Aggregation is a context-dependent constraint on protein evolution

Michele Monti*,[1,2] Alexandros Armaos,[1,2] Marco Fantini,[3] Annalisa Pastore,[4] and Gian Gaetano Tartaglia*[5,6,7]

[1] *Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, 08003 Barcelona*

[2] *RNA System Biology Lab, Department of Neuroscience and Brain Technologies, Istituto Italiano di Tecnologia (IIT), Via Morego 30, 16163, Genoa, Italy*

[3] *Department of Chemistry, Columbia University, 1208 Northwest Corner Building, 12th Floor, 550 West 120th Street, New York, NY 10027, USA*

[4] *3UK-DRI Centre at the Maurice Wohl Institute, Department of Clinical and Basic Neuroscience, King's College London, 125 Coldharbour Lane, SE5 9NU, London, UK*

[5] *Centre for Genomic Regulation (CRG) and ICREA, The Barcelona Institute for Science and Technology, Dr. Aiguader 88, 08003 Barcelona*

[6] *RNA System Biology Lab, Centre for Human Technologies, Istituto Italiano di Tecnologia (IIT), Via Morego 30, 16163, Genoa, Italy*

[7] *La Sapienza University, Dipartimento di Biologia e Biotecnologie, Piazzale Aldo Moro 5, Rome, Italy*

Solubility is a requirement for many cellular processes. Loss of solubility and aggregation can lead to the partial or complete abrogation of protein function. Thus, understanding the relationship between protein evolution and aggregation is an important goal. Here, we analysed two deep mutational scanning experiments to investigate the role of protein aggregation in molecular evolution. In one data set, mutants of a protein involved in RNA biogenesis and processing, human TAR DNA binding protein 43 (TDP-43), were expressed in *S. cerevisiae*. In the other data set, mutants of a bacterial enzyme that controls resistance to penicillins and cephalosporins, TEM-1 beta-lactamase, were expressed in *E. coli* under the selective pressure of an antibiotic treatment. We found that aggregation differentiates the effects of mutations in the two different cellular contexts. Specifically, aggregation was found to be associated with increased cell fitness in the case of TDP-43 mutations, as it protects the host from aberrant interactions. By contrast, in the case of TEM-1 beta-lactamase mutations, aggregation is linked to a decreased cell fitness due to inactivation of protein function. Our study shows that aggregation is an important context-dependent constraint of molecular evolution and opens up new avenues to investigate the role of aggregation in different cellular contexts.

I. INTRODUCTION

The majority of proteins function as monodispersed ordered species dissolved in intra- or extra-cellular aqueous fluids. There are, however, abnormal or pathologic states in which proteins aggregate, often becoming insoluble and dropping out of solution [1, 2]. In other cases, protein assembly leads to a liquid-liquid phase separation producing a reversible soluble coacervate [3, 4]. A liquid-to-solid phase transition is often detrimental to the cell as it is associated with a loss of the function that the protein has evolved to attain in the specific context [5]. In some cases, aggregates could acquire aberrant functionalities, which can have consequences for the onset and development of disorders [6]. Although these principles are valid for a large part of the proteome [7, 8], aggregation can be required in specific cases to increase the functionality of a biological system. This occurs, for instance, for Pmel17, a protein found in mammalian melanosomes, whose aggregation is necessary for the production and storage of melanin [9]. Given the double valence of protein aggregation that can result either in the formation of toxic assemblies or physiologically required organelles, it has long been debated what could determine whether aggregation is beneficial or detrimental from an evolutionary perspective.

Here, we set to clarify this complex matter and question of whether aggregation is an important constraint in protein evolution that can be the discriminant between beneficial and detrimental situations by considering two cases: we studied molecular evolution of proteins in an endogenous versus exogenous host. One of the important factors that reduce the danger of aggregation is in fact the presence of interaction partners that protect the regions that would otherwise be involved in self-assembly [10, 11]. Indeed, the interfaces through which proteins interact are often highly aggregation prone, so that their engagement in interactions reduces propensity to self-assembly [12]. This means that regions under the evolutionary pressure of protein function may be the same that lead to aberrant

* Corresponding authors: michele.monti@iit.it, Annalisa.Pastore@crick.ac.uk, gian.tartaglia@iit.it

