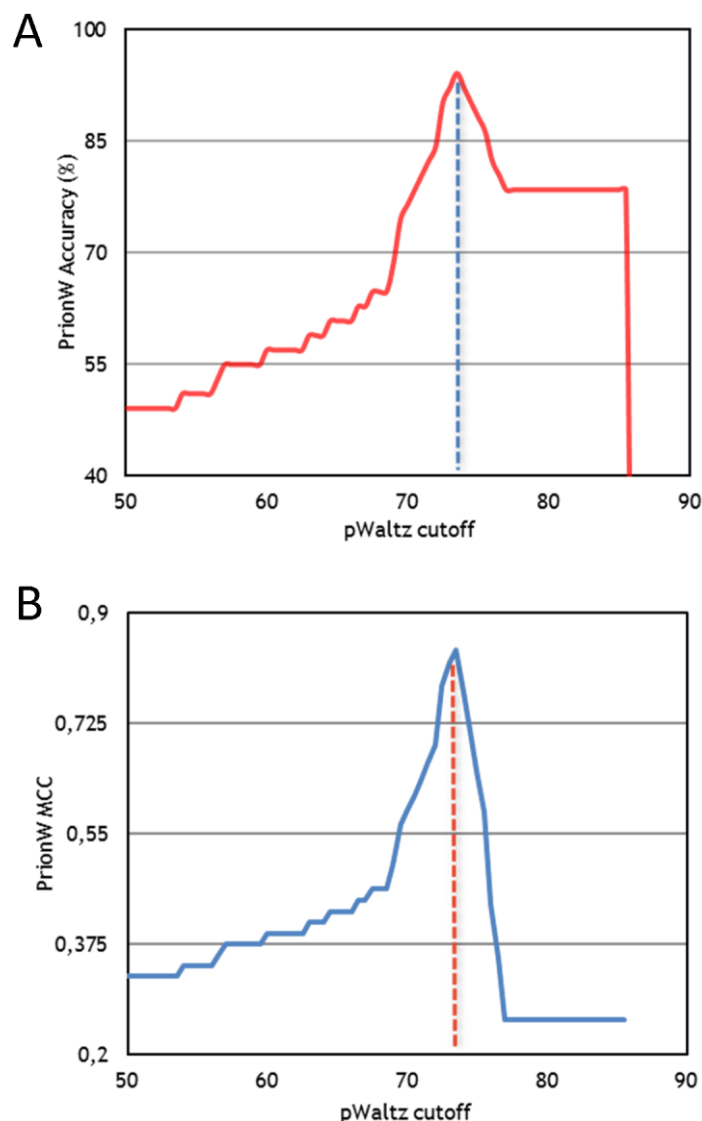


Figure 5.1 – Accuracy cut-off plot for PrionW. The Accuracy obtained for the correct classification of TP and TN is graphed against increasing pWALTZ cut-offs. We highlighted the highest accuracy of the assay, used to set the predictive cut-off of 73.55. TP and TN correspond to true positives and true negatives.

Two pioneering works addressed previously the discovery of potential novel prion-forming proteins exploiting their Q/N bias. Michelitsch and Weissman developed DIANA (Defined Interval Amino acid Numerating Algorithm), an algorithm aimed to identify proteins containing regions of consecutive amino acids with exceptionally high Q/N content (Michelitsch and Weissman, 2000). Harrison and Gerstein derived a method for identifying biased regions that relies on defining the lowest-probability subsequences (LPSs) for a given amino-acid composition and applied this formalism to analyse the prevalence of Q- and N-rich regions in different proteomes (Harrison and Gerstein, 2003). A comparison of the performance of PrionW, with that of the DIANA and LPSs approaches (**Table 5.1**), illustrates the usefulness of evaluating the presence and potency of soft-amyloidogenic regions in the context of Q/N rich sequences to discriminate prionogenic sequences in complete proteomes.

The ability to perform predictions in complete proteomes allows using Gene Ontology (GO) annotations to classify proteins containing PrLDs according cellular locations, functional classes and processes, uncovering the role played by these polypeptides in the cell. According to the GO classification in the

Sacharomyces Genome Database (SGD) (Cherry, et al., 2012) the detected proteins are associated to cytoplasmatic ribonucleoprotein granules ($P = 4.1 \times 10^{-05}$) and nucleus ($P = 6.1 \times 10^{-05}$), their preferential function is mRNA binding ($P = 3.0 \times 10^{-05}$) and more generally nucleic acid binding ($P = 6.3 \times 10^{-03}$) and they work in the regulation of biological processes ($P = 5.9 \times 10^{-07}$) and more specifically in the regulation of gene expression ($P = 7.7 \times 10^{-06}$). This analysis highlights the important role played by PrLDs-containing proteins in the yeast physiology, a role that might be also exerted in higher organisms.

According to FoldIndex and other disorder predictors like RONN (Yang, et al., 2005) or FoldUnfold (Galzitskaya, et al., 2006), in most of the 62 hits, the detected PrLDs are accompanied by at least a folded domain, which are likely the responsible of the protein activity and probably widely offset from the fibril backbone in the amyloid state (Baxa, et al., 2011). As expected, in contrast to pWALTZ, PrionW can identify genuine prions even when their PrDs represent a small fraction in the complete sequence of an essentially folded protein (**Figure 5.2**).

The requirement to adjust the Q+N content and pWALTZ parameters when using PrionW to screen for prion-like proteins in proteomes different from yeast is best illustrated by the fact that the algorithm is not able to identify a set of human proteins which have been proposed to display prion-like behaviour (Malinowska, et al., 2013), including hnRNPA1, hnRNPA2, hnRNPA3, HNRDL, FUS, EWS, TAF15 and TPD43 with the default settings. However, setting the Q/N content at $\geq 15\%$ and pWALTZ cut-off at 64.00 allows retrieving them, except TDP43, and identifying their putative soft-amyloid cores. The overall lower amyloidogenic potential of the nucleating cores of those human prion-like proteins likely respond to the fact they are not actual prions, but rather proteins able to self-assemble reversibly for functional purposes, and even if they have been shown to form intracellular aggregates upon mutation (Kim, et al., 2013), it is not evident that they can be propagated as *bona fide* prions.

5.1.5 SERVER DESCRIPTION

The PrionW web server does not require any user registration or identification. The interface can process up to 10000 sequences at a time.

Input Interface

PrionW is presented as an application running in a single web page (**Figure 5.2A**). One or more sequences in FASTA format must be pasted in the text box or uploaded as a file. Two algorithm parameters can be tuned by the user: "Q+N richness" defines the minimum proportion of Q and N residues a region should have to be considered disordered; "pWaltz cut off" defines the minimum pWaltz score for a soft-amyloid core to be considered positive. Default values are otherwise assigned to these parameters (see methods for more details). The web page displays four links in its upper margin: i) reference publications of methods and web application, ii) a contact mail, iii) a help with a short description of the algorithm, input instructions, output explanation and information on examples and iv) examples that will populate the input text area with full-length sequences of the well-characterized yeast prions NEW1, RNQ1, SWI1,

SUP35 and URE2 and a set of prion positive and negative control synthetic sequences proposed by Toombs and co-workers (Toombs, et al., 2010).

Table 5.1 – Performance of DIANA, LPSs and PrionW approaches in the prediction of experimental yeast prion-like proteins (protein dataset in Supplementary Material S5.1). The best value for each parameter is indicated in bold.

	<i>DIANA</i>	<i>LPSs</i>	<i>PrionW</i>
Sensitivity	0.917	1	0.917
Specificity	0.385	0.128	0.949
Precision	0.314	0.261	0.846
FDR¹	0.686	0.739	0.154
Accuracy	0.510	0.333	0.941
MCC²	0.275	0.183	0.842

¹False Discovery Rate

²Matthews correlation coefficient

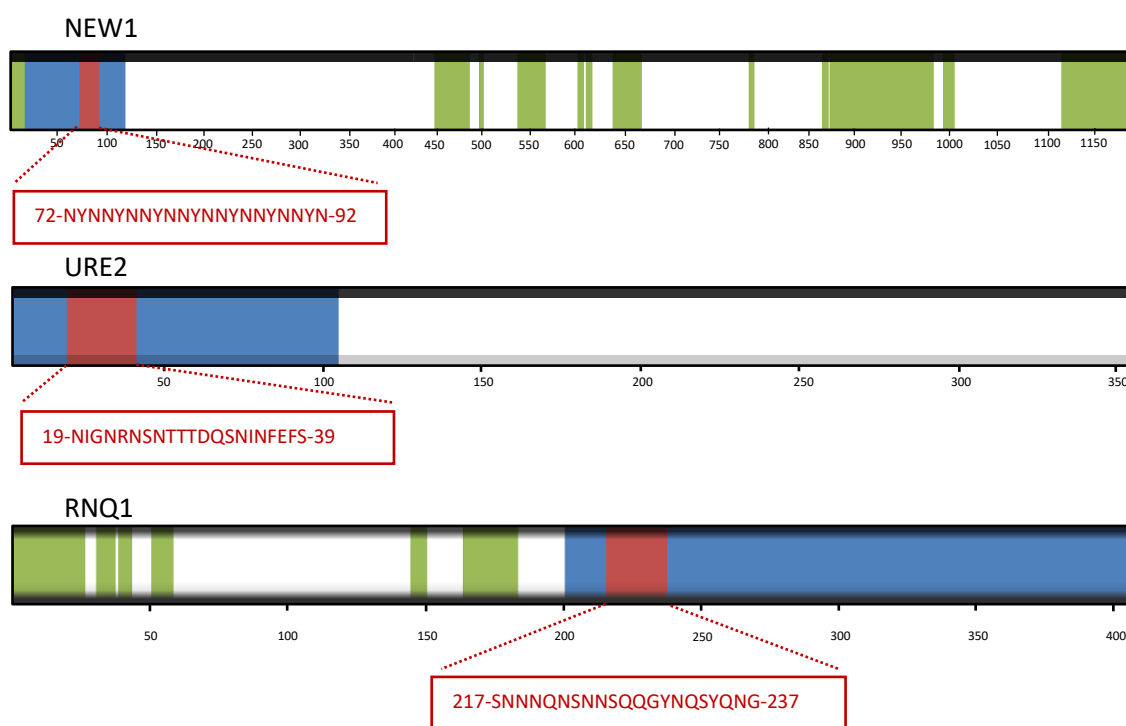


Figure 5.2 – PrionW predictions of prion-like domains and soft-amyloid cores in the sequences of the yeast prions **NEW1**, **URE2** and **RNQ1**. Folded domains, disordered regions, PrLDs and soft-amyloid cores are shown in white, green, blue and red, respectively.

Output

When clicking the submit button the input frame changes. After checking for the correct FASTA format, a header showing the number of interpreted sequences, input parameters and job identifier appears. After the calculation has finished, a link to a CSV file containing the output data is provided. Below the header, positive hits are printed in the same order as they were submitted. For each hit, the name, the predicted 21-residue soft-amyloid core, the pWaltz score and the predicted Q/N rich PrLD with the soft-amyloid core highlighted in red are presented (**Figure 5.2B**). If no positive sequences are detected in the input

We have described PrionW, a web server for the prediction of Q/N rich prion-like domains and their soft-amyloid cores in large sequence datasets. The algorithm should find application in the discovery of new candidates in different organisms for further experimental characterization, in the identification of mutations endorsing wild type proteins with prion-like properties, in the design of synthetic prion or prion-like domains for different purposes or in the design and synthesis of short peptides corresponding to PrLDs soft-amyloid cores able to seed the aggregation of the complete protein and, more generally, in understanding prion function and regulation in different species.

5.1.7 REFERENCES

- Alberti, S., *et al.* (2009) A systematic survey identifies prions and illuminates sequence features of prionogenic proteins, *Cell*, **137**, 146-158.
- Ashe, K.H. and Aguzzi, A. (2012) Prions, prionoids and pathogenic proteins in Alzheimer disease, *Prion*, **7**.
- Baxa, U., *et al.* (2011) In Sup35p filaments (the [PSI⁺] prion), the globular C-terminal domains are widely offset from the amyloid fibril backbone, *Mol. Microbiol.*, **79**, 523-532.
- Eisenberg, D. and Jucker, M. (2012) The amyloid state of proteins in human diseases, *Cell*, **148**, 1188-1203.
- Espinosa Angarica, V., Ventura, S. and Sancho, J. (2013) Discovering putative prion sequences in complete proteomes using probabilistic representations of Q/N-rich domains, *BMC Genomics*, **14**, 316.
- Fowler, D.M., *et al.* (2006) Functional amyloid formation within mammalian tissue, *PLoS biology*, **4**, e6.
- Fowler, D.M., *et al.* (2007) Functional amyloid--from bacteria to humans, *Trends Biochem Sci*, **32**, 217-224.
- Galzitskaya, O.V., Garbuzynskiy, S.O. and Lobanov, M.Y. (2006) FoldUnfold: web server for the prediction of disordered regions in protein chain, *Bioinformatics*, **22**, 2948-2949.
- Greenwald, J. and Riek, R. (2010) Biology of amyloid: structure, function, and regulation, *Structure*, **18**, 1244-1260.
- Harrison, P.M. and Gerstein, M. (2003) A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes, *Genome Biol*, **4**, R40.
- Kim, H.J., *et al.* (2013) Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS, *Nature*, **495**, 467-473.
- Malinowska, L., Kroschwald, S. and Alberti, S. (2013) Protein disorder, prion propensities, and self-organizing macromolecular collectives, *Biochim Biophys Acta*, **1834**, 918-931.
- Maurer-Stroh, S., *et al.* (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices, *Nat. Meth.*, **7**, 237-242.
- Michelitsch, M.D. and Weissman, J.S. (2000) A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions, *Proc Natl Acad Sci U S A*, **97**, 11910-11915.
- Nystrom, S., *et al.* (2012) Multiple substitutions of methionine 129 in human prion protein reveal its importance in the amyloid fibrillation pathway, *J Biol Chem*, **287**, 25975-25984.
- Prilusky, J., *et al.* (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded, *Bioinformatics*, **21**, 3435-3438.
- Rousseau, F., Serrano, L. and Schymkowitz, J.W. (2006) How evolutionary pressure against protein aggregation shaped chaperone specificity, *Journal of molecular biology*, **355**, 1037-1047.
- Sabate, R., *et al.* (2015) What makes a protein sequence a prion?, *PLoS Comput Biol*, **11**, e1004013.
- Toombs, J.A., McCarty, B.R. and Ross, E.D. (2010) Compositional determinants of prion formation in yeast, *Mol Cell Biol*, **30**, 319-332.
- Toombs, J.A., *et al.* (2012) De novo design of synthetic prion domains, *Proc Natl Acad Sci U S A*, **109**, 6519-6524.
- Westermarck, G.T. and Westermarck, P. (2010) Prion-like aggregates: infectious agents in human disease, *Trends in molecular medicine*, **16**, 501-507.
- Yang, Z.R., *et al.* (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins, *Bioinformatics*, **21**, 3369-3376.4

5.2 AMYCO: Evaluation of mutational impact on prion-like proteins aggregation propensity

Valentín Iglesias^{1†}, Oscar Conchillo-Sole^{1†}, Cistina Batlle¹ and Salvador Ventura^{1*}

¹ Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, 08193, Spain

†These authors contributed equally to this work. *To whom correspondence should be addressed.

Author Contribution: Software, validation, writing—original draft preparation.

5.2.1 ABSTRACT

Background: Around 1% of human proteins are predicted to contain a disordered and low complexity prion-like domain (PrLD). Mutations in PrLDs might promote a transition to an amyloid-like, aggregation-prone, state linked to disease.

Results: We have recently shown that an algorithm that considers both the effects of mutations on PrLDs composition and in localized amyloid propensity can approach their impact on intracellular protein aggregation. Here, we implement this concept into the AMYCO web server, an algorithm that forecasts the influence of amino acid changes in prion-like proteins aggregation propensity better than state-of-the-art predictors.

Conclusions: The AMYCO web server allows for a fast and automated evaluation of the effect of mutations on the aggregation properties of prion-like proteins. This might uncover novel disease-linked amino acid changes occurring in the sequences of the increasing number of prion-like proteins being identified in the human proteome. Additionally, it can find application in the *in silico* design of synthetic prion-like proteins with tuned aggregation propensities for different purposes.

Availability and Implementation: AMYCO does not require previous registration and is freely available to all users at: <http://bioinf.uab.cat/amyco/>.

5.2.2 BACKGROUND

Prions are proteins able to adopt multiple structural conformations from which at least one has self-propagating properties (Aguzzi and Calella, 2009). Yeast prions are the best understood subset of functional prions. A common feature of most yeast prions is the presence of an intrinsically disordered and low complexity prion domain (PrD), which is necessary and sufficient for prion conversion and propagation. Proteins bearing prion-like domains (PrLD) sharing these properties seem to exist in all kingdoms of life (Chakrabortee, et al., 2016; Iglesias, et al., 2015; Malinowska, et al., 2015; Pallares, et al., 2015; Yuan and Hochschild, 2017). In particular, around 1% of the human proteome has been predicted to correspond to prion-like proteins (King, et al., 2012). This human protein subset is enriched in nucleic

acid-binding proteins and involved in the formation of membraneless compartments through highly dynamic liquid-liquid demixing (King, et al., 2012; Patel, et al., 2015). A number of mutations in human PrLDs have shown to convert these liquid compartments into solid aggregates, abolishing their dynamic nature and leading to the onset of neurodegenerative disorders (Patel, et al., 2015; Polymenidou and Cleveland, 2012). The development of tools able to anticipate the impact of such pathogenic amino acid changes is attracting increasing interest.

The self-assembling properties of prion-like proteins are thought to ultimately rely on the biased amino acid composition of their PrLDs (Toombs, et al., 2012), whereas disease-linked mutations seem to act by enhancing or extending aggregation-prone regions, facilitating the transition to amyloid-like states (Ryan, et al., 2018; Sabate, et al., 2015). We have recently shown that the impact of point and multiple mutations or deletions on the aggregation of the model ALS-associated prion-like hnRNPA2 protein is best predicted by a function that takes into account both compositional features and amyloidogenic propensities (Batlle, et al., 2017b). Here we introduce the AMYCO (combined AMYloid and Composition based prediction of prion-like aggregation propensity) web server, which implements this approach to perform automated and fast predictions on top of prion-like protein sequences.

5.2.3 IMPLEMENTATION

AMYCO is written in Python and uses Python2.7 as the interpreter (Anaconda distribution). The web interface has been build using HTML/CSS and inputs and outputs are processed by a CGI written in Perl. It all runs in a CentOS 5 server with Apache 2.2.3 using Intel Xeon 'Clovertown' processors.

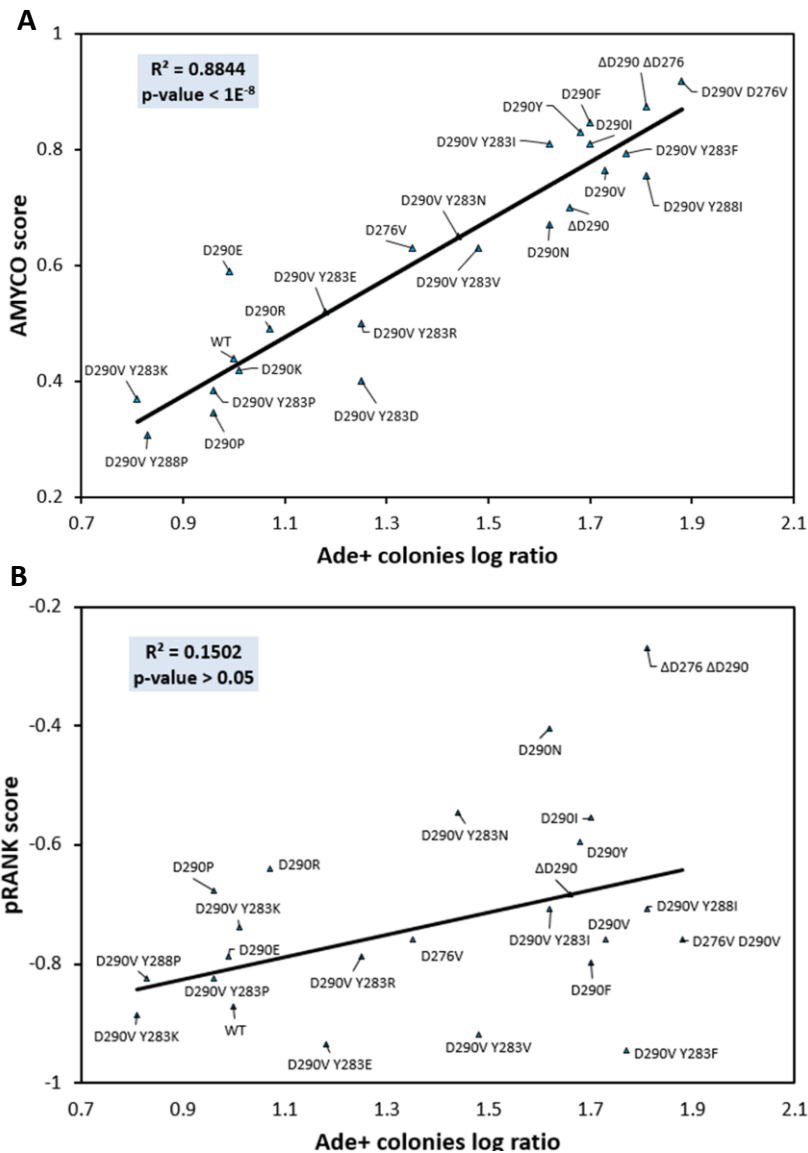


Figure 5.4 – Correlation between AMYCO and pRANK predictions and the aggregation propensity of human hnRNPA2 prion-like protein. Graphic representation of the correlation between the variants **A)** AMYCO and **B)** pRANK scores and their ability to form Ade+ colonies when expressed in yeast, a direct reporter for their aggregation propensity (Batlle, et al., 2017b).

AMYCO pipeline

AMYCO evaluates the impact of mutations on the aggregation propensity of PrLDs in prion-like proteins. They can be single or multiple residues substitutions, as well as deletions and insertions. It exploits the highly significant correlation between the scores obtained from a parameterized linear function, that balances the contribution of both PrLDs composition and amyloid propensity (Batlle, et al., 2017b), and the intracellular aggregation of hnRNPA2 variants; the unique prion-like protein for which a large set of mutations, both natural and artificial have been experimentally validated (**Figure 5.4**). The AMYCO web server is free and open to all users, and no previous login or registration is required.

The home page of AMYCO displays three clickable links in its upper margin: (i) a help page containing a brief description of the method, the output explanation and information on examples, (ii) references for

the methodology and the web application and (iii) a contact e-mail. Immediately below a link that switches between default mode in which multiple sequences can be compared, or single mutation mode in which all possible mutations for a given protein residue are evaluated,) and examples that fill the input text areas with the full-length sequences of wild type (*wt*) human hnRNPA2 protein and its aggregation-prone D290V mutant for compare mode, or all its possible mutants for position 290 for single mutation mode (Kim, et al., 2013).

The input interface allows two working modes (Figure 5.5). In *compare sequences* mode (default mode) (Figure 5.5A); the user should introduce a reference sequence and the mutated variants (one or several) in the left and right text boxes, respectively; all in FASTA format. In the *single mutation* mode (Figure 5.5B), the user should introduce a single sequence as well as the position to be scanned. Protein sequences should be at least 60-residues long and only the 20 standard proteinogenic amino acids are allowed.

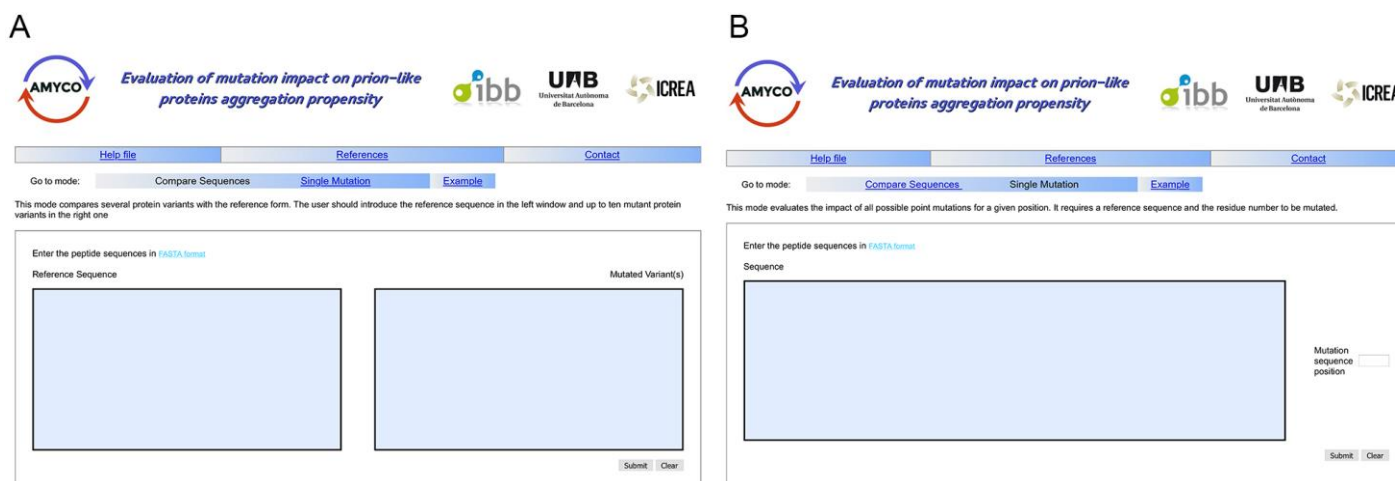


Figure 5.5 – AMYCO web server main page. The server presents two different working modes. On **A)** *compare sequences* mode, a single FASTA sequence must be pasted on the “Reference Sequence” box and one or more variants on the right “Mutated Variant(s)” box, while on **B)** *single mutation* mode, users must introduce one sequence and specify a position for all possible point mutations to be evaluated.

After submission, the output page will display a job identification number along with the names of the input sequences and the mutation position if applicable. The algorithm will return the AMYCO score for each sequence, together with a description of the mutations impact of the overall aggregation propensity. In addition, a graphical representation of the mutation/s effect will be displayed (Figure 5.6). We set two thresholds of low (< 0.45) and high (> 0.78) AMYCO scores. hnRNPA2 mutants scoring < 0.45 were shown to decrease or increase < 5 times the propensity of the non-aggregating wild type protein, whereas, mutants scoring > 0.78 increased its aggregation by > 50 times (Paul, et al., 2017). Mutations rendering an AMYCO score < 0.45 are considered of low aggregation propensity and labeled in blue. Mutations that increase the aggregation propensity of the protein, but whose AMYCO score is below 0.78 are labeled in red, whereas mutations above this threshold are considered to be of high aggregation propensity are labeled in red and bold. Sequences might display AMYCO scores > 1.0, indicating that they are predicted

to be more aggregation-prone than the highest scoring hnRNPA2 variant used in the parametrization of the prediction function. The output files can be downloaded for further analyses, as a ZIP file containing the resulting text explanation, a machine readable JSON file, the visualizations as PNG files and in the single mutation mode, a FASTA file with all mutants.

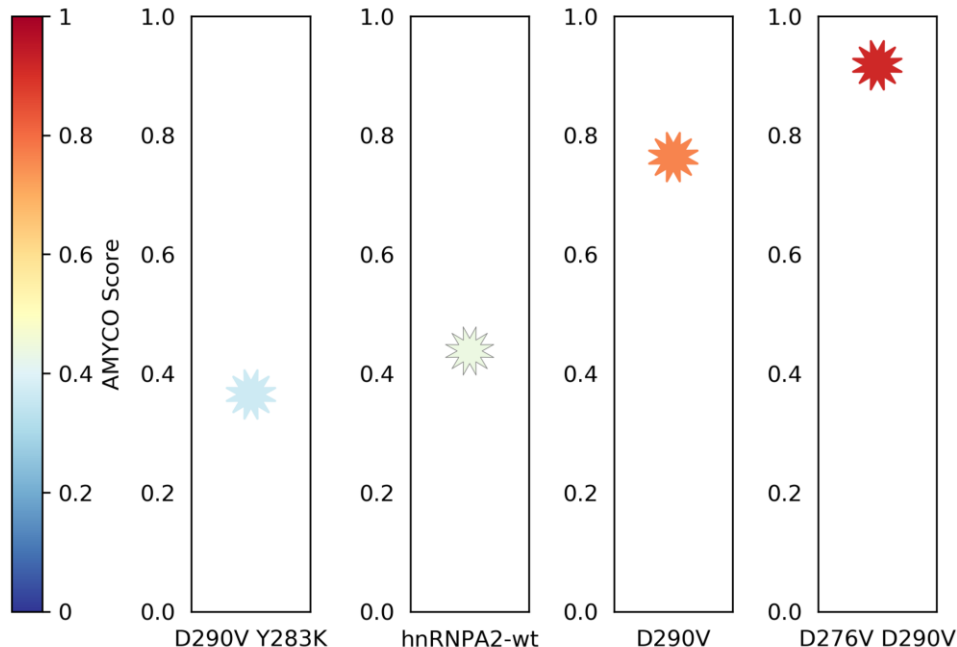


Figure 5.6 – Graphical representation of the AMYCO score. AMYCO output representation of a low aggregation-prone (D290V Y283K), the wild type, the natural pathogenic mutant D290V and a high aggregation -prone (D290V D276V) hnRNPA2 prion-like protein variants.

5.2.4 RESULTS

Performance

The AMYCO algorithm, which combines the predictions of the composition based prion domain predictor PAPA (Toombs, et al., 2012) with our previously developed pWALTZ program (Sabate, et al., 2015), which identifies sequences with amyloidogenic potential inside these domains, performs better than PAPA alone (**Table 5.2**). pRANK is a novel multiple-instance machine learning method aimed to predict prion propensity based on amino acid composition alone (Afsar Minhas, et al., 2017). We compared the performance of AMYCO and pRANK web servers in predicting the impact of mutations on human hnRNPA2 aggregation propensity (**Figure 5.4**). AMYCO clearly outperforms pRANK (**Table 5.2**), an observation which is consistent with the important influence that sequential features exert on protein aggregation (Sabate, et al., 2010).

Table 5.2 – Performance of pRANK and AMYCO approaches in the prediction of mutation impact upon the aggregation of the human prion-like protein hnRNPA2.

	pRANK	AMYCO
Sensitivity	0	1
Specificity	1	1
Precision	-	1
Accuracy	0.45	1
MCC	-	1
Mean % error	-7.08	-1.25
Standard Deviation (%)	37.71	12.19
SEM (%)	8.04	2.44
Coefficient of Determination	0.150	0.882
P-value (two tailed test)	0.468	< 1.00 x 10⁻⁸
Rho (ρ)	0.334	0.929

The best performance according to each particular parameter is shown in bold. The sensitivity, specificity, precision, accuracy and Matthews correlation coefficient (MCC) were calculated from point mutations in hnRNPA2 considering positive those mutations which increased the mutant/wild type prion Ade⁺ colony ratio (a reporter of their aggregation) by at least one order of magnitude (Paul, et al., 2017). Proline mutants score under pWALTZ threshold, so they are not taken into account. The final dataset was composed of 13 True positives (TN) and 8 true negatives (TN).

AMYCO was further assayed on known mutations promoting the apparition of a *de novo* prion-like behavior (**Table 5.3**). It was able to predict a large increase in aggregation propensity for mutations that convert the non-prionic PrLDs of PUF4, YLR177W, KC11 and PDC2 yeast proteins into prionic when expressed in yeast (Paul, et al., 2015) (**Table 5.3**). Importantly, according to AMYCO, five out of the eight variants were predicted to have acquired a very high aggregation propensity. These variants are exactly the ones experimentally shown to induce a prionic phenotype with basal protein levels, without a need for overexpression (Paul, et al., 2015) (**Table 5.3**).

Table 5.3 – AMYCO correctly predicts prion converting mutations on yeast proteins.

Protein variant	AMYCO score
PUF4 <i>wt</i>	0
<i>PUF4^{mut}</i>	<i>0.69</i>
*PUF4^{6PP,1N}	0.93
<i>PUF4^{4PP}</i>	<i>0.60</i>
YLR177W <i>wt</i>	0
*YLR177W^{mut}	0.85
*YLR177W^{4PP,1N}	1.23
*YLR177W^{4PP}	1.03
KC11 <i>wt</i>	0
*KC11^{mut}	0.97
PDC2 <i>wt</i>	0
<i>PDC2^{mut}</i>	<i>0.78</i>

AMYCO correctly predicts mutations that induce prionic phenotypes (Paul, et al., 2015). Mutations predicted to increase and highly increase aggregation propensity are shown in italics and bold, respectively. Variants that do not need overexpression to generate a prionic phenotype in yeast are indicated with an asterisk.

Finally, AMYCO is able to predict an increase in aggregation propensity for a series of disease-linked mutations occurring in different human prion-like proteins. In particular, mutations in hnRNPA1 associated to ALS (Kim, et al., 2013), mutations in hnRNP D0/AUF1 identified in familiar cases of Crohn Disease (Prakash, et al., 2017) and mutations in hnRNP DL causing limb-girdle muscular dystrophy 1G (Vieira, et al., 2014) (**Table 5.4**).

Table 5.4 – AMYCO predicts disease-causing mutations on human prion-like proteins

Protein variant	AMYCO score
hnRNPA1 wt	0.34
hnRNPA1 D314V	0.59
hnRNPA1 D314N	0.53
hnRNP DL wt	1.18
hnRNP DL D378H	1.26
hnRNP DL D378N	1.30
hnRNP D0 wt	1.13
hnRNP D0 D319V	1.33
hnRNP D0 isoform-2 D300V	1.33

AMYCO identifies multisystem proteinopathy and ALS causing mutations on hnRNP A1 (Kim, et al., 2013), Crohn Disease causing mutations on both isoforms of hnRNP D0/AUF1 (Prakash, et al., 2017) and limb-girdle muscular dystrophy 1G (LGMD1G) on hnRNP DL (Vieira, et al., 2014)

5.2.4 CONCLUSION

AMYCO has been developed as a web application to assess the impact of mutations on the aggregation propensity of prion-like proteins, allowing a fast and accurate evaluation of the effect of disease-associated mutations in these polypeptides; as well as engineering novel variants with designed aggregation propensities for different applications.

5.2.5 AVAILABILITY AND REQUIREMENTS

Project name: AMYCO

Project home page: <http://bioinf.uab.cat/amyco/>

Operating system(s): Platform independent

Programming language: A computing core coded in Python and a front end written in a combination of HTML and Perl CGI.

Other requirements: A web browser with a working internet connection.

License: None

Any restrictions to use by non-academics: None

5.2.6 REFERENCES

- Afsar Minhas, F.U., Ross, E.D. and Ben-Hur, A. (2017) Amino acid composition predicts prion activity, *PLoS Comput Biol*, **13**, e1005465.
- Aguzzi, A. and Calella, A.M. (2009) Prions: protein aggregation and infectious diseases., *Physiological reviews*, **89**, 1105-1152.
- Battle, C., et al. (2017) Perfecting prediction of mutational impact on the aggregation propensity of the ALS-associated hnRNPA2 prion-like protein, *FEBS letters*.
- Chakrabortee, S., et al. (2016) Luminidependens (LD) is an Arabidopsis protein with prion behavior, *Proceedings of the National Academy of Sciences*, 201604478.
- Hunter, J.D. (2007) Matplotlib: A 2D Graphics Environment, *Computing in Science & Engineering*, **9**, 90-95.
- Iglesias, V., de Groot, N.S. and Ventura, S. (2015) Computational analysis of candidate prion-like proteins in bacteria and their role, *Frontiers in Microbiology*, **6**, 1-13.

- Kim, H.J., *et al.* (2013) Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS, *Nature*, **495**, 467-473.
- King, O.D., Gitler, A.D. and Shorter, J. (2012) The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease, *Brain Research*, **1462**, 61-80.
- Malinowska, L., *et al.* (2015) Dictyostelium discoideum has a highly Q/N-rich proteome and shows an unusual resilience to protein aggregation, *Proc Natl Acad Sci U S A*, **112**, E2620-2629.
- Pallarès, I., Iglesias, V. and Ventura, S. (2016) The Rho Termination Factor of Clostridium botulinum Contains a Prion-Like Domain with a Highly Amyloidogenic Core, *Frontiers in microbiology*, **6**, 1-12.
- Patel, A., *et al.* (2015) A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation, *Cell*, **162**, 1066-1077.
- Paul, K.R., *et al.* (2015) Generating new prions by targeted mutation or segment duplication, *Proc Natl Acad Sci U S A*, **112**, 8584-8589.
- Paul, K.R., *et al.* (2017) Effects of Mutations on the Aggregation Propensity of the Human Prion-Like Protein hnRNPA2B1, *Mol Cell Biol*, **37**.
- Polymenidou, M. and Cleveland, D.W. (2012) Prion-like spread of protein aggregates in neurodegeneration, *The Journal of experimental medicine*, **209**, 889-893.
- Prakash, T., Veerappa, A. and N, B.R. (2017) Complex interaction between HNRNPD mutations and risk polymorphisms is associated with discordant Crohn's disease in monozygotic twins, *Autoimmunity*, **50**, 275-276.
- Ryan, V.H., *et al.* (2018) Mechanistic View of hnRNPA2 Low-Complexity Domain Structure, Interactions, and Phase Separation Altered by Mutation and Arginine Methylation, *Mol Cell*, **69**, 465-479 e467.
- Sabate, R., *et al.* (2010) The role of protein sequence and amino acid composition in amyloid formation: scrambling and backward reading of IAPP amyloid fibrils, *J. Mol. Biol.*, **404**, 337-352.
- Sabate, R., *et al.* (2015) What Makes a Protein Sequence a Prion?, *PLoS Computational Biology*, **11**, e1004013.
- Toombs, J.A., *et al.* (2012) De novo design of synthetic prion domains, *Proc Natl Acad Sci U S A*, **109**, 6519-6524.
- Vieira, N.M., *et al.* (2014) A defect in the RNA-processing protein HNRPDL causes limb-girdle muscular dystrophy 1G (LGMD1G), *Hum Mol Genet*, **23**, 4103-4110.
- Yuan, A.H. and Hochschild, A. (2017) A bacterial global regulator forms a prion, *Science*, **355**, 198-201.

Chapter IV – Characterization of prion-like proteins

6.1 Computational analysis of candidate prion-like proteins in Bacteria and their role

Valentín Iglesias¹, Natalia Sanchez de Groot^{1*} and Salvador Ventura^{1,*}

¹ Institut de Biotecnologia i Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

† These authors contributed equally to this work. *To whom correspondence should be addressed.

Author Contributions: Software, validation, data curation, writing—original draft preparation.

6.1.1 ABSTRACT

Prion proteins were initially associated with mammalian transmissible spongiform encephalopathies such as Creutzfeldt Jakob or kuru. However, deeper research revealed them as versatile tools, exploited by the cells to execute diverse functions, acting as epigenetic elements or building membrane free compartments in eukaryotes. One of the most intriguing properties of prion proteins is their ability to propagate a conformational assembly, even across species. In this context, it has been observed that bacterial amyloids can trigger the formation of protein aggregates by interacting with host proteins. As our life is closely linked to bacteria, either through a parasitic or symbiotic relationship, prion-like proteins produced by bacterial cells might play a role in this association. Bioinformatics is helping us to understand the factors that determine conformational conversion and infectivity in prion-like proteins. We have used PrionScan to detect prion domains in 839 different bacteria proteomes, detecting 2200 putative prions in these organisms. We studied this set of proteins in order to try to understand their functional role and structural properties. Our results suggest that these bacterial polypeptides are associated to peripheral rearrangement, macromolecular assembly, cell adaptability and invasion. Overall, these data could reveal new threats and therapeutic targets associated to infectious diseases.

6.1.2 INTRODUCTION

An diverse number of human diseases are associated with amyloid forming proteins (Chiti and Dobson, 2006). Despite these polypeptides are diverse in function, sequence and origin, all share the propensity to form β -sheet aggregates (Karran, et al., 2011). Amyloid fibril forming proteins appear to be highly conserved and have been detected in all kingdoms of life, suggesting that, despite they are usually thought to be involved in pathogenic processes, they might indeed provide selective advantages (Espinosa Angarica, et al., 2013; Malinovska, et al., 2013; Sanchez de Groot, et al., 2015; Sanchez de Groot, et al., 2012). In fact, cells exploit the formation of amyloid fibrils for diverse purposes (Coustou, et al., 1997; Chapman, et al., 2002; Fowler, et al., 2006; Graether, et al., 2003; Ionomidou, et al., 2000; Maji, et al., 2009; Podrabsky, et al., 2001), from structure scaffolding, such as the melanin at the skin, to heritable information transmission, such as the yeast prions (Chien and Weissman, 2001; Liebman and Chernoff, 2012; Shorter and Lindquist, 2005; Staniforth and Tuite, 2012). Because amyloid fibers and their unstable

intermediates can be highly cytotoxic (e.g. by disrupting the membrane integrity), the assembly of functional amyloids is a process tightly regulated by the organisms, which involves the assistance of chaperones and a spatiotemporal control (Blanco, et al., 2012; Evans, et al., 2015; Gsponer and Babu, 2012; Taylor and Matthews, 2015).

Prions are a singular subset of proteins able to change from one conformational state to another, often an amyloid aggregate, and transmit it to other homologous polypeptide sequences. Importantly, recent results suggest that amyloid proteins involved in Alzheimer's and Parkinson's diseases could be infectious and act as prion-like proteins in the brain (Chiti and Dobson, 2006; Stohr, et al., 2012). Most prions (with the exception of the mammalian prion protein PrP), constitute a subset of aggregation-prone proteins with special sequential composition. Whereas classical amyloid proteins contain specific regions rich in hydrophobic residues that lead the protein self-assembly, yeast prions exhibit domains that are commonly enriched in asparagine and glutamine (Q/N) (Dorsman, et al., 2002; Fandrich and Dobson, 2002; Halfmann, et al., 2011) but also in glycine, serine and tyrosine residues (Kato, et al., 2012) which are generally known as prion domains (PrD). This pattern has been found in human prion-like proteins associated to neurodegenerative diseases, such as FUS (linked to dementia) or TDP43 (related to amyotrophic lateral sclerosis) (Kato, et al., 2012). This special bias results in low complexity sequences displaying disordered structures, a crucial property that ensures conformational flexibility, permits self-assembly without a requirement for conformational unfolding and allows conversion between species (Fuxreiter, 2012; Fuxreiter and Tompa, 2012; Malinowska, et al., 2013; Tompa and Fuxreiter, 2008). In fact, one of the main evolutionary strategies to control protein aggregation is to ensure a stable globular structure preventing, in this way, the exposition of aggregation prone stretches (Lim and Sauer, 1991; Monsellier, et al., 2007; Sanchez, et al., 2006). However, a polypeptide sequence requires more than just low complexity to behave as a prion (Espinosa Angarica, et al., 2013; Malinowska, et al., 2013). Hence, it has been found that the propagation of amyloid aggregation depends on characteristics such as the degree of over/under representation of specific residues and the length of the considered low complexity region (Ross, et al., 2004; Ross, et al., 2005; Toombs, et al., 2010).

