# Investigation of factors influencing routes of *in vivo* protein aggregation in *Escherichia coli*

**Ádám Györkei**

**Summary of the PhD thesis**


**Supervisors: Dr. Balázs Papp and Dr. Bálint Kintses**
**Biológia Doktori Iskola**

**ELKH Biological Research Center of Szeged**
**University of Szeged, Faculty of Science and Informatics, Ph. D. School of Biology**
**2022**
**Szeged**

# Summary

In the past decades, the notion that proteins inside aggregates are macromolecules lacking their native structure and function has been widely accepted. This was partly due to the circumstances of the discovery of protein aggregates and their pivotal role in human diseases. However, in recent years, more and more results were presented showing that protein aggregation may play an important physiological role in both bacteria and eukaryotes. Later, protein aggregates with enzymes retaining their native activity were identified. This prompted a major change in how we view protein aggregation. Despite the mounting evidence of the existence of aggregates with proteins maintaining their native or native-like structure with activity and intense investigations of proteins in regards to foring aggregates, no systematic analyses were conducted in order to identify the underlying factors for their formation.

My doctoral work focused on systematically investigating the aggregation pathways of overexpressed *E. coli* proteins using an image-based *in vivo* aggregation assay coupled with machine learning. We have confirmed the presence of two characteristically different phenotypes of protein aggregates. Furthermore, we identified on a proteome-wide scale those proteins that form inclusion bodies slowly, after folding and those that aggregate quickly, before reaching the folded state. We find that the frequency of the two types of aggregates are comparable on a proteome scale, however, their subcellular localization differs substantially. While dark aggregates are mainly proteins present in the membrane under

normal circumstances, the majority of aggregation-prone cytoplasmic proteins form aggregates slowly and after folding. We have collected a large-scale, systematic dataset of protein parameters in order to identify the forces affecting the fate of an aggregation-prone protein. We see that compared to proteins that aggregate before adopting a folded state, fluorescent aggregate proteins exhibit faster folding kinetics and lower aggregation kinetics. Interaction with the chaperone network, especially the *DnaK* chaperone also helps to maintain the folded state in the translation milieu before, and most likely after aggregation. Our findings indicate that rapid folding and slow aggregation of a protein allows its folding before aggregation. Overall, these results support the notion that the formation of active, native-like aggregates compared to inactive ones is governed by the kinetic competition between folding and aggregation. Several lines of evidence also reinforce the notion that these principles operate in proteins expressed at their expression levels and in their native genetic contexts. For example, under normal expression levels, the chaperon system plays a fundamental role in preventing aggregation of newly synthesized as well as folded proteins. Thus, two classes of DnaK client proteins have been identified, depending on whether the chaperone aids in its folding after translation, or maintains its native conformation. Additionally, while the proximity of the translational apparatus is highly enriched in nascent polypeptides susceptible to aggregation during translation, some proteins with their native structure intact has been shown to aggregate when their concentration exceeds the critical concentration of their solubility, a phenomenon termed supersaturation. Proteins showing supersaturation have shown a substantial enrichment in

biological processes related to neurodegenerative diseases, suggesting their role in these pathologies, and others related to aging and stress. Future studies deciphering the links between the aggregation route taken by proteins upon overexpression and environmental changes are thus strongly warranted.

It has been established, that natively abundant proteins have evolved a substantially low surface stickiness in order to avoid non-specific interactions that would result in interference with other proteins (Levy et al., 2012). Our work has demonstrated that proteins with low surface stickiness are also depleted in both proteins forming aggregates through either pathway. Based on these results, we propose that avoidance of aggregation is the driving force behind the reduction of non-specific interaction propensity of natively highly expressed proteins. Interestingly, we find that dark aggregate proteins have a distinctly higher surface stickiness than fluorescent aggregate proteins, however, no such difference can be found for amino acids in the protein core. This suggests that aggregation occurs before the GFP fusion protein adopts its native structure, but after the fast collapse of the nascent polypeptide chain into a structure resembling a partially folded, compact intermediate that already has similar surface characteristics of the native protein (Hartl és Hayer-Hartl, 2009; Kim et al., 2013). Further investigations will most probably shed some light on the sequence of events leading to the aggregation of such proteins.

Our work also provides several new insights into the role protein disorder plays in aggregation. First, we show that protein *in vivo* solubility in *E. coli* is strongly driven by protein disorder content, and this effect is

independent of surface stickiness. Given that the association between protein disorder and solubility is also independent of both charge and hydrophobicity, we reinforce the notion that the flexibility provided by the disordered residues enhances solubility (Santner et al., 2014; Simone et al., 2012). Disordered protein segments have been shown to act as enthropic bristles and providing favourable surface area for protein solubility. It is important to note that soluble proteins also contain a larger frequency of disordered segments, supporting their positive enthropic effects. Also, while proteins of two aggregation classes do not differ in the frequency of disordered segments, dark aggregates contain more disordered residues outside these segments. Future studies should shed more light on the particular effects of disordered residues and segments and their contribution to the maintenance of solubility or contribution to aggregation. Second, or results shed a new light on how natural selection affects protein disorder on a proteome scale. Although it has been shown that highly expressed proteins have a higher disorder content in *E. coli* (Tartaglia et al., 2009), the underlying mechanisms have been largely unidentified so far. Our findings suggest that the increased disordered content of highly expressed proteins is a means of avoiding aggregation under physiological conditions. Thus, protein disorder and low surface stickiness are both adaptations to maintain high solubility of abundant proteins in *E. coli* and possibly in other bacteria. And third, our results highlight the substantial difference of protein disorder on solubility in prokaryotes and eukaryotes. Eukaryotes are known to contain substantially higher amounts of disordered proteins as well as longer disordered segments. This is due to the role of these proteins in signaling (Peng et