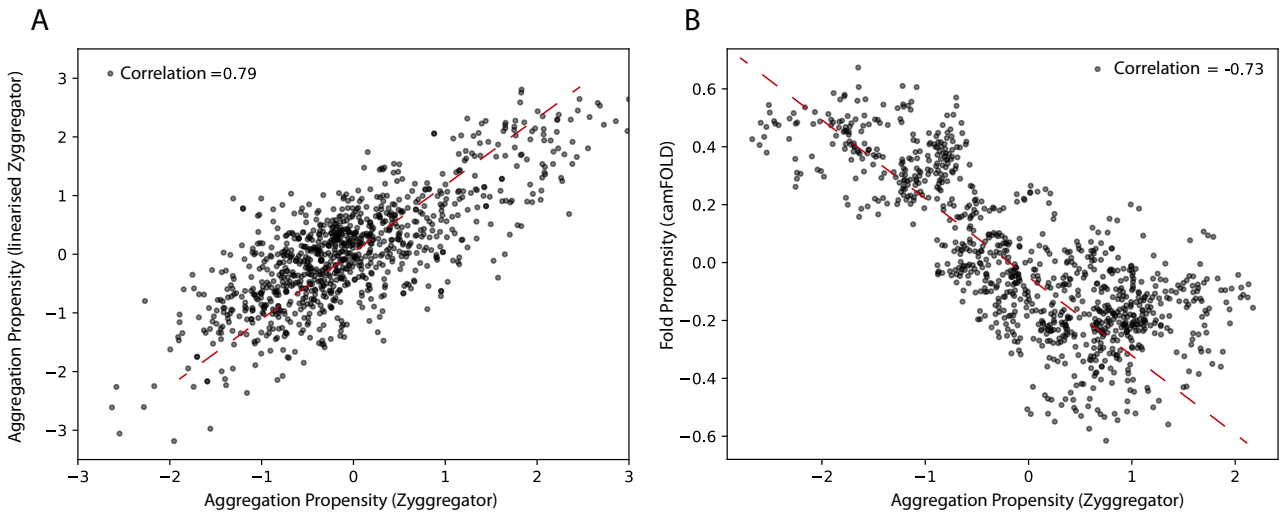

FIG. 1: Aggregation and folding propensity of the human proteome. A) Scatter plot of Zyggregator algorithm and 'linearized' version computed on 5000 random proteins from the human proteome. The correlation among the two methods is 0.79, which justifies the use of the linear approximation of Zyggregator. B) Scatter plot of Zyggregator and camFOLD scores. Folding and aggregation propensities are anti-correlated with a Pearson correlation coefficient of -0.73.

aggregation [13–15]. Any situation, such as mutations or anomalous expression, could result in disruption of these beneficial interactions with consequent aggregation and depletion of normal function.

Expression of a protein in a host different from the natural one is of course an artificial case but represents a prime example in which natural interactions cannot occur. This situation could induce a completely unnatural environment in which proteostasis (or maintainance of protein homeostasis) is impeded [16]: if not promptly degraded, an exogenous protein could establish aberrant interactions and aggregate [3, 17].

To evaluate the effect of exogenous versus endogenous expression, we analysed two large-scale data sets in which variants of an exogenous and an endogenous protein were expressed in different host systems. We focused on molecular evolution data of human TAR DNA-binding protein 43 (TDP-43) exogenously expressed in yeast and bacterial TEM-1 beta-lactamase endogenously expressed in *E. coli*. TDP-43 is involved in various steps of RNA biogenesis and processing, including alternative splicing [18]. TEM-1 beta-lactamase is an enzyme that hydrolyzes the beta-lactam bond in susceptible beta-lactam antibiotics, conferring resistance to penicillins and cephalosporins in bacteria [19]. The mutational library of TDP-43 was designed to introduce mutations specifically in the intrinsically unstructured region near the C-terminus of the protein (290–331 or 332–373). The library of beta-lactamase mutants targeted instead the whole stably structured enzyme.

With these two systems, we investigated whether and how aggregation acts as a constraint of the evolutionary process by finely tuning the selection of proteins to fit in a specific environment [8]. While aggregation leads always to protein depletion, we aimed to determine if the detrimental or beneficial effects sensitively depend on the context and nature of the selection sieve.

II. MATERIAL AND METHODS

A. Random Mutations and Algorithms

In our analyses, we used mutation of TDP-43 and TEM-1 beta-lactamase sequences that have the same length as their respective wild types. Each amino acid was selected for a mutation with the same frequency. Changing the amino acid frequencies from uniform to weighted according to the natural occurrences does not change the results (Supplementary Material). As for the negative sets used in TDP-43 and TEM-1 beta-lactamase analyses, the amount of random mutations was matched with the experimental occurrences of the positive sets.

Predictions were carried out with Zyggregator [2]. This algorithm estimates whether a protein aggregates into solid-like aggregates (fibrillar or protofibrillar structures) on the basis of combinations of physico-chemical properties such as hydrophobicity, helical and beta-propensities [20]. Since the physico-chemical properties of the amino acids change with the environment in which the aggregation processes take place, the coefficients in the Zyggregator algorithm were fitted on a database of experimental rates collected for homogeneous conditions and the predictions were only performed in the vicinity of such conditions. Here, we used a linearized version of Zyggregator [21] to speed up the analysis of the large-scale sets of mutants. By linearization we mean a procedure in which the aggregation propensity was computed considering the amino acid composition of aggregation-prone regions predicted by Zyggregator. This calculation transforms each amino acid and close neighbours (7 amino acids as defined in previous works [2, 21]) in a numerical value to estimate the aggregation propensity.

The camFOLD method [22] was used to calculate the folding propensity profiles of proteins according to a model that exploits the same physicochemical properties exploited by Zyggregator but combined in a different way.

B. Data generation description

The experimental datasets were retrieved from two publications in which error-prone oligonucleotide synthesis was adopted to comprehensively generate the mutational data sets of human TDP-43 [23] and TEM-1 beta-lactamase [24].

Briefly, human TDP-43 was expressed in *S. cerevisiae* [23]. Deep sequencing before and after TDP-43 mutant library induction was used to quantify the relative effects caused by each variant on growth (the input culture was split into two cultures and fitness was calculated from changes of output to input variant frequencies relative to wild type). After quality control and filtering (three replicates of deep mutational scanning experiments, minimum mean input read count threshold of ten) [23], the approach quantified the toxicity of 1,266 single and 56,730 double amino acid changes with high reproducibility.