The knowledge acquired in the last decade has allowed the design of approaches to predict prion-like proteins. The first predictive algorithms were based on the properties of the primary sequence responsible for the formation of the classical amyloid aggregates (e.g. high hydrophobicity and intrinsic β -sheet propensity). However, they failed to detect Q/N-rich stretches since these are polar residues that do not fulfil the typical requirements associated with classical β -sheet-amyloid aggregation (Pawar, et al., 2005). Then, the algorithms focused on localising Q/N rich segments in the primary sequence (Harrison and Gerstein, 2003; Michelitsch and Weissman, 2000), disregarding the contribution of the rest of residues (Ross, et al., 2005), but being unable to score the proteins in terms of their relative prionogenicity. A big improvement was achieved by combining computational approaches with the experimental validation of new proteins displaying *in vitro* prion properties. This strategy enlarged the set of prion sequences and permitted the refinement of the available theoretical models. Alberti and co-

workers employed a hidden Markov model (HMM), based on the 4 *bona fide* yeast prions identified to that moment, obtaining 200 yeast protein candidates carrying prion-like domains (PrLDs) (Alberti, et al., 2009). The *in vivo* and *in vitro* analysis of the top 100 candidates rendered 29 proteins that proved heritable switch and significant *in vivo* amyloid formation. We have recently exploited this experimentally curated dataset to develop a probabilistic model of PrLDs able to discover prionogenic proteins in complete proteomes (Espinosa Angarica, et al., 2013). We have implemented this model in a web-based algorithm called PrionScan able to handle with large sequence databases and predict prion-like sequence stretches in the proteomes annotated in UniprotKB (Espinosa Angarica, et al., 2014). In a previous work, we employed this predictor to analyse all the proteomes reported until that moment (1536 organisms) (Espinosa Angarica, et al., 2014). We discovered 20540 new prion candidates present in 10 different taxonomic divisions, supporting prions' universality. We also observed that in most cases the ratio of proteins with prion-like domains is less than 1% of the whole proteome. Thus, in Archaea and Viruses the number is less than 10 per proteome, while in Bacteria, Fungi, Plantae and Animalia the range is from few tens to few hundreds, depending on the organisms. Interestingly, we observed that, in different organisms, the predicted prion-like proteins are associated with different cellular components and biological processes, thus supporting prion properties being employed for diverse biological purposes.

Bacteria are ubiquitous in the world, adapted to multiple environments and able to growth in the most extreme conditions. Moreover, bacterial infection remains a leading cause of death in both Western and developing world (WorldHealthOrganisation, WHO). Understanding which bacteria proteins display prionic properties could help to deepen our understanding of bacterial biology and pathogenesis. Indeed, despite no genuine prion has been characterized so far in prokaryotes, it is clear that at least *E. coli* can generate infectious conformations of heterologous fungal prions (Espargaro, et al., 2012; Garrity, et al., 2010; Sabate, et al., 2009; Yuan, et al., 2014). In an analogous manner, the formation of amyloids was initially thought to be restricted to eukaryotic cells, but after the first report demonstrating that the curli fibers that emerge from the surfaces of *E. coli* cells had the same physical properties as human amyloids (Chapman, et al., 2002), the number of discovered bacterial proteins displaying this ability is steadily increasing (Blanco, et al., 2012; Otzen and Nielsen, 2008; Schwartz and Boles, 2013). Moreover, it has been observed that bacterial amyloids can initiate the formation of amyloid aggregates upon interaction with diverse host proteins (Friedland, 2015; Hill and Lukiw, 2015; Hufnagel, et al., 2013; Otzen and Nielsen, 2008). With the aim to understand better the potential relevance of bacterial PrLDs, here we focus on study the 2200 putative prion proteins predicted by PrionScan within the taxon domain bacteria, as derived from the study of 839 bacterial proteomes. Specifically, we analyse the functions and structures associated to these proteins and discuss the possible advantages that they could provide, ensuring their evolutionary conservation.

6.1.3 MATERIAL AND METHODS

Sequence Dataset

Our database was comprised of Uniprot Knowledgebase (UniProt, 2015) entries included both in Swissprot and TrEMBL (update 2012_03) under the taxon domain bacteria in order to track the prion like domains present in bacterial proteomes.

Discovering Prion-Like Domains

PrionScan, an algorithm developed by our group and described previously (Espinosa Angarica, et al., 2014), was used in order to predict prion-like domains. Employing a cutoff of 50 bits, we identified 2200 PrLD. Further analysis was made *a posteriori* in order to identify common traits including the Gene Ontology GO terms for the molecular functions, biological processes and cellular components and relevant domains according to Pfam database. Pfam domains and GO terms were manually annotated and counted in the 2200 positive PrLD containing bacterial proteins according to the UniprotKB annotations (UniProt, 2015). Due to the large amount of individual Pfam domains, only those ones represented over 5 times were considered in the analysis. Then, domains were manually clustered by similarity in their cellular function or process. Pathogenic bacteria (n = 18) were manually annotated by stringent bibliographical search for evidences of human pathogenic association at the NCBI. Then we calculated the GO terms and Pfam domains enrichment.

Statistics analysis

The enrichment analysis was performed with GOSTat (Beissbarth and Speed, 2004) against the goa_uniprot database (UniProt, 2015). Out of 2200 initial proteins, 244 (11.09%) were annotated. A p-value of 0.1 was set as a cut-off and a false discovery rate (Benjamini) test was performed to obtain it. The initial clustering was performed by classifying the obtained Gene Ontologies according to their category: biological process, cellular component or molecular functions. We calculated the enrichment factors (EF) for every GO term to show how much higher is the proportion of hits in relation to the background sample (the total number of proteins). Accordingly, the EF is the number of hits among PrLDs (nl) divided by the number of annotated proteins in our list (pl) and subsequently divided by the ratio between the hits of that GO term in goa_annotation (nb) and the total number of proteins (pb) in this specific GO term:

$$EF = \frac{\frac{n^l}{p^l}}{\frac{n^b}{p^b}} = \frac{n^l p^b}{n^b p^l}$$

Only those GO terms with a log₂ fold enrichment > 0.5 were considered to be significant for their subsequent analysis.

6.1.4 RESULTS

Identifying PrLDs in Bacteria proteomes

We have analysed 839 bacterial proteomes containing a total of 860337 proteins with PrionScan (Espinosa Angarica, et al., 2013), from which we detected 2200 putative prion proteins, accounting for a 0.3% of the complete protein dataset. Interestingly, in the 18 selected pathogenic bacteria, proteins

containing PrLDs are significantly more abundant (2.4%) and indeed they constitute 40% of all the detected PrLDs (891 PrLDs). Moreover, some specific pathogenic organisms appear to be specially enriched in PrLDs: *Staphylococcus aureus* (18%), *Enterococcus faecalis* (10%), *Enterococcus faecium* (5%) or *Staphylococcus epidermidis* (3%). These data show the diverse distribution of predicted PrLDs in bacterial species, suggesting certain associated functionality.

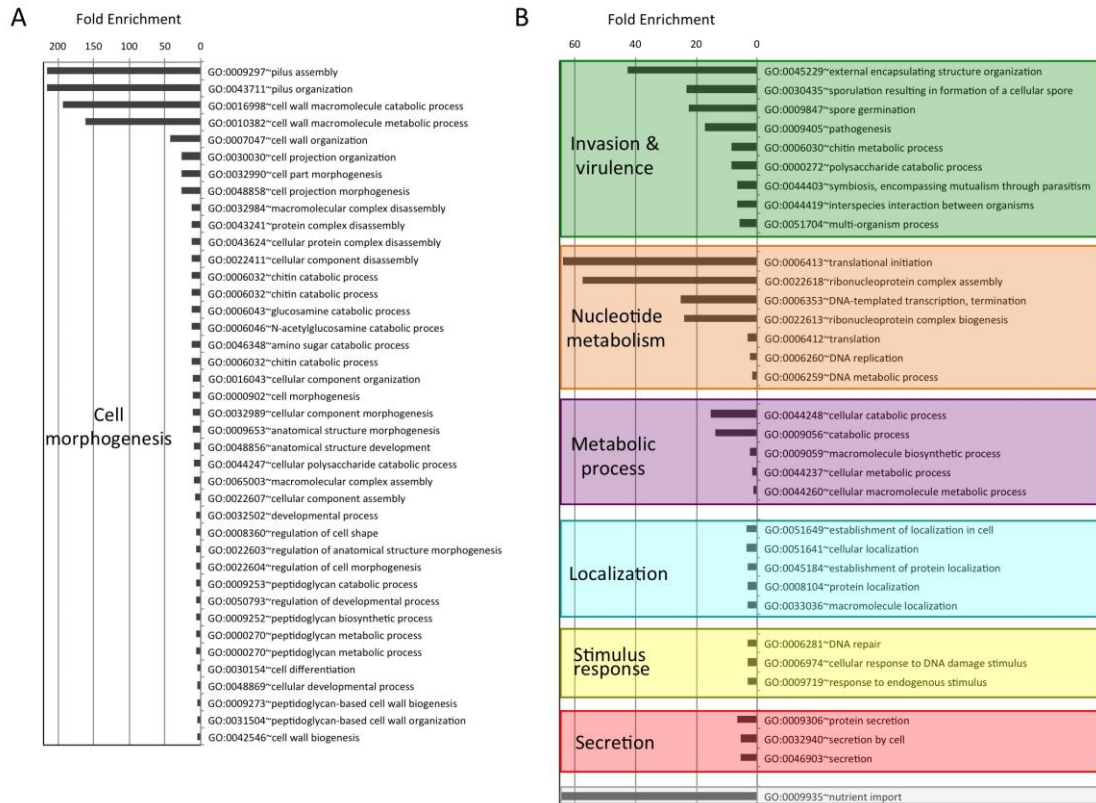


Figure 6.1 – Enrichment and clustering of PrLDs-containing proteins in bacteria accordingly to their biological process GO terms. The enrichment analysis was performed with Gostat against the goa_uniprot database. **A)** Proteins with GO terms associated with cell morphogenesis. **B)** Proteins with GO terms associated to other biological processes.

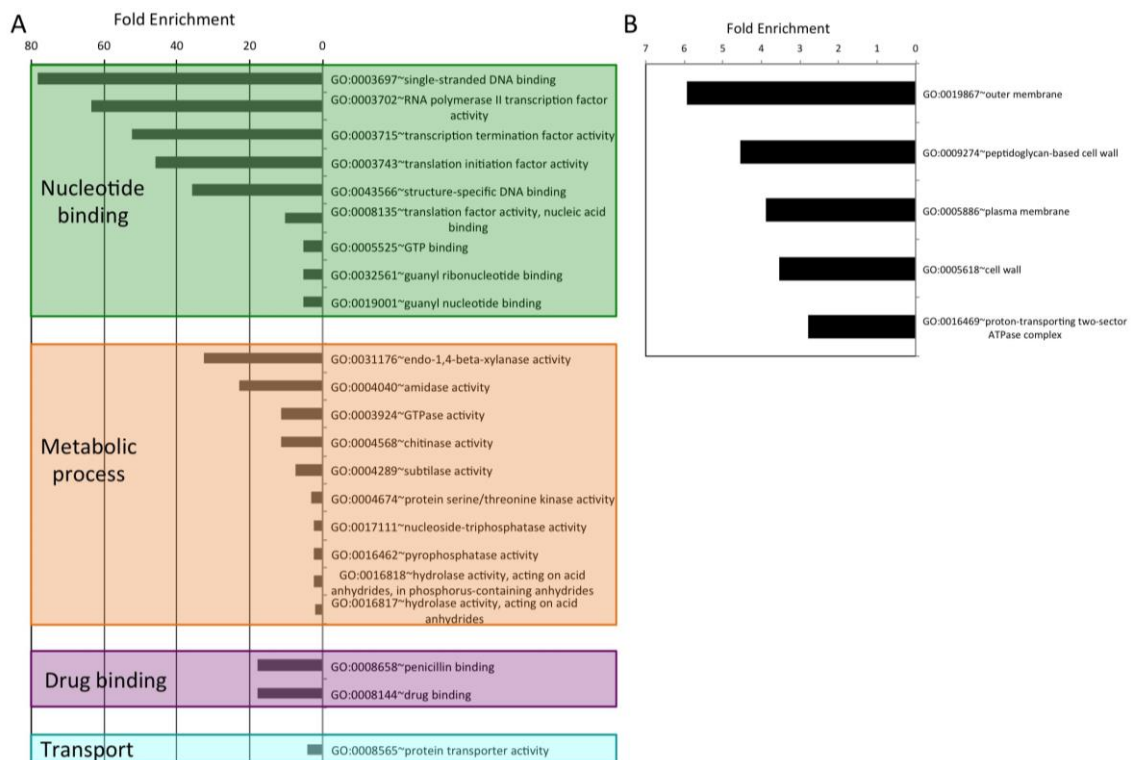


Figure 6.2 – Enrichment and clustering of PrLDs-containing proteins in bacteria according to their GO terms. A) Molecular function GO terms and **B)** Cellular component GO terms ontologies represented.

As an attempt to understand the biological purpose of these PrLDs we analysed the Gene Ontology enrichment of the corresponding proteins. To facilitate the data interpretation, we grouped the enriched GO terms by similar cellular function or process. We found the largest cluster of GO terms corresponds to Biological Processes involved in cell morphogenesis, such as cell projection or cell wall dynamics. This group contains 40 different terms, some of them with fold enrichments above 200 (pilus assembly) (**Figure 6.1A**). We also found several enriched Biological Processes involved in secretion, nutrient import, invasion and virulence; all of them involved in interaction with the surrounding environment. Interestingly, in invasion and virulence we find processes associated to encapsulation, sporulation and interaction with other organisms. Between the Biological Processes, the metabolic ones are particularly involved in the assembly of macromolecules such as polysaccharides and peptidoglycan (**Figure 6.1B**). The other three Biological Processes clusters are nucleotide metabolism, stimulus to response and localisation, which are associated to cellular adaptation and the formation of contacts between molecules. When we analyse the Molecular Functions (**Figure 6.2A**), the enriched GO terms correspond to nucleic acid binding, metabolic processes, drug binding and transport. These groups correspond to activities associated with the formation of functional interactions. Additionally, the clusters of metabolic process and drug binding perform functions related to cell wall such as peptidoglycan synthesis or chitin production. Moreover, nucleic acid binding functions could be associated to mechanisms of cellular adaptation. The proteins in this cluster are strongly associated to two essential functions: translation initiation and DNA templated transcription. Surprisingly, the enriched GO terms of the Cell Component do not include any inside part of the cell, just terms associated to the external part:

outer membrane, peptidoglycan based cell wall, plasma membrane, cell wall and proton transport in flagella (**Figure 6.2B**) (Namba, 2001). It is clear that many of the detected proteins, and specifically those involved in nucleotide binding, are located at the cytosol; however due to the large majority of bacterial proteins are categorized as cytosolic, this may result in a poor enrichment factor for this compartment. Overall, the most remarkable characteristics of the bacteria proteins containing PrLDs are their role in contact formation (e.g. macromolecular assembly), their relationship with the cell periphery and their involvement in nucleic acid mediated processes.

Structural domains linked to Bacteria PrLDs proteins

To learn more about the bacterial proteins that possess putative PrLDs we examined their constituent functional domains (Finn, et al., 2014) (**Figure 6.3**). After clustering the Pfam domains we obtained eight functional groups: nucleotide binding, cell wall dynamics, invasion and virulence, protein-protein interaction, iron transport, heat-shock and domains of unknown function.

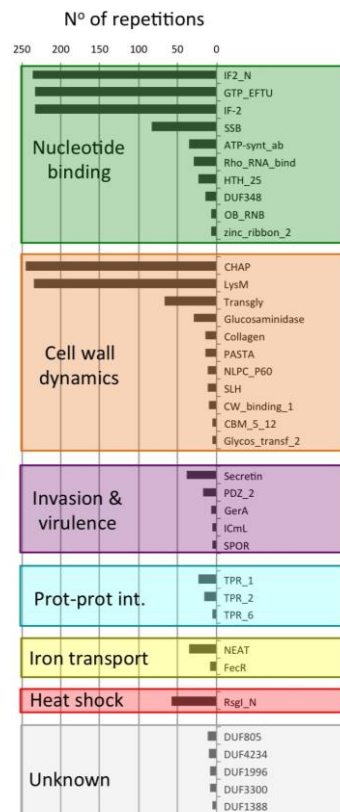


Figure 6.3 – Number of different Pfam domains found in PrLDs-containing proteins. The domains are indicated by their Pfam ID. This plot only shows the domains with > 5 repetitions in the dataset.

The most abundant group of Pfam domains is the one involved in nucleotide binding (1183 domains). Among them, several domains are associated to translation such as GTP-binding elongation factors (GTP_EFTU), Rho termination factors (Rho_RNA_bind and Rho_N) and translation initiation factors (IF2 and IF2-N). Canonical nucleotide binding domains are also be found such as the single stranded binding protein (SSB), the single zinc ribbon domain (zinc_ribbon_2), the major structural motif helix-turn-helix (HTHth-25) and the S1 RNA binding domain. Finally, in this group we can also find an ATP synthase domain,

associated with Rho termination factors (ATP-synt_ab), and the Ribonuclease B OB domain (Finn, et al., 2014).

The second most abundant group of Pfam domains is, as seen in the GO functional enrichment, associated to cell wall dynamics (978 domains). This group clusters domains involved in cell wall metabolism (including biosynthesis and degradation) and proteins that bind the wall to build functional structures. For instance, the lysine motif (Lysm) is involved in bacterial cell wall degradation and may also have peptidoglycan binding function (Bateman and Bycroft, 2000). The Glucosaminidase, Glycosyl transferase family 2 (Glycos_transf_2) and Transpeptidase are three domains associated with the biosynthesis of polysaccharides and peptidoglycan. We also found 67 proteins with a transglycosylase domain (Transgly) that catalyse the polymerisation of murein glycan chains as well as 12 proteins with a SLH domain that is associated with the assembly of (glyco)proteins that coat the bacteria surface. The PASTA domain is involved in cell wall biosynthesis and can bind the β -lactam rings enclosed in antibiotics. The most abundant domain from this group is the CHAP domain (245 proteins) with an amidase activity implicated in cell wall metabolism. Other domains also linked to cell wall are: the collagen domain (connective structures), the NlpC/P60 family (Anantharaman and Aravind, 2003) (peptidases associated to lipoproteins), the G5 domain (adhesion), the fibronectin type III (fn3, adhesion), the cell wall binding motif 1 (CW_binding_1, a repeat similar to some clostridia toxins) and the carbohydrate-binding module (CBM_5_12, enriched in chitinases and associated to cellulose scaffolding). Additionally, the unknown domain DUF1388 has also been associated with surface lipoproteins.

The third group contains 130 proteins with domains associated to secretion and invasion. Here we have several domains associated to sporulation (SPOR) and spore germination (GerA). The secretin domains are involved in protein export via pore formation in a signal sequence-dependent manner (Tosi, et al., 2014; Van der Meer, et al., 2013). The PDZ domains maintain together and organize signalling complexes located throughout the cellular membranes. Finally, the macrophage killing protein domain (ICmL) and the Endotoxin_N are domains involved in the formation of pores at the host cell membrane (Finn, et al., 2014).

Between the PrLDs containing proteins we have also found three different tetratricopeptide repeat domains (46 repetitions), which scaffold protein-protein interactions and mediate the assembly of multi-protein complexes. In addition, we also obtained 54 domains linked to iron binding and transport (Metallophos, NEAT and FecR) and 58 proteins involved in heat shock response (Anti-sigma factor N-terminus), both types of domains aimed to interact with or to transduce signals coming from the cell external microenvironment.

Overall, the functional families of the PrLDs containing proteins (**Figure 6.3**) resemble their GO enrichment classifications (**Figure 6.1** and **6.2**), supporting the idea that these proteins are predominantly associated to the external part of the cell (e.g. cell wall) and in interactions with other molecules (e.g. nucleotide binding).

Structure composition of bacteria PrLDs containing proteins

As expected, the detected PrLDs are located inside low complexity regions (e.g. disordered, coiled coil, etc) (**Figure 6.4** and **6.5**). These regions are abundant in prion-like proteins and connect different structural domains (**Figure 6.4**) and elements with secondary structure (**Figure 6.5**).

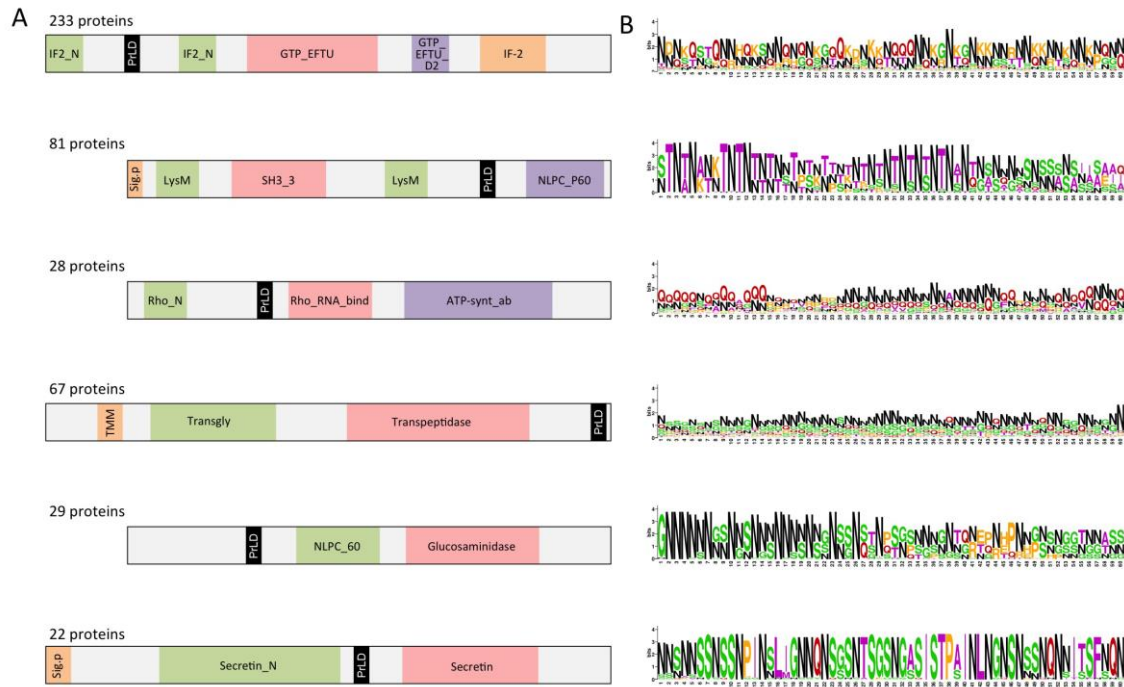


Figure 6.4 – PrLDs-containing proteins also contain multiple domains. A) Diagrams showing a consensus distribution and size of the most common domain combinations as collected in Pfam. The light grey spaces represent low complexity regions (coiled coil, disordered, etc). The domains are indicated by their Pfam ID. **B)** PrLDs sequence conservation measured in bits. The symbol height reflects the relative frequency of the corresponding amino acid at that position. Colour code: N in black; G in red; G, S and Y in green; the other residues in purple.

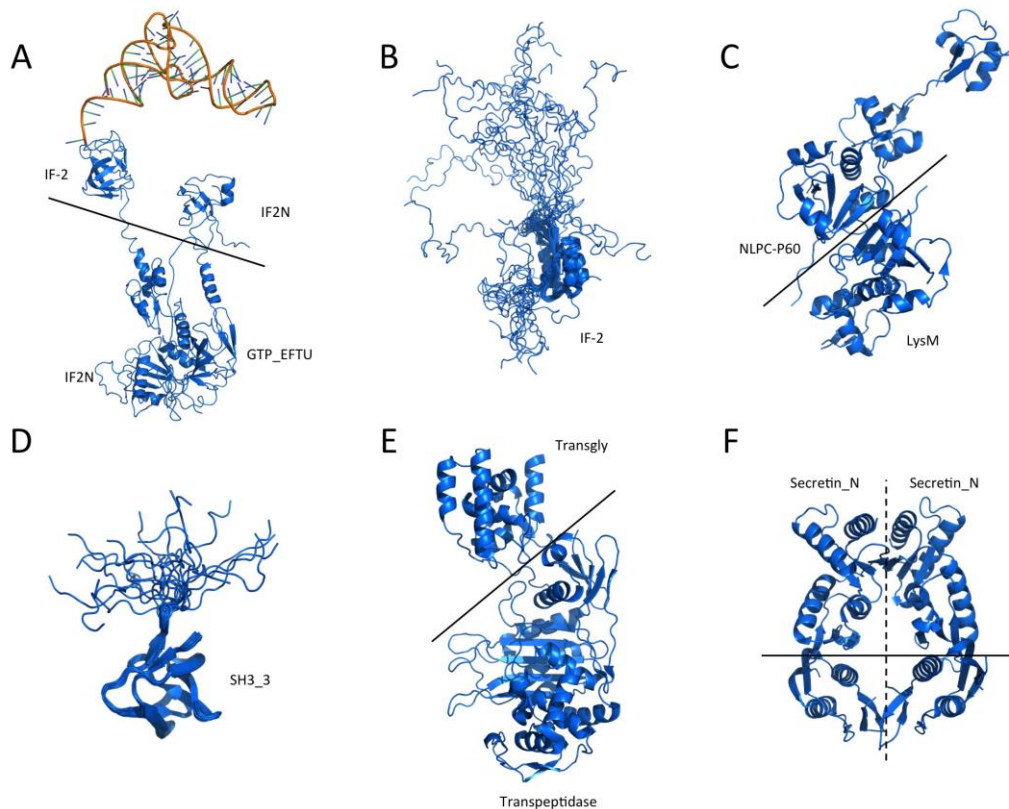


Figure 6.5 – Structure of the domains located in the PrLDs-containing proteins. Representative structures of the domains and domain combinations enclosed in the PrLD-containing proteins. The domains are indicated by their Pfam ID. **A)** Example of quaternary complex where a multi-domain structure, composed by IF-2, IF2_N and GTP_EFTU domains, interact with a tRNA. The image shows partial information from the PDB structure 1MJ1. Fitting the ternary complex of EF-TU/tRNA/GTP and ribosomal proteins into a 13 Å cryo-EM map of the coli 70S ribosome. **B)** Example of IF-2 domain structure and the different states of the disordered region located in front of it. PDB structure 1Z9B. Solution NMR structure of the C1-subdomain of *Bacillus stearothermophilus* translation initiation factor IF2 (fragment 515 - 635). **C)** Example of multi-domain structure composed by a Nlpc-P60 and a Lysm domains. PDB structure 4XCM. Crystal structure of the putative Nlpc/P60 D,L endopeptidase from *Thermus thermophilus*. **D)** Example of SH3_3 domain structure and the different states of the disordered region located after it. PDB structure 2KRS. Solution NMR structure of SH3 domain from CPF_0587 (fragment 415-479) from *Clostridium perfringens*. **E)** Example of multi-domain structure composed by a transglycosylase and a transpeptidase domain. PDB structure 3ZG7 Crystal Structure of Penicillin-Binding Protein 4 from *Listeria monocytogenes* in the apo form. **F)** Structure showing a homodimer constituted by Secretin_N domains. PDB structure 4E9J. Crystal structure of the N-terminal domain of the secretin XcpQ from *Pseudomonas aeruginosa*. Notice that at the multi-domain structures (**B** and **D**) the low complexity regions are abundant.

From 2200 PrLDs containing proteins, 1514 have at least one defined Pfam domain (69%). Additionally, 612 of these sequences (40%) have more than one structural domain (Ekman, et al., 2005). When we focus on the prion-like proteins from pathogenic bacteria (**Figure 6.6**), we observe that they have a lower number of designated Pfam domains (only 301 proteins) suggesting they could be less structured proteins or, more likely, carry still unknown domains and functions. Despite this, the proteins from pathogenic bacteria with reported Pfam domains tend to contain more than one structural domain family. The percentage of proteins with multiple domains appears to be higher in these proteomes (60%) than in the complete protein dataset (Ekman, et al., 2005).

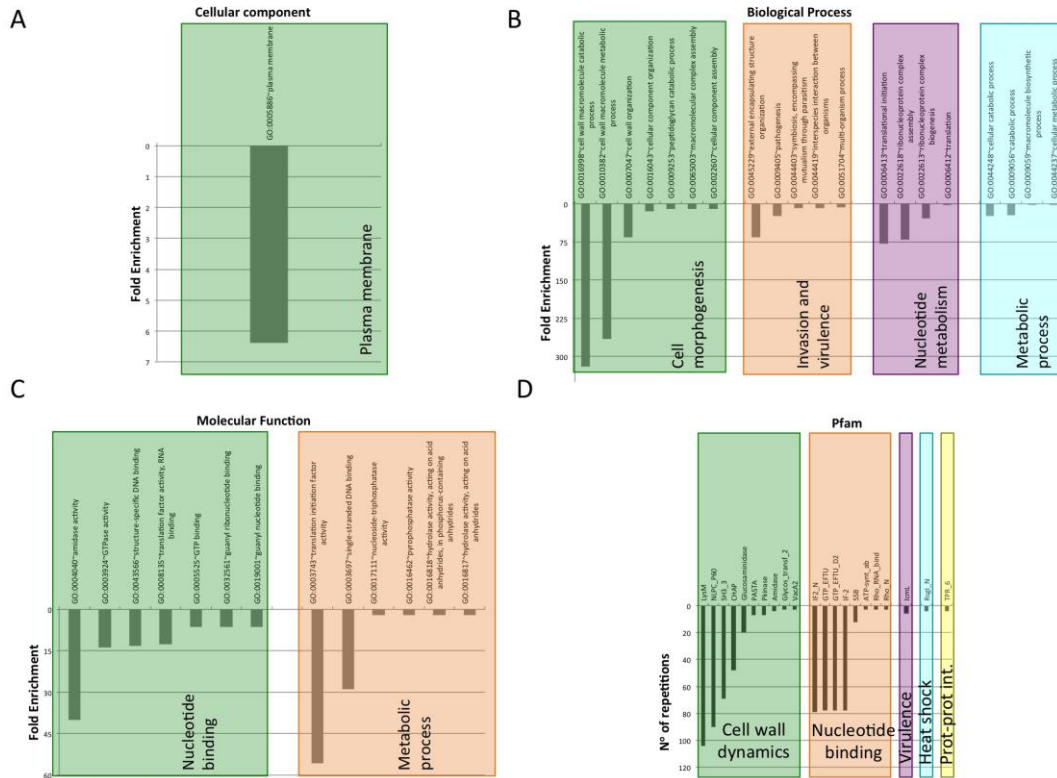


Figure 6.6 – Clustering of GO terms and Pfam domains associated to PrLDs-containing proteins in pathogen bacteria. A) Cellular component GO terms. B) Biological Process GO terms. C) Molecular Function GO terms. D) Pfam domains associated.

When the proteins have multiple structural domains, the PrLDs regions can be located either close to an end or between structures (**Figure 6.4A**). Interestingly, the amino acid composition of the PrLDs regions is similar between proteins sharing similar domain arrangement but different between proteins with distinct domains composition (**Figure 6.4B**). In agreement with the data reported for yeast prion PrDs, we observe that the bacterial PrLDs are abundant in N (30%), Q (21%), S (11%) and G (11%).

The domain combinations tend to be functionally associated. For example, we found 233 protein sequences containing two GTP-binding elongation factor domains and two translation initiation factor domains that are related with nucleotide binding and translation (**Figure 6.4**). During protein synthesis the initiation factors (IF2) form a ternary complex with GTP and the initiator Met-tRNA (Wienk, et al., 2005). This complex binds the ribosome to interact with the AUG-codon of the starting methionine, once the codon is found IF2 has to hydrolyse its GTP to be released (**Figure 6.5A** and **6.5B**).

P60 domain is a cell-wall-associated peptidase domain essential for adherence and invasion in some *Listeria* species. In agreement with previous studies (Anantharaman and Aravind, 2003; Ponting, et al., 1999), we observed the P60 domain associated with SH3 and LysM domains (**Figure 6.4**, **Figure 6.5C** and **6.5D**). It has been hypothesized that this combination facilitates the interaction with peptides, carbohydrates and lipids from the bacterial cell wall and thus their functionality (Anantharaman and Aravind, 2003; Ponting, et al., 1999).

Rho factor proteins tend to be accompanied with an RNA-binding domain and an ATP-hydrolysis domain (**Figure 6.4**). The Rho termination factor disengages newly transcribed RNA from its DNA template. Rho catalyses the 3' endpoint formation and the release of mRNA molecules from DNA templates (Skordalakes and Berger, 2003). The hydrolysis of ATP provides the energy required to get the RNA-DNA region and break the hybrid structure.

Another example of functional domain combination that contains PrLDs are the penicillin-binding proteins. They are bifunctional proteins involved in the final stages of the peptidoglycan synthesis (**Figure 6.4** and **Figure 6.5E**). At the N-terminus there is a transglycosylase domain involved in the formation of linear glycan strands. And at the C-terminus there is a transpeptidase domain involved in the cross-linking of peptide subunits and drug binding, which is also responsible of the penicillin-sensitivity (Contreras-Martel, et al., 2011; Macheboeuf, et al., 2005; Sauvage, et al., 2008).

NLPC/P60 and Glucosaminidase are two cell wall endopeptidase domains, which emerged together and that we have found accompanied with a PrLDs (**Figure 6.4**). These two domains are commonly employed to cleave the septa connecting the daughter cells during cell separation (Anantharaman and Aravind, 2003; Ruggiero, et al., 2010).

The secretins are another example of domain combination found in our set of PrLDs bacteria containing proteins (**Figure 6.4** and **6.5F**). Particularly it is the most abundant combination of two domains (67 times) found in the PrLDs containing proteins. The secretin domains detected take part in protein secretion systems type II and III. They build multimeric pores to transport macromolecules either to the periplasm or to inject them into eukaryotic cells (Tosi, et al., 2014). In general, secretin proteins consist of two domains: an N-terminal periplasmic domain responsible of the pore formation and a C-terminal domain responsible of the attachment to the outer membrane (Tosi, et al., 2014; Van der Meeren, et al., 2013). Interestingly, the PrLDs detected are located between these two secretin domains (**Figure 6.4**).

6.1.5 DISCUSSION

Bacterial PrLDs are associated to cellular adaptability

We observed that a significant fraction of the bacteria PrLDs containing proteins are located at the cell periphery and are involved in cell wall metabolism, especially peptidoglycan biogenesis. Peptidoglycan is the major component of bacterial cell walls; it is essential for growth, cell division and maintenance of the cellular shape, enabling the bacteria to resist intracellular pressures of several atmospheres. In some particular cases, the proteins present in the peptidoglycan can be anchored to the biofilm amyloid network and, more interesting, assist its assembly. This is the case of the TapA protein from *Bacillus subtilis*, which is present in the peptidoglycan, where it functions as an anchor point for TasaA fibres (Friedland, 2015; Romero, et al., 2011; Sauvage, et al., 2008). The formation of biofilms is a powerful strategy that protects a bacterial community from chemicals and antibiotics and facilitates the attachment to different surfaces even host cells. Interestingly, *Staphylococcus aureus*, a biofilm forming pathogen, is the bacteria specie with the highest content in PrLDs. In this organism we found PrLDs-containing proteins

linked to cell wall, proteins involved in secretion and proteins associated to virulence. These data point to a possible relationship between the identified proteins and the biofilm formation. In fact, preliminary data shows the *S. aureus* PrLDs-containing protein staphylococcal secretory antigen ssaA2 (UniprotKB accession number Q2G2J2) is able to form amyloid fibrils *in vitro*. Thus, a more exhaustive analysis of these proteins might confirm their association to biofilm formation and their possible role as a druggable targets.

The other processes enriched in the PrLDs containing proteins can also provide versatility and adaptability to different environments. For instance, the proteins involved in stimulus response and invasion and in virulence have a clear role in supporting the bacteria development under variable conditions. From inside the cell the nucleotide binding proteins can be involved in functions that support cell adjustment such as transcription and translation (i.e. change the expression levels) or DNA repair that can enhance cell survival in stress conditions. Interestingly, most of the novel prion-like proteins discovered recently in humans play a role in RNA/DNA binding (King, et al., 2012). In bacteria, we also found proteins involved in cellular localization that can rearrange different compounds adapting the cell to new requirements. Overall, as previously proposed for yeast prions, bacterial prions might serve as bet-hedging devices for diversifying microbial phenotypes.

Bacterial PrLDs are associated to functional and interacting proteins

The 69% of PrLDs containing proteins have defined Pfam domains and 40% of them carry multiple domains. Since domains come together to increase proteins functionality (Alberti, et al., 2009; Anantharaman and Aravind, 2003), our data suggest that the proteins with PrLDs tend to be functional. Moreover, in pathogenic bacteria PrLDs are associated to higher percentage of proteins with multiple domains, more than the average of the proteomes from this taxon (Ekman, et al., 2005). This data suggests that, in pathogenic bacteria, PrLDs containing proteins might have a versatile character.

The detected PrLDs are located in proteins rich in low complexity regions. These regions are important to provide the structural flexibility required to form interactions between proteins. This flexibility also allows the formation of reversible interactions, which are essential to build dynamic macromolecular assemblies. In fact, the GO terms associated to the PrLDs detected by PrionScan comprise functions and processes linked to interaction and assembly. Many of these GO terms involve binding proteins, nucleotides or other cellular compounds. Human RNA/DNA binding proteins use their PrLDs to attain functional macromolecular assemblies that regulate transcription and translation. In many cases these functions are exerted in the so called ribonucleoprotein granules (Malinowska, et al., 2013). Many of the proteins containing DNA/RNA binding domains identified in the present also work by forming large complexes and indeed are implied in ribonucleoprotein complex biogenesis and assembly suggesting that this property can be conserved across species. In addition, the association to cell wall dynamics suggests that certain proteins can be implied in the assembly and disassembly of peptidoglycans and polysaccharides. Overall, our data supports that, as previously suggested for eukaryotic PrLDs, bacteria PrLDs could play an important role in the arrangement of macromolecular structures (Malinowska, et al., 2013).

Prions in other proteomes

Saccharomyces cerevisiae is the organism from which more information about its prion proteins has been so far collected (Alberti, et al., 2009; Malinovska, et al., 2013). These works showed for the first time that proteins could be employed for relevant functions such as epigenetic elements essential to adapt the cellular metabolism and increase the cell survival in front of environmental changes (Alberti, et al., 2009; Newby and Lindquist, 2013). In *S. cerevisiae* the prion proteins are associated to functions that involve the formation of contacts such as RNA-binding, membrane-interacting, DNA binding and protein interaction domains (Malinovska, et al., 2013). These proteins are located at the cytoskeleton, nucleus, ribonucleoprotein complexes and chromatin. Comparing *S. cerevisiae* with other eukaryotic proteomes shows PrLDs-containing proteins with similar function and location. For example, in human and fruit fly these proteins are also involved in transcription, chromatin remodelling, ribonucleoprotein complex formation, and cytoskeleton (Malinovska, et al., 2013). As a general trend, PrLDs in Animalia tend to be involved in the regulation of central biological processes and organism development, which in vertebrates includes the development of the neural crest. Hence, many human PrLDs are found in RNA-binding proteins, whose deregulation has previously been associated with several neurodegenerative diseases (King, et al., 2012).

Eukaryote PrLDs-containing proteins show less functional diversity than bacteria. In fact, here we have collected all the enriched eukaryote functions (i.e. transcription, RNA binding and DNA binding) in just one cluster (nucleotide binding). Despite this difference, it appears that, independently of the considered taxon, PrLDs-containing proteins mostly appear to be involved in a similar regulatory purpose: adapting the cell to a variable environment. This is basically achieved through the control of the gene expression in eukaryotes, but in prokaryotes this is also reached by interacting with the environment. This difference could be due to the different surrounding conditions as microorganisms face the constant challenge of fluctuating conditions in their natural environments. These strategies may have facilitated the invasion of new environments (e.g. water, air) and the coexistence or exploitation of diverse life forms (e.g. host cells).

Bacteria PrLDs and human diseases

Our life is closely linked to bacteria, either through a parasitic or symbiotic relationship. On one hand, human microbiota is required to assist many processes and ensure a healthy body. On the other hand, many common pathogenic bacteria are acquiring antibiotic resistance in all regions of the world (e.g. urinary tract infections, pneumonia, bloodstream infections) (WorldHealthOrganisation, WHO). These bacteria cause many hospital-acquired infections, such as the methicillin-resistant *Staphylococcus aureus*, with an associated high mortality rate (Contreras-Martel, et al., 2011; WorldHealthOrganisation, WHO). To the already intricate scenario where bacteria and host interact, the risk of their amyloid proteins concurring and altering their conformational states adds an extra level of complexity (Otzen and Nielsen, 2008). Additionally, the long periods that bacteria stay in the body, due to chronic infection or microbiota coexistence, enhances the chances of this event. In fact, recent studies have demonstrated that bacterial

amyloids can initiate the formation of amyloid aggregates upon interaction with host proteins (Hill and Lukiw, 2015; Hufnagel, et al., 2013; Otzen and Nielsen, 2008; Zhou, et al., 2012). Moreover, it has been reported that the injection of bacteria amyloids in mice causes the development of amyloidosis (Lundmark, et al., 2005). Overall, these data reminds the conformational template process associated to prion transmission and suggest that bacterial infection could be linked to neurodegenerative diseases (Friedland, 2015).