TEM-1 beta-lactamase was expressed in *E. coli* and twelve libraries of variants were generated [24]. These libraries underwent cell phenotypic selection in a growth functional assay to retain only the variants coding for a functional protein. Differently from TDP-43, only positive fitness variants were sequenced. The number of transformants were in the same order of magnitude of the sequencing capacity of the next-generation sequencing platform (between 100 thousand and 1 million) to guarantee good library complexity. The different generations (or evolutionary cycles) were produced serially, which indicates that the number of mutations increase with respect to the wild-type. The sequencing was performed at the 1st, 5th and 12th generation. While in the 1st generation the dose of antibiotics for phenotypic selection was at 25ug/mL, in the 5th and 12th generation the level was raised to 100 ug/mL to allow stricter selection. In the last generation library a total of 177,253 protein variants with the same length of the wild-type protein (i.e., 286 amino acids without inclusion of premature stop codons) were retrieved[24].
.

C. Data Analysis

For the analysis of the TDP-43 data set, we used both the negative and positive fitness scores measured relatively to the toxicity of wild-type protein in *S. cerevisae*. By contrast, in the TEM-1 beta-lactamase case, we only had variants selected through the evolutionary process and comparison with an antagonistic set was not directly possible. To carry out the analysis, we generated a negative set that was as close as possible to the pool of mutations in the experiments of deep mutational scanning. We tested the randomized negative set using the TDP-43 case: we designed a set of sequences with one or two random mutations depending on the original composition of the pool and performed a comparative analysis. Our results indicated similar performances using a negative or random set, which suggests that they are interchangeable. Once the procedure to generate the random set was tested, we proceeded to build a negative set for the TEM-1 beta-lactamase case. We designed a number of mutations following the rule of keeping the number of mutations present in the positive set. To avoid statistical biases, we analyzed only the sequences from the positive set that have the same length of the wild-type. We analyzed the 1st, 5th and 12th generation sets for which sequencing data were available. To quantify how aggregation changes along the data sets, we computed the Receiver Operating Characteristics (ROC) curves and their relative Area Under the Curve (AUC) for each sub-data set using the corresponding random set as a negative control. We measured the AUC vs the number of mutations keeping track of the evolutionary cycles. Some sub-data sets with specific mutations were poorly populated in the later cycles (as the number of mutations increase in each evolutionary cycle) and

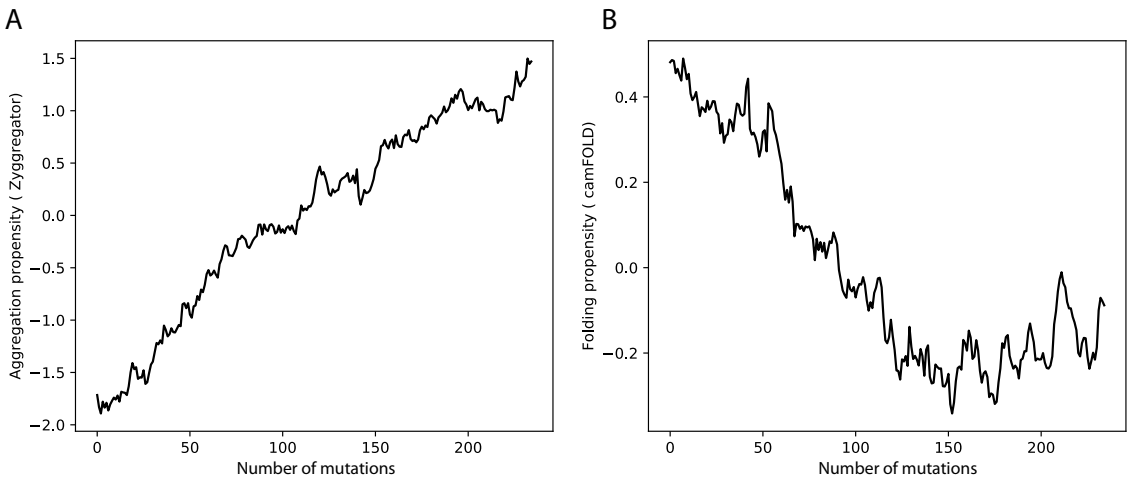

FIG. 2: Aggregation and folding propensities changes upon amino acid mutations in human proteome. A) The aggregation propensity increases while the protein accumulate random mutations. By contrast, B) the folding propensity decreases when the protein is subjected to mutation.