6.1.6 GENERAL CONCLUSIONS

Despite PrLDs-containing proteins seem to be ubiquitous (Espinosa Angarica, et al., 2013; Malinovska, et al., 2013) they play distinct functional roles across different organisms. In this background, the mechanisms underlying host-bacteria relationship are just starting to be elucidated and, as a result, also the interplay between their amyloid proteins (Schwartz and Boles, 2013; Seviour, et al., 2015; Zhou, et al., 2012). The studies on bacterial amyloids are showing us that these organisms rely on amyloid aggregates to execute a wide range of physiological functions (Blanco, et al., 2012; DePas and Chapman, 2012; Evans, et al., 2015; Gsponer and Babu, 2012; Schwartz and Boles, 2013; Seviour, et al., 2015; Taylor and Matthews, 2015; Zhou, et al., 2012). Although because the formation of amyloids comes at expenses of the presence of transient toxic species, cells tightly control the assembly of these macromolecular structures and how they can interact with proteins from other species (Evans, et al., 2015; Schwartz and Boles, 2013; Taylor and Matthews, 2015; Zhou, et al., 2012). Most of the bacterial amyloids described so far play a structural role and work extracellularly. Indeed, some of the PrLDs containing proteins with potential amyloidogenic properties were be linked to biofilms, structures that favour chronic human infections and, consequently, could increase the chances of a potential bacterial prion to alter the conformation of host proteins. However, despite their *in vitro* amyloid potential and *in vivo* prionic behaviour should be validated, the data in the present work suggest that, as it happens in yeast and humans, also in bacteria amyloid-like assemblies might play a regulatory role, since some of the detected candidates are linked to fundamental cellular functions such as transcription, translation or DNA repair. Intriguingly, linking the fact that we found at the same time association with extracellular environment and nucleic acid binding function, it has been reported recently that extracellular DNA is bound tightly by bacterial amyloid fibrils during biofilm formation and that amyloid/DNA composites are immune stimulators when injected into mice, leading to autoimmunity (Gallo, et al., 2015; Spaulding, et al., 2015). Overall, it becomes clear that a more exhaustive analysis of the putative bacterial prion proteins identified here is required in order to attain a better understand of their functional role and their relationship with human diseases. We envision the presented data could help to identify new drug targets and develop new potential therapeutic approaches.

6.1.6 REFERENCES

- Alberti, S., *et al.* (2009) A systematic survey identifies prions and illuminates sequence features of prionogenic proteins, *Cell*, **137**, 146-158.
- Anantharaman, V. and Aravind, L. (2003) Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes, *Genome Biol*, **4**, R11.
- Bateman, A. and Bycroft, M. (2000) The structure of a LysM domain from *E. coli* membrane-bound lytic murein transglycosylase D (MltD), *J Mol Biol*, **299**, 1113-1119.
- Beissbarth, T. and Speed, T.P. (2004) Gostat: find statistically overrepresented Gene Ontologies within a group of genes, *Bioinformatics*, **20**, 1464-1465.
- Blanco, L.P., *et al.* (2012) Diversity, biogenesis and function of microbial amyloids, *Trends in microbiology*, **20**, 66-73.
- Contreras-Martel, C., *et al.* (2011) Structure-guided design of cell wall biosynthesis inhibitors that overcome beta-lactam resistance in *Staphylococcus aureus* (MRSA), *ACS Chem Biol*, **6**, 943-951.
- Coustou, V., *et al.* (1997) The protein product of the het-s heterokaryon incompatibility gene of the fungus *Podospira anserina* behaves as a prion analog, *Proc Natl Acad Sci U S A*, **94**, 9773-9778.
- Chapman, M.R., *et al.* (2002) Role of *Escherichia coli* curli operons in directing amyloid fiber formation, *Science*, **295**, 851-855.
- Chien, P. and Weissman, J.S. (2001) Conformational diversity in a yeast prion dictates its seeding specificity, *Nature*, **410**, 223-227.
- Chiti, F. and Dobson, C.M. (2006) Protein misfolding, functional amyloid, and human disease, *Annu Rev Biochem*, **75**, 333-366.
- DePas, W.H. and Chapman, M.R. (2012) Microbial manipulation of the amyloid fold, *Research in microbiology*, **163**, 592-606.
- Dorsman, J.C., *et al.* (2002) Strong aggregation and increased toxicity of poly-leucine over poly-glutamine stretches in mammalian cells, *Hum Mol Genet*, **11**, 1487-1496.
- Ekman, D., *et al.* (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions, *J Mol Biol*, **348**, 231-243.
- Espargaro, A., *et al.* (2012) Yeast prions form infectious amyloid inclusion bodies in bacteria, *Microbial cell factories*, **11**, 89.
- Espinosa Angarica, V., *et al.* (2014) PrionScan: an online database of predicted prion domains in complete proteomes, *BMC Genomics*, **15**, 102.
- Espinosa Angarica, V., Ventura, S. and Sancho, J. (2013) Discovering putative prion sequences in complete proteomes using probabilistic representations of Q/N-rich domains, *BMC Genomics*, **14**, 316.
- Evans, M.L., *et al.* (2015) The bacterial curli system possesses a potent and selective inhibitor of amyloid formation, *Mol Cell*, **57**, 445-455.
- Fandrich, M. and Dobson, C.M. (2002) The behaviour of poly-amino acids reveals an inverse side chain effect in amyloid structure formation, *EMBO J*, **21**, 5682-5690.
- Finn, R.D., *et al.* (2014) Pfam: the protein families database, *Nucleic Acids Res*, **42**, D222-230.
- Fowler, D.M., *et al.* (2006) Functional amyloid formation within mammalian tissue, *PLoS biology*, **4**, e6.
- Friedland, R.P. (2015) Mechanisms of molecular mimicry involving the microbiota in neurodegeneration, *J Alzheimers Dis*, **45**, 349-362.
- Fuxreiter, M. (2012) Fuzziness: linking regulation to protein dynamics, *Mol Biosyst*, **8**, 168-177.
- Fuxreiter, M. and Tompa, P. (2012) Fuzzy complexes: a more stochastic view of protein function, *Adv Exp Med Biol*, **725**, 1-14.
- Gallo, P.M., *et al.* (2015) Amyloid-DNA Composites of Bacterial Biofilms Stimulate Autoimmunity, *Immunity*, **42**, 1171-1184.
- Garrity, S.J., *et al.* (2010) Conversion of a yeast prion protein to an infectious form in bacteria, *Proc Natl Acad Sci U S A*, **107**, 10596-10601.
- Graether, S.P., Slupsky, C.M. and Sykes, B.D. (2003) Freezing of a fish antifreeze protein results in amyloid fibril formation, *Biophys J*, **84**, 552-557.
- Gsponer, J. and Babu, M.M. (2012) Cellular strategies for regulating functional and nonfunctional protein aggregation, *Cell reports*, **2**, 1425-1437.
- Halfmann, R., *et al.* (2011) Opposing effects of glutamine and asparagine govern prion formation by intrinsically disordered proteins, *Mol Cell*, **43**, 72-84.
- Harrison, P.M. and Gerstein, M. (2003) A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes, *Genome Biol*, **4**, R40.
- Hill, J.M. and Lukiw, W.J. (2015) Microbial-generated amyloids and Alzheimer's disease (AD), *Front Aging Neurosci*, **7**, 9.
- Hufnagel, D.A., Tukul, C. and Chapman, M.R. (2013) Disease to dirt: the biology of microbial amyloids, *PLoS Pathog*, **9**, e1003740.
- Iconomidou, V.A., Vriend, G. and Hamodrakas, S.J. (2000) Amyloids protect the silkworm oocyte and embryo, *FEBS letters*, **479**, 141-145.
- Karran, E., Mercken, M. and De Strooper, B. (2011) The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics, *Nat Rev Drug Discov*, **10**, 698-712.
- Kato, M., *et al.* (2012) Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels, *Cell*, **149**, 753-767.

King, O.D., Gitler, A.D. and Shorter, J. (2012) The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease, *Brain Res*, **1462**, 61-80.

Kushnirov, V.V., *et al.* (2007) Prion and nonprion amyloids: a comparison inspired by the yeast Sup35 protein, *Prion*, **1**, 179-184.

Liebman, S.W. and Chernoff, Y.O. (2012) Prions in yeast, *Genetics*, **191**, 1041-1072.

Lim, W.A. and Sauer, R.T. (1991) The role of internal packing interactions in determining the structure and stability of a protein, *J Mol Biol*, **219**, 359-376.

Lundmark, K., *et al.* (2005) Protein fibrils in nature can enhance amyloid protein A amyloidosis in mice: Cross-seeding as a disease mechanism, *Proc Natl Acad Sci U S A*, **102**, 6098-6102.

Macheboeuf, P., *et al.* (2005) Active site restructuring regulates ligand recognition in class A penicillin-binding proteins, *Proc Natl Acad Sci U S A*, **102**, 577-582.

Maji, S.K., *et al.* (2009) Functional amyloids as natural storage of peptide hormones in pituitary secretory granules, *Science*, **325**, 328-332.

Malinowska, L., Kroschwald, S. and Alberti, S. (2013) Protein disorder, prion propensities, and self-organizing macromolecular collectives, *Biochim Biophys Acta*, **1834**, 918-931.

Michelitsch, M.D. and Weissman, J.S. (2000) A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions, *Proc Natl Acad Sci U S A*, **97**, 11910-11915.

Monsellier, E., *et al.* (2007) The distribution of residues in a polypeptide sequence is a determinant of aggregation optimized by evolution, *Biophys J*, **93**, 4382-4391.

Namba, K. (2001) Roles of partly unfolded conformations in macromolecular self-assembly, *Genes Cells*, **6**, 1-12.

Newby, G.A. and Lindquist, S. (2013) Blessings in disguise: biological benefits of prion-like mechanisms, *Trends Cell Biol*, **23**, 251-259.

Otzen, D. and Nielsen, P.H. (2008) We find them here, we find them there: functional bacterial amyloid, *Cell Mol Life Sci*, **65**, 910-927.

Pawar, A.P., *et al.* (2005) Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases, *J Mol Biol*, **350**, 379-392.

Podrabsky, J.E., Carpenter, J.F. and Hand, S.C. (2001) Survival of water stress in annual fish embryos: dehydration avoidance and egg envelope amyloid fibers, *Am J Physiol Regul Integr Comp Physiol*, **280**, R123-131.

Ponting, C.P., *et al.* (1999) Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer, *J Mol Biol*, **289**, 729-745.

Romero, D., *et al.* (2011) An accessory protein required for anchoring and assembly of amyloid fibres in *B. subtilis* biofilms, *Molecular microbiology*, **80**, 1155-1168.

Ross, E.D., Baxa, U. and Wickner, R.B. (2004) Scrambled prion domains form prions and amyloid, *Mol Cell Biol*, **24**, 7206-7213.

Ross, E.D., *et al.* (2005) Primary sequence independence for prion formation, *Proc Natl Acad Sci U S A*, **102**, 12825-12830.

Ruggiero, A., *et al.* (2010) Structure and functional regulation of RipA, a mycobacterial enzyme essential for daughter cell separation, *Structure*, **18**, 1184-1190.

Sabate, R., *et al.* (2009) Characterization of the amyloid bacterial inclusion bodies of the HET-s fungal prion, *Microbial cell factories*, **8**, 56.

Sabate, R., *et al.* (2015) What makes a protein sequence a prion?, *PLoS Comput Biol*, **11**, e1004013.

Sanchez de Groot, N., *et al.* (2015) Proteome response at the edge of protein aggregation, *Open Biol*, **5**, 140221.

Sanchez de Groot, N., *et al.* (2012) Evolutionary selection for protein aggregation, *Biochem Soc Trans*, **40**, 1032-1037.

Sanchez, I.E., *et al.* (2006) Point mutations in protein globular domains: contributions from function, stability and misfolding, *J Mol Biol*, **363**, 422-432.

Sauvage, E., *et al.* (2008) The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis, *FEMS Microbiol Rev*, **32**, 234-258.

Schwartz, K. and Boles, B.R. (2013) Microbial amyloids--functions and interactions within the host, *Current opinion in microbiology*, **16**, 93-99.

Seviour, T., *et al.* (2015) Functional amyloids keep quorum-sensing molecules in check, *J Biol Chem*, **290**, 6457-6469.

Shorter, J. and Lindquist, S. (2005) Prions as adaptive conduits of memory and inheritance, *Nat Rev Genet*, **6**, 435-450.

Skordalakes, E. and Berger, J.M. (2003) Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading, *Cell*, **114**, 135-146.

Spaulding, C.N., *et al.* (2015) Fueling the Fire with Fibers: Bacterial Amyloids Promote Inflammatory Disorders, *Cell host & microbe*, **18**, 1-2.

Staniforth, G.L. and Tuite, M.F. (2012) Fungal prions, *Prog Mol Biol Transl Sci*, **107**, 417-456.

Stohr, J., *et al.* (2012) Purified and synthetic Alzheimer's amyloid beta (A β) prions, *Proc Natl Acad Sci U S A*, **109**, 11025-11030.

Taylor, J.D. and Matthews, S.J. (2015) New insight into the molecular control of bacterial functional amyloids, *Frontiers in cellular and infection microbiology*, **5**, 33.

Tompa, P. and Fuxreiter, M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions, *Trends Biochem Sci*, **33**, 2-8.

Toombs, J.A., McCarty, B.R. and Ross, E.D. (2010) Compositional determinants of prion formation in yeast, *Mol Cell Biol*, **30**, 319-332.

- Tosi, T., *et al.* (2014) Structural similarity of secretins from type II and type III secretion systems, *Structure*, **22**, 1348-1355.
- UniProt, C. (2015) UniProt: a hub for protein information, *Nucleic Acids Res*, **43**, D204-212.
- Van der Meeren, R., *et al.* (2013) New insights into the assembly of bacterial secretins: structural studies of the periplasmic domain of XcpQ from *Pseudomonas aeruginosa*, *J Biol Chem*, **288**, 1214-1225.
- Ventura, N.S.d.G.a.S. (2005) Amyloid fibril formation by bovine cytochrome c, *Spectroscopy*, **19**, 6.
- Wienk, H., *et al.* (2005) Solution structure of the C1-subdomain of *Bacillus stearothermophilus* translation initiation factor IF2, *Protein Sci*, **14**, 2461-2468.
- WorldHealthOrganisation (WHO) <http://www.who.int/mediacentre/factsheets/fs194/en/>.
- Yuan, A.H., *et al.* (2014) Prion propagation can occur in a prokaryote and requires the ClpB chaperone, *eLife*, **3**, e02949.
- Zhou, Y., *et al.* (2012) Promiscuous cross-seeding between bacterial amyloids promotes interspecies biofilms, *J Biol Chem*, **287**, 35092-35103.

6.2 Discovering putative prion-like proteins in *Plasmodium falciparum*: A computational and experimental analysis

Valentín Iglesias^{1,2†}, Irantzu Pallarès^{1,2†}, Natalia S. de Groot^{3,4†}, Sant'Anna Ricardo^{1,2}, Arnau Biosca^{5,6,7}, Xavier Fernández-Busquets^{5,6,7} & Salvador Ventura^{1,2,*}

¹ Institut de Biotecnologia i de Biomedicina and ²Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, E-08193 Bellaterra (Barcelona), Spain.

³ Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, E-08003 Barcelona, Spain and ⁴ Universitat Pompeu Fabra (UPF), Barcelona, Spain.

⁵ Nanomalaria Group, Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology, Baldiri Reixac 10-12, E-08028 Barcelona, Spain. And ⁶ Barcelona Institute for Global Health (ISGlobal), Barcelona Center for International Health Research (CRESIB, Hospital Clínic-Universitat de Barcelona), Rosselló 149-153, E-08036 Barcelona, Spain. And ⁷ Nanoscience and Nanotechnology Institute (IN2UB), University of Barcelona, Martí i Franquès 1, E-08028 Barcelona, Spain.

†These authors contributed equally to this work. *To whom correspondence should be addressed.

Author Contributions: Software, *in silico* validation, data curation, writing—original draft preparation.

6.2.1 ABSTRACT

Prions are a singular subset of proteins able to switch between a soluble conformation and a self-perpetuating amyloid state. Traditionally associated with neurodegenerative diseases, increasing evidence indicates that organisms exploit prion-like mechanisms for beneficial purposes. The ability to transit between conformations is encoded in the so-called prion domains, long disordered regions usually enriched in glutamine/asparagine residues. Interestingly, *Plasmodium falciparum*, the parasite that causes the most virulent form of malaria, is exceptionally rich in proteins bearing long Q/N-rich sequence stretches, accounting for roughly 30% of the proteome. This biased composition suggests that these protein regions might correspond to prion-like domains (PrLDs) and potentially form amyloid assemblies. To investigate this possibility, we performed a stringent computational survey for Q/N-rich PrLDs on *P. falciparum*. Our data indicates that ~10% of *P. falciparum* protein sequences have prionic signatures, and that this subproteome is enriched in regulatory proteins, such as transcription factors and RNA-binding proteins. Furthermore, we experimentally characterize that despite their disordered nature, several of the identified PrLDs contain inner short sequences able to spontaneously self-assemble into amyloid-like structures. Although the ability of these sequences to nucleate the conformational conversion of the respective full-length proteins should still be demonstrated, our analysis suggests that, as previously described for other organisms, prion-like proteins might play functional roles in *P. falciparum* physiology.

6.2.2 INTRODUCTION

Malaria caused approximately 445000 deaths in 2016 and in the latest World Malaria Report (November 2017) the number of cases was estimated to be as many as 216 million. Although the global response to malaria is considered one of the world's great public health achievements, the spread of resistance against anti-malarial drugs and insecticides, has stalled the incidence and mortality decline since 2014.

Plasmodium falciparum (*P. falciparum*) is the species responsible for 85% of the malaria cases, causing the most severe form of the disease. The complete sequencing of *P. falciparum* genome has revealed some specific features that may shed light onto the biology and biochemistry of this deadly parasite (Gardner, et al., 2002). A striking biased composition of its DNA was observed, with an overall AT content of 80.6%, a comparable AT enrichment only being observed in the social amoeba *Dictyostelium discoideum* (Eichinger, et al., 2005). In *P. falciparum*, AT-rich codons present a significant preference towards encoding asparagines (N) over lysines (K), which explains why ~30% of its proteome is rich in long low complexity regions exceptionally enriched in N (Aravind, et al., 2003; Singh, et al., 2004).

Glutamine (Q)- and asparagine (N)-rich sequences have been shown to increase the propensity of a protein to form amyloids, and indeed the expansion of CAG trinucleotide repeats, encoding for Q, in different human proteins, results in developmental and neuromuscular disorders such as Huntington's disease, Kennedy disease, and several ataxias caused by the accumulation of intracellular protein aggregates in specific neuron types (Orr and Zoghbi, 2007; Williams and Paulson, 2008). Proteins with long N repeats have been shown to have an aggregation propensity even higher than poly-Q stretches (Tartaglia, et al., 2005). Intriguingly, in spite of their inherent risk to promote aggregation, sequences with such amino acid compositions are common in yeast prions, and, thus are often referred to as prion domains (PrDs).

Among amyloids, prions are proteins with the unusual ability to adopt different structures and functional states, at least one of which is transmissible between individuals. In yeast, PrDs have been proved to be both sufficient and necessary for prion conformational conversion (Masison and Wickner, 1995). The detailed characterization of the prion phenomenon in yeast has provided important insights on the structural and sequential determinants of PrDs (Alberti, et al., 2009). This knowledge has fuelled the development of computational algorithms able to identify prion-like domains (PrLDs) in a genome-wide level in different organisms (Espinosa Angarica, et al., 2014; Espinosa Angarica, et al., 2013; Harrison and Gerstein, 2003; Lancaster, et al., 2014; Michelitsch and Weissman, 2000; Zambrano, et al., 2015), highlighting the existence of proteins bearing such intriguing sequences in all kingdoms of life (Espinosa Angarica, et al., 2014; Espinosa Angarica, et al., 2013).

It is now clear that evolution purges out proteins containing aggregation-prone regions, unless these sequences are beneficial, serving functional purposes (Chen and Dokholyan, 2008; Monsellier and Chiti, 2007). Given the intrinsic amyloid potential of PrLDs, their biological persistence suggests an

evolutionary determination to maintain these regions. The word prion is usually associated with neurodegenerative diseases. However, the recent identification of protein prion-like states executing physiological functions is rapidly changing this notation (Si, 2015). In higher eukaryotes, the earliest examples of functional prion-like polypeptides were described in *Aplysia* and *Drosophila*, where members of the CPEB protein family undergo prion conversion that facilitates memory formation (Heinrich and Lindquist, 2011; Majumdar, et al., 2012). Cai and co-workers have revealed that the human proteins MAVS and ASC propagate respective downstream signals through prion conversion, and that this signal amplification strategy is crucial for the initiation of the innate immune response (Cai, et al., 2014). More recently, non-pathogenic prion-like proteins have been described in plants and bacteria: Luminidipendens, an *Arabidopsis* protein, involved in flowering and plant memory regulation (Chakrabortee, et al., 2016) and the transcription terminator Rho factor of the *Clostridium botulinum* pathogen (Pallares, et al., 2015; Yuan and Hochschild, 2017), respectively. These findings suggest that the conformational conversion and subsequent self-assembly that characterize prion-like proteins might be indeed an evolutionary conserved phenomenon (Maji, et al., 2009; Pallares and Ventura, 2017; Tariq, et al., 2013).

The enrichment of *P. falciparum* in N-rich low complexity sequences, soon suggested that this organism might contain a significant number of prion-like proteins, whose identification might contribute to understand its particular biology (Singh, et al., 2004). Bioinformatic analysis found a correlation between the over-representation of homorepeats-containing proteins and the abundance of proteins with putative PrLDs, which were proposed to account for as much as 25% of the parasite proteome (Singh, et al., 2004). The biological significance of these protein domains is not clear (Muralidharan and Goldberg, 2013), but *P. falciparum* has evolved a very efficient proteostatic system to cope with such an aggregation-prone proteome (Muralidharan and Goldberg, 2013; Muralidharan, et al., 2012; Przyborski, et al., 2015).

In order to address the potential biological role for prion-like proteins in *P. falciparum*, we analysed its proteome using a highly stringent computational approach, searching for the presence of Q/N-rich long regions displaying compositional similitude to *bona fide* prions (Toombs, et al., 2010; Toombs, et al., 2012) and bearing specific amyloidogenic regions able to promote their self-assembly (Fernandez, et al., 2017; Pallares, et al., 2015; Sabate, et al., 2015). This is the same approach that recently allowed us to propose the transcription terminator Rho factor as a first prion-like protein in bacteria (Pallares, et al., 2015; Pallares and Ventura, 2017). Applying this strategy, we have identified 503 PrLDs-containing proteins in *P. falciparum*, accounting for ~10% of its proteome. An analysis of the gene ontology terms enriched in this subproteome indicates that the proteins it contains might participate in important biological processes, such as regulation of gene expression or vesicle-mediated transport. Several of the functions assigned to the putative *P. plasmodium* prion-like proteins are common to those reported in other eukaryotes, including humans, while other appear to be specific for this protozoan parasite.

Among all the identified prion-like candidates, we have selected three unrelated proteins and experimentally validated that their PrLDs contain specific short N-rich sequences able to form amyloid fibrils; having thus the potential to trigger the conformational conversion of the entire protein.

Collectively, our study suggest that prion-like proteins may play a functional role in the complex parasite's biology.

6.2.3 MATERIAL AND METHODS

Dataset construction

The *P. falciparum* (isolate 3D7) reference proteome (Proteome ID UP000001450, released 2015_04, published on April 1, 2015) consisting of 5353 proteins was downloaded from UniprotKB (UniProt, 2015). This was first screened for the presence of Q/N-rich domains using an in-house developed Python script. Briefly, it scans for consecutive 80-residue windows and retrieves those with at least 30 Q/Ns. Once applied to *P. falciparum*'s proteome, it rendered 1300 proteins with at least one Q/N-rich domain. These were further scanned with PAPA (using default parameters) for intrinsically disordered regions and compositional bias resembling yeast prions, rendering 581 proteins. A final scan in search for soft-amyloid cores within these PrLD was performed using pWALTZ (using default parameters), resulting in a prion-like dataset of 503 proteins.

Functional and structural enrichment analysis

Gene Ontology (GO) (Gene Ontology, 2015) terms (at the GO FAT category) and Pfam domain (Finn, et al., 2016) enrichment were analysed and clustered with the Functional Annotation Tool of DAVID 6.7 (Database for Annotation, Visualization and Integrated Discovery) (Huang da, et al., 2009). The GO term clustering was performed with a high classification stringency and a p-value ≤ 0.05 . The Pfam list was obtained with a p-value ≤ 0.05 and final clustering was manually curated. From the 503 proteins in the prion-like proteins dataset, 487 were identified and processed by DAVID. The translation rates of the 10% highest scoring prion-like proteins at the different stages of *P. falciparum* life cycle were retrieved from (Le Roch, et al., 2003). For every protein entry, the developmental stage with the highest translation rate was considered.

Peptide prediction, synthesis and preparation

The sequences of Sec24 (UniprotKB accession number COH489), the translation initiation factor-like protein IF2c (UniprotKB accession number Q8IBA3) and the protein kinase PK4 (UniprotKB accession number C6KTB8) were further analysed with prion predictor PLAAC (Lancaster, et al., 2014). The resulting sequences, their position in the full-length protein and their scores are shown in **Figure 6.8**. The 21-residue peptides corresponding to the soft-amyloid cores predicted by pWALTZ were purchased from CASLO ApS (Scion Denmark Technical University). Peptide stock solutions were prepared solubilizing the lyophilized peptides at a final concentration of 5 mM in 100% dimethyl sulfoxide and stored at -80 °C. Before each analysis, the samples were diluted to 150 μ M in PBS buffer. For aggregation assays the diluted samples were incubated for 48 h at 25 °C.

Binding to amyloid dyes

The fluorescence spectra of the binding of 25 μM Thioflavin-T (ThT) to peptide fibrils were recorded using a Cary Eclipse spectrofluorometer (Varian, Palo Alto, CA, USA) with an excitation wavelength of 440 nm and emission range from 460 to 600 nm at 25 °C in PBS buffer. Peptides were equilibrated at room temperature for 2 min before the measurement and solutions without peptide were employed as negative controls. Excitation and emission slit widths of 10 nm were used. For the Thioflavin-S (ThS) staining assays, aggregated peptides were incubated for 1 h in the presence of 125 μM ThS in PBS. Then, the samples were centrifuged (14000 $\times g$ for 5 min) and the precipitated fraction washed twice with PBS and finally placed on a microscope slide and sealed. Images of the aggregated peptides bound to ThS were obtained at 40-fold magnification under UV light or phase contrast in a Leica fluorescence microscope (Leica DMRB, Heidelberg, Germany).

Secondary structure determination

Attenuated total reflectance Fourier transform infrared (ATR FT-IR) spectroscopy analysis of peptide fibrils were performed using a Bruker Tensor FT-IR Spectrometer (Bruker Optics, Berlin, Germany) with a Golden Gate MKII ATR accessory. Each spectrum consisted of 16 independent scans, measured at spectral resolution of 1 cm^{-1} within the 1800-1500 cm^{-1} range. All spectral data were acquired and normalized using the OPUS MIR Tensor 27 software. Infrared spectra between 1725 and 1575 cm^{-1} were fitted through overlapping Gaussian curves, and the amplitude and area for each Gaussian function were calculated employing the nonlinear peak-fitting program (PeakFit package, Systat Software, San Jose, CA). Aggregated peptides were prepared at 150 μM in PBS buffer and incubated at 25 °C for 48 h. PBS buffer without peptide was used as a control and subtracted from the absorbance signal before deconvolution.

Transmission electron microscopy

Samples of aggregated peptides obtained as described previously were placed onto carbon-coated copper grids and incubated for 5 min. The grids were washed with distilled water and negatively stained with 2% (w/v) uranyl acetate for 2 min. Micrographs were obtained in a JEM-1400 (JEOL, Japan) transmission electron microscope (TEM) operated at 80 kV accelerating voltage.

***In vivo* amyloid-like detection**

Cultures of *P. falciparum* strain 3D7 were grown *in vitro* in group B human red blood cells (RBCs), purchased from the Banc de Sang i Teixits (<http://www.bancsang.net>), using previously described conditions (Cranmer, et al., 1997). Briefly, parasites (thawed from glycerol stocks) were cultured at 37 °C in T-Flasks containing RBCs in Roswell Park Memorial Institute (RPMI) complete medium under a gas mixture of 92% N_2 , 5% CO_2 , and 3% O_2 . Synchronized cultures were obtained by 5% sorbitol lysis (Lambros and Vanderberg, 1979) and the medium was changed every 2 days maintaining 3% hematocrit and a parasitemia below 5%. Staining with PROTEOSTAT[®] protein aggregation assay (Enzo Life Sciences, Inc.) was performed according to the manufacturer's instructions. Briefly, 200 μl of *P. falciparum* culture were harvested and washed twice with 1 ml of 7.5 mg BSA/ml PBS (PBS/BSA); the resulting cell pellet was taken

up in 200 μ l of PBS/BSA containing 2 μ g/ml Hoechst 33342 and PROTEOSTAT[®] (1:3000 stock dilution), and incubated for 30 min at room temperature in the dark before being washed again twice with 1 ml of PBS/BSA. 10 μ l of the washed cell suspension were transferred into a Lab-Tek chambered coverglass (Nunc, Thermo Fisher Scientific) containing 180 μ l of PBS/BSA and finally analysed with a Leica TCS SP5 laser scanning confocal microscope, using a 63 \times immersion oil objective with 1.4 numeric aperture. Hoechst 33342 and PROTEOSTAT[®] were detected, respectively, by excitation through 405 nm and 488 nm lasers. Emission was collected between 415 nm and 500 nm for Hoechst 33342, and between 590 and 670 nm for PROTEOSTAT[®].

6.2.4 RESULTS

The *P. falciparum* proteome is enriched in proteins with PrLDs

The *P. falciparum* proteome contains an unusually high amount of low complexity regions; long domains enriched in certain amino acids and without a defined secondary structure. Low complexity regions are present in 30% to 90% of *P. falciparum* proteins, depending on the detection stringency (DePristo, et al., 2006; Singh, et al., 2004), and they are specially enriched in N residues. It has been proposed that these disordered protein regions might share certain properties with the classical yeast Q/N-rich PrDs (Faux, et al., 2005; Fuxreiter, 2012; Fuxreiter and Tompa, 2012; Malinowska, et al., 2013; Pallares, et al., 2015; Tompa and Fuxreiter, 2008), and potentially support the formation of prion-like macromolecular assemblies (Espinosa Angarica, et al., 2013; Singh, et al., 2004).

In order to evaluate the presence of Q/N-rich prion-like proteins in *P. falciparum*, we examined its proteome combining the detection of local Q/N-enrichment together with PAPA and pWALTZ predictions (Pallares, et al., 2015; Toombs, et al., 2012; Zambrano, et al., 2015) (see **Section 6.2.3 Material and methods**). Thus, any predicted PrLD in our subproteome would fulfil the following requirements: being Q/N-rich, disordered (PAPA includes the disorder predictor FoldIndex (Prilusky, et al., 2005)), compositionally similar to yeast PrDs and contain a short sequence stretch able to facilitate its conversion into an amyloid-like state; we have generically named these stretches “soft-amyloid cores”, because their amyloid propensity is significantly lower than the classical amyloid regions of pathogenic proteins, but still enough to promote protein self-assembly (Batlle, et al., 2017c).

The analysis shows that 1300 proteins (24.3% of the proteome) bear at least one Q/N-rich domain, in excellent agreement with previous studies (Singh, et al., 2004). 581 of these Q/N-rich domains (44.7%) also display an amino acid composition similar to that of yeast PrDs and are disordered, as predicted by PAPA, and among them, 503 domains (86.6%) contain a soft-amyloid core as predicted by pWALTZ. Overall, we conclude that 9.4% of the *P. falciparum* proteome may have the physicochemical properties and the aggregation potential to behave as a prion. This value is lower than previously estimated applying other computational approaches based only in Q/N richness (Singh, et al., 2004) or than the one estimated using only compositional similitude to yeast PrDs with PAPA, which predicts 22.5% of the parasite proteins as prion-like. The Q/N rich only dataset contains many long poly-N stretches, without any inner hydrophobic residue, a requisite to act as a prion (Toombs, et al., 2010), whereas the

PAPA only dataset includes polypeptides unlike to behave as prions *in vivo*, like membrane integral proteins, only because they display sequence stretches enriched in certain hydrophobic residues. In any case, even with the stringent approach used here, roughly 10% of the *P. falciparum* proteome seems to correspond to proteins displaying PrLDs and may thus have a high intrinsic aggregation propensity. This generates several important questions: Why are *P. falciparum* proteins so rich in PrLDs? Which are these putative prions? What are their roles in *P. falciparum*?

Previous works addressed these questions by analysing the distribution of Q/N- or N-rich regions in the *P. falciparum* proteome (Singh, et al., 2004). They detected such stretches in all protein families and all developmental stages of *P. falciparum*, without an evident association with any specific biological process. However, an increasing number of studies connect PrLD to specific functions and processes in other species (Espinosa Angarica, et al., 2014; Iglesias, et al., 2015). We hypothesized that focusing the analysis in our curated subproteome may help to unravel the functional purpose, if any, of PrLDs in the protozoan.

Computational analysis of the role of prion-like proteins in *P. falciparum*

The DAVID Functional Annotation Clustering Tool was employed to identify enriched gene ontology (GO) categories (Gene Ontology, 2015; Huang da, et al., 2009) in the previously identified *P. falciparum* PrLD-containing proteins ($p\text{-value} \leq 0.05$). It is worth to mention that this analysis is constrained by the fact that 60% of *P. falciparum* genes have unknown functions (Gardner, et al., 2002), most of them have no clear homolog in other eukaryotes, and that the mechanisms underlying the main processes related to malaria pathogenesis in *P. falciparum* are still poorly understood. However, 487 out of the 503 proteins in our dataset were identified and processed by DAVID, with result in a coverage of 94.8% of our subproteome.

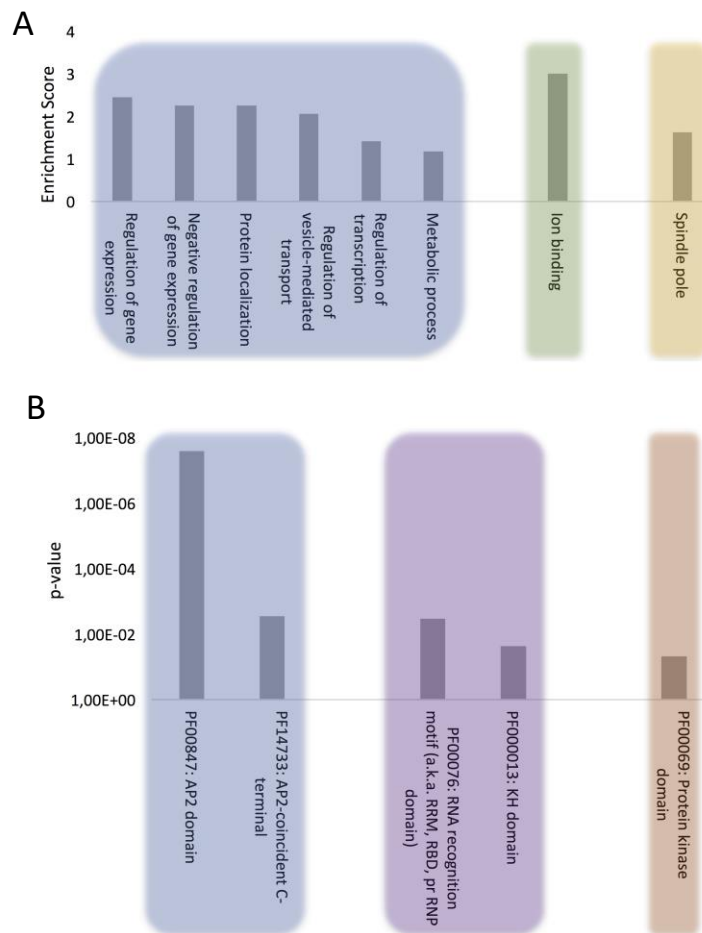


Figure 6.7 – Computational analysis of the role of *P. falciparum* PrLD-containing proteins. A) GO biological process, molecular function and cellular component terms enriched in *P. falciparum* PrLD-containing proteins. **B)** Pfam structural domains enriched in PrLD-containing proteins. The enrichment analysis was performed with Functional Annotation Tool of DAVID 6.7 using high stringency, p-value ≤ 0.05 for GO and Pfam terms.

The proteins were clustered according to the following ontologies: biological process, molecular function and cellular component (**Figure 6.7A**). The most significant biological process gene clusters include ‘regulation of gene expression’, ‘negative regulation of gene expression’ and ‘regulation of transcription’. The abundance of DNA and RNA binding proteins with PrLDs (**Figure 6.7B**) is consistent with the observation that many of the prion-like proteins discovered initially in yeast (Alberti, et al., 2009) and more recently in humans (King, et al., 2012), plants (Chakrabortee, et al., 2016) and bacteria proteomes (Iglesias, et al., 2015) are proteins associated with gene expression and translation regulation such as transcription factors and RNA-binding proteins.

Other biological process terms could also be arranged into enriched clusters: ‘protein localization’, ‘regulation of vesicle-mediated transport’ and ‘metabolic process’. Importantly, the vesicle-mediated transport system and the trafficking of parasite proteins to diverse locations in the host cell are essential to promote new parasite phenotypes, playing a crucial role in host-pathogen interactions, as well as in disease pathogenesis and susceptibility (Hiller, et al., 2004; Marti, et al., 2005; Miller, et al., 2002). Indeed, extracellular vesicles have been shown to act as delivery agents for prion-like proteins (Liu, et al., 2017).

The analysis of molecular function domains in the set of *P. falciparum* prion-like proteins revealed that the only significantly enriched cluster was ion binding. A deeper analysis of the GO annotations indicates that ~33% of these proteins function in DNA/RNA interaction and ~40% of them also contain structural domains related to nucleotide binding, such as Zinc fingers. In fact, the functions associated to nucleotide binding, especially RNA binding, appear to be associated to proteins containing PrLDs, regardless of the organism (Espinosa Angarica, et al., 2014; Iglesias, et al., 2015; Pallares, et al., 2015).

At this point, to dig a bit more on the functional role of our protein subset, we reanalysed the MF category setting a p-value cut off of 0.1. Several new GO terms came to light that could be grouped into two interesting MF subclusters: (i) chromatin remodelling, which is consistent with recent studies demonstrating that the physical properties of prion-like domains can retarget critical chromatin regulatory complexes (Boulay, et al., 2017) and facilitate heterochromatin assembly (Kataoka and Mochizuki, 2017) and (ii) GTPase regulatory activity, which is also detected in the PrLD-containing proteins of several other organisms (bacteria, plants, fungi and invertebrates) (Espinosa Angarica, et al., 2013); indeed, the canonical and best characterized yeast prion, Sup35, is a GTPase (Glover, et al., 1997).

Analysis of the cellular component ontology category shows a specific enrichment at the spindle pole. In yeast, prion proteins have been shown to interact specifically with spindle pole proteins (Treasch and Lindquist, 2012) and spindle-associated proteins have been shown to be involved in self-assembly mediated phase separation in *Xenopus* (Jiang, et al., 2015).

All the proteins in our Q/N-rich dataset have in common the presence a disordered region of 80 amino acids in which at least the 37.5% of the residues (30/80) correspond to Q or N. This compositional similitude might imply a certain overlap of functions between proteins bearing PrLDs and those devoid of them. A gene ontology analysis of Q/N-rich proteins without PrLDs (Supplementary Material S6.1), shows that the molecular function and cellular location terms are different, but related, to those found in the Q/N-rich protein subset bearing PrLDs; nucleotide binding and cytoskeleton being the most enriched terms for these two categories, respectively. In contrast, Q/N-rich proteins without PrLDs are poorly represented in specific biological processes, being the most enriched one DNA repair (Supplementary Material S6.1). This suggests that the compositional/sequential features of PrLDs might be important to specify the biological context in which the proteins act, whereas their generic molecular function depends mostly on the local enrichment in Q/N residues.

Prion-like proteins display a modular architecture in which one or several long and disordered PrLDs are adjacent to conventional globular domains and, accordingly, they tend to be large. We compared the average size of our protein subset with the one of the complete plasmodium proteome, confirming that proteins bearing PrLDs are effectively significantly longer (Supplementary Material S6.2). To discard that the GO terms identified for PrLD-containing proteins would respond only to their differential size, we selected the subset of the largest 503 proteins in the proteome and performed a gene ontology analysis. The resulting enriched terms did not coincide with those in our subset in any of the

categories. The most enriched biological processes in large proteins were pathogenesis and single organismal cell-cell adhesion; the most enriched compartments were infected host cell surface knob and host cell plasma membrane and the most enriched molecular functions were receptor activity and cell adhesion molecular binding.

In order to address if the expression of PrLD-containing proteins occurs preferentially at a given parasite stage, we analysed the expression levels of the 10% top ranking proteins in our dataset at each of the different life cycle stages, as reported by Winzeler and co-workers (Le Roch, et al., 2003). It turns out that, on the average, the highest translation rates for these proteins correspond to those at the merozoite and early ring stages (Supplementary Material S6.3).

Protein domains in *P. falciparum* prion-like proteins

To further evaluate the role of our collection of PrLD-containing proteins we examined in detail their constituent functional domains (Finn, et al., 2016) (**Figure 6.7B**). As expected, after clustering, the Pfam domains that were most often found in combination with PrLDs were involved in DNA/RNA binding, among which, the ApiAP2 stands out. The ApiAP2 family is homologous to the plant *Apetala2*/ethylene response factor (AP2/ERF) DNA-binding proteins, which comprise the second largest class of transcription factors in *Arabidopsis thaliana*. Balaji and co-workers described that ApiAP2 proteins are likely to function as a family of apicomplexan parasite-specific transcription factors (Balaji, et al., 2005) and that their amino acid sequences are highly conserved among orthologues. Strikingly, our data reveals that at least the 50% of the members composing this family in *P. falciparum* contain a PrLD. Several studies support their major role in mediating the regulation of stage-specific gene expression profiles in the parasite's development (Modrzynska, et al., 2017; Painter, et al., 2011; Yuda, et al., 2010) and suggest their crucial contribution to *P. falciparum* complexity and growth since very few ApiAP2 genes have been successfully knocked out (Behnke, et al., 2010; Yuda, et al., 2010).

The RNA recognition motif (RRM) is the most enriched RNA-binding domain (RBD) in our dataset. RRM is by far the most versatile and abundant RBDs, their fold being conserved from bacteria to higher eukaryotes (Reddy, et al., 2015). This result is consistent with the observation that many of the human proteins with PrLDs contain an RRM motif and are involved in liquid-liquid phase transitions facilitating the formation of dynamic membraneless intracellular compartments, such as ribonucleoprotein (RNP) granules. They allow material exchange and fast assemblage and adaptation to different environments and cell states (Malinowska, et al., 2013), the PrLDs in RNA binding proteins provide the special physicochemical properties that allow contacts between RNAs and proteins that sustain the liquid-like assemblies (Han, et al., 2012; Kato, et al., 2012). Indeed, the second most abundant RBD linked to our protein subproteome is the KH domain, a protein domain that was first identified in the human heterogeneous nuclear proteins (hnRNP) (Siomi, et al., 1993) and, together with RRM, constitutes the most abundant domain in RNA granule forming proteins (Kato, et al., 2012). In *P. falciparum* RNPs are involved in translation repression and posttranscriptional regulation of gene expression, critical for some stages of the parasite (Kramer, 2014).