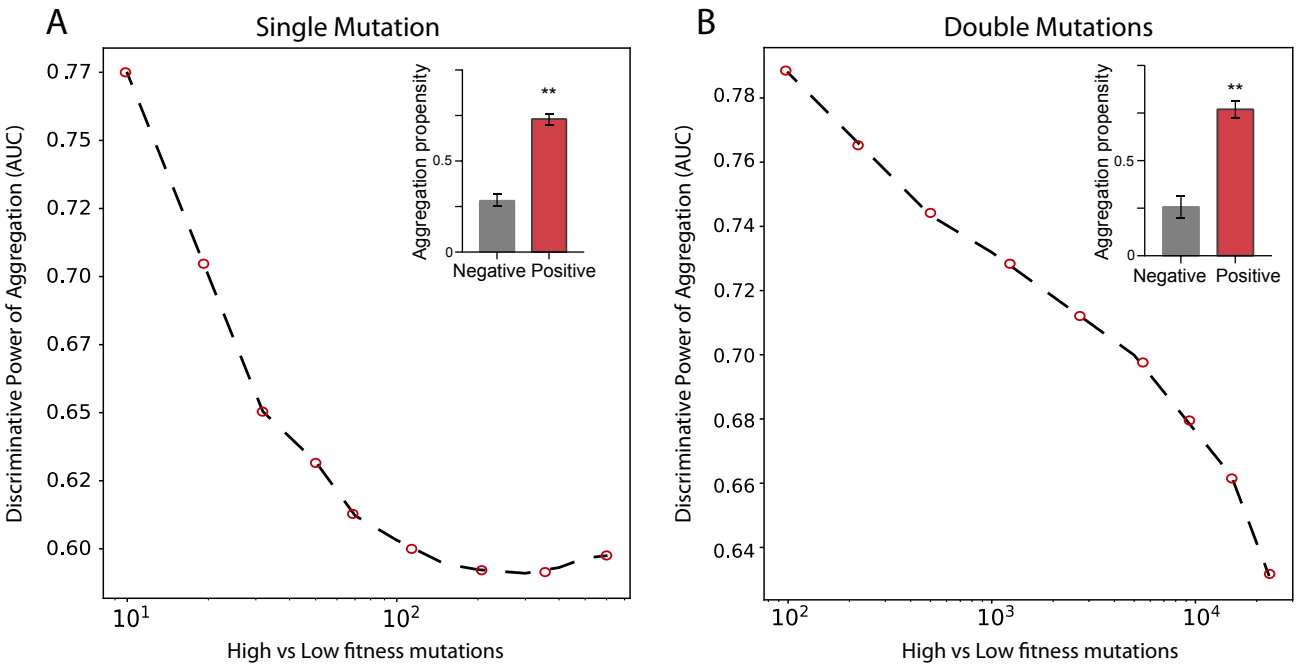

FIG. 3: Deep mutational scanning experiments of TDP-43 mutants expressed in *S. cervisiae*. AUC scores computed on the top and bottom sub-data sets. Panel A and B show the rise of the AUC values as function of the high vs low fitness data sets for single and double mutations, respectively.In the insets, we show the aggregation propensities of mutations associated with the highest (top 100) and lowest (bottom 100) fitness mutations (**, p-value<0.01, Kolmogorov-Smirnoff test)

therefore could not be used for our statistical analysis. To avoid the issue and maintain consistency, we analyzed only sub-data sets that had a size at least 5% that of the largest sub-data set at each evolutionary cycle. We then tested how aggregation and folding propensities changed while a protein accumulated random mutations. We generated a control data set selecting mutations randomly from the 20 natural aminoacids. At each step of evolution, we ran the linearised Zyggregator algorithm to get a score to quantify the aggregation and folding propensities. We did this for all the variants and checked if the sequences increased or decreased their propensities at each random mutation.

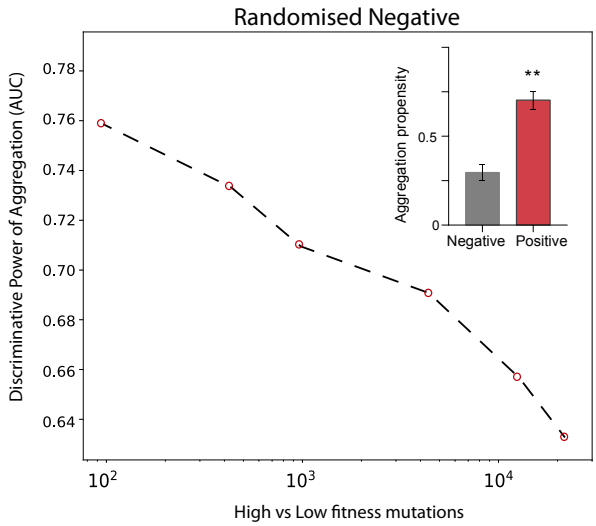

FIG. 4: Deep mutational scanning experiments of TDP-43 mutants expressed in *S. cervisiae* using a random negative set. The AUC values increase progressively with the experimental signal strength. Double mutants are analysed. In the insets, we show the aggregation propensities of mutations associated with the highest (top 100) and random (100 cases) fitness scores (**, p-value<0.01, Kolmogorov-Smirnoff test).

III. RESULTS

A. Aggregation and folding predictions

We first tested if a linearized version of the Zyggregator approach could be used to perform proteome-wide analyses, which comes with the advantage of having faster calculations (see Materials and Methods).

As shown in Figure 1, the linearized version of the algorithm shows a Pearson correlation of 0.78 with the original Zyggregator algorithm. This result indicates that the linearization procedure is able to capture the fundamental features of the Zyggregator algorithm and can be used in our analysis.

In agreement with previous work [22], we found that the Zyggregator and camFOLD scores (see Material and Methods) are functionally related for a large set of proteins taken from the human interactome (Figure 1, panel B; correlation of -0.73). In other words, proteins with a high propensity to fold show poor aggregation propensity.

B. Random mutations increase the aggregation propensity

Can evolution drive sequences to aggregate ? Clearly much depends on the context and functions of the proteins of interest, but it is possible to make some general considerations.

When we mutated sequences in the human proteome, the aggregation propensity rose and the folding propensities decreased (Figure 2, panels A and B). This result indicates that in many cases, including enzymes acting in aqueous solutions [25, 26], protein function requires a stable native fold. This is in agreement with the hypothesis that proteins "live on the edge" due to opposing effects on their sequences caused by the evolutionary pressure to prevent aggregation at the concentrations needed by the cell and the random mutational processes that inherently increase the aggregation propensities [8].

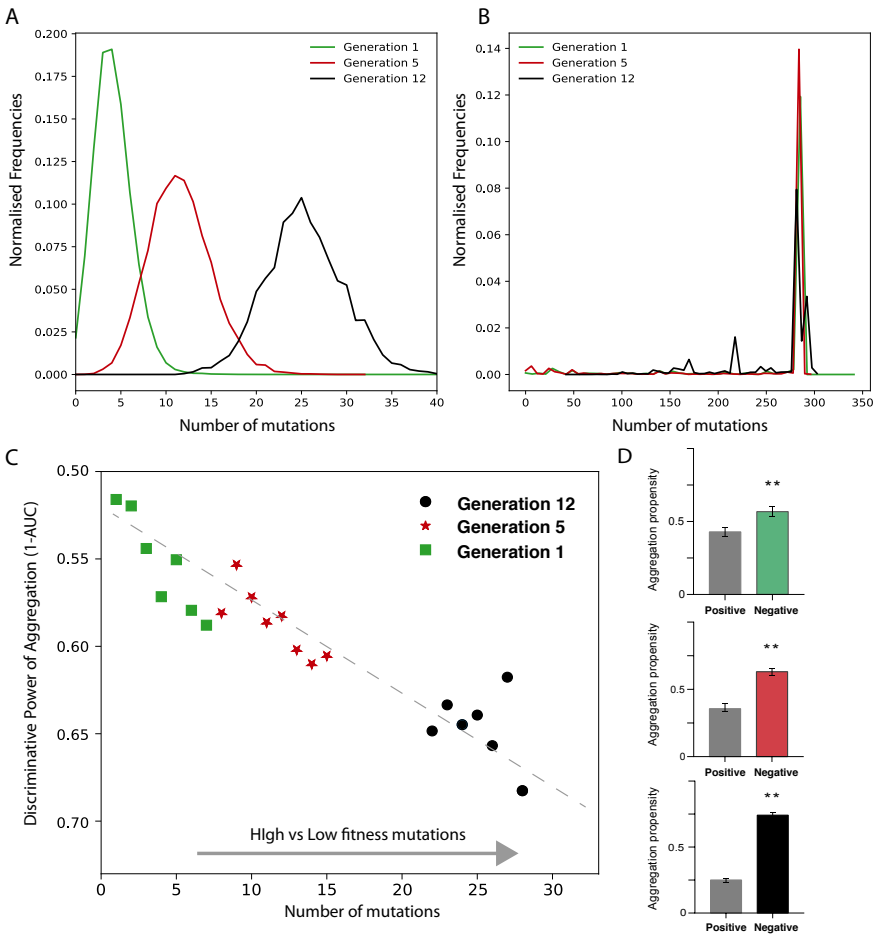

FIG. 5: Deep scanning experiments of TEM-1 beta-lactamase mutants expressed in *E. coli*. Distribution of A) mutation number and B) Distribution of sequence length for different generation of evolution. C) AUC value as a function of high vs low fitness data sets (data computed only for sub data set with at least 1000 data points). D) Aggregation propensities of mutations associated with the highest (top 100) and random (100 cases) fitness scores for generation 1,5, 12 (**, p-value<0.01, Kolmogorov-Smirnoff test). The colors match the points in the panel D.

C. Aggregation as a context-dependent constraint on protein evolution

To address the question on how much the environment influences the role that aggregation has in the evolution of proteins, we analyzed data from two different molecular evolution experiments. In the first experiment, TDP-43 was heterologously expressed in *S. cerevisiae* and the toxicity of variants collected. In the second experiment, variants of TEM-1 beta-lactamase were selected in *E. coli* while antibiotics were given to the cell culture.

In this analysis we used the ROC analysis to measure the strength of the link between the toxicity and the aggregation propensity [2]. The ROC is highly appropriate as it allows to simultaneously analyse the true and false positive rates.

Cell fitness and aggregation propensity of TDP-43 mutants

In the case of TDP-43 expression in *S. cerevisiae*, we divided the experimental sets in two partitions of equal size corresponding to the highest (positive) and lowest (negative) fitness values associated with the mutants. As shown in Figure 3, the Area Under the ROC Curve (AUC) rises progressively with the high vs low fitness mutations score. Our

result indicates that the aggregation is an excellent predictor of fitness, especially for those mutants that have the strongest experimental signal (AUC of 0.78 or more). In the insets of Figure 3 we show statistics for the aggregation propensities of single and double mutations associated with the highest and lowest fitness scores (corresponding to AUCs of 0.78 and 0.77, respectively).
We then swapped the negative set with a list of mutants generated randomly (Figure 4; Materials and Methods). We observed that the AUC increases monotonically with the signal strength encoded in the aggregation propensity of the high fitness score mutants (statistics of the mutations associated with the 100 highest vs 100 random mutations are shown in Figure 4B). This result indicates that the negative and random sets are interchangeable for our analyses.

Cell fitness and aggregation propensity of TEM-1 beta-lactamase mutants

In the case of TEM-1 beta-lactamase in *E. coli*, only the positive fitness set was available from the experiments. To generate a negative set following the approach used for TDP-43, we first analysed the number of mutations and sequence lengths of variants. We observed that the number of mutations increases on average with the generations (Figure 5, panels A and B). Focusing on the variants that have the same length as the reference TEM-1 beta-lactamase, we generated the negative set respecting the frequencies of mutations occurring at each cycle. In Figure 5C we show the values corresponding to the AUC computed using the aggregation propensities for the high (positive) vs low (negative) fitness score mutants. There is a clear trend towards a decrease of the AUC score (i.e., increase of 1-AUC) with the number of mutations and evolutionary cycles (generation). Thus, differently from TDP-43, we observed that the aggregation propensity decreases. In Figure 5D we show that the aggregation propensity of mutants associated with high fitness score decreases progressively through evolutionary cycles. In this analysis, we selected variants by ranking the number of mutations occurring at each evolutionary cycle and compared their aggregation propensities with their random counterpart (100 sequences for each set; the AUCs are 0.6, 0.65 and 0.70, respectively).