The last enriched Pfam family includes the protein kinase domain. It is well-known that phosphorylation/dephosphorylation is the major control mechanism for many cellular functions. Consistently, recent studies carried out in *P. falciparum* reveal stage-specific profiles of protein phosphorylation, suggesting that reversible protein phosphorylation plays a key role in the regulation of the *Plasmodium* life cycle (Pease, et al., 2013; Wu, et al., 2009). So far, no PrLD-containing protein kinase has been characterized experimentally, but it is obvious that the ability to control the activity of these enzymes by modulating their assembly would have important physiological consequences.

Predicted PrLD soft-amyloid cores in *P. falciparum* proteins

Based on the above computational results, we selected three PrLD-containing proteins for their experimental characterization: the putative transport protein Sec24 (UniprotKB Accession number COH489), the translation initiation factor-like protein IF2c (UniprotKB accession number Q8IBA3) and the protein kinase PK4 (UniprotKB accession number C6KT8). These proteins are associated with functions (nucleotide binding, Q8IBA3), cellular components (vesicle mediated-transport, COH489) and structural domains (kinase, C6KT8) that are enriched in our dataset. The selected candidates have no functional or sequential relationship and have not been previously suggested to act as prions.

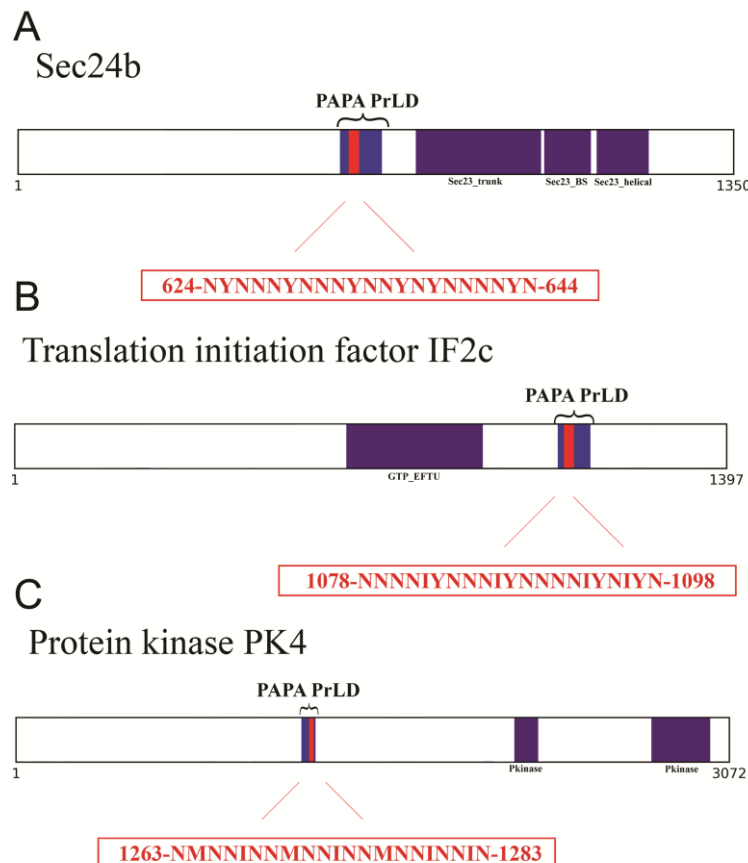


Figure 6.8 – Soft-amyloid cores prediction in the three candidate proteins. A) Sec24b diagram showing the location of the PrLD predicted by PLAAC (in blue, residues 608-686) or PAPA (between brackets, residues 607-687). **B)** IF2c location of the PrLD predicted by PLAAC (in blue, residues 1066-1134) or PAPA (between brackets, residues 1057-1137). PK4 location of the PrLD predicted by PLAAC (in blue, residues 1230-1290) or PAPA (between brackets, residues 1220-1300). Pfam domains and the soft-amyloid cores are shown in purple and red respectively, the exact position and the sequence of the predicted soft-amyloid cores are presented in the red box.

As a first candidate we chose Sec24b, a member of the Sec24 family. Within this family, Sec24b is by far the most enriched in Q/N, which constitute 21.5% of the amino acids in the complete sequence and 51.9% of the PrLD. Sec24b is closely related to the mammalian Sec24C/D family and the yeast Sec24 homologue Lst1 (Lee, et al., 2008). These proteins play a key role in shaping the vesicle, as well as in cargo selection and concentration (Roberg, et al., 1999). They have a scaffolding function required to generate vesicles that can accommodate difficult cargo proteins, including large oligomeric assemblies.

As a second candidate we selected IF2, one of the essential components for the initiation of protein synthesis. IF2 is a translation initiator factor acting as a GTPase that recruits the charged fMet-initiator tRNA onto the 30S ribosomal initiation complex (Antoun, et al., 2003). From the three IF2 homologues described in *P. falciparum* (Haider, et al., 2015), IF2c is the only one that holds a PrLD. It is worth to note that the IF2c C-terminal domain, where the PrLD maps, has the largest identity with bacterial IF2, a family of translation initiation factors rich in putative PrLDs (Iglesias, et al., 2015). 20.8% of IF2c residues are Q/N and this proportion raises to 49.4% in the PrLD.

The third selected protein was the kinase PK4, a protein that is essential for completion of the blood stage of the disease (Zhang, et al., 2012). Thus PK4 has been suggested as a novel target for the next generation of antimalarial compounds (Kahrstrom, 2012). PK4 phosphorylates eIF2 α and arrests global protein synthesis in schizonts (mature form of the blood cycle) and gametocytes (sexual form that infects the mosquitoes). 17.7% of PK4 amino acids are Q or N, with 55.5% of its PrLD corresponding to these polar residues.

Table 6.1 – Predicted *Plasmodium falciparum* PrLD soft-amyloid cores. For each PrLD-containing protein it is shown the UniprotKB accession number (UniprotKB Ac.), the 21 residue-long soft-amyloid core with its respective position in the sequence, pWALTZ score, PAPA score and PLAAC score (COREscore) with a cutoff of 73.55, 0.05 and >0 respectively. PAPA and PLAAC search for compositional similarity to yeast prion domains, defining the predicted PrLD, while pWALTZ scans for soft-amyloid cores within them.

Protein	UniprotKB Ac.	Soft-amyloid core	N (%)	pWALTZ	PAPA	PLAAC
Sec24b	COH489	624-NYNNNNYNNNNYNNNNNNYN-644	71	84.62	0.20	45.28
IF2c	Q8IBA3	1078-NNNNIYNNNIYNNNNIYNIYN-1098	62	87.71	0.07	36.18
PK4	C6KTB8	1263-NMNNINNMNNINNMNNININ-1283	67	77.34	0.25	47.97

To further confirm the presence of PrLDs in these proteins, and to define more precisely their boundaries we used PLAAC (Lancaster, et al., 2014), yet another composition-based predictor, in which, in contrast to PAPA, the length of the predicted PrLD also depends on the protein composition. PLAAC detected PrLDs overlapping with the regions previously identified by PAPA, in the three polypeptides (**Figure 6.8**).

We analysed these three putative prion-like proteins using the same computational approach we employed previously to detect and validate the soft-amyloid cores present in *bona fide* yeast prions (Sant'Anna, et al., 2016), in the pathogenic bacteria *C. botulinum* (Pallares, et al., 2015) and in human

prion-like proteins (Batlle, et al., 2017a). The predicted soft-amyloid cores for these *P. falciparum* proteins are shown in **Table 6.1**. Not surprisingly, these stretches are highly enriched in N residues, all containing > 60% of this polar residue. Interestingly enough, well-validated aggregation predictors like AGGRESCAN, Tango and Zyggregator (Conchillo-Sole, et al., 2007; Fernandez-Escamilla, et al., 2004; Tartaglia and Vendruscolo, 2008), all failed to classify these stretches as aggregation-prone (Supplementary Material S6.4), likely because of their much lower hydrophobicity, when compared with the classical amyloid stretches present in pathogenic amyloidogenic proteins. One of the restraints in our prediction scheme is that the identified PrLD should be essentially disordered, as predicted with FoldIndex (Prilusky, et al., 2005). In this structural context the identified soft-amyloid cores will be mostly exposed to solvent and able to establish intermolecular contacts, if they have this ability. Orthogonal analysis with alternative disorder prediction algorithms confirms that this is likely the case for the three proteins herein (Supplementary Material S6.5).

We synthesized 21-residue-long peptides corresponding to the detected soft-amyloid cores and characterized their amyloid properties experimentally.

Predicted PrLDs soft-amyloid cores assemble into β -sheet rich structures

As a first evaluation of the assembling properties of the selected peptides, we measured their ability to adopt a β -sheet-enriched structure, a hallmark of amyloid fibril formation (Nelson, et al., 2005). To this aim the peptides were prepared at 150 μ M in phosphate buffered saline (PBS) and incubated during 48 h at 25 °C. We used Fourier-transform infrared (FTIR) spectroscopy and recorded the amide I region of the spectrum (1700–1600 cm^{-1}) (**Figure 6.9**). This spectral region corresponds to the absorption of the carbonyl peptide bond group of the protein main chain and is sensitive to the protein conformation. Deconvolution of the spectra allowed us to assign the secondary structure elements and their relative contribution to the main absorbance. In the three cases, the main peaks mapped in the 1620–1630 cm^{-1} region of the spectra, accounting for 50% or more of the absorbance signals, indicating that the peptides have acquired significant intermolecular β -sheet structure. Interestingly, no anti-parallel β -sheet band was detected ($\sim 1690 \text{ cm}^{-1}$) in any of the samples; thus, suggesting that the detected β -strands in the self-assembled peptides would adopt preferentially a parallel disposition. The other detected signals were associated with the presence of disordered structure and turns (**Figure 6.9**).

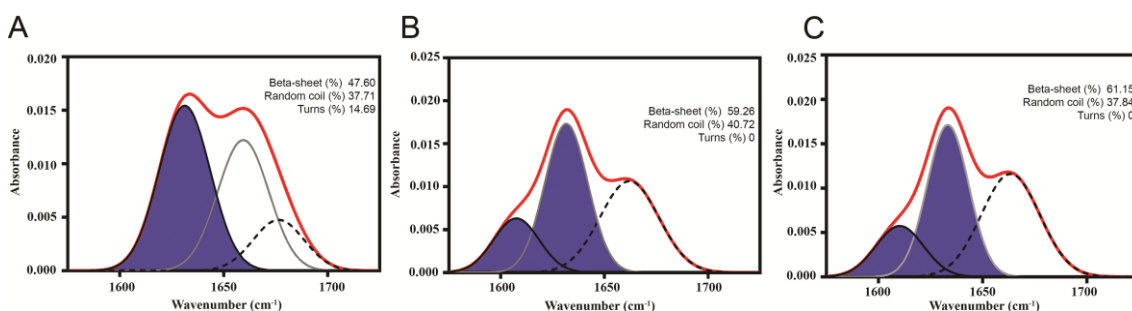


Figure 6.9 – Predicted PrLD soft-amyloid cores secondary structure. Secondary structure determined by ATR FT-IR in the amide I region. The red line corresponds to the absorbance spectrum; the blue area indicates the contribution of the inter-molecular β -sheet signal to the total area upon Gaussian deconvolution. **A)** Sec24b, **B)** IF2c and **C)** PK4.

Predicted PrLD soft-amyloid cores form amyloid-like fibrillar structures

To assess if the identified β -sheet-rich assemblies correspond to amyloid-like structures, we used the amyloid-specific dyes Thioflavin-T (ThT) and Thioflavin-S (ThS). After incubation at a concentration of 150 μ M in PBS during 48 h at 25 $^{\circ}$ C, all the peptides were able to promote a large increase in the intensity of ThT fluorescence emission (**Figure 6.10**). In addition, areas rich in fibrous material were stained by ThS to yield a bright green-yellow fluorescence against a dark background (**Figure 6.10**).

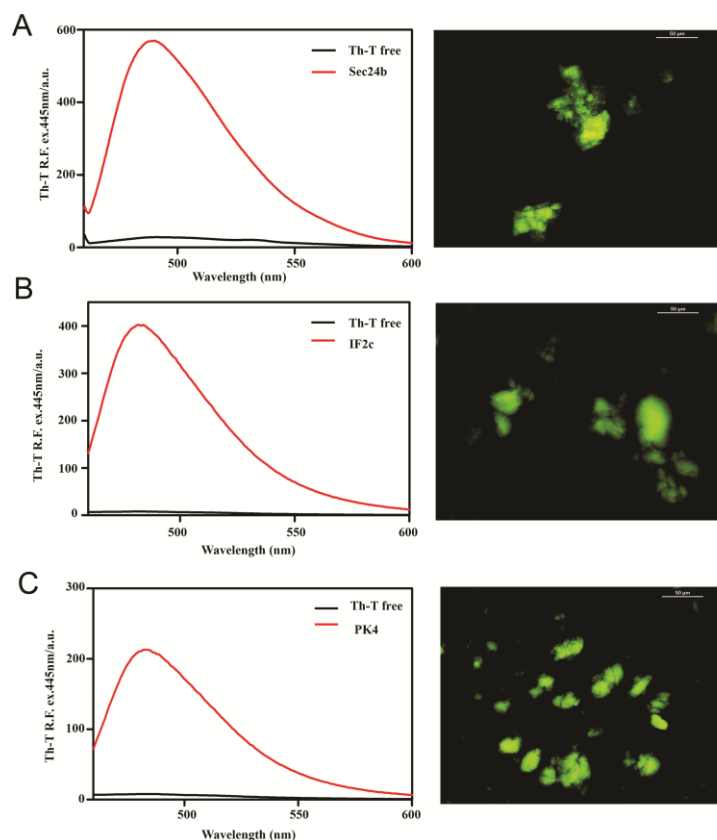


Figure 6.10 – Binding of the predicted PrLD soft-amyloid cores to amyloid specific dyes. Fluorescence emission spectrum of ThT when excited at 440 nm; note the characteristic fluorescence enhancement at \sim 480 nm when the dye is bound to amyloid-like aggregates. On the right side of the panel, ThS binding of aggregated peptides at 150 μ M in PBS after 48 h of incubation at 25 $^{\circ}$ C. The typical green fluorescence can be observed under the fluorescence microscope, images were obtained at 40X magnification. **A)** Sec24b, **B)** IF2c and **C)** PK4.

Transmission electron microscopy (TEM) examination of the morphological features of the incubated peptide solutions (**Figure 6.11**) revealed that they effectively assemble into supramolecular structures. Sec24b formed long and straight fibrils, whereas IF2c and PK4 formed short and curly fibrillar structures.

Overall, biophysical analysis of the three predicted peptides demonstrates the ability of the candidate *P. falciparum* soft-amyloid cores to nucleate the formation of β -sheet-rich amyloid-like structures.

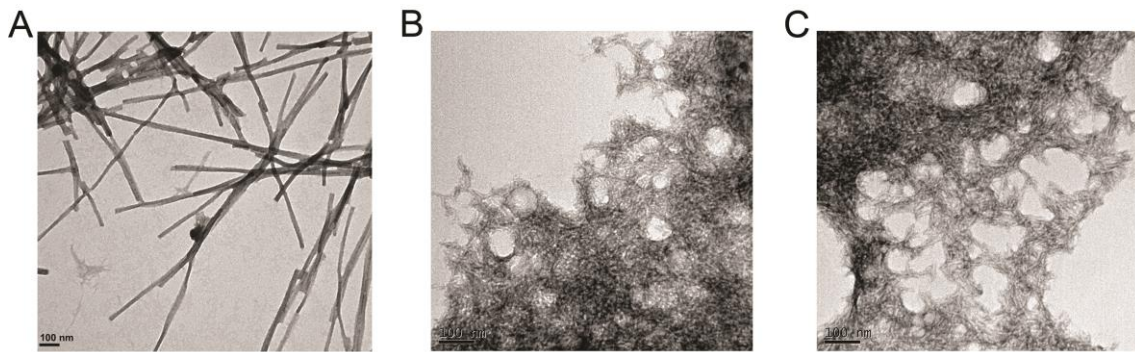


Figure 6.11 – Fibrillar structures formed by the predicted PrLD soft-amyloid cores. Representative TEM images for **A) Sec24b, B) IF2c and C) PK4** aggregated peptides at 150 μ M in PBS after 48 h of incubation at 25 $^{\circ}$ C.

Detection of intracellular protein aggregates in *P. falciparum*

The above experimental data suggest that \sim 10% of the *P. falciparum* proteome might possess the ability to establish amyloid-like contacts, at least transiently, *in vivo*, and thus; that at any time, a significant number of proteins might potentially aggregate in the parasite. We employed a permeable amyloid-specific dye (PROTEOSTAT[®]) to track the *in vivo* presence of intracellular amyloid-like aggregates in *P. falciparum*.

P. falciparum was grown in red blood cells (RBCs) and then the culture was incubated with the amyloid dye. We observed colocalization between PROTEOSTAT[®] fluorescence and the cytosol of *P. falciparum*-infected RBCs, whose nuclei stained with Hoechst 33342. The images evidenced the lack of structures able to bind the dye in non-infected erythrocytes, and that, accordingly, only upon infection by *Plasmodium*, red fluorescent amyloid foci are evident inside parasitized RBCs (**Figure 6.12**), demonstrating the high amyloid load that this parasite supports at this stage.

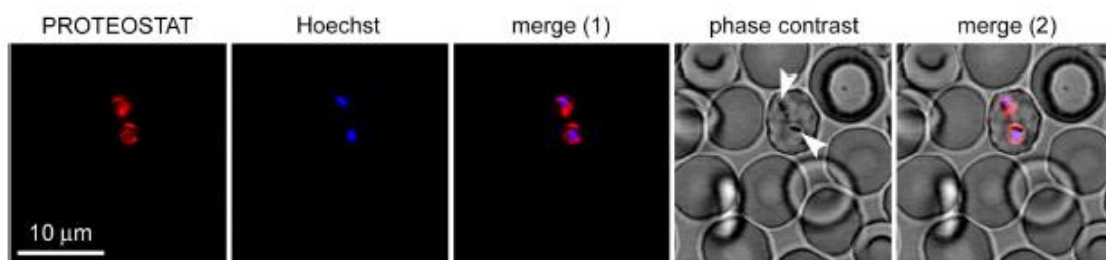


Figure 6.12 – Fluorescence microscopy analysis of the presence of protein aggregates in *P. falciparum*-infected RBCs (pRBCs). The selected field shows a single pRBC in early trophozoite stage, indicated by its characteristic nuclear Hoechst blue fluorescence among enucleated non-parasitized erythrocytes. The amyloid-specific dye PROTEOSTAT[®] reveals protein aggregates in the cytosol of the two parasite cells present in the pRBC. The arrowheads indicate nascent hemozoin crystals in the food vacuole of *Plasmodium*.

6.2.5 DISCUSSION

Many lines of evidence suggest that prion-like proteins can be both harmful and beneficial for the cell. The propensity of a protein to behave as a prion is encoded in its amino acid sequence (Sabate, et al., 2015). In particular, long and disordered N/Q-rich sequences seem to facilitate conformational conversion into functional amyloid-like states (Alberti, et al., 2009). The occurrence of N/Q-rich sequence

stretches varies substantially between organisms, with *P. falciparum* having one of the most enriched proteomes in this kind of regions, and specifically in N-rich sequences (Aravind, et al., 2003). Accordingly, it has been assumed that prion-like proteins would be common in this organism (Michelitsch and Weissman, 2000). Long Q- and N- homorepeats are inherently aggregation-prone (Halfmann, et al., 2011). However, these sequence stretches alone are not sufficient to sustain a prion-like behaviour (Toombs, et al., 2010). Here, using a stringent computational approach that considers that PrLDs should not be only Q/N-rich, but also display compositional similitude to *bona fide* yeast prion domains and encode for at least one specific short sequence stretch with moderate, but significant, amyloid propensity (Sabate, et al., 2015), we concluded that 503 polypeptides in *P. plasmodium* fulfil the requirements to potentially behave as prion-like proteins. This accounts for ~10% of the proteome, which despite being a lower fraction than previously proposed (~25%) (Singh, et al., 2004), still might constitute a high prionic load for the parasite.

A priori, the presence of PrLDs might be dangerous for *Plasmodium*, since prion-like proteins have an intrinsic propensity to aggregate and, in humans, disease-linked mutations occur preferentially in the PrLDs of these polypeptides (Kim, et al., 2013). On the other hand, these PrLDs might act as conformational switches that control protein assembly and thus protein function to allow adaptation to the changing environment that *P. falciparum* faces during its life cycle. Importantly, the PrLD-containing subproteome we identify here is associated with defined domains and functionalities in the parasite, which suggests that Q/N-rich PrLDs do not occur randomly in the *P. falciparum* proteome. This assumption is supported by the fact that PrLDs are associated with similar GO-clusters in organisms as divergent as *Plasmodium*, yeast, *Dictyostelium*, *Drosophila* and humans (Malinovska, et al., 2013; Malinovska, et al., 2015). For instance, the role of PrLDs-containing proteins in DNA and RNA binding is well conserved, with the RRM domain being among the most enriched PrLD-associated domains in these organisms. Indeed, 25% of the *P. falciparum* proteins bearing an RRM domain also contain a predicted PrLD. We found that this domain association is also conserved in *Plasmodium vinckei* and *Plasmodium yoelii*, with 13% and 15% of RRM-containing proteins having a Q/N-rich PrLD.

P. falciparum PrLDs exhibit specific associations with domains and functions not detected in other organisms, such as the ApiAP2 proteins, with 50% of their members displaying a PrLD. These proteins have been postulated as the main transcriptional regulators in *Plasmodium* parasites and the other Apicomplexa (Balaji, et al., 2005). Importantly, according to our analysis, the presence of PrLDs within AP2 transcription factors also seems to be evolutionary, with 39%, 29% and 24% of the AP2 proteins in *Plasmodium vinckei*, *Plasmodium yoelii* and *Plasmodium berghei* displaying Q/N-rich PrLDs, respectively. The association between PrLDs and the regulation of vesicle-mediated transport is also a specific feature of *Plasmodium*. This process allows the trafficking of some parasite proteins to the erythrocyte membrane (Hiller, et al., 2004; Marti, et al., 2005; Miller, et al., 2002).

Plasmodium is an obligate parasite that has evolved to survive in different hosts and cell types. It has a complex life cycle with cellular stages that differ in shape, size, metabolic activity and resource

requirements. Hence, to sustain this complexity, *Plasmodium* requires an efficient regulation, to which the conformational conversion of regulatory proteins bearing PrLDs might contribute. Changes in local protein concentration, binding to nucleic acids and posttranslational modifications have been shown to modulate the assembly of PrLDs, the functional outcome depending on the particular assembled protein (Alberti, 2017).

We and others have suggested that certain short amyloidogenic sequence stretches embedded in PrLDs contribute significantly to prion formation, maintenance, and transmission, at least in yeast (Batlle, et al., 2017c; Crow, et al., 2011; Sabate, et al., 2015; Sant'Anna, et al., 2016). The computational search for such regions in the putative PrLDs of a large number of bacterial proteomes (Iglesias, et al., 2015), previously thought to lack prions, and a subsequent experimental validation, allowed us to propose that the Rho transcription terminator might constitute a first bacterial prion (Pallares, et al., 2015; Pallares and Ventura, 2017). Soon after, Yuan and Hochschild confirmed the ability of this protein to adopt an infectious state, leading to global changes in the transcriptome (Yuan and Hochschild, 2017). Here we used the same approach to study the amyloidogenic potential of three *P. falciparum* PrLD-containing proteins: the translation initiation factor 2c, the kinase PK4, both involved in gene expression regulation (Haider, et al., 2015; Zhang, et al., 2012) and Sec24b, involved in vesicle trafficking (Lee, et al., 2008). Our data provides compelling evidence that, *in vitro*, all three candidate proteins contain short nucleating regions embedded in the PrLDs able to spontaneously self-assemble into amyloid-like structures. The presence of such stretches does not necessarily imply that the correspondent large full-length proteins would behave in a prion-like manner, and this behavior should be experimentally validated. However, several indirect evidences suggest that this could be the case: i) we have shown that when a predicted short amyloid sequence is administered to cells in its amyloid state it is able to seed the conformational conversion of the complete endogenous protein and its subsequent aggregation into a prionic form (Sant'Anna, et al., 2016), ii) the soft-amyloid stretch we identified in the PrLD of the bacterial Rho terminator factor has been shown to be absolutely essential for its self-assembly and prion activity (Yuan and Hochschild, 2017), iii) Vorberg and co-workers have shown that the soft-amyloid sequence we pinpointed in the PrLD of a model prion protein is the only region required for the induction, propagation and inheritance of the prion state in the mammalian cytosol (Duernberger, et al., 2018).

Overall, we identified a subset of putative Q/N-rich prion-like proteins in *P. falciparum* associated with specific biological processes and validated experimentally that their highly polar and disordered PrLDs contain cryptic sequences able to self-assemble into amyloids. The structural characterization and *in vivo* validation of the properties of the identified proteins is challenging, but it is worth the effort, since it might uncover a first *bona fide* prion in *Plasmodium*.

6.2.6 REFERENCES

Alberti, S. (2017) Phase separation in biology, *Current biology : CB*, **27**, R1097-R1102.

Alberti, S., et al. (2009) A systematic survey identifies prions and illuminates sequence features of prionogenic proteins, *Cell*, **137**, 146-158.

- Antoun, A., *et al.* (2003) The roles of initiation factor 2 and guanosine triphosphate in initiation of protein synthesis, *EMBO J*, **22**, 5593-5601.
- Aravind, L., *et al.* (2003) Plasmodium biology: genomic gleanings, *Cell*, **115**, 771-785.
- Balaji, S., *et al.* (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains, *Nucleic Acids Res*, **33**, 3994-4006.
- Battle, C., *et al.* (2017) Characterization of Soft Amyloid Cores in Human Prion-Like Proteins, *Sci Rep*, **7**, 12134.
- Behnke, M.S., *et al.* (2010) Coordinated progression through two subtranscriptomes underlies the tachyzoite cycle of *Toxoplasma gondii*, *PloS one*, **5**, e12354.
- Boulay, G., *et al.* (2017) Cancer-Specific Retargeting of BAF Complexes by a Prion-like Domain, *Cell*, **171**, 163-178 e119.
- Cai, X., *et al.* (2014) Prion-like polymerization underlies signal transduction in antiviral immune defense and inflammasome activation, *Cell*, **156**, 1207-1222.
- Conchillo-Sole, O., *et al.* (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides, *BMC Bioinformatics*, **8**, 65.
- Cranmer, S.L., *et al.* (1997) An alternative to serum for cultivation of *Plasmodium falciparum* in vitro, *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **91**, 363-365.
- Crow, E.T., Du, Z. and Li, L. (2011) A small, glutamine-free domain propagates the [SWI(+)] prion in budding yeast, *Molecular and cellular biology*, **31**, 3436-3444.
- Chakrabortee, S., *et al.* (2016) Luminidependens (LD) is an Arabidopsis protein with prion behavior, *Proc Natl Acad Sci U S A*, **113**, 6065-6070.
- Chen, Y. and Dokholyan, N.V. (2008) Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm, *Molecular biology and evolution*, **25**, 1530-1533.
- DePristo, M.A., Zilversmit, M.M. and Hartl, D.L. (2006) On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins, *Gene*, **378**, 19-30.
- Duernberger, Y., *et al.* (2018) Prion replication in the mammalian cytosol: Functional regions within a prion domain driving induction, propagation and inheritance, *Mol Cell Biol*.
- Eichinger, L., *et al.* (2005) The genome of the social amoeba *Dictyostelium discoideum*, *Nature*, **435**, 43-57.
- Espinosa Angarica, V., *et al.* (2014) PrionScan: an online database of predicted prion domains in complete proteomes, *BMC Genomics*, **15**, 102.
- Espinosa Angarica, V., Ventura, S. and Sancho, J. (2013) Discovering putative prion sequences in complete proteomes using probabilistic representations of Q/N-rich domains, *BMC Genomics*, **14**, 316.
- Faux, N.G., *et al.* (2005) Functional insights from the distribution and role of homopeptide repeat-containing proteins, *Genome Res*, **15**, 537-551.
- Fernandez-Escamilla, A.M., *et al.* (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins, *Nature biotechnology*, **22**, 1302-1306.
- Fernandez, M.R., *et al.* (2017) Amyloid cores in prion domains: Key regulators for prion conformational conversion, *Prion*, **11**, 31-39.
- Finn, R.D., *et al.* (2016) The Pfam protein families database: towards a more sustainable future, *Nucleic Acids Res*, **44**, D279-285.
- Fuxreiter, M. (2012) Fuzziness: linking regulation to protein dynamics, *Mol Biosyst*, **8**, 168-177.
- Fuxreiter, M. and Tompa, P. (2012) Fuzzy complexes: a more stochastic view of protein function, *Adv Exp Med Biol*, **725**, 1-14.
- Gardner, M.J., *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*, *Nature*, **419**, 498-511.
- Gene Ontology, C. (2015) Gene Ontology Consortium: going forward, *Nucleic Acids Res*, **43**, D1049-1056.
- Glover, J.R., *et al.* (1997) Self-seeded fibers formed by Sup35, the protein determinant of [PSI⁺], a heritable prion-like factor of *S. cerevisiae*, *Cell*, **89**, 811-819.

- Haider, A., *et al.* (2015) Targeting and function of proteins mediating translation initiation in organelles of *Plasmodium falciparum*, *Molecular microbiology*, **96**, 796-814.
- Halfmann, R., *et al.* (2011) Opposing effects of glutamine and asparagine govern prion formation by intrinsically disordered proteins, *Mol Cell*, **43**, 72-84.
- Han, T.W., *et al.* (2012) Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies, *Cell*, **149**, 768-779.
- Harrison, P.M. and Gerstein, M. (2003) A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes, *Genome Biol*, **4**, R40.
- Heinrich, S.U. and Lindquist, S. (2011) Protein-only mechanism induces self-perpetuating changes in the activity of neuronal Aplysia cytoplasmic polyadenylation element binding protein (CPEB), *Proc Natl Acad Sci U S A*, **108**, 2999-3004.
- Hiller, N.L., *et al.* (2004) A host-targeting signal in virulence proteins reveals a secretome in malarial infection, *Science*, **306**, 1934-1937.
- Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nature protocols*, **4**, 44-57.
- Iglesias, V., de Groot, N.S. and Ventura, S. (2015) Computational analysis of candidate prion-like proteins in bacteria and their role, *Frontiers in microbiology*, **6**, 1123.
- Jiang, H., *et al.* (2015) Phase transition of spindle-associated protein regulate spindle apparatus assembly, *Cell*, **163**, 108-122.
- Kahrstrom, C.T. (2012) Parasite physiology: Plasmodium gets the PK4 blood test, *Nat Rev Microbiol*, **10**, 237.
- Kataoka, K. and Mochizuki, K. (2017) Heterochromatin aggregation during DNA elimination in *Tetrahymena* is facilitated by a prion-like protein, *Journal of cell science*, **130**, 480-489.
- Kato, M., *et al.* (2012) Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels, *Cell*, **149**, 753-767.
- Kim, H.J., *et al.* (2013) Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS, *Nature*, **495**, 467-473.
- King, O.D., Gitler, A.D. and Shorter, J. (2012) The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease, *Brain Research*, **1462**, 61-80.
- Kramer, S. (2014) RNA in development: how ribonucleoprotein granules regulate the life cycles of pathogenic protozoa, *Wiley interdisciplinary reviews. RNA*, **5**, 263-284.
- Lambros, C. and Vanderberg, J.P. (1979) Synchronization of *Plasmodium falciparum* erythrocytic stages in culture, *The Journal of parasitology*, **65**, 418-420.
- Lancaster, A.K., *et al.* (2014) PLAAC: a web and command-line application to identify proteins with Prion-Like Amino Acid Composition., *Bioinformatics (Oxford, England)*, **30**, 2-3.
- Le Roch, K.G., *et al.* (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle, *Science*, **301**, 1503-1508.
- Lee, M.C., *et al.* (2008) *Plasmodium falciparum* Sec24 marks transitional ER that exports a model cargo via a diacidic motif, *Molecular microbiology*, **68**, 1535-1546.
- Liu, S., *et al.* (2017) Prions on the run: How extracellular vesicles serve as delivery vehicles for self-templating protein aggregates, *Prion*, **11**, 98-112.
- Maji, S.K., *et al.* (2009) Functional amyloids as natural storage of peptide hormones in pituitary secretory granules, *Science*, **325**, 328-332.
- Majumdar, A., *et al.* (2012) Critical role of amyloid-like oligomers of *Drosophila* Orb2 in the persistence of memory, *Cell*, **148**, 515-529.
- Malinowska, L., Kroschwald, S. and Alberti, S. (2013) Protein disorder, prion propensities, and self-organizing macromolecular collectives, *Biochim Biophys Acta*, **1834**, 918-931.
- Malinowska, L., *et al.* (2015) *Dictyostelium discoideum* has a highly Q/N-rich proteome and shows an unusual resilience to protein aggregation, *Proc Natl Acad Sci U S A*, **112**, E2620-2629.

- Marti, M., *et al.* (2005) Signal-mediated export of proteins from the malaria parasite to the host erythrocyte, *The Journal of cell biology*, **171**, 587-592.
- Masison, D.C. and Wickner, R.B. (1995) Prion-inducing domain of yeast Ure2p and protease resistance of Ure2p in prion-containing cells, *Science*, **270**, 93-95.
- Michelitsch, M.D. and Weissman, J.S. (2000) A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions, *Proc Natl Acad Sci U S A*, **97**, 11910-11915.
- Miller, L.H., *et al.* (2002) The pathogenic basis of malaria, *Nature*, **415**, 673-679.
- Modrzynska, K., *et al.* (2017) A Knockout Screen of ApiAP2 Genes Reveals Networks of Interacting Transcriptional Regulators Controlling the Plasmodium Life Cycle, *Cell host & microbe*, **21**, 11-22.
- Monsellier, E. and Chiti, F. (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution, *EMBO reports*, **8**, 737-742.
- Muralidharan, V. and Goldberg, D.E. (2013) Asparagine repeats in Plasmodium falciparum proteins: good for nothing?, *PLoS pathogens*, **9**, e1003488.
- Muralidharan, V., *et al.* (2012) Plasmodium falciparum heat shock protein 110 stabilizes the asparagine repeat-rich parasite proteome during malarial fevers, *Nature communications*, **3**, 1310.
- Nelson, R., *et al.* (2005) Structure of the cross-beta spine of amyloid-like fibrils, *Nature*, **435**, 773-778.
- Orr, H.T. and Zoghbi, H.Y. (2007) Trinucleotide repeat disorders, *Annual review of neuroscience*, **30**, 575-621.
- Painter, H.J., Campbell, T.L. and Llinas, M. (2011) The Apicomplexan AP2 family: integral factors regulating Plasmodium development, *Molecular and biochemical parasitology*, **176**, 1-7.
- Pallarès, I., Iglesias, V. and Ventura, S. (2016) The Rho Termination Factor of Clostridium botulinum Contains a Prion-Like Domain with a Highly Amyloidogenic Core, *Frontiers in Microbiology*, **6**, 1-12.
- Pallares, I. and Ventura, S. (2017) The Transcription Terminator Rho: A First Bacterial Prion, *Trends in microbiology*, **25**, 434-437.
- Pease, B.N., *et al.* (2013) Global analysis of protein expression and phosphorylation of three stages of Plasmodium falciparum intraerythrocytic development, *Journal of proteome research*, **12**, 4028-4045.
- Prilusky, J., *et al.* (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded, *Bioinformatics*, **21**, 3435-3438.
- Przyborski, J.M., Diehl, M. and Blatch, G.L. (2015) Plasmodial HSP70s are functionally adapted to the malaria parasite life cycle, *Frontiers in molecular biosciences*, **2**, 34.
- Reddy, B.P., *et al.* (2015) A bioinformatic survey of RNA-binding proteins in Plasmodium, *BMC Genomics*, **16**, 890.
- Roberg, K.J., *et al.* (1999) LST1 is a SEC24 homologue used for selective export of the plasma membrane ATPase from the endoplasmic reticulum, *The Journal of cell biology*, **145**, 659-672.
- Sabate, R., *et al.* (2015) Amyloids or prions? That is the question, *Prion*, **9**, 200-206.
- Sabate, R., *et al.* (2015) What makes a protein sequence a prion?, *PLoS Comput Biol*, **11**, e1004013.
- Sant'Anna, R., *et al.* (2016) Characterization of Amyloid Cores in Prion Domains, *Sci Rep*, **6**, 34274.
- Sant'Anna, R., *et al.* (2016) Characterization of Amyloid Cores in Prion Domains, *Scientific Reports*, **6**, 34274.
- Si, K. (2015) Prions: what are they good for?, *Annual review of cell and developmental biology*, **31**, 149-169.
- Singh, G.P., *et al.* (2004) Hyper-expansion of asparagines correlates with an abundance of proteins with prion-like domains in Plasmodium falciparum, *Molecular and biochemical parasitology*, **137**, 307-319.
- Siomi, H., *et al.* (1993) The pre-mRNA binding K protein contains a novel evolutionarily conserved motif, *Nucleic Acids Res*, **21**, 1193-1198.
- Tariq, M., *et al.* (2013) Drosophila GAGA factor polyglutamine domains exhibit prion-like behavior, *BMC Genomics*, **14**, 374.
- Tartaglia, G.G., *et al.* (2005) Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences, *Protein science : a publication of the Protein Society*, **14**, 2723-2734.

- Tartaglia, G.G. and Vendruscolo, M. (2008) The Zyggregator method for predicting protein aggregation propensities, *Chemical Society reviews*, **37**, 1395-1401.
- Tompa, P. and Fuxreiter, M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions, *Trends Biochem Sci*, **33**, 2-8.
- Toombs, J.A., McCarty, B.R. and Ross, E.D. (2010) Compositional determinants of prion formation in yeast, *Mol Cell Biol*, **30**, 319-332.
- Toombs, J.A., *et al.* (2012) De novo design of synthetic prion domains, *Proc Natl Acad Sci U S A*, **109**, 6519-6524.
- Treusch, S. and Lindquist, S. (2012) An intrinsically disordered yeast prion arrests the cell cycle by sequestering a spindle pole body component, *The Journal of cell biology*, **197**, 369-379.
- UniProt, C. (2015) UniProt: a hub for protein information, *Nucleic Acids Res*, **43**, D204-212.
- Williams, A.J. and Paulson, H.L. (2008) Polyglutamine neurodegeneration: protein misfolding revisited, *Trends in neurosciences*, **31**, 521-528.
- Wu, Y., *et al.* (2009) Identification of phosphorylated proteins in erythrocytes infected by the human malaria parasite *Plasmodium falciparum*, *Malaria journal*, **8**, 105.
- Yuan, A.H. and Hochschild, A. (2017) A bacterial global regulator forms a prion, *Science*, **355**, 198-201.
- Yuda, M., *et al.* (2010) Transcription factor AP2-Sp and its target genes in malarial sporozoites, *Molecular microbiology*, **75**, 854-863.
- Zambrano, R., *et al.* (2015) PrionW: a server to identify proteins containing glutamine/asparagine rich prion-like domains and their amyloid cores, *Nucleic Acids Research*, 1-7.
- Zhang, M., *et al.* (2012) PK4, a eukaryotic initiation factor 2alpha(eIF2alpha) kinase, is essential for the development of the erythrocytic cycle of *Plasmodium*, *Proc Natl Acad Sci U S A*, **109**, 3956-3961.

6.3 *In silico* characterization of human prion-like proteins: beyond neurological diseases

Valentín Iglesias^{1†}, Lisanna Paladin^{2†}, Teresa Juan-Blanco³, Irantzu Pallarès¹, Patrick Aloy^{3,4}, Silvio CE Tosatto² & Salvador Ventura^{1,*}

¹Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, 08193, Spain.

²Department of Biomedical Sciences, University of Padua, viale G. Colombo 3, 35121 Padova, Italy.

³Joint IRB-BSC-CRG Program in Computational Biology. Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology. Barcelona, Catalonia, Spain.

⁴Institució Catalana de Recerca i Estudis Avançats (ICREA). Barcelona, Catalonia, Spain.

†These authors contributed equally to this work. *To whom correspondence should be addressed.

Author Contributions: Software, validation, data curation, writing—original draft preparation.

6.3.1 ABSTRACT

Prion-like behavior has been in the spotlight since it was first associated with the onset of mammalian neurodegenerative diseases. However, a growing body of evidence suggests that this mechanism could be behind the regulation of processes such as transcription and translation in multiple species. Here, we perform a stringent computational survey to identify prion-like proteins in the human proteome. We detected 242 candidate polypeptides and computationally assessed their function, protein-protein interaction networks, tissular expression and their link to disease. Human prion-like proteins constitute a subset of modular polypeptides broadly expressed across different cell types and tissues, significantly associated with disease, embedded in highly connected interaction networks and involved in the flow of genetic information in the cell. Our analysis suggests that these proteins might play a relevant role not only in neurological disorders, but also in different types of cancer and viral infections.

6.3.2 INTRODUCTION

Prions were first reported in the context of mammalian neurodegenerative disorders (Harrison, et al., 2010; Prusiner, 1982; Sikorska and Liberski, 2012; van Rheede, et al., 2003), but it is now clear that different organisms exploit prion conformational conversion for functional purposes (Halfmann and Lindquist, 2010). The most studied organism is *Saccharomyces cerevisiae*, with up to 11 functional prions identified so far (Batlle, et al., 2017c; Cascarina and Ross, 2014). Initially, yeast prions were proposed to

be pathological agents (McGlinchey, et al., 2011; Nakayashiki, et al., 2005), but nowadays they are widely recognized to provide beneficial advantages in changing environments, predominantly by regulating transcription, translation or RNA processing (Halfmann, et al., 2012; Newby and Lindquist, 2013). Yeast prions switch from an initially soluble state through a structural conversion towards an aggregated amyloid conformation. This conversion is encoded in prion domains (PrDs); long intrinsically disordered regions of low complexity.

A significant number of proteins sharing most, but not all, prion characteristics have been identified in different organisms, and generically named prion-like proteins (Chakrabortee, et al., 2016; Pallares, et al., 2015; Si, 2015). In higher eukaryotes, prion-like structural conversion plays a central role in diverse functions such as viral response (Franklin, et al., 2014; Hou, et al., 2011; Xu, et al., 2014) or long-term memory acquisition and maintenance (Majumdar, et al., 2012; Si, et al., 2010; Si and Kandel, 2016). Even though multiple beneficial functions have been assigned to prion-like mechanisms across all kingdoms of life, aggregated proteins in human neurodegenerative diseases such as Alzheimer's and Parkinson's diseases and amyotrophic lateral sclerosis also share certain prion-like properties (Aguzzi and Rajendran, 2009; Gitler and Shorter, 2011; Kim, et al., 2013; Luk, et al., 2012; Nomura, et al., 2014; Stohr, et al., 2012).