IV. DISCUSSION

This study aimed to investigate the influence of protein aggregation on molecular evolution. Mutational scanning approaches have been previously used in the literature and demonstrated to provide a powerful tool to investigate the ability of a protein to aggregate and cause toxicity [27]. We also have previously observed that, under physiological conditions, expression levels are fine-tuned to the specific protein solubilities [8]. This finding suggests that mutations altering the aggregation propensity can impact cell fitness and thus influence evolution [7, 8].

The computational analysis presented in this paper revealed that aggregation and folding propensities anti-correlate at the proteomic level. These results agree with and expand studies that had suggested similar conclusions from completely different premises [22, 28]. Our data indicate that, for the large majority of cases, naturally-occurring protein sequences have been optimized by evolution to have lower aggregation and higher folding propensities [14]. Our calculations predict that random mutations render proteins more prone to aggregate at the expenses of their folding ability, even though this variability could help exploring new functionalities depending on the cellular context.

To estimate the impact of aggregation in different cellular contexts, we analysed molecular evolution experiments in which human TDP-43 and TEM-1 beta-lactamase mutants were expressed in *S. cerevisiae* [23] and in *E. coli* [24] respectively. Notably, TDP-43 contains $> 30\%$ of structural disorder in the C-terminus [3], whereas TEM-1 beta-lactamase has a well-defined native state [24]. When prion-like regions, as in the case of TDP-43 [23], are not protected by the native fold, aggregation takes place because the exposure to the solvent favours self-assembly [2, 3, 29]. For TDP-43 mutants, both negative and positive fitness scores were directly compared with the aggregation propensities. By contrast, for TEM-1 beta-lactamase, we had only positive fitness variants selected through the evolutionary process, which made the comparison with a negative set not directly possible. To carry out the analysis, we artificially generated a negative set as close as possible to the experimental one.

Our analysis revealed two quite different behaviors. With TEM-1 beta-lactamase, the system was under antibiotic selection pressure. Thus, the protein had to remain active to protect against penicillins and avoid aggregation. The variants were in fact selected by evolution to maintain enzyme activity and thus antibiotics resistance while minimizing aggregation [24]. While the negative set maintains a relative high aggregation propensity through the evolutionary cycles, the positive set significantly reduces it, showing existence of a purifying pressure

. Interestingly, it has been previously observed that M180T, E195D, L196I, and S281T were the most frequent mutations in the system. These residues are all exposed and their mutations should, with the exception of the well documented M180, only have minor effects on protein fold as indicated by visual inspection and semi-quantitative analysis with the DUET software, a web server for an integrated computational approach to study missense mutations in proteins [30]. Thus, these mutations would represent a "neutral" passage that do not impinge on fitness.

With TDP-43, we observed the opposite trend. We have recently stressed the importance of protein interactions with chaperones to prevent aggregation [10, 13, 15]. Expression of a protein in a heterologous system will lead by definition to a situation in which natural interactions cannot occur [3]. Accordingly, we observed that TDP-43 has a high toxicity in yeast, as expected for a protein that has a strong tendency to aggregate also in its natural host [23]. In this case, we observed that mutations that increased hydrophobicity and aggregation also decreased toxicity by eliminating from solution proteins that could not have functional partners. This hypothesis would explain why solid aggregates formed by TDP-43 A238V, Q360Y and G335I mutants are non-toxic, while the soluble W334K, M332K and A328P have a higher toxicity, as described in Bolognesi ed al. [23]. This would also be consistent with the generally accepted idea that soluble aggregates are more toxic than the insoluble end-points [31], although in this work we did not analyse liquid-liquid phase separations.

It is worth mentioning that the TDP-43 mutational library was designed to target a specific region of the protein, the intrinsically unstructured C-terminus. This meant no evolution pressure to preserve structure. Accordingly, our results showed that the aggregation propensity of TDP-43 increased with the fitness score.

The two scenarios indicate a clear ambivalence that protein aggregation can acquire depending on the system in which it occurs and suggest aggregation as an important constraint for evolution. Our approach demonstrates how different selective pressures may completely change the effect of aggregation on a specific system: an exogenous protein with no functional role, such as TDP-43 in *S. cerevisiae* is less toxic if compartimentalised in an aggregate, while an endogenous proteins, such as TEM-1 beta lactamase in *E. coli*, is functional when aggregation is avoided. Further analyses will be important to shed light onto the selective pressure against aggregation in specific cellular environments. Indeed, it has been previously suggested that the volume of the organelles imposes stringent limitations on protein evolution to avoid high aggregation propensities in confined compartments [32, 33]. This has implications for protein variability in different organisms [34]. In line with this observation, it should be mentioned that poly-anionic molecules such as RNA can reduce aggregation propensity [35] and [36]. On the contrary, mutations that abolish RNA-binding in these proteins can force them to change their cellular localization, accumulate and aggregate. This is the case of FUS and TDP-43 [35, 37]. Making large-scale predictions of protein aggregation can reveal important properties of biological systems, as demonstrated by the elegant study by Khodaparast et al. [38] who showed that identifying insoluble regions of a protein can be instrumental to block the process of bacterial infection.