The accumulated knowledge on the determinants of yeast prions conformational conversion has provided strong stimuli for the development of bioinformatics tools to uncover new prion-like domains (PrLDs) in other organisms (Afsar Minhas, et al., 2017; Batlle, et al., 2017c; Espinosa Angarica, et al., 2014; Harrison and Gerstein, 2003; Lancaster, et al., 2014; Michelitsch and Weissman, 2000; Toombs, et al., 2012). Previous screenings for PrLDs in the human proteome have targeted the characteristic compositional bias of these protein regions (An and Harrison, 2016). We have recently proposed that, in addition to a distinctive amino acidic composition, PrLDs contain soft-amyloidogenic sequence stretches that would contribute to trigger the initial protein self-assembly reaction (Sabate, et al., 2015; Sabate, et al., 2015). These cryptic amyloids were not only shown to be present and promote conformational conversion in *bona fide* yeast prions (Sant'Anna, et al., 2016), but they also exist in human prion-like proteins (Batlle, et al., 2017a) and appear to play key role in the induction, propagation and inheritance of the prion state in the mammalian cytosol (Duernberger, et al., 2018). The amyloid stretches embedded within PrLDs can be identified computationally (Sabate, et al., 2015; Zambrano, et al., 2015).

Here we applied to the human proteome the same prediction scheme that allowed us to uncover the first *bona fide* prion-like protein in a bacterial proteome (Pallares, et al., 2015; Yuan and Hochschild, 2017). Human proteins were first analysed for the presence of regions with compositional similitude to yeast prion domains using the PLAAC algorithm (Alberti, et al., 2009; Lancaster, et al., 2014) and afterwards these protein domains were individually screened for the presence of soft-amyloidogenic sequences using the pWALTZ program (Sabate, et al., 2015). Indeed, we have recently shown that such a combination of compositional and sequential PrLDs prediction, provides the best accuracy when forecasting the aggregation propensities of individual human prion-like proteins (Batlle, et al., 2017b).

In the present work, we computationally characterized the function, location, expression, protein-protein interaction networks and the connection to disease of the human prion-like subproteome. The picture that emerges from this analysis is that prion-like proteins are widespread expressed proteins that function in biological processes tightly associated to disease.

6.3.3 MATERIAL AND METHODS

Data acquisition

The human reference proteome dataset was obtained from UniprotKB (UniProt, 2015) (Proteome ID UP000005640; release 2016_09) and scanned for PrLDs with PLAAC using as background probability the frequency of human proteome. From the initial 70940 proteins in the proteome, 431 PrLD containing candidates were identified. Their predicted PrLDs were further evaluated with pWALTZ applying a cutoff of 60.00, as in (Batlle, et al., 2017a), which resulted in 242 final positive predictions.

Prion-like domain localization within the protein sequence

Each prion-like protein sequence was divided into 3 segments, the N- and C- terminal, accounted for 25% of the residues each, whereas the resting 50% of the sequence was considered as internal. Each predicted PrLD was located in the sequence and the number of residues mapping in each of the segments counted.

Functional Annotation

The GO annotation of all proteins in the prion-like dataset were collected, excluding the terms Inferred from Electronic Annotation (IEA) and filtering through the Generic GO slim developed by GO Consortium (Gene Ontology, 2015). All UniProtKB human proteins were used as background set to infer enrichment. A Fisher's exact test of GO term distributions was performed in the three ontologies separately, to calculate the enrichment/depletion of dataset proteins with respect to the whole UniProtKB. The Bonferroni correction was applied in performing all the tests. The results are shown in **Figure 6.15** applying the formula:

$$E = \log \frac{\text{GO freq. in } P_{PR}}{\text{Tot GO in } P_{PR}} - \log \frac{\text{GO freq. in } P_{Back}}{\text{Tot GO in } P_{Back}}$$

Where GO is the GO term, P_{PR} and P_{Back} are the datasets of prion-like proteins and the whole proteome, respectively. The abbreviations freq. and Tot stay for frequency and total.

Pfam domains

Pfam (Finn, et al., 2016) domains annotation in the dataset proteins were collected and compared to the human proteome (from UniProtKB). Fisher's exact test was used to assess significance.

Tissue and cellular localization

Tissue and cellular localization data of human proteins were retrieved from Human Protein Atlas (Uhlen, et al., 2015). The prion-like proteins identifiers were converted to Ensembl Gene Ids. Human Protein Atlas reports a textual ranking of protein expression of each coding gene. This ranking (“none”, “low”, “medium”, “high”) was converted to numerical expressions, from 0 to 3, and each gene value for each particular tissue was collected. The expression of the complete gene set for the tissue was then averaged.

Association to diseases

OMIM disease annotation were extracted from the field “diseases” of the UniProtKB description (Amberger, et al., 2015). All information regarding the associated diseases was collected from the OMIM FTP site. DisGeNET data was retrieved from DisGeNET download section (Pinero, et al., 2015). For both databases, the number of proteins associated to at least one disease ID was divided by the total number of proteins, obtaining the fraction of disease-associated proteins. The results were compared to 100 random sampling of sets with same number of proteins than the one in the database.

Human network analysis

The human prion-like protein dataset was curated for duplicities and scanned for protein-protein interactions (PPI) with Interactome3D (2017_06 version) (Mosca, et al., 2013). Out of the 121 unique identities, 100 had annotated physical binary interactions. The degree and the number of interactions between prion-like proteins were analysed and compared to a random distribution by sampling the complete human binary interactome in Interactome3D. Moreover, the size of the largest connected component (LCC) and the mean shortest distance (MSD) were measured (Menche, et al., 2015). The subnetwork of prion-like proteins and their interactors were functionally characterized with DAVID database (Huang da, et al., 2009) for Gene Ontology and KEGG pathways enrichment (n=1542). The significance of the differences was assessed by Wilcoxon p-value or empirical p-value.

6.3.4 RESULTS

Human prion-like proteins prevalence and modularity

A combination of prion-like compositional bias (PLAAC) and sequential soft-amyloid propensity (pWALTZ) analysis was applied to the complete human proteome. This resulted in the identification of a total of 242 polypeptides (unique UniProtKB entries) bearing PrLDs. Our list of candidates included all human prion-like proteins shown to behave as such both *in vitro* and *in vivo*: FUS (Ju, et al., 2011), TDP-43 (Wang, et al., 2012), EWS (Couthouis, et al., 2012), hnRNP A1 and hnRNP A2 (Kim, et al., 2013), TIA1 (Li, et al., 2014) and TAF15 (Couthouis, et al., 2011) proteins, reinforcing the suitability of our dataset for the further evaluation of the global properties of human prion-like sequences.

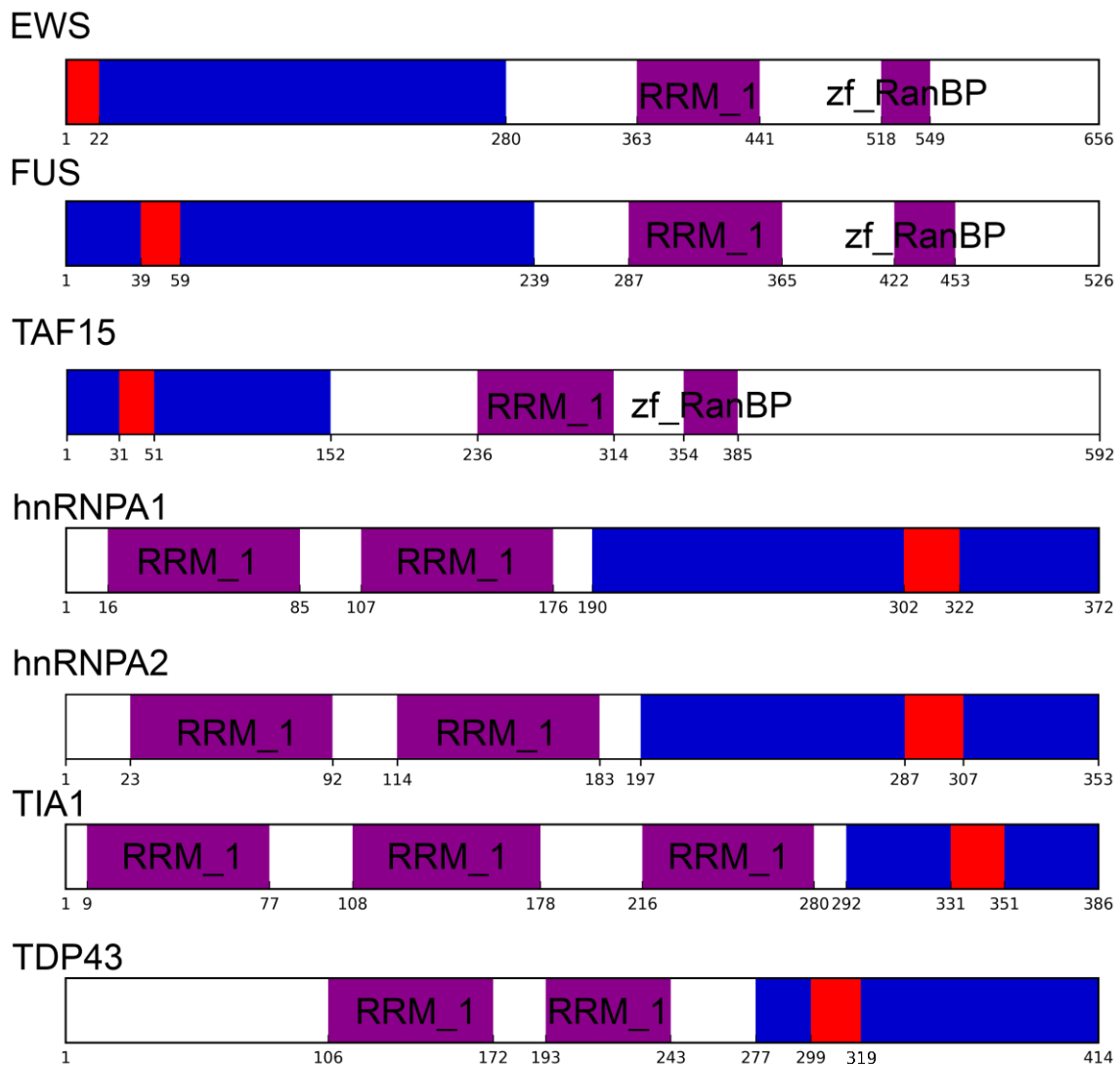


Figure 6.13 – Human prion-like proteins modularity. Well-characterized prion-like human proteins have their PrLD (as identified by PLAAC in blue) and soft-amyloid core (as identified by pWALTZ in red) at the protein edges, separated from their respective globular domains (retrieved from Pfam database in violet).

According to our predictions, prion-like proteins account for a 0.34% of the human proteome. This is in line with two previous independent surveys for human prion-like proteins that exploited compositional bias alone for their detection; both studies predicting that the prevalence of these proteins is < 1% (An and Harrison, 2016). Despite the percentage of proteins with PrLDs in the proteomes of different organisms seems to differ significantly (Chakrabortee, et al., 2016; Espinosa Angarica, et al., 2013; Malinovska, et al., 2015; Michelitsch and Weissman, 2000; Pallares, et al., 2018), their presence in all evolutionary lineages analysed so far suggests that these regions might play conserved functional roles (Batlle, et al., 2017a; Malinovska, et al., 2015; Michelitsch and Weissman, 2000).

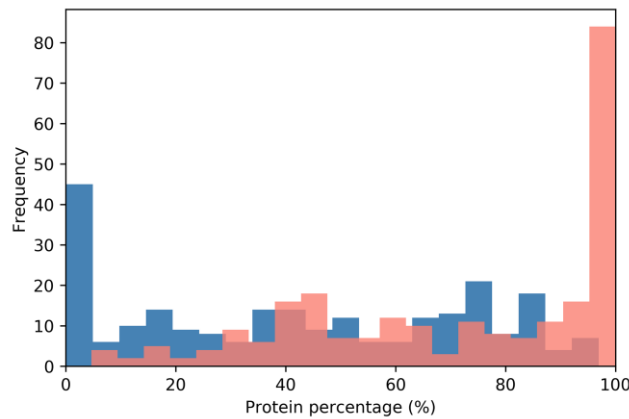


Figure 6.14 – PrLD distribution along the protein sequence. The relative position of PrLDs in the sequences of the complete protein dataset is plotted. Protein sequences were divided into 20 bins corresponding to 5 % of their length and the PrLDs start position represented in blue and the end in red.

Yeast prion proteins tend to be modular (Alberti, et al., 2009; Li and Lindquist, 2000). Prion domains being generally located near the N- or C- terminal ends of the sequence (Baxa, et al., 2007; Zambrano, et al., 2015). In our dataset, 195 proteins; an 80.6% of the putative human prion-like proteins, presented their PrLDs located in any of the protein's ends (**Figures 6.13, 6.14**). PrLDs were 1.67 times more frequent at the protein C-terminus. This was the case for 122 proteins, while in 73 of them the PrLDs were located at the N-terminus. This statistically significant imbalance between the presence of PrLDs at C- and N- in human proteins (p-value < 0.005, Z-test), contrasts with that found in *bona fide* yeast prion domains. In SUP35, URE2, NEW1, MOT3 and SWI1 proteins the prion domain is placed at the N-terminus, whereas only in RNQ1 it is located near the carboxyl end (Baxa, et al., 2007; Zambrano, et al., 2015). The modular architecture of prion-like proteins would allow the self-assembly of the PrLDs without disturbing the structure and productive associations of the adjacent globular moieties. This is likely facilitated by the predicted disordered nature of these protein segments.

Human prion-like proteins play a major role in nucleic acid binding

As a first step to gain insights into the biological role of the candidate human prion-like proteins we used a Gene Ontology (GO) enrichment analysis. GO terms were collected for biological process, molecular function, and cellular component categories and their enrichment with respect to the human proteome calculated (**Figure 6.15**). When we analysed the 'biological process' category for the set of candidate proteins, we found a statistically significantly enriched cluster of GO terms related to RNA and DNA associated processes, including positive regulation of transcription from RNA polymerase II promoter (p-value < 1.20×10^{-16} , 30 proteins), positive regulation of transcription DNA-templated (p-value < 6.92×10^{-14} , 22 proteins), mRNA splicing (p-value < 2.27×10^{-9} , 13 proteins), transcription DNA-templated (p-value < 5.26×10^{-8} , 36 proteins), RNA processing (p-value < 7.5×10^{-8} , 10 proteins) and negative regulation of transcription from RNA polymerase II promoter (p-value < 6.28×10^{-4} , 11 proteins) (**Figure 6.15A**). This result is consistent with the observation that the prion-like subproteomes identified in organisms

belonging to different taxonomic divisions are usually enriched in proteins associated to the regulation of the flux of genetic information in the cell (Iglesias, et al., 2015; Pallares, et al., 2018).

With respect to the 'molecular function', the most enriched GO terms are all involved in essential activities related with nucleic acid binding and transcription processes, such as transcription coactivator activity (p-value < 5.63×10^{-17} , 20 proteins), nucleotide binding (p-value < 4.96×10^{-17} , 37 proteins), poly(A)RNA-binding (p-value < 3.94×10^{-15} , 30 proteins), RNA-binding (p-value < 2.99×10^{-14} , 31 proteins), chromatin binding (p-value < 3.34×10^{-14} , 14 proteins), transcription factor activity-sequence-specific DNA binding (p-value < 9.79×10^{-6} , 29 proteins) and ATP binding (p-value < 1.14×10^{-4} , 13 proteins) (**Figure 6.15B**). The conformational plasticity of PrLDs has been shown to be behind certain transcription factors ability to bind to many different targets and to play a role in the formation of chromatin regulatory complexes (Boulay, et al., 2017; Cho, et al., 2018; Kataoka and Mochizuki, 2017). Moreover, it is becoming increasingly clear that PrLDs are crucial for the formation of membraneless organelles, since they enable RNA-binding proteins (RBPs) to undergo liquid-liquid transition, confining their RNA cargos (Villarroya-Beltri, et al., 2013; Wang, et al., 2018).

When we analysed the cellular components populated by our protein subset, the most enriched GO terms were the nucleoplasm, nucleus and the intracellular ribonucleoprotein complex (**Figure 6.15C** and **6.15D**). As expected, all these compartments correspond to locations where the binding between nucleic acids and proteins occur frequently. Of particular interest is the so-called ribonucleoprotein complex which includes cellular structures like the stress granules, or P-bodies, which are sites for mRNA decay as well as for mRNA storage and therefore act as important cell regulatory centers in determining levels of gene expression (Anderson, et al., 2015). The RBPs associated to those membrane-less organelles are key determinants in the control of the organelle function and have been implicated not only in adaptation to stress but also in tumor biology and the pathogenesis of neurodegenerative, immunological and infectious diseases (Anderson, et al., 2015; Harrison and Shorter, 2017; Loomis, et al., 1990; Villarroya-Beltri, et al., 2013).

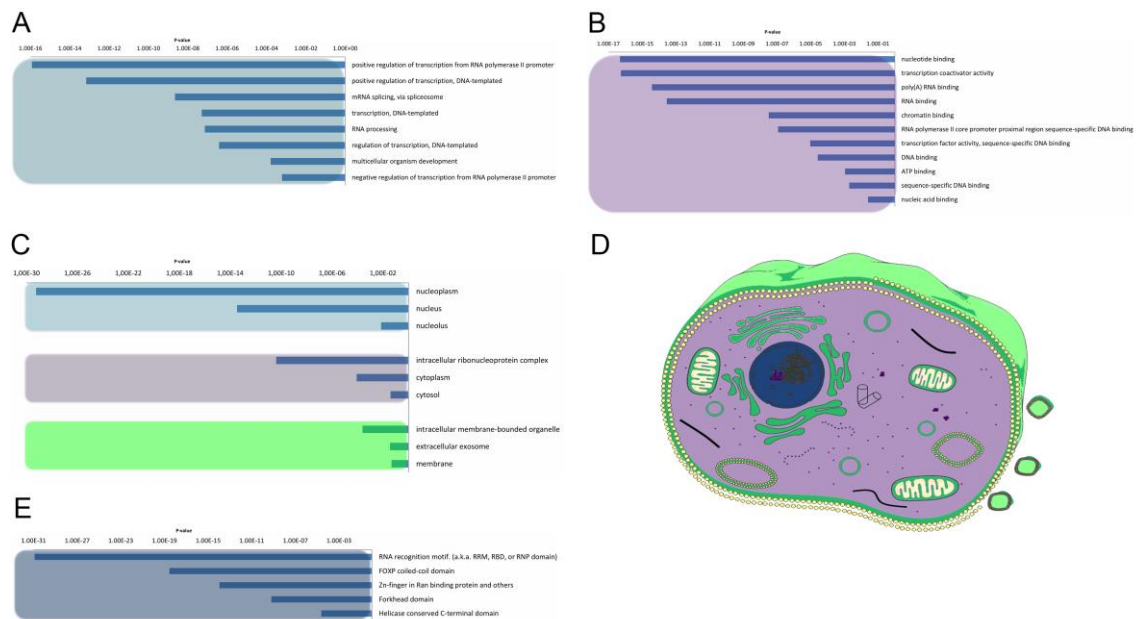


Figure 6.15 – Human prion-like proteins GO enrichment analysis. The prion-like proteins GO enrichment was performed and separated into its three ontologies. **A)** Biological Process **B)** Molecular function **C)** and **D)** Cellular component. Clusters were grouped by color and represented with the same color-code in a mammalian cell in **D**. **E)** Pfam structural domains enriched in prion-like proteins were computed against the human proteome background.

We extended our analysis to look for the role of the constituent functional domains in the collection of PrLDs containing proteins. In agreement with the above presented results, Pfam domain clustering rendered DNA/RNA binding as the most enriched functional group (**Figure 6.15E**). Among them, the canonical RNA recognition motif (RRM) is by far the most statistically enriched, with 14% of the detected proteins harboring an RRM. This observation is line with previous studies (King, et al., 2012) and consistent with the fact that the RRM is the most abundant domain in RBPs, conserved from bacteria to higher eukaryotes (Reddy, et al., 2015). This set of RRM-bearing prion-like proteins includes FUS, TDP-43, TIA1 or hnRNP A1, all involved in the formation of dynamic membraneless intracellular compartments and associated to disease (Cascarina and Ross, 2014; March, et al., 2016; Wang, et al., 2018).

The second most enriched domain in our data set is the FoxP coiled-coil (p-value < 2.95×10^{-19} , 10 proteins). It corresponds to a coiled-coil domain involved in the modulation of the dimeric associations of the forkhead box family of transcription factors FoxP. There are multiple lines of evidence suggesting the biological relevance of domain swapping in FoxP functionality being important not only for their function regulation but also linked to disease onset (Hafner-Bratkovic, et al., 2011; Medina, et al., 2016).

The other two enriched Pfam families include Zinc-fingers in Ran binding proteins (Zn_RanBP) (p-value < 1.14×10^{-14} , 9 proteins) and the Helicase conserved C-terminal domain (p-value < 2.47×10^{-5} , 7 proteins). Zinc Finger domains are a very versatile group of small protein domains which are evolutionary conserved. Interestingly, RBPs with PrLDs such as FUS or EWS accommodate in their structure a Zn_RanBP domain in close proximity to an RRM domain. The Helicase conserved C-terminal domain is found at the C-terminus of DEAD-box helicases. Helicases function in the separation of double-stranded RNA, DNA, and RNA/DNA structures in an energy-dependent manner and therefore it is clear their role in RNA metabolism.

Interestingly, the first prion-like protein identified in bacteria corresponds to the transcription terminator Rho, a helicase that can undergo a prion-state that results in genome-wide changes at the transcriptome level, contributing to rapid bacterial adaptation to fluctuating environments (Pallares, et al., 2015; Yuan and Hochschild, 2017). The multitasking transcriptional regulators DDX5 and DDX17 included in our dataset contain an helicase domain in their structure reported to be associated with cancer development and cell proliferation (Fuller-Pace, 2013; Mazurek, et al., 2012).

Prion-like proteins are widespread among tissues

The histological localization of human prion-like proteins was assayed by retrieving data from the Human Protein Atlas. To compare the expression levels, proteins were mapped to Ensemble gene annotations (121 genes). The expression data was collected for each cell type and averaged by tissue and organ. The result illustrates that prion-like proteins are widely distributed in human tissues (**Figure 6.16**). Importantly, the data indicates that, globally, the expression of these proteins in the brain is not higher than in most organs or tissues, being more represented in endocrine tissues, in the gastrointestinal tract, the kidney or the lung.

In order to identify interesting cases, we clustered the dataset by representing each gene as a vector of the variance of its expression with respect to the proteome-level tissue average ($V_g = [(E - \bar{E})_1 \dots (E - \bar{E})_n]$ where: V_g : the vector of gene expressions; E : gene expression in tissue n and \bar{E} : average expression of all human proteome in tissue n). The clustering was performed through k-means algorithm implementation of scikit-learn Python module, which uses Euclidean distances by default. We tested cluster numbers from 3 to 10 and chose 6 as the most discriminative one. Thus, the highest expression level cluster represents a group of prion-like proteins that are generally over-expressed and remarkably includes most of the human prion-like proteins for which it has been already demonstrated their direct involvement in disease: FUS, TDP-43, hnRNP A1 hnRNP A2/B1, hnRNP A3, hnRNP U, hnRNP H1 and EWS. Many of these proteins have already been described to be spread throughout most tissues and identified at different developmental stages (Bastian, et al., 2008; Uhlen, et al., 2015).

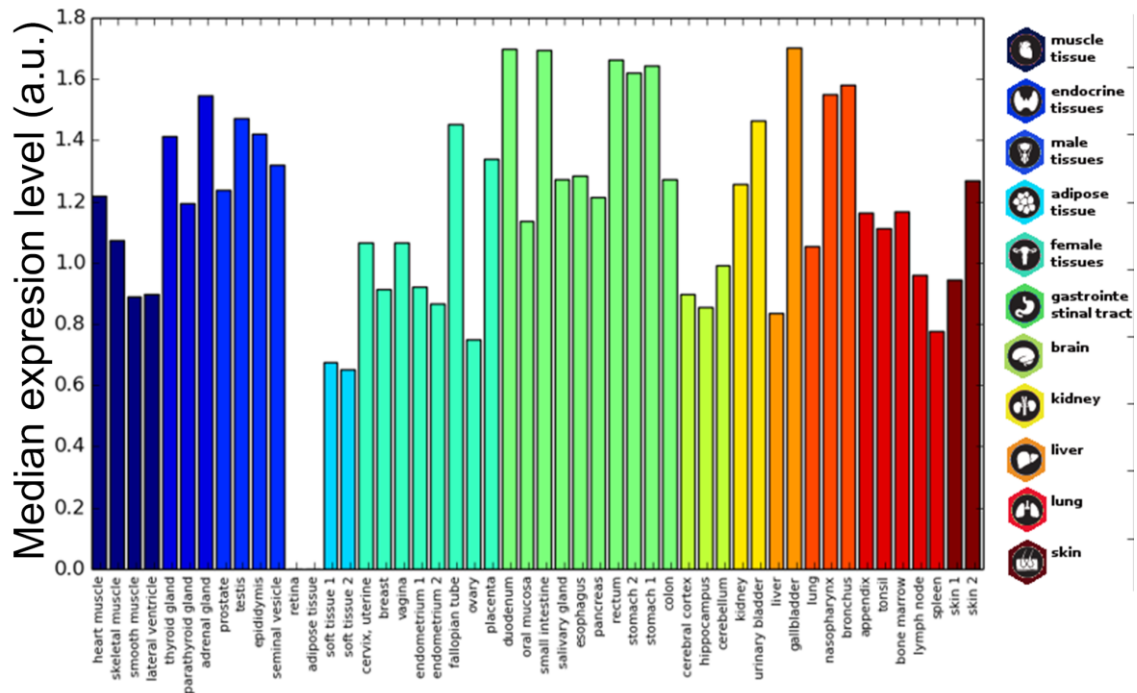


Figure 6.16 – Prion-like proteins expression in human tissues. The average expression of prion-like proteins dataset is plotted for different tissues. The tissue bars are colored based on the corresponding organ/tissue. Values range from 0 to 3 corresponding to Human Protein Atlas annotation “not detected”, “low”, “medium” and “high”.

Prion-like proteins are disease related

Given the widespread tissue distribution of the prion-like proteins and the link to disease of proteins in the most expressed cluster, we explored whether, globally, genes encoding for these polypeptides were connected to pathological processes. Their association to diseases was retrieved separately from the Online Mendelian Inheritance in Man (OMIM) (Amberger, et al., 2015) and the database of gene-disease association (DisGeNET) (Pinero, et al., 2017). The percentage of genes with disease annotations was calculated and compared with that in the complete human UniProtKB dataset, which was used as background. According to the OMIM database, 13.22% of the prion-like proteins encoding genes are disease-related against a 2.39% for the UniProtKB dataset, whereas values of 33.47% and 9.49% were obtained in the case of DisGeNET (p -value $< 1.0 \times 10^{-5}$ for both databases, Z-test). Thus, the association with disease of prion-like proteins was three-fold and five-fold higher than the one in the complete human proteome, according to DisGeNET and OMIM, respectively. To assess the significance of this enrichment, 100 random samples with the same size that the prion-like proteins dataset were selected from the background, the percentage of proteins associated to a disease in each sample was counted and the distribution of the percentages calculated (**Figure 6.17**). For both OMIM and DisGeNET, the prion-like dataset proportion is clearly above the 95 percentile of the distribution, which implies a significant over-representation of disease-associated proteins among human prion-like proteins. At this point, it is important to underline that the prion-like protein identification pipeline is sequence-based and totally blind with respect to the protein annotation.

Prion-like proteins have been associated to the onset of neurological disorders (Harrison, and Shorter, 2017). The 9% of genes encoding for prion-like proteins, 11 out of 121, are linked to neurological diseases, according to OMIM. This constitutes a significant enrichment, relative to the complete proteome (p -value $< 1.5 \times 10^{-8}$). However, it is important to note that, despite proteins connected with neurological disorders are over enriched by 1.4-fold within the disease associated prion-like protein subgroup, this enrichment is not statistically significant (p -value > 0.11). It is clear from the results presented above, that many of the detected proteins are ubiquitous regulators involved in a wide range of signaling pathways; which suggests that perturbations affecting their function may have a great impact in multiple disorders and not exclusively in neurological diseases, as it is usually assumed.

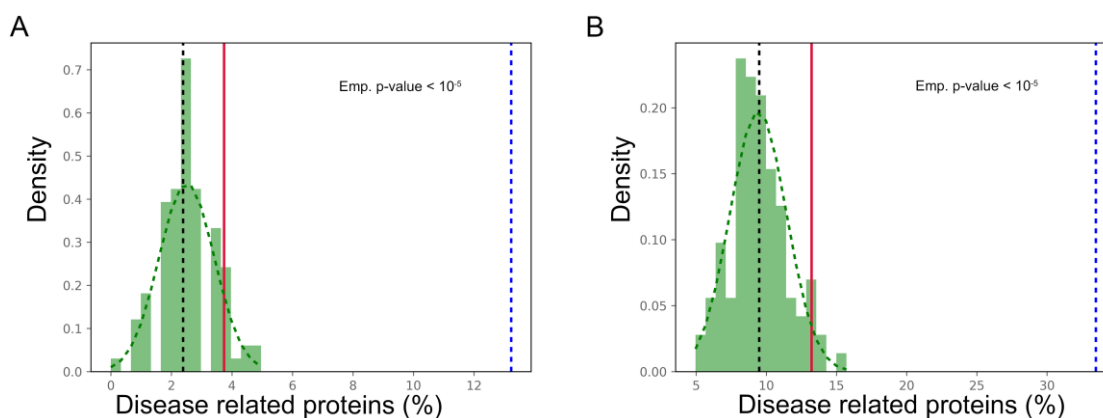


Figure 6.17 – Human prion-like proteins disease association. Number of disease-associations for prion-like proteins (dotted blue line) compared to 100 random sampling of the human UniProtKB from **A)** OMIM and **B)** DisGeNET databases. The median of the background sample is plotted as a dotted black line, while the red line refers to the 95 percentile of the distribution (p -value < 0.05).

Prion-like proteins' role in highly interconnected subnetworks

Proteins rarely perform their functions independently; but mostly rely on complexes to carry them out. The connectivity of human prion-like proteins and the properties of their interactors were analysed. As above, prion-like proteins were first mapped to genes to obtain unique entities. Out of the 121 resulting genes, 100 had annotated physical binary interactions (physical interactions between two individual proteins). Overall, prion-like dataset and the proteins they interact with establish a subnetwork of 1544 proteins with 2079 protein-protein interactions (PPI) between them. Both the prion-like dataset and the complete subnetwork have higher average interaction degrees than the human interactome (**Figure 6.18A**). To uncover whether prion-like proteins interact more than expected by chance, the average degree of interactions of the prion-like protein set was compared with 1000 random sets of proteins of the same size (**Figure 6.18B**). This analysis confirms that prion-like proteins exhibit a significant higher number of interactions than the average human interactome. Next, we assessed whether prion-like proteins interact more between them than expected by chance, by comparing the number of intra-set interactions with that in 1000 random sets, as before. The results showed that prion-like proteins establish more interactions -one order of magnitude higher- between them than expected randomly (**Figure 6.18C**). To further describe the human prion-like subnetwork, it was tested to what extent prion-like proteins

cluster into specialized interactome neighborhoods. The size of the largest connected component (LCC) and the mean shortest distance (MSD) was measured and compared to 1000 random sets (**Table 6.2**). The results clearly show that prion-like proteins share a higher interactomic vicinity than expected randomly, providing support to the concept that they exist well-defined interaction networks for human prion-like proteins.

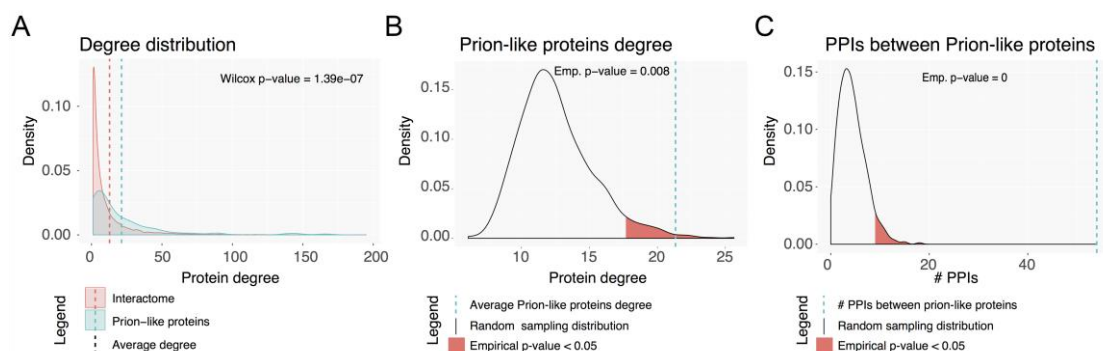


Figure 6.18 – Human prion-like interactome. **A)** Degree distribution for the complete interactome in red and the prion-like proteins network (first neighbors) in blue. **B)** Prion-like proteins average interaction degree (dotted blue line) compared to a random sampling of 1000 sets. **C)** Number of PPIs between prion-like proteins (dotted blue line) compared to a random sampling of 1000 sets.

Table 6.2 – Prion-like proteins are located nearer in the network than expected by chance.

	<i>Prion-like protein set</i>	<i>Random expectation</i>	<i>Z-score</i>	<i>P-value</i>
<i>LCC size</i>	32	3.16	16.8	<1 x 10 ⁻⁵
<i>MSD</i>	1.64	2.2	-6.78	<1 x 10 ⁻⁵

To functionally characterize this subnetwork of prion-like proteins and their interactors, the 1544 proteins were analysed for Gene Ontology and KEGG pathways enrichment. GO enrichment analysis are consistent with the results obtained for the prion-like proteins dataset alone, as it highlights regulation of gene expression through DNA and RNA binding as the main biological role played by this protein subset.

When we examined the statistically enriched pathways obtained from KEGG analysis, we observed that they can be grouped into two main clusters. Remarkably, the largest cluster collects pathways involved in different types of cancer, such as transcriptional misregulation in cancer (p-value < 9.86 x 10⁻¹⁵, 53 proteins), pancreatic cancer (p-value < 2.01 x 10⁻¹², 29 proteins), prostate cancer (p-value < 1.31 x 10⁻¹¹, 33 proteins) or colorectal cancer (p-value < 1.88 x 10⁻⁷, 22 proteins) among others. 12 prion-like proteins (10% of the total unique entries) and 122 (8.4%) of their interactors were found in these cancer related pathways. These interactors include cornerstones in mitogenesis, growth factor signaling, apoptotic attenuation, cell cycle progression, angiogenesis, cell invasion, immune regulation, and microenvironment alterations.

The second group encompasses pathways associated with viral infection, such as viral carcinogenesis (p-value < 1.42×10^{-15} , 61 proteins), Epstein-Barr virus infection (p-value < 3.91×10^{-14} , 56 proteins), herpes simplex infection (p-value < 6.6×10^{-12} , 51 proteins) or Hepatitis C (p-value < 1.07×10^{-6} , 38 proteins). This is consistent with the involvement of RNA-binding proteins, helicases and splicing-related proteins in the control of viral assembly and trafficking of the viral genomic RNA from the nucleus. Prion-like candidates such as DDX17 (Moy, et al., 2014), DDX5 (Cheng, et al., 2018) and hnRNP A2B2 (Levesque, et al., 2006) have been already described to play key roles in these processes.

6.3.5 DISCUSSION

In the present work we used a stringent computational approach that considers that PrLDs should not be only disordered and compositionally biased, but also encode for short sequences with moderate, but significant, amyloid propensity (Sabate, et al., 2015). We concluded that 242 polypeptides in the human proteome fulfil the requirements to potentially behave as prion-like proteins. This accounts for less than 1% of the human proteins, which implies that, compared with organisms like *Plasmodium* or *Dictyostelium* where 10 to 25% of their proteins are predicted be prionogenic (Singh, et al., 2004), the prionic load of the human proteome is low. The dataset included several widely studied proteins with prion-like behavior, such as FUS, TIA1, TDP-43, EWS, and several hnRNPs, but also previously undescribed proteins with very important cellular functions: members of the mediator complex, nucleoporins, chromatin remodeling proteins and transcription factors.

As their counterparts in yeast (Alberti, et al., 2009; Santoso, et al., 2000), human prion-like proteins, locate their PrLDs mostly at their ends; with a slight preference for the amino terminus. This might imply that the position of the PrLD within the protein sequence might be relevant for its function. Indeed, previous analyses on proteins containing low complexity regions, already suggested that these terminal positions would allow them to act as promiscuous interfaces for protein binding, without steric interferences by the adjacent globular domains (Coletta, et al., 2010). In a similar manner, prion-like modularity and the preference for terminal regions are likely maintained in order to delimit a flexible region which can switch its conformation and assemble, modulating in this way the activity of folded domains without impacting their native 3D structure.

According to the GO terms analysis, a highly significant fraction of prion-like proteins are involved in functions related to nucleic acid binding and transcription and translation activities. This include proteins of the Mediator complex, implicated in the regulated transcription of nearly all RNA polymerase II-dependent genes (Cho, et al., 2018; Zhu, et al., 2015), proteins recruited in chromatin-remodeling complexes (Boulay, et al., 2017; Kataoka and Mochizuki, 2017), and a significant number of transcription factors. The dataset also includes the large majority of RNA-binding proteins already described to behave as prion-like in humans, such as FUS which is implicated in transcription, DNA repair, and RNA biogenesis (Patel, et al., 2015), TIA1 which functions in mRNA turnover and regulation of translation (Li, et al., 2014), TDP-43 which is involved in transcriptional regulation and RNA processing (Buratti and Baralle, 2008; King,

et al., 2012), EWS which is implicated in RNA binding and processing or diverse hnRNPs involved in the packaging of pre-mRNA into RNP particles (He and Smith, 2009). Not surprisingly, we found that a high proportion of these proteins map into the nucleus and intracellular ribonucleoprotein complex. This last observation is consistent the extensive literature identifying prion-like sequences as drivers of liquid-liquid phase separation in membrane-less cellular compartments (Banani, et al., 2017; Patel, et al., 2015).

Our data reveal that human prion-like proteins are multifunctional proteins involved in important regulatory processes. Indeed, 50% of the proteins in our dataset carry at least two different Pfam domains. As expected from the molecular functions in which these proteins are involved, the most statistically enriched domains correspond to RNA and DNA binding domains such as the canonical RNA recognition motif, the Zn finger domain, the forkhead domain or the helicase domain. All them present in well characterized transcription factors and RNPs. These are evolutionary conserved domains in which, because of their functional relevance, genetic mutations are often linked to disease (Cascarina and Ross, 2014; King, et al., 2012).

We assessed the expression of genes coding for prion-like proteins for each human tissue, to try to rationalize why, so far, these proteins have been mostly related to neurological diseases. Human prion-like protein expression was not restricted to nervous tissue but ubiquitously spread among tissues; also, they are not especially abundant in the brain, relative to other organs of the human body. This suggests that they play a physiological role in different cellular types, although it raises the question of why most prion-like proteins related diseases are tissue-specific. This situation is not unique for prion-like proteins but common to other proteins involved in neurodegenerative disorders, i. e. α -synuclein the protein responsible for Parkinson's disease, is abundantly expressed in both the cerebral cortex and the bone marrow, but only aggregates in the brain (Barbour, et al., 2008; Spillantini, et al., 1997). The protein quality control machinery has an active role in managing protein misfolding and aggregation. Cellular aging impacts cell homeostasis and leads to proteostatic-compromised cells in which misfolding and aggregation events cannot be compensated (Aguzzi and Altmeyer, 2016). It has been proposed that the low efficacy of replacing dying neurons, relative to other cells types, could be one of the underlying reasons why the malfunction of prion-like proteins is more often associated to neurological conditions. One important finding here is that many of the human prion-like proteins that have been convincingly associated to disease are among the most expressed polypeptides in the dataset. This fits very well with the so called "life at the edge" hypothesis, which states that, because protein aggregation is extremely dependent on concentration, abundant proteins are, on the average, at highest risk of misfolding and aggregation (Tartaglia, et al., 2007).

Independently of their tissue distribution, what becomes clear from the analysis of the OMIM and DisGeNet databases is that human prion-like proteins are strongly connected to disease. Two complementary properties might explain, at least in part, this strong association. First, the propensity of PrLDs to establish intermolecular interactions together with the presence of regions with significant amyloid propensity, exposed to solvent within large disordered regions, impose an inherent risk to

aggregate to these polypeptides. In fact, genetic mutations that increase the aggregation propensity of PrLDs have been shown to be directly associated with disease (Harrison and Shorter, 2017). Second, according to the “centrality-lethality rule” (Jeong, et al., 2001) the highest the number of interactions for a protein is, the largest is the impact of its disruption on cell function. Thus, the high connectivity of prion-like proteins networks might well account for their strong link to human diseases. Importantly, KEGG pathway enrichment analysis of the prion-like proteins interactome allowed us to uncover a highly significant association with two previously undescribed set of devastating pathological processes: cancer and viral infections.

Overall, despite the present study constitutes only a first theoretical approach to the function of human prion-like proteins, our results indicate that this subproteome exert important regulatory functions in different biological pathways, thanks to both their protein-protein and protein-nucleic acids binding capabilities, two properties that seem to be favored by their modular architecture. The analysis suggests that in the forthcoming years we can expect the discovery of a connection between prion-like proteins malfunction and other pathologies apart from neurological disorders.

6.3.6 REFERENCES

- Afsar Minhas, F.U., Ross, E.D. and Ben-Hur, A. (2017) Amino acid composition predicts prion activity, *PLoS Comput Biol*, **13**, e1005465.
- Aguzzi, A. and Altmeyer, M. (2016) Phase Separation: Linking Cellular Compartmentalization to Disease, *Trends Cell Biol*, **26**, 547-558.
- Aguzzi, A. and Rajendran, L. (2009) The transcellular spread of cytosolic amyloids, prions, and prionoids, *Neuron*, **64**, 783-790.
- Alberti, S., et al. (2009) A systematic survey identifies prions and illuminates sequence features of prionogenic proteins, *Cell*, **137**, 146-158.
- Amberger, J.S., et al. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders, *Nucleic Acids Res*, **43**, D789-798.
- An, L. and Harrison, P.M. (2016) The evolutionary scope and neurological disease linkage of yeast-prion-like proteins in humans, *Biology direct*, **11**, 32.
- Anderson, P., Kedersha, N. and Ivanov, P. (2015) Stress granules, P-bodies and cancer, *Biochim Biophys Acta*, **1849**, 861-870.
- Banani, S.F., et al. (2017) Biomolecular condensates: organizers of cellular biochemistry, *Nature reviews. Molecular cell biology*, **18**, 285-298.
- Barbour, R., et al. (2008) Red blood cells are the major source of alpha-synuclein in blood, *Neuro-degenerative diseases*, **5**, 55-59.
- Bastian, F., et al. (2008) Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. In Bairoch, A., Cohen-Boulakia, S. and Froidevaux, C. (eds), *Data Integration in the Life Sciences: 5th International Workshop, DILS 2008, Evry, France, June 25-27, 2008. Proceedings*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 124-131.
- Battle, C., et al. (2017) Characterization of Soft Amyloid Cores in Human Prion-Like Proteins, *Sci Rep*, **7**, 12134.
- Battle, C., et al. (2017) Perfecting prediction of mutational impact on the aggregation propensity of the ALS-associated hnRNPA2 prion-like protein, *FEBS letters*.
- Battle, C., et al. (2017) Prion-like proteins and their computational identification in proteomes, *Expert review of proteomics*, **14**, 335-350.
- Baxa, U., et al. (2007) Characterization of beta-sheet structure in Ure2p1-89 yeast prion fibrils by solid-state nuclear magnetic resonance, *Biochemistry*, **46**, 13149-13162.
- Boulay, G., et al. (2017) Cancer-Specific Retargeting of BAF Complexes by a Prion-like Domain, *Cell*, **171**, 163-178 e119.
- Buratti, E. and Baralle, F.E. (2008) Multiple roles of TDP-43 in gene expression, splicing regulation, and human disease, *Frontiers in bioscience : a journal and virtual library*, **13**, 867-878.
- Cascarina, S.M. and Ross, E.D. (2014) Yeast prions and human prion-like proteins: sequence features and prediction methods, *Cellular and molecular life sciences : CMLS*, **71**, 2047-2063.

Coletta, A., *et al.* (2010) Low-complexity regions within protein sequences have position-dependent roles, *BMC systems biology*, **4**, 43.

Couthouis, J., *et al.* (2012) Evaluating the role of the FUS/TLS-related gene EWSR1 in amyotrophic lateral sclerosis, *Hum Mol Genet*, **21**, 2899-2911.

Couthouis, J., *et al.* (2011) A yeast functional screen predicts new candidate ALS disease genes, *Proc Natl Acad Sci U S A*, **108**, 20881-20890.

Chakrabortee, S., *et al.* (2016) Luminidependens (LD) is an Arabidopsis protein with prion behavior, *Proc Natl Acad Sci U S A*, **113**, 6065-6070.

Cheng, W., *et al.* (2018) DDX5 RNA Helicases: Emerging Roles in Viral Infection, *International journal of molecular sciences*, **19**.

Cho, W.K., *et al.* (2018) Mediator and RNA polymerase II clusters associate in transcription-dependent condensates, *Science*, **361**, 412-415.

Duernberger, Y., *et al.* (2018) Prion replication in the mammalian cytosol: Functional regions within a prion domain driving induction, propagation and inheritance, *Mol Cell Biol*.

Espinosa Angarica, V., *et al.* (2014) PrionScan: an online database of predicted prion domains in complete proteomes, *BMC Genomics*, **15**, 102.

Espinosa Angarica, V., Ventura, S. and Sancho, J. (2013) Discovering putative prion sequences in complete proteomes using probabilistic representations of Q/N-rich domains, *BMC Genomics*, **14**, 316.

Finn, R.D., *et al.* (2016) The Pfam protein families database: towards a more sustainable future, *Nucleic Acids Res*, **44**, D279-285.

Franklin, B.S., *et al.* (2014) The adaptor ASC has extracellular and 'prionoid' activities that propagate inflammation, *Nature immunology*, **15**, 727-737.

Fuller-Pace, F.V. (2013) The DEAD box proteins DDX5 (p68) and DDX17 (p72): multi-tasking transcriptional regulators, *Biochim Biophys Acta*, **1829**, 756-763.

Gene Ontology, C. (2015) Gene Ontology Consortium: going forward, *Nucleic Acids Res*, **43**, D1049-1056.

Gitler, A.D. and Shorter, J. (2011) RNA-binding proteins with prion-like domains in ALS and FTL-D-U, *Prion*, **5**, 179-187.

Hafner-Bratkovic, I., *et al.* (2011) Globular domain of the prion protein needs to be unlocked by domain swapping to support prion protein conversion, *The Journal of biological chemistry*, **286**, 12149-12156.

Halfmann, R., *et al.* (2012) Prions are a common mechanism for phenotypic inheritance in wild yeasts, *Nature*, **482**, 363-368.

Halfmann, R. and Lindquist, S. (2010) Epigenetics in the extreme: prions and the inheritance of environmentally acquired traits, *Science*, **330**, 629-632.

Harrison, A.F. and Shorter, J. (2017) RNA-binding proteins with prion-like domains in health and disease, *The Biochemical journal*, **474**, 1417-1438.

Harrison, P.M. and Gerstein, M. (2003) A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes, *Genome Biol*, **4**, R40.

Harrison, P.M., Khachane, A. and Kumar, M. (2010) Genomic assessment of the evolution of the prion protein gene family in vertebrates, *Genomics*, **95**, 268-277.

He, Y. and Smith, R. (2009) Nuclear functions of heterogeneous nuclear ribonucleoproteins A/B, *Cellular and molecular life sciences : CMLS*, **66**, 1239-1256.

Hou, F., *et al.* (2011) MAVS forms functional prion-like aggregates to activate and propagate antiviral innate immune response, *Cell*, **146**, 448-461.

Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nature protocols*, **4**, 44-57.

Iglesias, V., de Groot, N.S. and Ventura, S. (2015) Computational analysis of candidate prion-like proteins in bacteria and their role, *Frontiers in microbiology*, **6**, 1123.

Jeong, H., *et al.* (2001) Lethality and centrality in protein networks, *Nature*, **411**, 41-42.

Ju, S., *et al.* (2011) A yeast model of FUS/TLS-dependent cytotoxicity, *PLoS biology*, **9**, e1001052.

Kataoka, K. and Mochizuki, K. (2017) Heterochromatin aggregation during DNA elimination in Tetrahymena is facilitated by a prion-like protein, *Journal of cell science*, **130**, 480-489.

Kim, H.J., *et al.* (2013) Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS, *Nature*, **495**, 467-473.

King, O.D., Gitler, A.D. and Shorter, J. (2012) The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease, *Brain Research*, **1462**, 61-80.

Lancaster, A.K., *et al.* (2014) PLAAC: a web and command-line application to identify proteins with Prion-Like Amino Acid Composition., *Bioinformatics (Oxford, England)*, **30**, 2-3.

Levesque, K., *et al.* (2006) Trafficking of HIV-1 RNA is mediated by heterogeneous nuclear ribonucleoprotein A2 expression and impacts on viral assembly, *Traffic*, **7**, 1177-1193.

Li, L. and Lindquist, S. (2000) Creating a protein-based element of inheritance, *Science*, **287**, 661-664.

Li, X., *et al.* (2014) Functional role of Tia1/Pub1 and Sup35 prion domains: directing protein synthesis machinery to the tubulin cytoskeleton, *Mol Cell*, **55**, 305-318.

Loomis, P.A., *et al.* (1990) Identification of nuclear tau isoforms in human neuroblastoma cells, *Proc Natl Acad Sci U S A*, **87**, 8422-8426.

Luk, K.C., *et al.* (2012) Pathological alpha-synuclein transmission initiates Parkinson-like neurodegeneration in nontransgenic mice, *Science*, **338**, 949-953.

Majumdar, A., *et al.* (2012) Critical role of amyloid-like oligomers of *Drosophila* Orb2 in the persistence of memory, *Cell*, **148**, 515-529.

Malinowska, L., *et al.* (2015) Dictyostelium discoideum has a highly Q/N-rich proteome and shows an unusual resilience to protein aggregation, *Proc Natl Acad Sci U S A*, **112**, E2620-2629.

March, Z.M., King, O.D. and Shorter, J. (2016) Prion-like domains as epigenetic regulators, scaffolds for subcellular organization, and drivers of neurodegenerative disease, *Brain Res*, **1647**, 9-18.

Mazurek, A., *et al.* (2012) DDX5 regulates DNA replication and is required for cell proliferation in a subset of breast cancer cells, *Cancer discovery*, **2**, 812-825.

McGlinchey, R.P., Kryndushkin, D. and Wickner, R.B. (2011) Suicidal [PSI⁺] is a lethal yeast prion, *Proc Natl Acad Sci U S A*, **108**, 5337-5341.

Medina, E., *et al.* (2016) Three-Dimensional Domain Swapping Changes the Folding Mechanism of the Forkhead Domain of FoxP1, *Biophysical journal*, **110**, 2349-2360.

Menche, J., *et al.* (2015) Disease networks. Uncovering disease-disease relationships through the incomplete interactome, *Science*, **347**, 1257601.

Michelitsch, M.D. and Weissman, J.S. (2000) A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions, *Proc Natl Acad Sci U S A*, **97**, 11910-11915.

Mosca, R., Ceol, A. and Aloy, P. (2013) Interactome3D: adding structural details to protein networks, *Nature methods*, **10**, 47-53.

Moy, R.H., *et al.* (2014) Stem-loop recognition by DDX17 facilitates miRNA processing and antiviral defense, *Cell*, **158**, 764-777.

Nakayashiki, T., *et al.* (2005) Yeast prions [URE3] and [PSI⁺] are diseases, *Proc Natl Acad Sci U S A*, **102**, 10575-10580.

Newby, G.A. and Lindquist, S. (2013) Blessings in disguise: biological benefits of prion-like mechanisms, *Trends Cell Biol*, **23**, 251-259.

Nomura, T., *et al.* (2014) Intranuclear aggregation of mutant FUS/TLS as a molecular pathomechanism of amyotrophic lateral sclerosis, *The Journal of biological chemistry*, **289**, 1192-1202.

Pallares, I., *et al.* (2018) Discovering Putative Prion-Like Proteins in *Plasmodium falciparum*: A Computational and Experimental Analysis, *Frontiers in microbiology*, **9**, 1737.

Pallares, I., Iglesias, V. and Ventura, S. (2015) The Rho Termination Factor of *Clostridium botulinum* Contains a Prion-Like Domain with a Highly Amyloidogenic Core, *Frontiers in microbiology*, **6**, 1516.

Patel, A., *et al.* (2015) A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation, *Cell*, **162**, 1066-1077.

Pinero, J., *et al.* (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, *Nucleic Acids Res*, **45**, D833-D839.

Pinero, J., *et al.* (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes, *Database : the journal of biological databases and curation*, **2015**, bav028.

Prusiner, S.B. (1982) Novel proteinaceous infectious particles cause scrapie, *Science*, **216**, 136-144.

Reddy, B.P., *et al.* (2015) A bioinformatic survey of RNA-binding proteins in *Plasmodium*, *BMC Genomics*, **16**, 890.

Sabate, R., *et al.* (2015) Amyloids or prions? That is the question, *Prion*, **9**, 200-206.

Sabate, R., *et al.* (2015) What Makes a Protein Sequence a Prion?, *PLoS Computational Biology*, **11**, e1004013.

Sabate, R., *et al.* (2015) What makes a protein sequence a prion?, *PLoS Comput Biol*, **11**, e1004013.

Sant'Anna, R., *et al.* (2016) Characterization of Amyloid Cores in Prion Domains, *Scientific Reports*, **6**, 34274.

Santoso, A., *et al.* (2000) Molecular basis of a yeast prion species barrier, *Cell*, **100**, 277-288.

Si, K. (2015) Prions: what are they good for?, *Annual review of cell and developmental biology*, **31**, 149-169.

Si, K., *et al.* (2010) Aplysia CPEB can form prion-like multimers in sensory neurons that contribute to long-term facilitation, *Cell*, **140**, 421-435.

Si, K. and Kandel, E.R. (2016) The Role of Functional Prion-Like Proteins in the Persistence of Memory, *Cold Spring Harbor perspectives in biology*, **8**, a021774.

Sikorska, B. and Liberski, P.P. (2012) Human prion diseases: from Kuru to variant Creutzfeldt-Jakob disease, *Subcellular biochemistry*, **65**, 457-496.

Singh, G.P., *et al.* (2004) Hyper-expansion of asparagines correlates with an abundance of proteins with prion-like domains in *Plasmodium falciparum*, *Molecular and biochemical parasitology*, **137**, 307-319.

Spillantini, M.G., *et al.* (1997) Alpha-synuclein in Lewy bodies, *Nature*, **388**, 839-840.

Stohr, J., *et al.* (2012) Purified and synthetic Alzheimer's amyloid beta (A β) prions, *Proc Natl Acad Sci U S A*, **109**, 11025-11030.

Tartaglia, G.G., *et al.* (2007) Life on the edge: a link between gene expression levels and aggregation rates of human proteins, *Trends Biochem Sci*, **32**, 204-206.

Toombs, J.A., *et al.* (2012) De novo design of synthetic prion domains, *Proc Natl Acad Sci U S A*, **109**, 6519-6524.

Uhlen, M., *et al.* (2015) Proteomics. Tissue-based map of the human proteome, *Science*, **347**, 1260419.

UniProt, C. (2015) UniProt: a hub for protein information, *Nucleic Acids Res*, **43**, D204-212.

van Rheede, T., *et al.* (2003) Molecular evolution of the mammalian prion protein, *Molecular biology and evolution*, **20**, 111-121.

Villarroya-Beltri, C., *et al.* (2013) Sumoylated hnRNP A2B1 controls the sorting of miRNAs into exosomes through binding to specific motifs, *Nature communications*, **4**, 2980.

Wang, I.F., *et al.* (2012) The self-interaction of native TDP-43 C terminus inhibits its degradation and contributes to early proteinopathies, *Nature communications*, **3**, 766.

- Wang, J., *et al.* (2018) A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins, *Cell*, **174**, 688-699 e616.
- Xu, H., *et al.* (2014) Structural basis for the prion-like MAVS filaments in antiviral innate immunity, *Elife*, **3**, e01489.
- Yuan, A.H. and Hochschild, A. (2017) A bacterial global regulator forms a prion, *Science*, **355**, 198-201.
- Zambrano, R., *et al.* (2015) PrionW: a server to identify proteins containing glutamine/asparagine rich prion-like domains and their amyloid cores, *Nucleic Acids Research*, 1-7.
- Zhu, X., *et al.* (2015) Mediator tail subunits can form amyloid-like aggregates in vivo and affect stress response in yeast, *Nucleic Acids Res*, **43**, 7306-7314.

7. Concluding Remarks

Chapter I – Globular protein aggregation

- Structural aggregation predictors are widely used to study the aggregation landscape of globular proteins or protein complexes. Since its publication, A3D has assisted a variety of research topics such as the study of disease and non-disease related proteins' aggregation propensities or to help in the design of biotechnological products.
- STAP has shown to be a very useful parameter for identifying aggregation prone surfaces in folded proteins or protein complexes. Additionally, stability has shown to play a major role in protein structural integrity. Taking these evidences into account we have updated A3D 2.0 with FoldX force field calculations to compute the effect of mutations on protein stability and how they might impact STAP.
- A3D 2.0 includes a dynamic mode which is able to model flexibility for big proteins or multimeric complexes. These transient conformations can conceal STAPs with high influence on the overall aggregation propensity. This was the case for most of the assayed complexes: 69 homodimers, 54 heterodimers and 60 antibodies. Therefore, this update provides a more precise view of protein aggregation landscape in real-life scenarios.
- Protein aggregation is an economic limitation for the development of protein-based products. We implemented a novel tool aimed to easily redesign protein solubility. Automated mutations widget identifies the most aggregation-prone regions and virtually mutates them to charged residues that will presumably act as gatekeepers and assesses the mutations' impact on solubility and stability. STAPs can be required for protein function; therefore we allow users to protect selected functional residues. The tool should allow users to obtain soluble, yet functional, variants of their proteins, as shown for the redesign of GFP/KKK.
- High throughput bioinformatic analyses increasingly rely on automated pipelines to process large amounts of data. To improve A3D's applicability, we have enabled A3D 2.0's fully access to functionalities via the command line through RESTful web services.
- A3D 2.0 improves *in situ* output data visualization with additions such as the possibility to tag certain amino acids, to take and store pictures of the protein, to compare A3D scores in interactive graphs or to visualise larger protein complexes. The aforementioned updates make A3D 2.0, a powerful application, intended to help in the analysis of pathogenic mutations in conformational disorders and in redesign of soluble proteins for biotechnological and biomedical applications in a cost and time-efficient manner.

Chapter II – Effect of pH in protein compaction

- The effect of protein environment on its aggregation dynamics has been hitherto neglected or at least only partially addressed. Previous research has shown pH affects amino acids at two main levels: changing their hydrophobicity and net charge. These properties modulate in turn important protein aspects such as folding and aggregation.
- IDPs are proteins that do not require a folded 3D structure to develop cellular functions. Their compositional bias allows preservation of their structural plasticity. IDPs lack structural constraints that could blur the effect of hydrophobicity and net charge changes on aggregation propensity. Therefore, IDPs constituted a good protein set to disentangle particular physicochemical contributions to aggregation. PNTs were a perfect starting point, as previous research from our collaborators had modelled how changes in net charge and pH could affect protein solubility.
- Plotting net charge and lipophilicity against experimental solubility revealed the dispersion resembled a plane; a flat 2D surface on the 3D plot.
- A combination of hydrophobicity and net charge can predict aggregation in disease-related IDPs. By modelling how these two physicochemical aspects change with pH, our approach showed robust enough to correctly forecast protein aggregation on a wide range of experiments and different aggregation reporters. Finally, this phenomenological approach was consistent enough to anticipate changes in protein aggregation on human and yeast functional amyloids.
- We developed SolupHred web server to incorporate the aforementioned calculations in a fast and easy to use way.
- SolupHred performs predictions in individual proteins and large datasets. Moreover, it allows users to select whether to predict aggregation propensity over a range of pHs or just at a specific pH.
- SolupHred web server was designed to be easy to use and its results as intuitive as possible. The output includes machine-interpretable JSON file to allow implementation of SolupHred calculations into bioinformatics pipelines.
- IDPs can undergo conditional folding. SolupHred is limited to proteins which remain disordered along the calculated pH range.
- IDPs are enriched in polar and ionizable amino acids compared to folded proteins. These compositional determinants allowed to accurately discriminate disordered and ordered proteins in a charge-hydrophobicity phase space diagram. As of today, several state-of-the-art prediction methods apply this principle to calculate protein disorder.
- Taking into account that protein disorder prediction could be anticipated by hydrophobicity and net charge, we revisited the C-H concept applying pH as an additional variable, to infer if we could model pH-driven order-disorder transitions.
- Available data of IDPs which underwent pH-conditional folding was used to model the influence of hydrophobicity and net charge in protein order. To obtain the maximal separation between the two populations, we applied a machine learning strategy specially designed to obtain the best binary separator (SVM). This strategy allowed us to identify a linear boundary condition similar to the one described for pH independent order-disorder transitions which is able to anticipate the effect of pH on IDPs conditional folding for diverse experiments from different authors.
- We developed DispHred, a first computational approach to predict protein disorder as a function of the pH. DispHred uses the pH dependent C-H plot analysis to discriminate between folded and disordered states in a user specified pH range or at a selected pH.
- DispHred also tackles SolupHred's most important limitation, as it can predict the range of pHs where the IDP will be disordered. As SolupHred, DispHred was designed to be user-friendly and of ease interpretation, allowing the incorporation of its calculations into computational pipelines.
- Protein environment is important for processes such as conditional disorder and protein aggregation. pH affects both hydrophobicity and net charge. Modelling how pH affects these two

physicochemical properties we were able to anticipate IDPs compaction and self-assembly. We next implemented the derived algorithms into publicly accessible web servers. These tools might help in further understanding the dynamic nature of IDPs, the mechanisms by which they convert into pathogenic forms, the design of synthetic IDPs which could transition at a certain pH, but also increase our understanding of conditional folding across species or the adaptations of organisms living under extreme pH conditions. We expect similar approximations to be incorporated into state-of-the art prediction methods in the following years, portraying them into more real-life scenarios.

Chapter III – Prediction of prion-like behaviour

- Research on yeast prions has allowed the identification and characterization of several proteins undergoing conformational conversion. Yeast prions have an intrinsic compositional bias: Q/N-rich PrD in disordered regions. Inside these domains, yeast prions have shown to present soft-amyloidogenic cores.
- Bioinformatic tools rely on these biases to identify prion-like proteins.
- We have developed PrionW, the first server to consider both the Q/N-rich composition bias in disordered regions and amyloidogenic propensity inside them. This approach outperformed previously available algorithms.
- PrionW server allows fast and accurate predictions and was intended to be useful for individual proteins and for large, proteomic-wide datasets.
- Different organisms have shown distinct compositional bias on their proteomes. This conundrum is addressed by PrionW. Users are allowed to tune the Q/N- richness of their query PrLD.
- A number of mutations mapped to PrLDs of human prion-like proteins have been shown to enhance their aggregation, which often results in the onset of degenerative disorders.
- Previous work assessed the impact of a large set of mutations (point and multiple mutations or deletions; natural and artificial) on the aggregation of the model ALS-associated prion-like hnRNPA2 protein. We showed their effect is best predicted by a function that takes into account both compositional features and amyloidogenic propensities.
- We have developed AMYCO (combined AMYloid and Composition based prediction of prion-like aggregation propensity) web server to implement the aforementioned function and perform automated and fast calculations on the aggregational impact of mutations on prion-like proteins.
- AMYCO is an intuitive web server able to assess specific user-defined mutations or predict the impact of all possible mutants at a specific position. The input screen, the output figures and tables were designed to be easily readable and interpretable. Moreover, AMYCO was configured to have its calculations portable to large scale bioinformatic pipelines, for which a machine-readable JSON file with all calculations can be retrieved.
- The linear function behind AMYCO was parametrized on a dataset of mutants for prion-like protein hnRNPA2. Further testing showed this algorithm was able to predict increase in aggregation propensity identified in disease-causing mutants of human prion-like proteins.
- The methodology showed robust enough to predict increase in aggregation propensity for mutations of yeast prion-like proteins that convert them into prions when expressed in yeast.
- All in all, the prediction accuracy achieved denotes that both a biased composition and a certain amyloidogenic propensity play a role in prion and prion-like conversion.
- Eventually, the progressive identification of novel proteins which undergo prion-like conversion will help decipher the underlying mechanism behind this transition.

Chapter IV – Characterization of prion-like proteins

- Prion-like proteins with aggregation potential are widespread in different kingdoms; which hints at a potentially conserved functionality. Each species faces different selective pressures which made them evolve particular compositional bias in their proteomes, for which different detection strategies are needed.
- Functional characterization of prion-like proteins in bacteria show these proteins mediate the cells' interaction with the environment, remodelling and nucleic acid-related processes. Overall, this suggests prion-like proteins could be a way to rapidly adapt to changing conditions.
- Bacterial prion-like proteins show similar modularity as yeast prions. A significant number of these proteins have one or multiple globular domains, with their PrLD be located at their N- or C-terminus.
- Pathogenic bacteria have significant more prion-like proteins than their non-pathogenic counterparts. These proteins could be linked to their pathogenicity and infectivity, as it is suggested from the analysis of *S. aureus* and other amyloid biofilm forming bacteria.
- These results preceded the recognition of Rho, a major transcription terminator factor in botulism-causing agent *C. botulinum*, as the first identified bacterial prion.
- *Plasmodium falciparum*, the species responsible for most cases of malaria in humans, has one of the most compositionally biased proteomes, with up to 30% of it being LC, especially rich in N. To counteract such an aggregation-prone proteome, the parasite has evolved efficient proteostatic systems. A more stringent methodology was used to analyse the presence of proteins which could display prion-like behaviour in the protozoan.
- The identified prion-like dataset was linked to functions previously seen in other species' prion-like subset, such as regulators of gene expression or nucleic acid binding proteins, but also in *Plasmodium f.* specific functions such as vesicle trafficking.
- We chose 3 soft-amyloid cores inside N-rich PrLDs representatives of these enriched functions. Our data provide compelling evidence that, all three candidate proteins contain short nucleating regions embedded in the PrLDs that *in vitro* are able to spontaneously self-assemble into amyloid-like structures. Finally, we tracked *in vivo* red blood cell-infecting trophozoite stage *P. falciparum* showing intracellular amyloid-like aggregates.
- In humans, a related approach allowed us to identify previous reported proteins that had experimentally shown to undergo prion-like transitions.
- Architecturally, human prion-like proteins have shown a similar modularity than for previous species; having their PrLD predominantly in the protein ends and accompanied by globular domains.
- As previously reported, prion-like proteins are linked to regulating gene transcription through binding nucleic acids. This function mirrors those found for other species' prion-like subsets; suggesting a possible conserved mechanism.
- Quite surprisingly, human prion like proteins were not restricted to any specific tissue but were found ubiquitously expressed among most cell types. Moreover, prion-like proteins were found strongly connected with disease. Taken both results together, we can anticipate prion-like proteins linked to non-neurological conditions will be soon identified.
- Human prion-like proteins establish highly interconnected networks in which they preferably interact between them. Importantly, functional analysis of this interactome reveals association with two previously undescribed set of diseases as cancer and viral infections.
- Overall, human prion-like proteins tend to be modular, interconnected, regulating gene transcription and its gain or loss-of-function can be directly or indirectly linked to diseases.
- All in all, prion-like proteins seem to act as a cellular tool to regulate gene expression (in multiple organisms); by taking advantage of its potential phenotypic conversion, as a fast response in front of changing conditions. The conformational switch would have an immediate effect on the nucleic acids they bind or regulate. This aspect seems clearer in organisms with higher degree of

annotation like human or yeast. Dysfunction of these proteins can originate protein aggregation, causing multiple pathologies either by the loss or the gain of function.

8. References

- Afsar Minhas, F.U., Ross, E.D. and Ben-Hur, A. (2017) Amino acid composition predicts prion activity, *PLoS Comput Biol*, **13**, e1005465.
- Aguzzi, A. and Altmeyer, M. (2016) Phase Separation: Linking Cellular Compartmentalization to Disease, *Trends Cell Biol*, **26**, 547-558.
- Aguzzi, A. and Calella, A.M. (2009) Prions: protein aggregation and infectious diseases., *Physiological reviews*, **89**, 1105-1152.
- Aguzzi, A., Heikenwalder, M. and Polymenidou, M. (2007) Insights into prion strains and neurotoxicity, *Nature reviews. Molecular cell biology*, **8**, 552-561.
- Aguzzi, A. and Rajendran, L. (2009) The transcellular spread of cytosolic amyloids, prions, and prionoids, *Neuron*, **64**, 783-790.
- Akter, R., et al. (2016) Islet Amyloid Polypeptide: Structure, Function, and Pathophysiology, *Journal of diabetes research*, **2016**, 2798269.
- Alberti, S. (2017) Phase separation in biology, *Current biology : CB*, **27**, R1097-R1102.
- Alberti, S., et al. (2009) A systematic survey identifies prions and illuminates sequence features of prionogenic proteins, *Cell*, **137**, 146-158.
- Alberti, S., et al. (2009) A systematic survey identifies prions and illuminates sequence features of prionogenic proteins., *Cell*, **137**, 146-158.
- Amberger, J.S., et al. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders, *Nucleic Acids Res*, **43**, D789-798.
- An, L. and Harrison, P.M. (2016) The evolutionary scope and neurological disease linkage of yeast-prion-like proteins in humans, *Biology direct*, **11**, 32.
- Anantharaman, V. and Aravind, L. (2003) Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes, *Genome Biol*, **4**, R11.
- Anderson, P., Kedersha, N. and Ivanov, P. (2015) Stress granules, P-bodies and cancer, *Biochim Biophys Acta*, **1849**, 861-870.
- Anfinsen, C.B. (1973) Principles that govern the folding of protein chains, *Science*, **181**, 223-230.
- Ansari, M.Z. and Swaminathan, R. (2020) Structure and dynamics at N- and C-terminal regions of intrinsically disordered human c-Myc PEST degron reveal a pH-induced transition, *Proteins*, **88**, 889-909.
- Antonets, K.S., et al. (2020) Accumulation of storage proteins in plant seeds is mediated by amyloid formation, *PLoS biology*, **18**, e3000564.
- Antoun, A., et al. (2003) The roles of initiation factor 2 and guanosine triphosphate in initiation of protein synthesis, *EMBO J*, **22**, 5593-5601.
- Aravind, L., et al. (2003) Plasmodium biology: genomic gleanings, *Cell*, **115**, 771-785.
- Ashe, K.H. and Aguzzi, A. (2012) Prions, prionoids and pathogenic proteins in Alzheimer disease, *Prion*, **7**.
- Auer, S., et al. (2008) Self-templated nucleation in peptide and protein aggregation, *Physical review letters*, **101**, 258101.
- Babu, M.M., et al. (2011) Intrinsically disordered proteins: regulation and disease, *Curr Opin Struct Biol*, **21**, 432-440.
- Balaji, S., et al. (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains, *Nucleic Acids Res*, **33**, 3994-4006.
- Balguerie, A., et al. (2003) Domain organization and structure-function relationship of the HET-s prion protein of *Podospora anserina*, *EMBO J*, **22**, 2071-2081.
- Banani, S.F., et al. (2017) Biomolecular condensates: organizers of cellular biochemistry, *Nature reviews. Molecular cell biology*, **18**, 285-298.
- Barbour, R., et al. (2008) Red blood cells are the major source of alpha-synuclein in blood, *Neuro-degenerative diseases*, **5**, 55-59.
- Bardwell, J.C. and Jakob, U. (2012) Conditional disorder in chaperone action, *Trends Biochem Sci*, **37**, 517-525.
- Bartlett, A.I. and Radford, S.E. (2009) An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms, *Nature structural & molecular biology*, **16**, 582-588.
- Bastian, F., et al. (2008) Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. In Bairoch, A., Cohen-Boulakia, S. and Froidevaux, C. (eds), *Data Integration in the Life Sciences: 5th International Workshop, DILS 2008, Evry, France, June 25-27, 2008. Proceedings*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 124-131.
- Bateman, A. and Bycroft, M. (2000) The structure of a LysM domain from *E. coli* membrane-bound lytic murein transglycosylase D (MltD), *J Mol Biol*, **299**, 1113-1119.
- Battle, C., et al. (2017a) Characterization of Soft Amyloid Cores in Human Prion-Like Proteins, *Sci Rep*, **7**, 12134.
- Battle, C., et al. (2017b) Perfecting prediction of mutational impact on the aggregation propensity of the ALS-associated hnRNPA2 prion-like protein, *FEBS letters*.
- Battle, C., et al. (2017c) Prion-like proteins and their computational identification in proteomes, *Expert review of proteomics*, **14**, 335-350.
- Baxa, U., et al. (2011) In Sup35p filaments (the [PSI⁺] prion), the globular C-terminal domains are widely offset from the amyloid fibril backbone, *Mol. Microbiol.*, **79**, 523-532.

Baxa, U., *et al.* (2007) Characterization of beta-sheet structure in Ure2p1-89 yeast prion fibrils by solid-state nuclear magnetic resonance, *Biochemistry*, **46**, 13149-13162.

Behnke, M.S., *et al.* (2010) Coordinated progression through two subtranscriptomes underlies the tachyzoite cycle of *Toxoplasma gondii*, *PLoS one*, **5**, e12354.

Beissbarth, T. and Speed, T.P. (2004) Gostat: find statistically overrepresented Gene Ontologies within a group of genes, *Bioinformatics*, **20**, 1464-1465.

Belli, M., Ramazzotti, M. and Chiti, F. (2011) Prediction of amyloid aggregation in vivo, *EMBO reports*, **12**, 657-663.

Bhandare, V.V. and Ramaswamy, A. (2018) The proteinopathy of D169G and K263E mutants at the RNA Recognition Motif (RRM) domain of tar DNA-binding protein (tdp43) causing neurological disorders: A computational study, *Journal of biomolecular structure & dynamics*, **36**, 1075-1093.

Blanco, L.P., *et al.* (2012) Diversity, biogenesis and function of microbial amyloids, *Trends in microbiology*, **20**, 66-73.

Boulay, G., *et al.* (2017) Cancer-Specific Retargeting of BAF Complexes by a Prion-like Domain, *Cell*, **171**, 163-178 e119.

Brookmeyer, R., *et al.* (2007) Forecasting the global burden of Alzheimer's disease, *Alzheimer's & dementia : the journal of the Alzheimer's Association*, **3**, 186-191.

Brown, J.C.S. and Lindquist, S. (2009) A heritable switch in carbon source utilization driven by an unusual yeast prion, *Genes & Development*, **23**, 2320-2332.

Bulawa, C.E., *et al.* (2012) Tafamidis, a potent and selective transthyretin kinetic stabilizer that inhibits the amyloid cascade, *Proc Natl Acad Sci U S A*, **109**, 9629-9634.

Buratti, E. and Baralle, F.E. (2008) Multiple roles of TDP-43 in gene expression, splicing regulation, and human disease, *Frontiers in bioscience : a journal and virtual library*, **13**, 867-878.

Cai, X., *et al.* (2014) Prion-like polymerization underlies signal transduction in antiviral immune defense and inflammasome activation, *Cell*, **156**, 1207-1222.

Camara-Almiron, J., *et al.* (2018) Beyond the expected: the structural and functional diversity of bacterial amyloids, *Critical reviews in microbiology*, **44**, 653-666.

Carr, C.M. and Kim, P.S. (1993) A spring-loaded mechanism for the conformational change of influenza hemagglutinin, *Cell*, **73**, 823-832.

Cascarina, S.M. and Ross, E.D. (2014) Yeast prions and human prion-like proteins: sequence features and prediction methods, *Cellular and molecular life sciences : CMLS*, **71**, 2047-2063.

Cascarina, S.M. and Ross, E.D. (2020) Natural and pathogenic protein sequence variation affecting prion-like domains within and across human proteomes, *BMC Genomics*, **21**, 23.

Castillo, V., *et al.* (2010) Deciphering the role of the thermodynamic and kinetic stabilities of SH3 domains on their aggregation inside bacteria, *Proteomics*, **10**, 4172-4185.

Castillo, V., *et al.* (2011) Prediction of the aggregation propensity of proteins from the primary sequence: aggregation properties of proteomes, *Biotechnology journal*, **6**, 674-685.

Castillo, V. and Ventura, S. (2009) Amyloidogenic regions and interaction surfaces overlap in globular proteins related to conformational diseases, *PLoS Comput Biol*, **5**, e1000476.

Coletta, A., *et al.* (2010) Low-complexity regions within protein sequences have position-dependent roles, *BMC systems biology*, **4**, 43.

Collaborators, G.B.D.P.s.D. (2018) Global, regional, and national burden of Parkinson's disease, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016, *The Lancet. Neurology*, **17**, 939-953.

Conchillo-Sole, O., *et al.* (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides, *BMC bioinformatics*, **8**, 65.

Contreras-Martel, C., *et al.* (2011) Structure-guided design of cell wall biosynthesis inhibitors that overcome beta-lactam resistance in *Staphylococcus aureus* (MRSA), *ACS Chem Biol*, **6**, 943-951.

Costa, D.C., *et al.* (2016) Aggregation and Prion-Like Properties of Misfolded Tumor Suppressors: Is Cancer a Prion Disease?, *Cold Spring Harbor perspectives in biology*, **8**.

Courtois, F., *et al.* (2016) Rational design of therapeutic mAbs against aggregation through protein engineering and incorporation of glycosylation motifs applied to bevacizumab, *mAbs*, **8**, 99-112.

Coustou, V., *et al.* (1997) The protein product of the het-s heterokaryon incompatibility gene of the fungus *Podospora anserina* behaves as a prion analog, *Proc Natl Acad Sci U S A*, **94**, 9773-9778.

Couthouis, J., *et al.* (2012) Evaluating the role of the FUS/TLS-related gene EWSR1 in amyotrophic lateral sclerosis, *Hum Mol Genet*, **21**, 2899-2911.

Couthouis, J., *et al.* (2011) A yeast functional screen predicts new candidate ALS disease genes, *Proc Natl Acad Sci U S A*, **108**, 20881-20890.

Cox, B.S. (1965) Ψ , A cytoplasmic suppressor of super-suppressor in yeast, *Heredity*, **20**, 505.

Cranmer, S.L., *et al.* (1997) An alternative to serum for cultivation of *Plasmodium falciparum* in vitro, *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **91**, 363-365.

Cromwell, M.E., Hilario, E. and Jacobson, F. (2006) Protein aggregation and bioprocessing, *The AAPS journal*, **8**, E572-579.

Crow, E.T., Du, Z. and Li, L. (2011) A small, glutamine-free domain propagates the [SWI(+)] prion in budding yeast, *Molecular and cellular biology*, **31**, 3436-3444.

Chakrabortee, S., *et al.* (2016) Luminidependens (LD) is an Arabidopsis protein with prion behavior, *Proc Natl Acad Sci U S A*, **113**, 6065-6070.

Chakravarty, A.K., *et al.* (2020) A Non-amyloid Prion Particle that Activates a Heritable Gene Expression Program, *Mol Cell*, **77**, 251-265 e259.

Chapman, M.R., *et al.* (2002) Role of Escherichia coli curli operons in directing amyloid fiber formation, *Science*, **295**, 851-855.

Chartier-Harlin, M.C., *et al.* (2004) Alpha-synuclein locus duplication as a cause of familial Parkinson's disease, *Lancet*, **364**, 1167-1169.

Chen, B., Newnam, G.P. and Chernoff, Y.O. (2007) Prion species barrier between the closely related yeast proteins is detected despite coaggregation, *Proc Natl Acad Sci U S A*, **104**, 2791-2796.

Chen, J. and Kriwacki, R.W. (2018) Intrinsically Disordered Proteins: Structure, Function and Therapeutics, *J Mol Biol*, **430**, 2275-2277.

Chen, Y. and Dokholyan, N.V. (2008) Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm, *Molecular biology and evolution*, **25**, 1530-1533.

Cheng, W., *et al.* (2018) DDX5 RNA Helicases: Emerging Roles in Viral Infection, *International journal of molecular sciences*, **19**.

Chennamsetty, N., *et al.* (2009) Design of therapeutic proteins with enhanced stability, *Proc Natl Acad Sci U S A*, **106**, 11937-11942.

Cheon, M., *et al.* (2007) Structural reorganisation and potential toxicity of oligomeric species formed during the assembly of amyloid fibrils, *PLoS Comput Biol*, **3**, 1727-1738.

Cherry, J.M., *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast, *Nucleic Acids Res*, **40**, D700-705.

Chicco, D. and Jurman, G. (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, *BMC Genomics*, **21**, 6.

Chien, P. and Weissman, J.S. (2001) Conformational diversity in a yeast prion dictates its seeding specificity, *Nature*, **410**, 223-227.

Chiti, F., *et al.* (2002) Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases, *Proc Natl Acad Sci U S A*, **99 Suppl 4**, 16419-16426.

Chiti, F. and Dobson, C.M. (2006) Protein misfolding, functional amyloid, and human disease, *Annu Rev Biochem*, **75**, 333-366.

Chiti, F. and Dobson, C.M. (2017) Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade, *Annu Rev Biochem*, **86**, 27-68.

Chiti, F., *et al.* (2002) Kinetic partitioning of protein folding and aggregation, *Nature structural biology*, **9**, 137-143.

Cho, W.K., *et al.* (2018) Mediator and RNA polymerase II clusters associate in transcription-dependent condensates, *Science*, **361**, 412-415.

Davenport, K.A., *et al.* (2015) Insights into Chronic Wasting Disease and Bovine Spongiform Encephalopathy Species Barriers by Use of Real-Time Conversion, *Journal of virology*, **89**, 9524-9531.

de Groot, N.S., *et al.* (2006) Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities, *FEBS J*, **273**, 658-668.

de Groot, N.S., *et al.* (2012) AGGRESKAN: method, application, and perspectives for drug design, *Methods in molecular biology*, **819**, 199-220.

de Oliveira, G.A.P., *et al.* (2020) The Status of p53 Oligomeric and Aggregation States in Cancer, *Biomolecules*, **10**.

Denroche, H.C. and Verchere, C.B. (2018) IAPP and type 1 diabetes: implications for immunity, metabolism and islet transplants, *Journal of molecular endocrinology*, **60**, R57-R75.

DePas, W.H. and Chapman, M.R. (2012) Microbial manipulation of the amyloid fold, *Research in microbiology*, **163**, 592-606.

DePristo, M.A., Zilversmit, M.M. and Hartl, D.L. (2006) On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins, *Gene*, **378**, 19-30.

Diaz-Caballero, M., *et al.* (2018) Prion-based nanomaterials and their emerging applications, *Prion*, **12**, 266-272.

Diaz-Caballero, M., *et al.* (2018) Minimalist Prion-Inspired Polar Self-Assembling Peptides, *ACS nano*, **12**, 5394-5407.

Dinkel, P.D., *et al.* (2011) Variations in filament conformation dictate seeding barrier between three- and four-repeat tau, *Biochemistry*, **50**, 4330-4336.

Dobson, C.M. (2003) Protein folding and misfolding, *Nature*, **426**, 884-890.

Dobson, C.M., Sali, A. and Karplus, M. (1998) Protein Folding: A Perspective from Theory and Experiment, *Angewandte Chemie*, **37**, 868-893.

Dorsman, J.C., *et al.* (2002) Strong aggregation and increased toxicity of poly-leucine over poly-glutamine stretches in mammalian cells, *Hum Mol Genet*, **11**, 1487-1496.

Dosztanyi, Z. (2018) Prediction of protein disorder based on IUPred, *Protein Sci*, **27**, 331-340.

DuBay, K.F., *et al.* (2004) Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains, *Journal of molecular biology*, **341**, 1317-1326.

Dudgeon, K., *et al.* (2012) General strategy for the generation of human antibody variable domains with increased aggregation resistance, *Proc Natl Acad Sci U S A*, **109**, 10879-10884.

Duernberger, Y., *et al.* (2018) Prion replication in the mammalian cytosol: Functional regions within a prion domain driving induction, propagation and inheritance, *Mol Cell Biol*.

Dunker, A.K. and Obradovic, Z. (2001) The protein trinity--linking function and disorder, *Nat Biotechnol*, **19**, 805-806.

Dyson, H.J. (2016) Making Sense of Intrinsically Disordered Proteins, *Biophys J*, **110**, 1013-1016.

Eichinger, L., *et al.* (2005) The genome of the social amoeba Dictyostelium discoideum, *Nature*, **435**, 43-57.

Eisenberg, D. and Jucker, M. (2012) The amyloid state of proteins in human diseases, *Cell*, **148**, 1188-1203.