In the future, it will be interesting to investigate the role of specific molecular partners, such as RNA, in regulating aggregation. Computational approaches such as the one presented in this study [39, 40] will be key to achieve a more complete understanding of molecular evolution, and to increase our ability to manipulate proteins for specific purposes.

V. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Mattia Miotto, Dr Lorenzo Di Rienzo, Dr Tommaso Muto and Dr. Claudia Giambartolomei for discussions. The research leading to these results was supported by the Dementia Research Institute (REI 3556) and AlzUK (ARUK-PG2019B-020), European Research Council (RIBOMYLOME 309545 and ASTRA 855923), the H2020 projects IASIS 727658 and INFORE 25080, the Spanish Ministry of Economy and Competitiveness BFU2017-86970-P as well as the collaboration with Peter St. George-Hyslop financed by the Wellcome Trust.

[1] C. M. Dobson, Cold Spring Harbor perspectives in biology **9**, a023648 (2017), ISSN 1943-0264, URL `https://pubmed.ncbi.nlm.nih.gov/28062560https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5453392/`.

[2] G. G. Tartaglia, A. P. Pawar, S. Campioni, C. M. Dobson, F. Chiti, and M. Vendruscolo, Journal of Molecular Biology **380**, 425 (2008), ISSN 0022-2836, URL https://www.sciencedirect.com/science/article/pii/S0022283608005676.

[3] N. L. Gotor, A. Armaos, G. Calloni, M. Torrent Burgas, R. Vabulas, N. S. De Groot, and G. G. Tartaglia, Nucleic Acids Research **48**, 9491 (2020), ISSN 0305-1048, URL https://doi.org/10.1093/nar/gkaa681.

[4] B. Bolognesi, N. L. Gotor, R. Dhar, D. Cirillo, M. Baldrighi, G. G. Tartaglia, and B. Lehner, Cell Reports **28**, 222 (2016).

[5] G. De Baets, L. Van Doorn, F. Rousseau, and J. Schymkowitz, PLOS Computational Biology **11**, 1 (2015), URL https://doi.org/10.1371/journal.pcbi.1004374.

[6] T. Murakami, S.-P. Yang, L. Xie, T. Kawano, D. Fu, A. Mukai, C. Bohm, F. Chen, J. Robertson, H. Suzuki, et al., Human Molecular Genetics **21**, 1 (2011), ISSN 0964-6906, https://academic.oup.com/hmg/article-pdf/21/1/1/17254842/ddr417.pdf, URL https://doi.org/10.1093/hmg/ddr417.

[7] G. Vecchi, P. Sormanni, B. Mannini, A. Vandelli, G. G. Tartaglia, C. M. Dobson, F. U. Hartl, and M. Vendruscolo, Proceedings of the National Academy of Sciences **117**, 1015 (2020), ISSN 0027-8424, https://www.pnas.org/content/117/2/1015.full.pdf, URL https://www.pnas.org/content/117/2/1015.

[8] G. Tartaglia, S. Pechmann, C. Dobson, and M. Vendruscolo, Trends in biochemical sciences **32 5**, 204 (2007).

[9] D. M. Fowler, A. Koulov, W. Balch, and J. Kelly, Trends in biochemical sciences **32 5**, 217 (2007).

[10] A. Louka, E. Zacco, P. A. Temussi, G. G. Tartaglia, and A. Pastore, Nucleic Acids Research **48**, 11880 (2020), ISSN 0305-1048, https://academic.oup.com/nar/article-pdf/48/21/11880/34658676/gkaa822.pdf, URL https://doi.org/10.1093/nar/gkaa822.

[11] E. Zacco, R. Graña-Montes, S. R. Martin, N. S. de Groot, C. Alfano, G. G. Tartaglia, and A. Pastore, Journal of Molecular Biology **431**, 1671 (2019), ISSN 0022-2836, URL https://www.sciencedirect.com/science/article/pii/S0022283618310854.

[12] S. Pechmann, E. D. Levy, G. G. Tartaglia, and M. Vendruscolo, Proceedings of the National Academy of Sciences **106**, 10159 (2009), ISSN 0027-8424, https://www.pnas.org/content/106/25/10159.full.pdf, URL https://www.pnas.org/content/106/25/10159.

[13] L. Masino, G. Nicastro, L. Calder, M. Vendruscolo, and A. Pastore, The FASEB Journal **25**, 45 (2011), ISSN 0892-6638, URL https://doi.org/10.1096/fj.10-161208.

[14] Y. Lee, T. Zhou, G. G. Tartaglia, M. Vendruscolo, and C. O. Wilke, PROTEOMICS **10**, 4163 (2010), https://onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.201000229, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.201000229.

[15] P. A. Temussi, G. G. Tartaglia, and A. Pastore, Bioessays **in press** (2021).

[16] G. G. Tartaglia, C. M. Dobson, F. U. Hartl, and M. Vendruscolo, Journal of Molecular Biology **400**, 579 (2010), ISSN 0022-2836, URL https://www.sciencedirect.com/science/article/pii/S0022283610003773.

[17] J. Beerten, S. Joost, and R. Frederic, *Aggregation Prone Regions and Gatekeeping Residues in Protein Sequences* (2012), URL http://www.eurekaselect.com/node/107892/article.

[18] A. Bhardwaj, M. P. Myers, E. Buratti, and F. E. Baralle, Nucleic Acids Research **41**, 5062 (2013), ISSN 0305-1048, URL https://doi.org/10.1093/nar/gkt189.

[19] R. A. Bonomo, Cold Spring Harbor perspectives in medicine **7**, a025239 (2017), ISSN 2157-1422, URL https://pubmed.ncbi.nlm.nih.gov/27742735https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5204326/.

[20] G. G. Tartaglia and M. Vendruscolo, Chem. Soc. Rev. **37**, 1395 (2008), URL http://dx.doi.org/10.1039/B706784B.

[21] P. Klus, B. Bolognesi, F. Agostini, D. Marchese, A. Zanzoni, and G. G. Tartaglia, Bioinformatics **30**, 1601 (2014), ISSN 1367-4803, URL https://doi.org/10.1093/bioinformatics/btu074.

[22] G. G. Tartaglia and M. Vendruscolo, Journal of Molecular Biology **402**, 919 (2010), ISSN 0022-2836, URL https://www.sciencedirect.com/science/article/pii/S0022283610008648.

[23] B. Bolognesi, A. J. Faure, M. Seuma, J. M. Schmiedel, G. G. Tartaglia, and B. Lehner, Nature Communications **10**, 4162 (2019), ISSN 2041-1723, URL https://doi.org/10.1038/s41467-019-12101-z.

[24] M. Fantini, S. Lisi, P. De Los Rios, A. Cattaneo, and A. Pastore, Molecular Biology and Evolution **37**, 1179 (2020), ISSN 0737-4038, URL https://doi.org/10.1093/molbev/msz256.

[25] Y.-H. Ho and A. P. Gasch, Current genetics **61**, 503 (2015), ISSN 1432-0983, URL https://pubmed.ncbi.nlm.nih.gov/25957506https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4800808/.

[26] J. Saarikangas and Y. Barral, Current Genetics **62**, 711 (2016), ISSN 1432-0983, URL https://doi.org/10.1007/s00294-016-0596-0.

[27] L. M. Luheshi, G. G. Tartaglia, A.-C. Brorsson, A. P. Pawar, I. E. Watson, F. Chiti, M. Vendruscolo, D. A. Lomas, C. M. Dobson, and D. C. Crowther, PLOS Biology **5**, 1 (2007), URL https://doi.org/10.1371/journal.pbio.0050290.

[28] T. R. Jahn and S. E. Radford, Archives of Biochemistry and Biophysics **469**, 100 (2008), ISSN 0003-9861, URL https://www.sciencedirect.com/science/article/pii/S0003986107002731.

[29] R. Linding, J. Schymkowitz, F. Rousseau, F. Diella, and L. Serrano, Journal of Molecular Biology **342**, 345 (2004), ISSN 0022-2836, URL https://www.sciencedirect.com/science/article/pii/S0022283604007715.

[30] D. E. V. Pires, D. B. Ascher, and T. L. Blundell, Nucleic Acids Research **42**, W314 (2014), ISSN 0305-1048, URL https://doi.org/10.1093/nar/gku411.

[31] H. A. Lashuel and P. T. Lansbury, Q Rev Biophys. **39**, 167 (2006).

[32] G. T. GG, R. Pellarin, A. Cavalli, and A. Caflisch, Protein Science **10**, 2735 (2005).

[33] G. G. Tartaglia and M. Vendruscolo, Mol. BioSyst. **5**, 1873 (2009), URL http://dx.doi.org/10.1039/B913099N.

[34] G. G. Tartaglia, R. Pellarin, A. Cavalli, and A. Caflisch, Protein Science **14**, 2735 (2005), ISSN 0961-8368, URL https://doi.org/10.1110/ps.051473805.

[35] H.-J. Chen, S. D. Topp, H. S. Hui, E. Zacco, M. Katarya, C. McLoughlin, A. King, B. N. Smith, C. Troakes, A. Pastore, et al., Brain **142**, 3753 (2019), ISSN 0006-8950, URL https://doi.org/10.1093/brain/awz313.

[36] N. Sanchez de Groot, A. Armaos, R. Graña-Montes, M. Alriquet, G. Calloni, R. M. Vabulas, and G. G. Tartaglia, Nature Communications **10**, 3246 (2019), ISSN 2041-1723, URL https://doi.org/10.1038/s41467-019-10923-5.

[37] L. Marrone, H. C. A. Drexler, J. Wang, P. Tripathi, T. Distler, P. Heisterkamp, E. N. Anderson, S. Kour, A. Moraiti, S. Maharana, et al., Acta Neuropathologica **138**, 67 (2019), ISSN 1432-0533, URL https://doi.org/10.1007/s00401-019-01998-x.

[38] L. Khodaparast, L. Khodaparast, R. Gallardo, N. N. Louros, E. Michiels, R. Ramakrishnan, M. Ramakers, F. Claes, L. Young, M. Shahrooei, et al., Nature Communications **9**, 866 (2018), ISSN 2041-1723, URL https://doi.org/10.1038/s41467-018-03131-0.

[39] N. Louros, G. Orlando, M. De Vleeschouwer, F. Rousseau, and J. Schymkowitz, Nature Communications **11**, 3314 (2020), ISSN 2041-1723, URL https://doi.org/10.1038/s41467-020-17207-3.

[40] F. Agostini, D. Cirillo, C. M. Livi, R. Delli Ponti, and G. G. Tartaglia, Bioinformatics **30**, 2975 (2014), ISSN 1367-4803, URL https://doi.org/10.1093/bioinformatics/btu420.