Ekman, D., *et al.* (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions, *J Mol Biol*, **348**, 231-243.

Eliezer, D., *et al.* (2005) Residual structure in the repeat domain of tau: echoes of microtubule binding and paired helical filament formation, *Biochemistry*, **44**, 1026-1036.

Emamzadeh, F.N. (2016) Alpha-synuclein structure, functions, and interactions, *Journal of research in medical sciences : the official journal of Isfahan University of Medical Sciences*, **21**, 29.

Emily, M., Talvas, A. and Delamarche, C. (2013) MetAmyl: a METa-predictor for AMYLoid proteins, *PLoS One*, **8**, e79722.

Espargaro, A., *et al.* (2008) The in vivo and in vitro aggregation properties of globular proteins correlate with their conformational stability: the SH3 case, *Journal of molecular biology*, **378**, 1116-1131.

Espargaro, A., *et al.* (2012) Yeast prions form infectious amyloid inclusion bodies in bacteria, *Microbial cell factories*, **11**, 89.

Espinosa Angarica, V., *et al.* (2014) PrionScan: an online database of predicted prion domains in complete proteomes, *BMC Genomics*, **15**, 102.

Espinosa Angarica, V., Ventura, S. and Sancho, J. (2013) Discovering putative prion sequences in complete proteomes using probabilistic representations of Q/N-rich domains, *BMC Genomics*, **14**, 316.

Evans, M.L., *et al.* (2015) The bacterial curli system possesses a potent and selective inhibitor of amyloid formation, *Mol Cell*, **57**, 445-455.

Familia, C., *et al.* (2015) Prediction of Peptide and Protein Propensity for Amyloid Formation, *PLoS one*, **10**, e0134679.

Fandrich, M. and Dobson, C.M. (2002) The behaviour of polyamino acids reveals an inverse side chain effect in amyloid structure formation, *EMBO J*, **21**, 5682-5690.

Faux, N.G., *et al.* (2005) Functional insights from the distribution and role of homopeptide repeat-containing proteins, *Genome Res*, **15**, 537-551.

Fernandez-Escamilla, A.M., *et al.* (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins, *Nature biotechnology*, **22**, 1302-1306.

Fernandez, M.R., *et al.* (2017) Amyloid cores in prion domains: Key regulators for prion conformational conversion, *Prion*, **11**, 31-39.

Fink, A.L. (1998) Protein aggregation: folding aggregates, inclusion bodies and amyloid, *Folding & design*, **3**, R9-23.

Finn, R.D., *et al.* (2014) Pfam: the protein families database, *Nucleic Acids Res*, **42**, D222-230.

Finn, R.D., *et al.* (2016) The Pfam protein families database: towards a more sustainable future, *Nucleic Acids Res*, **44**, D279-285.

Flemming, H.C. and Wingender, J. (2010) The biofilm matrix, *Nature reviews. Microbiology*, **8**, 623-633.

Floege, J. and Ketteler, M. (2001) beta2-microglobulin-derived amyloidosis: an update, *Kidney international. Supplement*, **78**, S164-171.

Fonin, A.V., *et al.* (2019) Folding of poly-amino acids and intrinsically disordered proteins in overcrowded milieu induced by pH change, *Int J Biol Macromol*, **125**, 244-255.

Fowler, D.M., *et al.* (2006) Functional amyloid formation within mammalian tissue, *PLoS biology*, **4**, e6.

Fowler, D.M., *et al.* (2007) Functional amyloid--from bacteria to humans, *Trends Biochem Sci*, **32**, 217-224.

Franklin, B.S., *et al.* (2014) The adaptor ASC has extracellular and 'prionoid' activities that propagate inflammation, *Nature immunology*, **15**, 727-737.

Franzmann, T.M., *et al.* (2018) Phase separation of a yeast prion protein promotes cellular fitness, *Science*, **359**.

Franzosa, E.A. and Xia, Y. (2009) Structural determinants of protein evolution are context-sensitive at the residue level, *Molecular biology and evolution*, **26**, 2387-2395.

Friedland, R.P. (2015) Mechanisms of molecular mimicry involving the microbiota in neurodegeneration, *J Alzheimers Dis*, **45**, 349-362.

Fuller-Pace, F.V. (2013) The DEAD box proteins DDX5 (p68) and DDX17 (p72): multi-tasking transcriptional regulators, *Biochim Biophys Acta*, **1829**, 756-763.

Fuxreiter, M. (2012) Fuzziness: linking regulation to protein dynamics, *Mol Biosyst*, **8**, 168-177.

Fuxreiter, M. and Tompa, P. (2012) Fuzzy complexes: a more stochastic view of protein function, *Adv Exp Med Biol*, **725**, 1-14.

Galzitskaya, O.V., Garbuzynskiy, S.O. and Lobanov, M.Y. (2006) FoldUnfold: web server for the prediction of disordered regions in protein chain, *Bioinformatics*, **22**, 2948-2949.

Gallo, P.M., *et al.* (2015) Amyloid-DNA Composites of Bacterial Biofilms Stimulate Autoimmunity, *Immunity*, **42**, 1171-1184.

Gamez, J., *et al.* (2019) Transthyretin stabilization activity of the catechol-O-methyltransferase inhibitor tolcapone (SOM0226) in hereditary ATTR amyloidosis patients and asymptomatic carriers: proof-of-concept study(), *Amyloid : the international journal of experimental and clinical investigation : the official journal of the International Society of Amyloidosis*, **26**, 74-84.

Garbuzynskiy, S.O., Lobanov, M.Y. and Galzitskaya, O.V. (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence, *Bioinformatics*, **26**, 326-332.

Gardner, M.J., *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*, *Nature*, **419**, 498-511.

Garner, E., *et al.* (1999) Predicting Binding Regions within Disordered Proteins, *Genome Inform Ser Workshop Genome Inform*, **10**, 41-50.

Garrity, S.J., *et al.* (2010) Conversion of a yeast prion protein to an infectious form in bacteria, *Proc Natl Acad Sci U S A*, **107**, 10596-10601.

Gasior, P. and Kotulska, M. (2014) FISH Amyloid - a new method for finding amyloidogenic segments in proteins based on site specific co-occurrence of aminoacids, *BMC bioinformatics*, **15**, 54.

Gene Ontology, C. (2015) Gene Ontology Consortium: going forward, *Nucleic Acids Res*, **43**, D1049-1056.

Gil-Garcia, M., *et al.* (2018) Combining structural aggregation propensity and stability predictions to re-design protein solubility, *Molecular pharmaceuticals*.

Gitler, A.D. and Shorter, J. (2011) RNA-binding proteins with prion-like domains in ALS and FTL-D, *Prion*, **5**, 179-187.

Glover, J.R., *et al.* (1997) Self-seeded fibers formed by Sup35, the protein determinant of [PSI⁺], a heritable prion-like factor of *S. cerevisiae*, *Cell*, **89**, 811-819.

Goedert, M., *et al.* (2013) 100 years of Lewy pathology, *Nature reviews. Neurology*, **9**, 13-24.

Graether, S.P., Slupsky, C.M. and Sykes, B.D. (2003) Freezing of a fish antifreeze protein results in amyloid fibril formation, *Biophys J*, **84**, 552-557.

Grana-Montes, R., Sant'anna de Oliveira, R. and Ventura, S. (2012) Protein aggregation profile of the human kinome, *Frontiers in physiology*, **3**, 438.

Graña-Montes, R., *et al.* (2017) Prediction of Protein Aggregation and Amyloid Formation. In Rigden, D.J. (ed), *From Protein Structure to Function with Bioinformatics*. Springer, pp. 205-263.

Greenwald, J. and Riek, R. (2010) Biology of amyloid: structure, function, and regulation, *Structure*, **18**, 1244-1260.

Gridelli, C., *et al.* (2018) Safety and Efficacy of Bevacizumab Plus Standard-of-Care Treatment Beyond Disease Progression in Patients With Advanced Non-Small Cell Lung Cancer: The AvaALL Randomized Clinical Trial, *JAMA oncology*, **4**, e183486.

Gsponer, J. and Babu, M.M. (2012) Cellular strategies for regulating functional and nonfunctional protein aggregation, *Cell reports*, **2**, 1425-1437.

Guerois, R., Nielsen, J.E. and Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations, *Journal of molecular biology*, **320**, 369-387.

Guy, H.R. (1985) Amino acid side-chain partition energies and distribution of residues in soluble proteins, *Biophys J*, **47**, 61-70.

Hafner-Bratkovic, I., *et al.* (2011) Globular domain of the prion protein needs to be unlocked by domain swapping to support prion protein conversion, *The Journal of biological chemistry*, **286**, 12149-12156.

Haider, A., *et al.* (2015) Targeting and function of proteins mediating translation initiation in organelles of *Plasmodium falciparum*, *Molecular microbiology*, **96**, 796-814.

Halfmann, R., *et al.* (2011) Opposing effects of glutamine and asparagine govern prion formation by intrinsically disordered proteins, *Mol Cell*, **43**, 72-84.

Halfmann, R., *et al.* (2012) Prions are a common mechanism for phenotypic inheritance in wild yeasts, *Nature*, **482**, 363-368.

Halfmann, R. and Lindquist, S. (2010) Epigenetics in the extreme: prions and the inheritance of environmentally acquired traits, *Science*, **330**, 629-632.

Halfmann, R., *et al.* (2012) Prion formation by a yeast GLFG nucleoporin, <http://dx.doi.org/10.4161/pri.20199>.

Hamrang, Z., Rattray, N.J. and Pluen, A. (2013) Proteins behaving badly: emerging technologies in profiling biopharmaceutical aggregation, *Trends in biotechnology*, **31**, 448-458.

Han, T.W., *et al.* (2012) Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies, *Cell*, **149**, 768-779.

Harmon, T.S., *et al.* (2016) GADIS: Algorithm for designing sequences to achieve target secondary structure profiles of intrinsically disordered proteins, *Protein Eng Des Sel*, **29**, 339-346.

Harrison, A.F. and Shorter, J. (2017) RNA-binding proteins with prion-like domains in health and disease, *The Biochemical journal*, **474**, 1417-1438.

Harrison, P.M. and Gerstein, M. (2003) A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes, *Genome Biol*, **4**, R40.

Harrison, P.M., Khachane, A. and Kumar, M. (2010) Genomic assessment of the evolution of the prion protein gene family in vertebrates, *Genomics*, **95**, 268-277.

Harvey, Z.H., *et al.* (2020) A Prion Epigenetic Switch Establishes an Active Chromatin State, *Cell*, **180**, 928-940 e914.

Hatos, A., *et al.* (2020) DisProt: intrinsic protein disorder annotation in 2020, *Nucleic Acids Res*, **48**, D269-D276.

He, B., *et al.* (2009) Predicting intrinsic disorder in proteins: an overview, *Cell Res*, **19**, 929-949.

He, Y. and Smith, R. (2009) Nuclear functions of heterogeneous nuclear ribonucleoproteins A/B, *Cellular and molecular life sciences : CMLS*, **66**, 1239-1256.

Heinrich, S.U. and Lindquist, S. (2011) Protein-only mechanism induces self-perpetuating changes in the activity of neuronal Aplysia cytoplasmic polyadenylation element binding protein (CPEB), *Proc Natl Acad Sci U S A*, **108**, 2999-3004.

Hill, J.M. and Lukiw, W.J. (2015) Microbial-generated amyloids and Alzheimer's disease (AD), *Front Aging Neurosci*, **7**, 9.

Hiller, N.L., *et al.* (2004) A host-targeting signal in virulence proteins reveals a secretome in malarial infection, *Science*, **306**, 1934-1937.

Hortschansky, P., *et al.* (2005) The aggregation kinetics of Alzheimer's beta-amyloid peptide is controlled by stochastic nucleation, *Protein science : a publication of the Protein Society*, **14**, 1753-1759.

Hosoda, N., *et al.* (2003) Translation Termination Factor eRF3 Mediates mRNA Decay through the Regulation of Deadenylation, *Journal of Biological Chemistry*, **278**, 38287-38291.

Hou, F., *et al.* (2011) MAVS forms functional prion-like aggregates to activate and propagate antiviral innate immune response, *Cell*, **146**, 448-461.

Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nature protocols*, **4**, 44-57.

Huang, F., *et al.* (2014) Improving protein order-disorder classification using charge-hydropathy plots, *BMC Bioinformatics*, **15 Suppl 17**, S4.

Hufnagel, D.A., Tukul, C. and Chapman, M.R. (2013) Disease to dirt: the biology of microbial amyloids, *PLoS Pathog*, **9**, e1003740.

Hunter, J. (2007) Matplotlib: A 2D Graphics Environment, *Computing in Science & Engineering*, **9**, 90-95.

Iakoucheva, L.M., *et al.* (2004) The importance of intrinsic disorder for protein phosphorylation, *Nucleic Acids Res*, **32**, 1037-1049.

Iconomidou, V.A., Vriend, G. and Hamodrakas, S.J. (2000) Amyloids protect the silkworm oocyte and embryo, *FEBS letters*, **479**, 141-145.

Iglesias, V., de Groot, N.S. and Ventura, S. (2015) Computational analysis of candidate prion-like proteins in bacteria and their role, *Frontiers in microbiology*, **6**, 1123.

Invernizzi, G., *et al.* (2012) Protein aggregation: mechanisms and functional consequences, *The international journal of biochemistry & cell biology*, **44**, 1541-1554.

Itakura, A.K., *et al.* (2020) Widespread Prion-Based Control of Growth and Differentiation Strategies in *Saccharomyces cerevisiae*, *Mol Cell*, **77**, 266-278 e266.

Jackson, M.P. and Hewitt, E.W. (2017) Why are Functional Amyloids Non-Toxic in Humans?, *Biomolecules*, **7**.

Jahn, T.R. and Radford, S.E. (2005) The Yin and Yang of protein folding, *FEBS J*, **272**, 5962-5970.

Jahn, T.R. and Radford, S.E. (2008) Folding versus aggregation: polypeptide conformations on competing pathways, *Archives of biochemistry and biophysics*, **469**, 100-117.

Jakob, U., Kriwacki, R. and Uversky, V.N. (2014) Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function, *Chem Rev*, **114**, 6779-6805.

Jamroz, M., Kolinski, A. and Kmiecik, S. (2013) CABS-flex: Server for fast simulation of protein structure fluctuations, *Nucleic Acids Res*, **41**, W427-431.

Jamroz, M., Kolinski, A. and Kmiecik, S. (2014) CABS-flex predictions of protein flexibility compared with NMR ensembles, *Bioinformatics*, **30**, 2150-2154.

Jamroz, M., *et al.* (2013) Consistent View of Protein Fluctuations from All-Atom Molecular Dynamics and Coarse-Grained Dynamics with Knowledge-Based Force-Field, *Journal of chemical theory and computation*, **9**, 119-125.

Janin, J. and Wodak, S.J. (1983) Structural domains in proteins and their role in the dynamics of protein function, *Progress in biophysics and molecular biology*, **42**, 21-78.

Jeganathan, S., *et al.* (2008) The natively unfolded character of tau and its aggregation to Alzheimer-like paired helical filaments, *Biochemistry*, **47**, 10526-10539.

Jeong, H., *et al.* (2001) Lethality and centrality in protein networks, *Nature*, **411**, 41-42.

Jha, R.K., *et al.* (2010) Computational design of a PAK1 binding protein, *Journal of molecular biology*, **400**, 257-270.

Jha, S., *et al.* (2014) pH dependence of amylin fibrillization, *Biochemistry*, **53**, 300-310.

Jiang, H., *et al.* (2015) Phase transition of spindle-associated protein regulate spindle apparatus assembly, *Cell*, **163**, 108-122.

Johansson, J., *et al.* (1998) Conformation-dependent antibacterial activity of the naturally occurring human peptide LL-37, *J Biol Chem*, **273**, 3718-3724.

Ju, S., *et al.* (2011) A yeast model of FUS/TLS-dependent cytotoxicity, *PLoS biology*, **9**, e1001052.

Kahrstrom, C.T. (2012) Parasite physiology: Plasmodium gets the PK4 blood test, *Nat Rev Microbiol*, **10**, 237.

Karran, E., Mercken, M. and De Strooper, B. (2011) The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics, *Nat Rev Drug Discov*, **10**, 698-712.

Kataoka, K. and Mochizuki, K. (2017) Heterochromatin aggregation during DNA elimination in *Tetrahymena* is facilitated by a prion-like protein, *Journal of cell science*, **130**, 480-489.

Katina, N.S., *et al.* (2017) sw ApoMb Amyloid Aggregation under Nondenaturing Conditions: The Role of Native Structure Stability, *Biophysical journal*, **113**, 991-1001.

Kato, M., *et al.* (2012) Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels, *Cell*, **149**, 753-767.

Kaufman, R.J., *et al.* (2002) The unfolded protein response in nutrient sensing and differentiation, *Nature reviews. Molecular cell biology*, **3**, 411-421.

Kauzmann, W. (1959) Some factors in the interpretation of protein denaturation, *Advances in protein chemistry*, **14**, 1-63.

Khemtemourian, L., *et al.* (2011) Low pH acts as inhibitor of membrane damage induced by human islet amyloid polypeptide, *Journal of the American Chemical Society*, **133**, 15598-15604.

Khurana, V. and Lindquist, S. (2010) Modelling neurodegeneration in *Saccharomyces cerevisiae*: why cook with baker's yeast?, *Nature reviews. Neuroscience*, **11**, 436-449.

Kim, C., *et al.* (2009) NetCSSP: web application for predicting chameleon sequences and amyloid fibril formation, *Nucleic Acids Res*, **37**, W469-473.

Kim, H.J., *et al.* (2013) Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS, *Nature*, **495**, 467-473.

Kim, Y.E., *et al.* (2013) Molecular chaperone functions in protein folding and proteostasis, *Annu Rev Biochem*, **82**, 323-355.

King, O.D., Gitler, A.D. and Shorter, J. (2012) The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease, *Brain Research*, **1462**, 61-80.

Knowles, T.P. and Mezzenga, R. (2016) Amyloid Fibrils as Building Blocks for Natural and Artificial Functional Materials, *Advanced materials*, **28**, 6546-6561.

Knowles, T.P., Vendruscolo, M. and Dobson, C.M. (2014) The amyloid state and its association with protein misfolding diseases, *Nature reviews. Molecular cell biology*, **15**, 384-396.

Kramer, S. (2014) RNA in development: how ribonucleoprotein granules regulate the life cycles of pathogenic protozoa, *Wiley interdisciplinary reviews. RNA*, **5**, 263-284.

Kraus, A., Groveman, B.R. and Caughey, B. (2013) Prions and the potential transmissibility of protein misfolding diseases., *Annual review of microbiology*, **67**, 543-564.

Kulkarni, V. and Kulkarni, P. (2019) Intrinsically disordered proteins and phenotypic switching: Implications in cancer, *Prog Mol Biol Transl Sci*, **166**, 63-84.

Kurcinski, M., *et al.* (2018) CABS-flex standalone: a simulation environment for fast modeling of protein flexibility, *Bioinformatics*.

Kuriata, A., *et al.* (2018) CABS-flex 2.0: a web server for fast simulations of flexibility of protein structures, *Nucleic Acids Res*, **46**, W338-W343.

Kuriata, A., *et al.* (2019) Aggrescan3D standalone package for structure-based prediction of protein aggregation properties, *Bioinformatics*, **35**, 3834-3835.

Kuriata, A., *et al.* (2019) Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility, *Nucleic Acids Res*, **47**, W300-W307.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein, *J Mol Biol*, **157**, 105-132.

Lacroix, E., Viguera, A.R. and Serrano, L. (1998) Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters, *Journal of molecular biology*, **284**, 173-191.

Lambros, C. and Vanderberg, J.P. (1979) Synchronization of Plasmodium falciparum erythrocytic stages in culture, *The Journal of parasitology*, **65**, 418-420.

Lancaster, A.K., *et al.* (2010) The spontaneous appearance rate of the yeast prion [PSI⁺] and its implications for the evolution of the evolvability properties of the [PSI⁺] system., *Genetics*, **184**, 393-400.

Lancaster, A.K., *et al.* (2014) PLAAC: a web and command-line application to identify proteins with Prion-Like Amino Acid Composition., *Bioinformatics (Oxford, England)*, **30**, 2-3.

Lane, C.A., Hardy, J. and Schott, J.M. (2018) Alzheimer's disease, *European journal of neurology*, **25**, 59-70.

Lashuel, H.A., *et al.* (2013) The many faces of alpha-synuclein: from structure and toxicity to therapeutic target, *Nature reviews. Neuroscience*, **14**, 38-48.

Lassen, L.B., *et al.* (2016) Protein Partners of alpha-Synuclein in Health and Disease, *Brain pathology*, **26**, 389-397.

Le Roch, K.G., *et al.* (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle, *Science*, **301**, 1503-1508.

Lee, M.C., *et al.* (2008) Plasmodium falciparum Sec24 marks transitional ER that exports a model cargo via a diacidic motif, *Molecular microbiology*, **68**, 1535-1546.

Lehninger, A.L., Nelson, D.L. and Cox, M.M. (2005) *Lehninger Principles of Biochemistry*. W. H. Freeman.

Levesque, K., *et al.* (2006) Trafficking of HIV-1 RNA is mediated by heterogeneous nuclear ribonucleoprotein A2 expression and impacts on viral assembly, *Traffic*, **7**, 1177-1193.

Levinthal, C. (1969) How to fold graciously.

Li, C., Adamcik, J. and Mezzenga, R. (2012) Biodegradable nanocomposites of amyloid fibrils and graphene with shape-memory and enzyme-sensing properties, *Nature nanotechnology*, **7**, 421-427.

Li, L. and Lindquist, S. (2000) Creating a protein-based element of inheritance, *Science*, **287**, 661-664.

Li, X., *et al.* (2014) Functional role of Tia1/Pub1 and Sup35 prion domains: directing protein synthesis machinery to the tubulin cytoskeleton, *Mol Cell*, **55**, 305-318.

Liebman, S.W. and Chernoff, Y.O. (2012) Prions in yeast, *Genetics*, **191**, 1041-1072.

Lieutaud, P., *et al.* (2016) How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe, *Intrinsically Disord Proteins*, **4**, e1259708.

Lim, W.A. and Sauer, R.T. (1991) The role of internal packing interactions in determining the structure and stability of a protein, *J Mol Biol*, **219**, 359-376.

Lin, J.J., *et al.* (2000) Stability of human serum albumin during bioprocessing: denaturation and aggregation during processing of albumin paste, *Pharmaceutical research*, **17**, 391-396.

Linding, R., *et al.* (2003) Protein disorder prediction: implications for structural proteomics, *Structure*, **11**, 1453-1459.

Linding, R., *et al.* (2004) A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins, *Journal of molecular biology*, **342**, 345-353.

Lindorff-Larsen, K., *et al.* (2005) Protein folding and the organization of the protein topology universe, *Trends Biochem Sci*, **30**, 13-19.

Liu, J.J., Sondheimer, N. and Lindquist, S.L. (2002) Changes in the middle region of Sup35 profoundly alter the nature of epigenetic inheritance for the yeast prion [PSI⁺], *Proc Natl Acad Sci U S A*, **99 Suppl 4**, 16446-16453.

Liu, S., *et al.* (2017) Prions on the run: How extracellular vesicles serve as delivery vehicles for self-templating protein aggregates, *Prion*, **11**, 98-112.

Loomis, P.A., *et al.* (1990) Identification of nuclear tau isoforms in human neuroblastoma cells, *Proc Natl Acad Sci U S A*, **87**, 8422-8426.

Lopez de la Paz, M. and Serrano, L. (2004) Sequence determinants of amyloid fibril formation, *Proc Natl Acad Sci U S A*, **101**, 87-92.

Loquet, A., Saupe, S.J. and Romero, D. (2018) Functional Amyloids in Health and Disease, *Journal of molecular biology*, **430**, 3629-3630.

Luk, K.C., *et al.* (2012) Pathological alpha-synuclein transmission initiates Parkinson-like neurodegeneration in nontransgenic mice, *Science*, **338**, 949-953.

Lundmark, K., *et al.* (2005) Protein fibrils in nature can enhance amyloid protein A amyloidosis in mice: Cross-seeding as a disease mechanism, *Proc Natl Acad Sci U S A*, **102**, 6098-6102.

MacCallum, J.L. and Tieleman, D.P. (2011) Hydrophobicity scales: a thermodynamic looking glass into lipid-protein interactions, *Trends Biochem Sci*, **36**, 653-662.

Macheboeuf, P., *et al.* (2005) Active site restructuring regulates ligand recognition in class A penicillin-binding proteins, *Proc Natl Acad Sci U S A*, **102**, 577-582.

Maji, S.K., *et al.* (2009) Functional amyloids as natural storage of peptide hormones in pituitary secretory granules, *Science*, **325**, 328-332.

Majumdar, A., *et al.* (2012) Critical role of amyloid-like oligomers of *Drosophila* Orb2 in the persistence of memory, *Cell*, **148**, 515-529.

Malinowska, L., Kroschwald, S. and Alberti, S. (2013) Protein disorder, prion propensities, and self-organizing macromolecular collectives, *Biochim Biophys Acta*, **1834**, 918-931.

Malinowska, L., *et al.* (2015) Dictyostelium discoideum has a highly Q/N-rich proteome and shows an unusual resilience to protein aggregation, *Proc Natl Acad Sci U S A*, **112**, E2620-2629.

March, Z.M., King, O.D. and Shorter, J. (2016) Prion-like domains as epigenetic regulators, scaffolds for subcellular organization, and drivers of neurodegenerative disease, *Brain Res*, **1647**, 9-18.

Marti, M., *et al.* (2005) Signal-mediated export of proteins from the malaria parasite to the host erythrocyte, *The Journal of cell biology*, **171**, 587-592.

Masison, D.C., Maddelein, M.L. and Wickner, R.B. (1997) The prion model for [URE3] of yeast: spontaneous generation and requirements for propagation, *Proc Natl Acad Sci U S A*, **94**, 12503-12508.

Masison, D.C. and Wickner, R.B. (1995) Prion-inducing domain of yeast Ure2p and protease resistance of Ure2p in prion-containing cells, *Science*, **270**, 93-95.

Maurer-Stroh, S., *et al.* (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices, *Nat. Meth.*, **7**, 237-242.

Mazurek, A., *et al.* (2012) DDX5 regulates DNA replication and is required for cell proliferation in a subset of breast cancer cells, *Cancer discovery*, **2**, 812-825.

McGlinchey, R.P., Kryndushkin, D. and Wickner, R.B. (2011) Suicidal [PSI⁺] is a lethal yeast prion, *Proc Natl Acad Sci U S A*, **108**, 5337-5341.

McGlinchey, R.P. and Lee, J.C. (2018) Why Study Functional Amyloids? Lessons from the Repeat Domain of Pmel17, *Journal of molecular biology*, **430**, 3696-3706.

Medina, E., *et al.* (2016) Three-Dimensional Domain Swapping Changes the Folding Mechanism of the Forkhead Domain of FoxP1, *Biophysical journal*, **110**, 2349-2360.

Menche, J., *et al.* (2015) Disease networks. Uncovering disease-disease relationships through the incomplete interactome, *Science*, **347**, 1257601.

Meng, F., *et al.* (2018) Highly Disordered Amyloid-beta Monomer Probed by Single-Molecule FRET and MD Simulation, *Biophysical journal*, **114**, 870-884.

Meng, F., Uversky, V.N. and Kurgan, L. (2017) Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions, *Cell Mol Life Sci*, **74**, 3069-3090.

Meric, G., Robinson, A.S. and Roberts, C.J. (2017) Driving Forces for Nonnative Protein Aggregation and Approaches to Predict Aggregation-Prone Regions, *Annual review of chemical and biomolecular engineering*, **8**, 139-159.

Meszaros, B., Erdos, G. and Dosztanyi, Z. (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding, *Nucleic Acids Res*, **46**, W329-W337.

Meszaros, B., *et al.* (2019) PhaSePro: the database of proteins driving liquid-liquid phase separation, *Nucleic Acids Res*.

Michelitsch, M.D. and Weissman, J.S. (2000) A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions, *Proc Natl Acad Sci U S A*, **97**, 11910-11915.

Miller, L.H., *et al.* (2002) The pathogenic basis of malaria, *Nature*, **415**, 673-679.

Minde, D.P., Halff, E.F. and Tans, S. (2013) Designing disorder: Tales of the unexpected tails, *Intrinsically disordered proteins*, **1**, e26790.

Modrzynska, K., *et al.* (2017) A Knockout Screen of ApiAP2 Genes Reveals Networks of Interacting Transcriptional Regulators Controlling the Plasmodium Life Cycle, *Cell host & microbe*, **21**, 11-22.

Monsellier, E. and Chiti, F. (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution, *EMBO reports*, **8**, 737-742.

Monsellier, E., *et al.* (2007) The distribution of residues in a polypeptide sequence is a determinant of aggregation optimized by evolution, *Biophys J*, **93**, 4382-4391.

Morris, A.M., Watzky, M.A. and Finke, R.G. (2009) Protein aggregation kinetics, mechanism, and curve-fitting: a review of the literature, *Biochim Biophys Acta*, **1794**, 375-397.

Mosca, R., Ceol, A. and Aloy, P. (2013) Interactome3D: adding structural details to protein networks, *Nature methods*, **10**, 47-53.

Moy, R.H., *et al.* (2014) Stem-loop recognition by DDX17 facilitates miRNA processing and antiviral defense, *Cell*, **158**, 764-777.

Mukherjee, A., *et al.* (2017) Induction of IAPP amyloid deposition and associated diabetic abnormalities by a prion-like mechanism, *The Journal of experimental medicine*, **214**, 2591-2610.

Munishkina, L.A., Fink, A.L. and Uversky, V.N. (2004) Conformational prerequisites for formation of amyloid fibrils from histones, *J Mol Biol*, **342**, 1305-1324.

Munoz, V. and Serrano, L. (1995) Elucidating the folding problem of helical peptides using empirical parameters. III. Temperature and pH dependence, *J Mol Biol*, **245**, 297-308.

Muralidharan, V. and Goldberg, D.E. (2013) Asparagine repeats in Plasmodium falciparum proteins: good for nothing?, *PLoS pathogens*, **9**, e1003488.

Muralidharan, V., *et al.* (2012) Plasmodium falciparum heat shock protein 110 stabilizes the asparagine repeat-rich parasite proteome during malarial fevers, *Nature communications*, **3**, 1310.

Nakayashiki, T., *et al.* (2005) Yeast prions [URE3] and [PSI⁺] are diseases, *Proc Natl Acad Sci U S A*, **102**, 10575-10580.

Namba, K. (2001) Roles of partly unfolded conformations in macromolecular self-assembly, *Genes Cells*, **6**, 1-12.

Nelson, R., *et al.* (2005) Structure of the cross-beta spine of amyloid-like fibrils, *Nature*, **435**, 773-778.

Newby, G.A. and Lindquist, S. (2013) Blessings in disguise: biological benefits of prion-like mechanisms, *Trends Cell Biol*, **23**, 251-259.

Nomura, T., *et al.* (2014) Intranuclear aggregation of mutant FUS/TLS as a molecular pathomechanism of amyotrophic lateral sclerosis, *The Journal of biological chemistry*, **289**, 1192-1202.

Nordlund, A. and Oliveberg, M. (2008) SOD1-associated ALS: a promising system for elucidating the origin of protein-misfolding disease, *HFSP journal*, **2**, 354-364.

Nuvolone, M. and Merlini, G. (2017) Systemic amyloidosis: novel therapies and role of biomarkers, *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association*, **32**, 770-780.

Nystrom, S., *et al.* (2012) Multiple substitutions of methionine 129 in human prion protein reveal its importance in the amyloid fibrillation pathway, *J Biol Chem*, **287**, 25975-25984.

O'Donnell, C.W., *et al.* (2011) A method for probing the mutational landscape of amyloid structure, *Bioinformatics*, **27**, i34-42.

Oldfield, C.J. and Dunker, A.K. (2014) Intrinsically disordered proteins and intrinsically disordered protein regions, *Annu Rev Biochem*, **83**, 553-584.

Oliva, A., Llabres, M. and Farina, J.B. (2014) Capability measurement of size-exclusion chromatography with a light-scattering detection method in a stability study of bevacizumab using the process capability indices, *Journal of chromatography. A*, **1353**, 89-98.

Ormo, M., *et al.* (1996) Crystal structure of the Aequorea victoria green fluorescent protein, *Science*, **273**, 1392-1395.

Orr, H.T. and Zoghbi, H.Y. (2007) Trinucleotide repeat disorders, *Annual review of neuroscience*, **30**, 575-621.

Otzen, D. (2010) Functional amyloid: turning swords into plowshares, *Prion*, **4**, 256-264.

Otzen, D. and Nielsen, P.H. (2008) We find them here, we find them there: functional bacterial amyloid, *Cell Mol Life Sci*, **65**, 910-927.

Otzen, D. and Riek, R. (2019) Functional Amyloids, *Cold Spring Harbor perspectives in biology*, **11**.

Painter, H.J., Campbell, T.L. and Llinas, M. (2011) The Apicomplexan AP2 family: integral factors regulating Plasmodium development, *Molecular and biochemical parasitology*, **176**, 1-7.

Pallares, I., *et al.* (2018) Discovering Putative Prion-Like Proteins in Plasmodium falciparum: A Computational and Experimental Analysis, *Frontiers in microbiology*, **9**, 1737.

Pallares, I., Iglesias, V. and Ventura, S. (2015) The Rho Termination Factor of Clostridium botulinum Contains a Prion-Like Domain with a Highly Amyloidogenic Core, *Frontiers in microbiology*, **6**, 1516.

Pallares, I., *et al.* (2004) Amyloid fibril formation by a partially structured intermediate state of alpha-chymotrypsin, *Journal of molecular biology*, **342**, 321-331.

Pallares, I. and Ventura, S. (2017) Advances in the prediction of protein aggregation propensity, *Current medicinal chemistry*.

Pallares, I. and Ventura, S. (2017) The Transcription Terminator Rho: A First Bacterial Prion, *Trends in microbiology*, **25**, 434-437.

Parrini, C., *et al.* (2005) Glycine residues appear to be evolutionarily conserved for their ability to inhibit aggregation, *Structure*, **13**, 1143-1151.

Patel, A., *et al.* (2015) A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation, *Cell*, **162**, 1066-1077.

Patzelt, H., *et al.* (2001) Binding specificity of Escherichia coli trigger factor, *Proc Natl Acad Sci U S A*, **98**, 14244-14249.

Paul, K.R., *et al.* (2015) Generating new prions by targeted mutation or segment duplication, *Proc Natl Acad Sci U S A*, **112**, 8584-8589.

Paul, K.R., *et al.* (2017) Effects of Mutations on the Aggregation Propensity of the Human Prion-Like Protein hnRNPA2B1, *Mol Cell Biol*, **37**.

Pawar, A.P., *et al.* (2005) Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases, *J Mol Biol*, **350**, 379-392.

Payliss, B.J., Vogel, J. and Mittermaier, A.K. (2019) Side chain electrostatic interactions and pH-dependent expansion of the intrinsically disordered, highly acidic carboxyl-terminus of gamma-tubulin, *Protein Sci*, **28**, 1095-1105.

Pease, B.N., *et al.* (2013) Global analysis of protein expression and phosphorylation of three stages of *Plasmodium falciparum* intraerythrocytic development, *Journal of proteome research*, **12**, 4028-4045.

Pedregosa, F., *et al.* (2011) Scikit-learn: Machine learning in Python, *the Journal of machine Learning research*, **12**, 2825-2830.

Perchiacca, J.M., Lee, C.C. and Tessier, P.M. (2014) Optimal charged mutations in the complementarity-determining regions that prevent domain antibody aggregation are dependent on the antibody scaffold, *Protein engineering, design & selection : PEDS*, **27**, 29-39.

Pfefferkorn, C.M., McGlinchey, R.P. and Lee, J.C. (2010) Effects of pH on aggregation kinetics of the repeat domain of a functional amyloid, Pmel17, *Proc Natl Acad Sci U S A*, **107**, 21447-21452.

Pham, C.L., Kwan, A.H. and Sunde, M. (2014) Functional amyloid: widespread in Nature, diverse in purpose, *Essays in biochemistry*, **56**, 207-219.

Pinero, J., *et al.* (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, *Nucleic Acids Res*, **45**, D833-D839.

Pinero, J., *et al.* (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes, *Database : the journal of biological databases and curation*, **2015**, bav028.

Pinheiro, F., *et al.* (2020) Tolcapone, a potent aggregation inhibitor for the treatment of familial leptomenigeal amyloidosis, *FEBS J*.

Podrabsky, J.E., Carpenter, J.F. and Hand, S.C. (2001) Survival of water stress in annual fish embryos: dehydration avoidance and egg envelope amyloid fibers, *Am J Physiol Regul Integr Comp Physiol*, **280**, R123-131.

Polymenidou, M. and Cleveland, D.W. (2012) Prion-like spread of protein aggregates in neurodegeneration, *The Journal of experimental medicine*, **209**, 889-893.

Ponting, C.P., *et al.* (1999) Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer, *J Mol Biol*, **289**, 729-745.

Prakash, T., Veerappa, A. and N, B.R. (2017) Complex interaction between HNRNPD mutations and risk polymorphisms is associated with discordant Crohn's disease in monozygotic twins, *Autoimmunity*, **50**, 275-276.

Prilusky, J., *et al.* (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded, *Bioinformatics*, **21**, 3435-3438.

Prince, M.J. (2015) *World Alzheimer Report 2015: the global impact of dementia: an analysis of prevalence, incidence, cost and trends*. Alzheimer's Disease International.

Prusiner, S.B. (1982) Novel proteinaceous infectious particles cause scrapie, *Science*, **216**, 136-144.

Przyborski, J.M., Diehl, M. and Blatch, G.L. (2015) Plasmodial HSP70s are functionally adapted to the malaria parasite life cycle, *Frontiers in molecular biosciences*, **2**, 34.

Pujols, J., *et al.* (2018) Small molecule inhibits alpha-synuclein aggregation, disrupts amyloid fibrils, and prevents degeneration of dopaminergic neurons, *Proc Natl Acad Sci U S A*, **115**, 10481-10486.

Pujols, J., *et al.* (2020) Chemical Chaperones as Novel Drugs for Parkinson's Disease, *Trends in molecular medicine*, **26**, 408-421.

Pujols, J., Pena-Diaz, S. and Ventura, S. (2018) AGGREGSCAN3D: Toward the Prediction of the Aggregation Propensities of Protein Structures, *Methods in molecular biology*, **1762**, 427-443.

Pulido, D., *et al.* (2016) Insights into the Antimicrobial Mechanism of Action of Human RNase6: Structural Determinants for Bacterial Cell Agglutination and Membrane Permeation, *International journal of molecular sciences*, **17**, 552.

Pulido, P., *et al.* (2016) Specific Hsp100 Chaperones Determine the Fate of the First Enzyme of the Plastidial Isoprenoid Pathway for Either Refolding or Degradation by the Stromal Clp Protease in Arabidopsis, *PLoS genetics*, **12**, e1005824.

Putnam, C. Protein Calculator.

Quintas, A., *et al.* (2001) Tetramer dissociation and monomer partial unfolding precedes protofibril formation in amyloidogenic transthyretin variants, *The Journal of biological chemistry*, **276**, 27207-27213.

Reddy, B.P., *et al.* (2015) A bioinformatic survey of RNA-binding proteins in *Plasmodium*, *BMC Genomics*, **16**, 890.

Reig, N., *et al.* (2015) SOM0226, a repositioned compound for the treatment of TTR amyloidosis, *Orphanet Journal of Rare Diseases*, **10**, P9.

Reumers, J., *et al.* (2009) Protein sequences encode safeguards against aggregation, *Human mutation*, **30**, 431-437.

Richardson, L.G., Jelokhani-Niaraki, M. and Smith, M.D. (2009) The acidic domains of the Toc159 chloroplast preprotein receptor family are intrinsically disordered protein domains, *BMC Biochem*, **10**, 35.

Riek, R. and Eisenberg, D.S. (2016) The activities of amyloids from a structural perspective, *Nature*, **539**, 227-235.

Roberg, K.J., *et al.* (1999) LST1 is a SEC24 homologue used for selective export of the plasma membrane ATPase from the endoplasmic reticulum, *The Journal of cell biology*, **145**, 659-672.

Roberts, B.T. and Wickner, R.B. (2003) Heritable activity: a prion that propagates by covalent autoactivation, *Genes & Development*, **17**, 2083-2087.

Roberts, C.J. (2014) Therapeutic protein aggregation: mechanisms, design, and control, *Trends in biotechnology*, **32**, 372-380.

Romero, D., *et al.* (2011) An accessory protein required for anchoring and assembly of amyloid fibres in *B. subtilis* biofilms, *Molecular microbiology*, **80**, 1155-1168.

Romero, P., *et al.* (1998) Thousands of proteins likely to have long disordered regions, *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 437-448.

Ross, E.D., Baxa, U. and Wickner, R.B. (2004) Scrambled prion domains form prions and amyloid, *Mol Cell Biol*, **24**, 7206-7213.

Ross, E.D., *et al.* (2005) Primary sequence independence for prion formation, *Proc Natl Acad Sci U S A*, **102**, 12825-12830.

Ross, E.D., Minton, A. and Wickner, R.B. (2005) Prion domains: sequences, structures and interactions, *Nature cell biology*, **7**, 1039-1044.

Rousseau, F., Schymkowitz, J. and Serrano, L. (2006) Protein aggregation and amyloidosis: confusion of the kinds?, *Current opinion in structural biology*, **16**, 118-126.

Rousseau, F., Serrano, L. and Schymkowitz, J.W. (2006) How evolutionary pressure against protein aggregation shaped chaperone specificity, *Journal of molecular biology*, **355**, 1037-1047.

Ruggiero, A., *et al.* (2010) Structure and functional regulation of RipA, a mycobacterial enzyme essential for daughter cell separation, *Structure*, **18**, 1184-1190.

Russell, R.B. (1994) Domain insertion, *Protein engineering*, **7**, 1407-1410.

Ryan, V.H., *et al.* (2018) Mechanistic View of hnRNPA2 Low-Complexity Domain Structure, Interactions, and Phase Separation Altered by Mutation and Arginine Methylation, *Mol Cell*, **69**, 465-479 e467.

Sabate, R., *et al.* (2010) The role of protein sequence and amino acid composition in amyloid formation: scrambling and backward reading of IAPP amyloid fibrils, *J. Mol. Biol.*, **404**, 337-352.

Sabate, R., *et al.* (2009) Characterization of the amyloid bacterial inclusion bodies of the HET-s fungal prion, *Microbial cell factories*, **8**, 56.

Sabate, R., *et al.* (2015) Amyloids or prions? That is the question, *Prion*, **9**, 200-206.

Sabate, R., *et al.* (2015) What makes a protein sequence a prion?, *PLoS Comput Biol*, **11**, e1004013.

Saksela, K. and Permi, P. (2012) SH3 domain ligand binding: What's the consensus and where's the specificity?, *FEBS letters*, **586**, 2609-2614.

Sanchez de Groot, N., *et al.* (2015) Proteome response at the edge of protein aggregation, *Open Biol*, **5**, 140221.

Sanchez de Groot, N., *et al.* (2005) Prediction of "hot spots" of aggregation in disease-linked polypeptides, *BMC structural biology*, **5**, 18.

Sanchez de Groot, N., *et al.* (2012) Evolutionary selection for protein aggregation, *Biochem Soc Trans*, **40**, 1032-1037.

Sanchez, I.E., *et al.* (2006) Point mutations in protein globular domains: contributions from function, stability and misfolding, *J Mol Biol*, **363**, 422-432.

Sant'Anna, R., *et al.* (2016) Characterization of Amyloid Cores in Prion Domains, *Sci Rep*, **6**, 34274.

Sant'Anna, R., *et al.* (2016) Repositioning tolcapone as a potent inhibitor of transthyretin amyloidogenesis and associated cellular toxicity, *Nature communications*, **7**, 10787.

Santos, J., *et al.* (2020a) Computational prediction of protein aggregation: Advances in proteomics, conformation-specific algorithms and biotechnological applications, *Computational and structural biotechnology journal*, **18**, 1403-1413.

Santos, J., Iglesias, V. and Ventura, S. (2020b) Computational prediction and redesign of aberrant protein oligomerization, *Progress in molecular biology and translational science*, **169**, 43-83.

Santos, J., *et al.* (2020c) pH-Dependent Aggregation in Intrinsically Disordered Proteins Is Determined by Charge and Lipophilicity, *Cells*, **9**.

Santos, J., *et al.* (2020d) DispHred: A Server to Predict pH-Dependent Order-Disorder Transitions in Intrinsically Disordered Proteins, *International journal of molecular sciences*, **21**, 5814.

Santos, J. and Ventura, S. (2020) Functional Amyloids Germinate in Plants, *Trends in plant science*.

Santoso, A., *et al.* (2000) Molecular basis of a yeast prion species barrier, *Cell*, **100**, 277-288.

Sauvage, E., *et al.* (2008) The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis, *FEMS Microbiol Rev*, **32**, 234-258.

Schramm, A., *et al.* (2017) InSiDDe: A Server for Designing Artificial Disordered Proteins, *Int J Mol Sci*, **19**.

Schwartz, K. and Boles, B.R. (2013) Microbial amyloids--functions and interactions within the host, *Current opinion in microbiology*, **16**, 93-99.

Schweers, O., *et al.* (1994) Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure, *The Journal of biological chemistry*, **269**, 24290-24297.

Schymkowitz, J., *et al.* (2005) The FoldX web server: an online force field, *Nucleic Acids Res*, **33**, W382-388.

Serio, T.R. and Lindquist, S.L. (2001) The yeast prion [PSI⁺]: molecular insights and functional consequences, *Advances in protein chemistry*, **59**, 391-412.

Seviour, T., *et al.* (2015) Functional amyloids keep quorum-sensing molecules in check, *J Biol Chem*, **290**, 6457-6469.

Shaw, K.L., *et al.* (2001) The effect of net charge on the solubility, activity, and stability of ribonuclease Sa, *Protein science : a publication of the Protein Society*, **10**, 1206-1215.

Shida, T., *et al.* (2020) Short disordered protein segment regulates cross-species transmission of a yeast prion, *Nat Chem Biol*, **16**, 756-765.

Shorter, J. and Lindquist, S. (2005) Prions as adaptive conduits of memory and inheritance, *Nat Rev Genet*, **6**, 435-450.

Si, K. (2015) Prions: what are they good for?, *Annual review of cell and developmental biology*, **31**, 149-169.

Si, K., *et al.* (2010) Aplysia CPEB can form prion-like multimers in sensory neurons that contribute to long-term facilitation, *Cell*, **140**, 421-435.

Si, K. and Kandel, E.R. (2016) The Role of Functional Prion-Like Proteins in the Persistence of Memory, *Cold Spring Harbor perspectives in biology*, **8**, a021774.

Siddiqua, A. and Margittai, M. (2010) Three- and four-repeat Tau coassemble into heterogeneous filaments: an implication for Alzheimer disease, *The Journal of biological chemistry*, **285**, 37920-37926.

Sidhu, S.S. (2000) Phage display in pharmaceutical biotechnology, *Current opinion in biotechnology*, **11**, 610-616.

Sikorska, B. and Liberski, P.P. (2012) Human prion diseases: from Kuru to variant Creutzfeldt-Jakob disease, *Subcellular biochemistry*, **65**, 457-496.

Simm, S., et al. (2016) 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification, *Biological research*, **49**, 31.

Singh, G.P., et al. (2004) Hyper-expansion of asparagines correlates with an abundance of proteins with prion-like domains in *Plasmodium falciparum*, *Molecular and biochemical parasitology*, **137**, 307-319.

Singleton, A.B., et al. (2003) alpha-Synuclein locus triplication causes Parkinson's disease, *Science*, **302**, 841.

Siomi, H., et al. (1993) The pre-mRNA binding K protein contains a novel evolutionarily conserved motif, *Nucleic Acids Res*, **21**, 1193-1198.

Skordalakes, E. and Berger, J.M. (2003) Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading, *Cell*, **114**, 135-146.

Smith, M.D. and Jelokhani-Niaraki, M. (2012) pH-induced changes in intrinsically disordered proteins, *Methods Mol Biol*, **896**, 223-231.

Soler, M.A., de Marco, A. and Fortuna, S. (2016) Molecular dynamics simulations and docking enable to explore the biophysical factors controlling the yields of engineered nanobodies, *Sci Rep*, **6**, 34869.

Soper, D.S. (2018) p-Value Calculator for Correlation Coefficients.

Sormanni, P., Aprile, F.A. and Vendruscolo, M. (2015) The CamSol method of rational design of protein mutants with enhanced solubility, *Journal of molecular biology*, **427**, 478-490.

Spaulding, C.N., et al. (2015) Fueling the Fire with Fibers: Bacterial Amyloids Promote Inflammatory Disorders, *Cell host & microbe*, **18**, 1-2.

Spillantini, M.G., et al. (1997) Alpha-synuclein in Lewy bodies, *Nature*, **388**, 839-840.

Staniforth, G.L. and Tuite, M.F. (2012) Fungal prions, *Prog Mol Biol Transl Sci*, **107**, 417-456.

Stohr, J., et al. (2012) Purified and synthetic Alzheimer's amyloid beta (Abeta) prions, *Proc Natl Acad Sci U S A*, **109**, 11025-11030.

Taglialegna, A., et al. (2016) Staphylococcal Bap Proteins Build Amyloid Scaffold Biofilm Matrices in Response to Environmental Signals, *PLoS pathogens*, **12**, e1005711.

Tank, E.M., et al. (2007) Prion protein repeat expansion results in increased aggregation and reveals phenotypic variability, *Mol Cell Biol*, **27**, 5445-5455.

Tariq, M., et al. (2013) Drosophila GAGA factor polyglutamine domains exhibit prion-like behavior, *BMC Genomics*, **14**, 374.

Tartaglia, G.G., et al. (2005) Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences, *Protein science : a publication of the Protein Society*, **14**, 2723-2734.

Tartaglia, G.G., et al. (2008) Prediction of aggregation-prone regions in structured proteins, *Journal of molecular biology*, **380**, 425-436.

Tartaglia, G.G., et al. (2007) Life on the edge: a link between gene expression levels and aggregation rates of human proteins, *Trends Biochem Sci*, **32**, 204-206.

Tartaglia, G.G. and Vendruscolo, M. (2008) The Zyggregator method for predicting protein aggregation propensities, *Chemical Society reviews*, **37**, 1395-1401.

Tartaglia, G.G. and Vendruscolo, M. (2009) Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations, *Mol Biosyst*, **5**, 1873-1876.

Taylor, J.D. and Matthews, S.J. (2015) New insight into the molecular control of bacterial functional amyloids, *Frontiers in cellular and infection microbiology*, **5**, 33.

Tedeschi, G., et al. (2017) Aggregation properties of a disordered protein are tunable by pH and depend on its net charge per residue, *Biochimica et biophysica acta. General subjects*, **1861**, 2543-2550.

Tepljakov, A., et al. (2016) Structural diversity in a human antibody germline library, *mAbs*, **8**, 1045-1063.

Ter-Avanesyan, M.D., et al. (1993) Deletion analysis of the SUP35 gene of the yeast *Saccharomyces cerevisiae* reveals two non-overlapping functional regions in the encoded protein, *Molecular microbiology*, **7**, 683-692.

Tomba, P. (2002) Intrinsically unstructured proteins, *Trends Biochem Sci*, **27**, 527-533.

Tomba, P. (2012) Intrinsically disordered proteins: a 10-year recap, *Trends Biochem Sci*, **37**, 509-516.

Tomba, P. and Fuxreiter, M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions, *Trends Biochem Sci*, **33**, 2-8.

Toombs, J.A., McCarty, B.R. and Ross, E.D. (2010) Compositional determinants of prion formation in yeast, *Mol Cell Biol*, **30**, 319-332.

Toombs, J.A., et al. (2012) De novo design of synthetic prion domains, *Proc Natl Acad Sci U S A*, **109**, 6519-6524.

Tosi, T., et al. (2014) Structural similarity of secretins from type II and type III secretion systems, *Structure*, **22**, 1348-1355.

Treusch, S. and Lindquist, S. (2012) An intrinsically disordered yeast prion arrests the cell cycle by sequestering a spindle pole body component, *The Journal of cell biology*, **197**, 369-379.

True, H.L. and Lindquist, S.L. (2000) A yeast prion provides a mechanism for genetic variation and phenotypic diversity, *Nature*, **407**, 477-483.

Tsolis, A.C., et al. (2013) A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins, *PLoS one*, **8**, e54175.

Uhlen, M., et al. (2015) Proteomics. Tissue-based map of the human proteome, *Science*, **347**, 1260419.

UniProt, C. (2015) UniProt: a hub for protein information, *Nucleic Acids Res*, **43**, D204-212.

Uptain, S.M. and Lindquist, S. (2002) Prions as protein-based genetic elements, *Annual review of microbiology*, **56**, 703-741.

Uversky, V.N. (2009) Intrinsically disordered proteins and their environment: effects of strong denaturants, temperature, pH, counter ions, membranes, binding partners, osmolytes, and macromolecular crowding, *The protein journal*, **28**, 305-325.

Uversky, V.N. and Fink, A.L. (2004) Conformational constraints for amyloid fibrillation: the importance of being unfolded, *Biochim Biophys Acta*, **1698**, 131-153.

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

Uversky, V.N., *et al.* (1999) Natively unfolded human prothymosin alpha adopts partially folded collapsed conformation at acidic pH, *Biochemistry*, **38**, 15009-15016.

Uversky, V.N., Li, J. and Fink, A.L. (2001) Evidence for a partially folded intermediate in alpha-synuclein fibril formation, *The Journal of biological chemistry*, **276**, 10737-10744.

van der Kant, R., *et al.* (2017) Prediction and Reduction of the Aggregation of Monoclonal Antibodies, *Journal of molecular biology*, **429**, 1244-1261.

van der Lee, R., *et al.* (2014) Classification of intrinsically disordered regions and proteins, *Chemical reviews*, **114**, 6589-6631.

Van der Meeren, R., *et al.* (2013) New insights into the assembly of bacterial secretins: structural studies of the periplasmic domain of XcpQ from *Pseudomonas aeruginosa*, *J Biol Chem*, **288**, 1214-1225.

Van Durme, J., *et al.* (2016) Solubis: a webserver to reduce protein aggregation through mutation, *Protein engineering, design & selection : PEDS*, **29**, 285-289.

van Rheede, T., *et al.* (2003) Molecular evolution of the mammalian prion protein, *Molecular biology and evolution*, **20**, 111-121.

Vapnik, V. (1998) *Statistical learning theory*. New York. Wiley.

Vapnik, V. (2013) *The nature of statistical learning theory*. Springer science & business media.

Ventura, S., *et al.* (2004) Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case, *Proc Natl Acad Sci U S A*, **101**, 7258-7263.

Vieira, N.M., *et al.* (2014) A defect in the RNA-processing protein HNRPDL causes limb-girdle muscular dystrophy 1G (LGMD1G), *Hum Mol Genet*, **23**, 4103-4110.

Villar-Pique, A., Lopes da Fonseca, T. and Outeiro, T.F. (2016) Structure, function and toxicity of alpha-synuclein: the Bermuda triangle in synucleinopathies, *Journal of neurochemistry*, **139 Suppl 1**, 240-255.

Villarroya-Beltri, C., *et al.* (2013) Sumoylated hnRNP A2B1 controls the sorting of miRNAs into exosomes through binding to specific motifs, *Nature communications*, **4**, 2980.

Walsh, I., *et al.* (2012) ESpritz: accurate and fast prediction of protein disorder, *Bioinformatics*, **28**, 503-509.

Walsh, I., *et al.* (2014) PASTA 2.0: an improved server for protein aggregation prediction, *Nucleic Acids Res*, **42**, W301-307.

Wang, I.F., *et al.* (2012) The self-interaction of native TDP-43 C terminus inhibits its degradation and contributes to early proteinopathies, *Nature communications*, **3**, 766.

Wang, J., *et al.* (2018) A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins, *Cell*, **174**, 688-699 e616.

Wang, W., *et al.* (2007) Antibody structure, instability, and formulation, *Journal of pharmaceutical sciences*, **96**, 1-26.

Wang, W. and Ventura, S. (2020) Prion domains as a driving force for the assembly of functional nanomaterials, *Prion*, **14**, 170-179.

Ward, J.J., *et al.* (2004) The DISOPRED server for the prediction of protein disorder, *Bioinformatics*, **20**, 2138-2139.

Watt, B., *et al.* (2013) PMEL: a pigment cell-specific model for functional amyloid formation, *Pigment cell & melanoma research*, **26**, 300-315.

Wei, G., *et al.* (2017) Self-assembling peptide and protein amyloids: from structure to tailored function in nanotechnology, *Chemical Society reviews*, **46**, 4661-4708.

Westermarck, G.T. and Westermarck, P. (2010) Prion-like aggregates: infectious agents in human disease, *Trends in molecular medicine*, **16**, 501-507.

Westermarck, G.T., *et al.* (2008) Widespread amyloid deposition in transplanted human pancreatic islets, *The New England journal of medicine*, **359**, 977-979.

Wickner, R.B. (1994) [URE3] as an altered URE2 protein: evidence for a prion analog in *Saccharomyces cerevisiae*, *Science*, **264**, 566-569.

Wickner, R.B. and Kelly, A.C. (2016) Prions are affected by evolution at two levels, *Cellular and molecular life sciences : CMLS*, **73**, 1131-1144.

Wickner, R.B., *et al.* (2015) Yeast Prions: Structure, Biology, and Prion-Handling Systems, *Microbiology and Molecular Biology Reviews*, **79**, 1-17.

Wienk, H., *et al.* (2005) Solution structure of the C1-subdomain of *Bacillus stearothermophilus* translation initiation factor IF2, *Protein Sci*, **14**, 2461-2468.

Williams, A.J. and Paulson, H.L. (2008) Polyglutamine neurodegeneration: protein misfolding revisited, *Trends in neurosciences*, **31**, 521-528.

Wood, S.J., *et al.* (1995) Prolines and amyloidogenicity in fragments of the Alzheimer's peptide beta/A4, *Biochemistry*, **34**, 724-730.

WorldHealthOrganisation (WHO) <http://www.who.int/mediacentre/factsheets/fs194/en/>.

- Wright, P.E. and Dyson, H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, *Journal of molecular biology*, **293**, 321-331.
- Wright, P.E. and Dyson, H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation, *Nat Rev Mol Cell Biol*, **16**, 18-29.
- Wu, Y., *et al.* (2009) Identification of phosphorylated proteins in erythrocytes infected by the human malaria parasite *Plasmodium falciparum*, *Malaria journal*, **8**, 105.
- Xia, X., *et al.* (2016) Engineering a Cysteine-Free Form of Human Fibroblast Growth Factor-1 for "Second Generation" Therapeutic Application, *Journal of pharmaceutical sciences*, **105**, 1444-1453.
- Xu, H., *et al.* (2014) Structural basis for the prion-like MAVS filaments in antiviral innate immunity, *Elife*, **3**, e01489.
- Xue, B., *et al.* (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids, *Biochim Biophys Acta*, **1804**, 996-1010.
- Yang, Z.R., *et al.* (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins, *Bioinformatics*, **21**, 3369-3376.
- Yee, A.W., *et al.* (2019) A molecular mechanism for transthyretin amyloidogenesis, *Nature communications*, **10**, 925.
- Yuan, A.H., *et al.* (2014) Prion propagation can occur in a prokaryote and requires the ClpB chaperone, *eLife*, **3**, e02949.
- Yuan, A.H. and Hochschild, A. (2017) A bacterial global regulator forms a prion, *Science*, **355**, 198-201.
- Yuda, M., *et al.* (2010) Transcription factor AP2-Sp and its target genes in malarial sporozoites, *Molecular microbiology*, **75**, 854-863.
- Zajkowski, T., *et al.* (2021) The hunt for ancient prions: Archaeal prion-like domains form amyloid-based epigenetic elements, *Molecular biology and evolution*.
- Zambrano, R., *et al.* (2015) PrionW: a server to identify proteins containing glutamine/asparagine rich prion-like domains and their amyloid cores, *Nucleic Acids Research*, 1-7.
- Zambrano, R., *et al.* (2015) AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures, *Nucleic Acids Res*, **43**, W306-313.
- Zamora, W.J., Campanera, J.M. and Luque, F.J. (2019) Development of a Structure-Based, pH-Dependent Lipophilicity Scale of Amino Acids from Continuum Solvation Calculations, *The journal of physical chemistry letters*, **10**, 883-889.
- Zeng, G., *et al.* (2015) Functional bacterial amyloid increases *Pseudomonas* biofilm hydrophobicity and stiffness, *Frontiers in microbiology*, **6**, 1099.
- Zerovnik, E. (2017) Putative alternative functions of human stefin B (cystatin B): binding to amyloid-beta, membranes, and copper, *Journal of molecular recognition : JMR*, **30**.
- Zhang, M., *et al.* (2012) PK4, a eukaryotic initiation factor 2alpha(eIF2alpha) kinase, is essential for the development of the erythrocytic cycle of *Plasmodium*, *Proc Natl Acad Sci U S A*, **109**, 3956-3961.
- Zhou, Y., *et al.* (2012) Promiscuous cross-seeding between bacterial amyloids promotes interspecies biofilms, *J Biol Chem*, **287**, 35092-35103.
- Zhu, X., *et al.* (2015) Mediator tail subunits can form amyloid-like aggregates in vivo and affect stress response in yeast, *Nucleic Acids Res*, **43**, 7306-7314.

9. Appendices

9.1 List of software used in the present thesis. *Developed software marked with an *.*

Adobe Photoshop/Acrobat Reader	MS Office suite
AGGRESCAN	NCBI BLAST
Aggrescan 3D *	Notepad++
Aggrescan 3D Standalone *	Numpy python module
AMYCO *	PAPA
Anaconda python distribution	Perl programming language
Bash programming language	PLAAC
Biopython	PrionScan
Bitbucket	pRANK
CABS-Flex	PrionW *
CSS style sheet language	PyMOL
DAVID	PythonAnywhere
DispHred *	Python programming language
FreeSasa	pWALTZ
FoldIndex	Scipy python module
FoldX	SolupHred
Github/Git bash	Sublime Text
GOSTat	TANGO
HTML markup language	UniprotKB BLAST/Retrieve-ID mapping tool
IUPred	WALTZ
JavaScript programming language	
Jupyter Notebook (iPython)	
Matplotlib python module	
MODELLER	

9.2 List of databases used in the present thesis.

DisGeNET	OMIM
Disprot	Pfam
Interactome3D	PDB
Malacards	STRING
NCBI	UniprotKB

9.3 List of Operative systems used in the present thesis.

Android

Ubuntu

iOS

Windows

MacOS

9.4 List of Web browsers used in the present thesis.

Chrome

Internet Explorer

Edge

Opera

Firefox

Safari

9.5 Supplementary Material

Chapter II – Effect of pH in protein compaction

Supplementary material S4.1 – Sequence of the PNTs variants

We used the N-terminus moiety of measles virus phosphoprotein (PNT) as a model IDP. Acidic and basic variants were obtained as described in Tedeschi, G., et al. (2017). Briefly, basic (H, K, R) residues from the wild type protein were substituted with acidic (E or D) in the acidic variant; while in the basic variant wild type acidic residues were almost all substituted by basic ones.

PNT wild type (N-terminus moiety of measles virus phosphoprotein)

MHHHHHHA^{EE}QAR^{HV}KNGL^{EC}IRALKA^{EP}IGSLAI^{EE}AMAAWS^{EIS}DNPGQ^{ER}ATC^RE^{EE}K
AGSSGLSK^PCLSAIGST^{EG}GAP^{RI}R^{IR}QGPG^{ES}DDDA^{ET}LGIPPNLQASSTGLQCHYVY^D
HSGEAV^KGIQ^{DA}DSIMVQSGLD^{GD}STLSGGD^{NE}SE^{NS}DVDIG^{EP}DT^{EG}YAIT^{DR}GSAPIS
MGRAS^DV^ETA^{EG}GG^{EI}HE^{LL}RLQS^{RG}NNF^{PK}L^{GK}TLNVPPPP^DPGRAS^TSGTPI^{KK}ENLY
FQGSHPGTMPGTM

PNT acidic variant

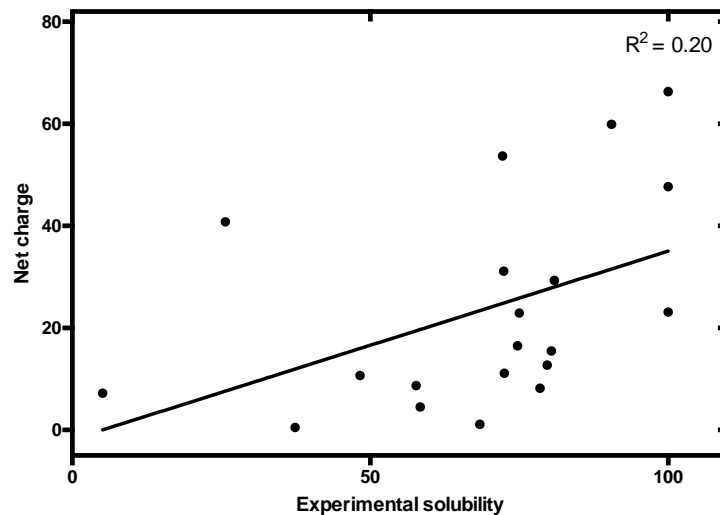
MHHHHHHA^{EE}QAD^{DD}VENGL^{EC}IEALDA^{EP}IGSLAI^{EE}AMAAWS^{EIS}DNPGQ^{ED}ATC^{EEEE}
AGSSGLSE^PCLSAIGST^{EG}GAP^{DI}D^DGGPG^{ES}DDDA^{ET}LGIPPNLQASSTGLQC^DYVY^D
HSGEAV^DGIQ^{DA}DSIMVQSGLD^{GD}STLSGGD^{NE}SE^{NS}DVDIG^{EP}DT^{EG}YAIT^{DE}GSAPIS
MGFD^{AS}D^VTA^{EG}GG^{EI}EE^{LL}ELQS^DGNNF^{EL}GD^TTLNVPPPP^DPGE^AASTSGTPI^{DD}ENLY
FQGSHPGTMPGTM

PNT basic variant

MHHHHHHA^{EE}QAR^{HV}KNGL^{EC}IRALKA^{EP}IGSLAI^{KE}AMAAWS^{EIS}RNPGQ^{KR}ATC^RE^{EE}K
AGSSGLSK^PCLSAIGST^{EG}GAP^{RI}R^{IR}QGPG^{ES}DRDA^{KT}LGIPPNLQASSTGLQCHYVY^R
HSGKAV^KGIQ^{DA}RSIMVQSGLD^{GR}STLSGG^{RN}ES^{RN}SR^{VD}IG^{KP}RT^{EG}YAIT^{DR}GSAPIS
MGRAS^DV^KTA^{EG}GG^{KI}HE^{LL}RLQS^{RG}NNF^{PK}L^{GK}TLNVPPPP^DPGRAS^TSGTPI^{KK}ENLY
FQGSHPGTMPGTM

Supplementary material S4.2 – Correlation between charge distribution and change in solubility in a range of pH for PNTs

Correlation between the experimental solubility and net charge variation. Solid line corresponds to the fit of the data to a linear regression with a p-value < 0.05.



Supplementary material S4.3 – SolupHred predictions correlation with experimental solubilities of disease-associated IDPs.

Protein	R ²	p-value
α-syn (K _{app})	0.82	0.013
α- syn (T _{lag})	0.87	0.0066
IAPP	0.95	< 0.00001
Aβ-40	0.99	0.000039
Tau K19	0.8	0.000037

α-syn: alpha-synuclein (K_{app}: apparent rate constant, T_{lag}: latency time)

IAPP: Islet Amyloid Polypeptide

Aβ-40: 40 residues beta amyloid-peptide

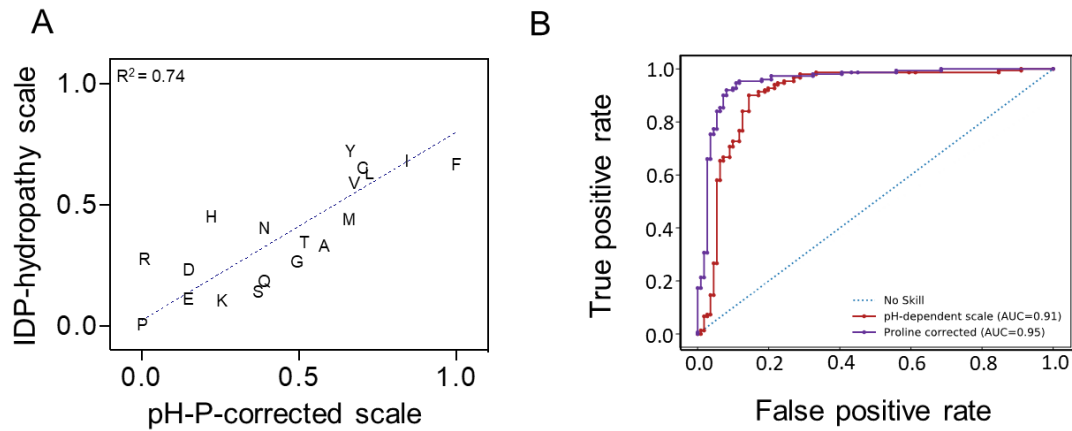
Tau K19: a truncated construct of the 3R the microtubule-binding protein Tau

Supplementary material S4.4 – Performance of SolupHred in predicting changes in the solubility of the disease-related proteins and functional amyloids upon deviation from neutral pH.

Measure	SolupHred predictions
Sensitivity	0.96
Specificity	0.85
Precision	0.88
False Discovery rate	0.12
Accuracy	0.91
F1 Score	0.92
Matthews Correlation Coefficient	0.81

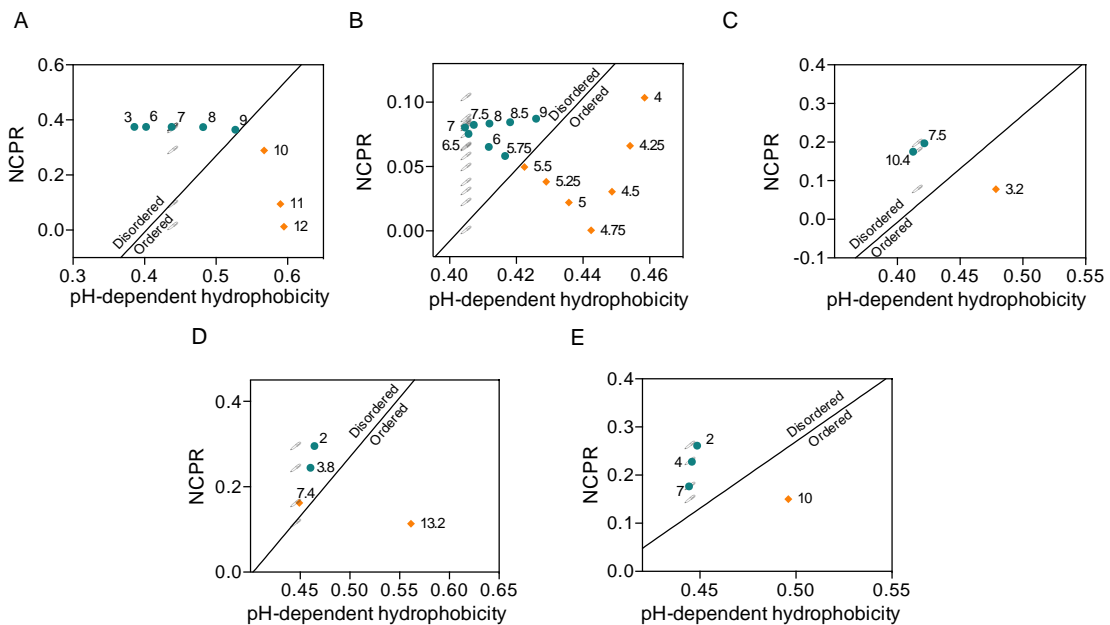
For each protein set, the experimental data obtained at a pHs closer to pH 7.0 was considered the aggregation at neutral pH. pHs in which the proteins showed increased experimental aggregation relative to neutral pH were considered positives, while less aggregative pHs were labeled as negative. SolupHred statistics were: TP (n=22), TN (n=17), FP (n=3) and FN (n=1).

Supplementary material S4.5 – Evaluation of the pH-P-corrected hydrophathy scale.



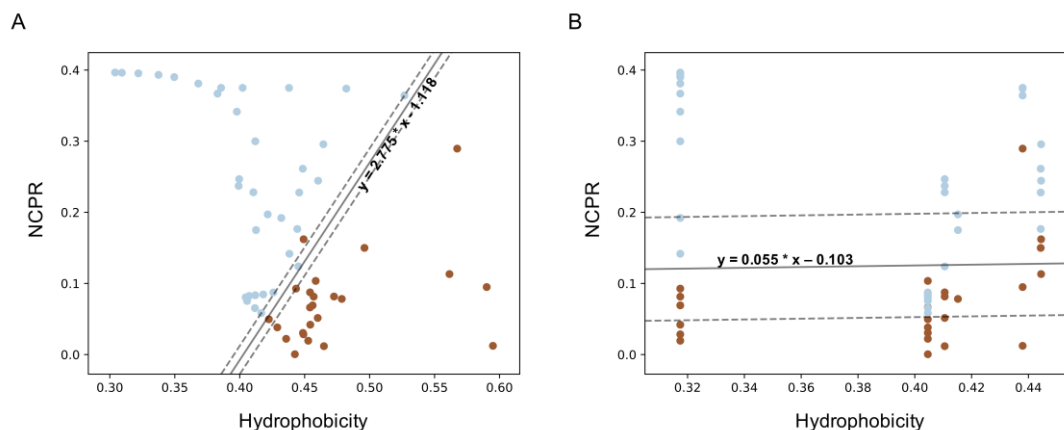
A) Correlation between pH-P-corrected hydrophathy scale and IDP-hydrophathy scale. **B)** ROC curve analysis of the performance of pH-dependent and pH-P-corrected hydrophathy scales in discriminating a dataset of a fully disordered ($n=111$) and single-chain folded ($n=150$) proteins.

Supplementary material S4.6 – C-H plots of disordered proteins and peptides



A) Ac-AKAAKAKAAKAAKA-NH2, **B)** a 36-loop region of the influenza hemagglutinin **C)** A-domain of the Toc132 receptor **D)** LL-37 **E)** and human histones. Solid line delimits folded-unfolded boundary condition. Blue and orange data points correspond to bibliographically unfolded and folded conditions, respectively. Open circles represent the same points considering a constant hydrophobicity (pH 7).

Supplementary material S4.7 – SVM-based classification of pH-conditioned ordered-disordered protein sequences based on their C-H relation.



A, B) C-H plots containing 59 datapoints; 35 labeled as disordered (blue) and 24 as folded (orange). Each point is defined by its calculated NCPR and its mean hydrophobicity at their **A)** experimental pH **B)** or neutral pH. The solid line represents the optimal boundary condition, whereas dashed lines delimitate the maximum margin.

Supplementary material S4.8 – Performance of the pH-independent hydrophobicity model derived by SVM in Supplementary material S4.7 in predicting order-disorder transitions in a C-H plot analysis

Measure	pH-independent hydrophobicity SVM analysis
Sensitivity	0.74
Specificity	0.88
Precision	0.90
False Discovery rate	0.08
Accuracy	0.80
F1 Score	0.81
Matthews Correlation Coefficient	0.60

Unfolded sequences correctly predicted to be unfolded were classified as true positives.

Chapter III – Prediction of prion-like behaviour

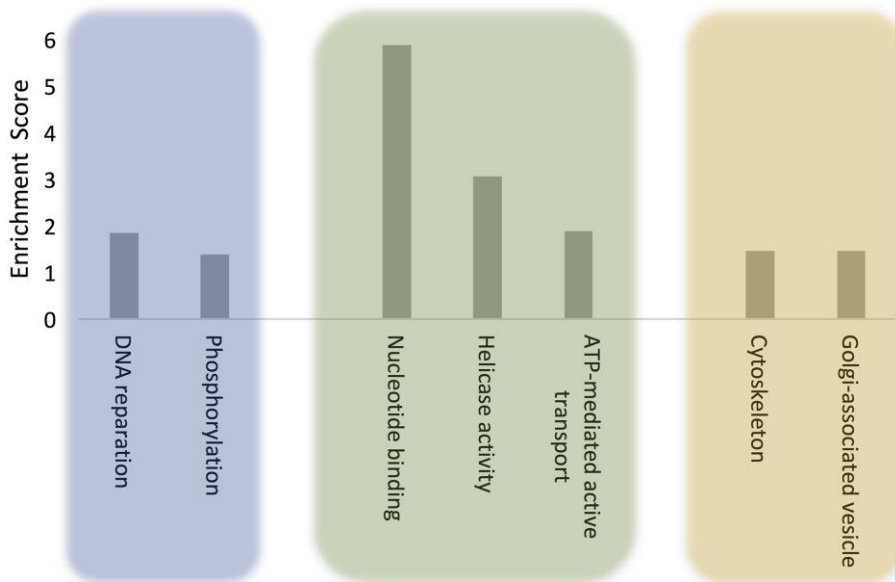
Supplementary material S5.1 – PrLD dataset used in the benchmarking of pWALTZ.

UniprotKB Ac.	Gene Name	Prion/Non-Prion	UniprotKB Ac.	Gene Name	Prion/Non-Prion
Q12221	PUF2	P	Q00772	SLT2	NP
P09547	SWI1	P	P41696	AZF1	NP
P38691	KSP1	P	P31384	CCR4	NP
Q05166	ASM4	P	P48837	NUP57	NP
P23202	URE2	P	P24276	SSD1	NP
P18494	GLN3	P	Q08831	VTS1	NP
P25367	RNQ1	P	P50896	PSP1	NP
Q08972	NEW1	P	P53309	YAP1802	NP
P32770	NRP1	P	P39523	YMR124W	NP
P40070	LSM4	P	Q05785	ENT2	NP
P38180	YBL081W	P	P32900	SKG6	NP
P05453	SUP35	P	Q06251	YLR177W	NP
P11746	MCM1	NP	Q06449	PIN3	NP
P32505	NAB2	NP	P43582	WWM1	NP
Q03761	TAF12	NP	P38996	NAB3	NP
P23291	YCK1	NP	P29295	HRR25	NP
Q12124	MED2	NP	P32896	PDC2	NP
P38080	AKL1	NP	Q02792	RAT1	NP
P25339	PUF4	NP	P32790	SLA1	NP
P39081	PCF11	NP	P22579	SIN3	NP
P43572	EPL1	NP	Q12151	UPC2	NP
P22082	SNF2	NP	P39936	TIF4632	NP
P45978	SCD6	NP	P48562	CLA4	NP
P14680	YAK1	NP	Q06315	SKG3	NP
P53829	CAF40	NP	P39935	TIF4631	NP
P53617	NRD1	NP			

Prion/Non-Prion (P/NP) classification according to Alberti et. al. scale of prion propensity (Alberti, et al., 2009). Sequences scoring ≤ 2 (1 positive assay as a maximum) were considered non-prions (NP) while sequences scoring ≥ 9 (all four assays positives) were considered prions (P). PFDs were as described in (2).

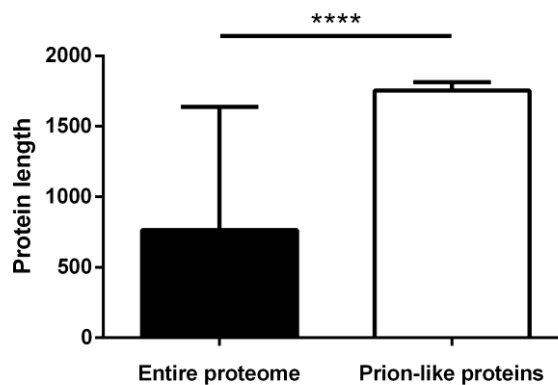
Chapter IV – Characterization of prion-like proteins

Supplementary material S6.1 – Computational analysis of the Q/N-rich proteins devoid of PrLDs in *P. falciparum* proteome.



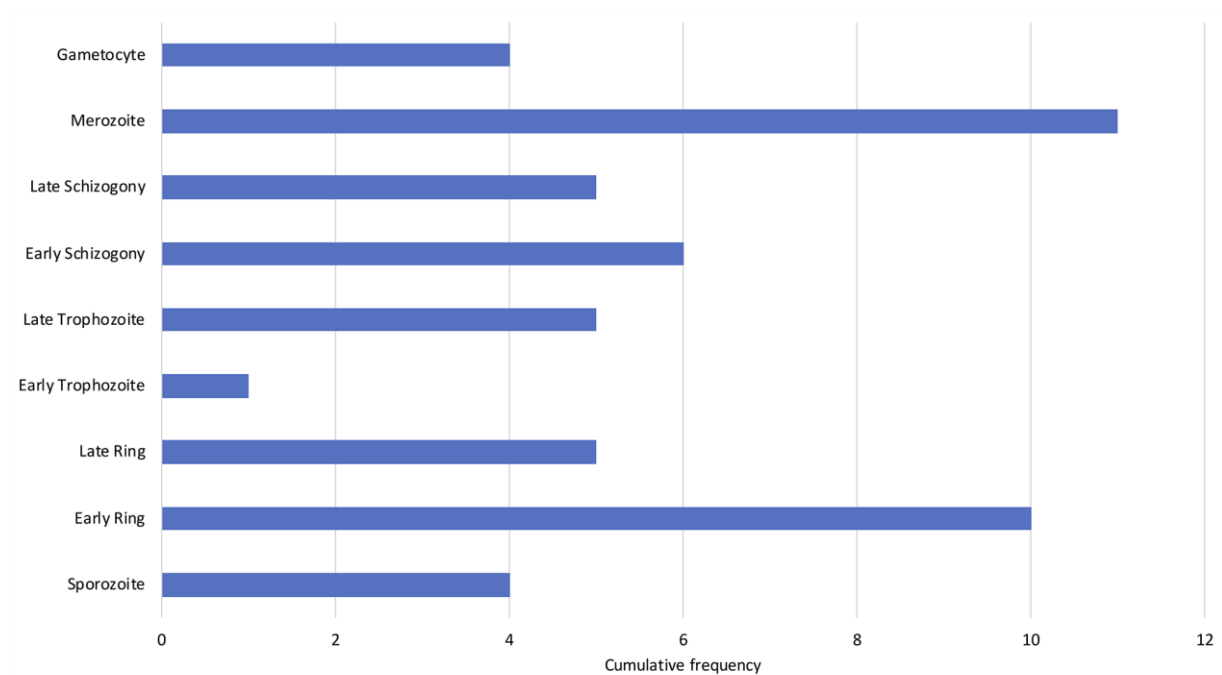
Clusters of GO enriched terms for each ontology; biological process in blue, cellular component in green and molecular function in orange. The enrichment analysis was performed with Functional Annotation Tool of DAVID 6.7 database using high stringency clusters, p-value ≤ 0.05 for GO terms.

Supplementary material S6.2 – Size of *P. falciparum* Q/N-rich prion-like proteins.



The size of 5353 and 503 proteins were analysed and averaged for the complete proteome and the prion-like subset, respectively. The mean size of the proteins in the entire proteome is 764 ± 1193 residues. The mean size of the proteins in the prion-like subset is 1755 ± 59 residues. The P value for the unpaired t test < 0.0001 .

Supplementary material S6.3 – Developmental stages with highest prion-like protein expression.



Highest expression stage for prion-like proteins in *P. falciparum*'s life cycle. Data corresponds to 10% highest scoring prion-like proteins (n=51).

Supplementary material S6.4 – *P. falciparum* soft-amyloid cores aggregation prediction.

Protein	Soft-amyloid core	AGGRESKAN	Tango (%)	Zyggregator
Sec24b	NYNNNYNNNYNNNYNNNNYN	-44.40	0	-4.14
IF2c	NNNNIYNNNIYNNNNIYNIYN	-27.00	2.55	-0.93
PK4	NMNNINNMNNINNMNNINNIN	-26.40	18.32	-3.23

Analysis of the aggregation tendencies of the *P. falciparum* soft-amyloid cores using AGGRESKAN (Conchillo-Sole, et al., 2007), Tango (Fernandez-Escamilla, et al., 2004) and Zyggregator (Tartaglia and Vendruscolo, 2008) prediction methods. None of them is able to correctly identify any significant amyloid propensity in the peptides.

Supplementary material S6.5 – Disorder context of the soft-amyloid cores.

Prediction method	Sec24b	IF2c	PK4
FoldIndex	100	100	100
PONDR-FIT	100	100	100
IUPRED	79	80	61
RONN	69	62	69
AVERAGE	87	86	83

To predict disorder, FoldIndex (Prilusky, et al., 2005), PONDR-FIT (Xue, et al., 2010), IUPRED (Dosztanyi, et al., 2005), RONN (Yang, et al., 2005) prediction methods were used. Disorder was analysed for the 21 residues-long peptides and 20 flanking residues at each end and expressed as the percentage of disordered residues in those 61 residues-long segments. Average disorder accounts for the mean of all disorder predictions for a given segment.