al., 2015) and other cellular functions requiring diverse and transient protein interactions. Indeed, the presence of intrinsically disordered regions have been associated with stressinduced aggregation in humans (Määttä et al., 2020). Also, contrarily to *E. coli*, highly expressed yeast proteins contain significantly lower levels of intrinsically disorder, potentially to reduce non-native, non-specific interactions that are mediated by linear motifs (MacossayCastillo et al., 2019). These difference can be attributed to either the proteins themselves, the cellular milieu or a combination of both. However, our results show that human proteins expressed in *E. coli* adhere to positive the correlation between solubility and disorder content, suggesting that the difference can be attribute to differences in the cellular context. The underlying mechanisms for such differences would be a great benefit to our understanding of both protein disorder and protein aggregation, as well as providing an important framework for future large scale studies.

Our results have may also have substantial biotechnological implications. Extraction of functional proteins from inclusion bodies as well as their other applications may offer novel solutions and design principles (Rinas et al., 2017; Singhvi et al., 2019; Villaverde és mtsai., 2012; Wu et al., 2011) to the protein production pipeline. Native-like enzymes can be applied as immobilized biocatalysts (Rinas et al., 2017; Wu et al., 2011) in industrial applications. Inclusion bodies may also function as drug delivery scaffolds in medical applications (Villaverde et al., 2012), with the benefit of greater stability and longer shelflife. We show that proteins of all enzyme classes tend to form aggregates after folding, suggesting the wide applicability of designed protein aggregates in industrial use. From a

methodological standpoint, high-content microscopy has proven to be a higly scalable, efficient and robust way of identifying aggregates with active proteins, with great protential in screening of recombinant proteins for industrial applications. It is important to mention, that some proteins in active aggregates may in time convert to misfolded, inactive forms (Elia et al., 2017), thus future investigations must put an emphasis on the forces that help maintain protein activity inside aggregates and the exact proportion of active and inactive proteins inside aggregates. Our high-content microscopy and machine learning method, combined with high-throughput protein purification (Jäger et al., 2020) and functional profiling(Huang et al., 2015; Kuznetsova et al., 2006) may provide and efficient way of unlocking the potential in active protein aggregates.

## Aims:

- Establishment of the frequency of the different protein aggregate phenotypes *in vivo*, comparison to previous *in vitro* findings.
- Investigation of the „Life on the edge" hypothesis, and if proteins are indeed at their solubility limits under native conditions.
- Establishment of the effectiveness of high-content microscopy and machine learning in the investigation of *in vivo* protein aggregation.
- Identification of factors influencing the formation of the different types of aggregates, understanding the mechanisms behind this process.
- Establishment of the role of kinetic competition in determining the type of protein aggregate formed.
- Identification of factors enhancing in vivo solubility, with particular focus on protein structure and surface.

- Expansion of our findings to heterologously expressed proteins.

## Methods used:

- Cell culturing (ASKA library)
- High-throughput fluorescent single-cell microscopy
- Image analysis with machine learning methods
- Experimental valudation of gene expression: gel electrophoresis and western blot
- Bioinformatic analysis of protein structures (Areaimol, CCP4)
- Webcrawlers (Perl)
- Statistical and regression analyses (R)
- Database creation and management (R and perl)

## Most important findings of the thesis:

1.) Upon overexpression, the majority of proteins (approximately 70%) form aggregates *in vivo*. These aggregates are either formed by proteins with native-like or misfolded structures with roughly equal frequencies. However, a significant fraction of proteins (30%) remain soluble despite overexpression, contradicting the notion that proteins are at their solubility limits under native conditions in the cell.

2.) The list of aggregation-prone proteins show a significant overlap with previous *in vitro* results, although our results show a more detailed picture. Given the results of our validation experiments, high-content microscopy and machine learning is a fast and efficient method to investigate protein

aggregation while also providing information on the folding state of the protein.

3.) Many features influencing protein aggregation were previously established and our results confirmed most of them *in vivo*. These include native expression levels, negative charge and hydrophobicity. In addition we found significant aggregation inhibiting effect of low promiscuous surface interactions due to low protein surface stickiness and high protein disorder content. The effects of the latter were previously contradictory in the literature, however, our results show a clear picture in the case of *E. coli*.

4.) The relative speed of protein aggregation and folding were found the be the most important determinants of the type of protein aggregate produced. This is consistent with the kinetic competition hypothesis, which was confirmed *in vivo* on a proteome level. Additionally surface stickiness was shown to play a role for partially folded proteins.

5.) We confirmed the solubilty enhancing properties of protein disorder for heterologously expressed proteins. This suggests that the contradictory role of protein disorder for protein aggregation may be the result of organism-specific effects.

## List of publications related to the thesis:

1. Györkei, Ádám; Daruka, Lejla; Balogh, Dávid; Őszi, Erika; Magyar, Zoltán; Szappanos, Balázs; Fekete, Gergely; Fuxreiter, Mónika; Horváth, Péter; Pál,

Csaba; Kintses, Bálint; Papp Balázs Proteome-wide landscape of solubility limits in a bacterial cell
SCIENTIFIC REPORTS (accepted, under publication, 2022)

2. Kintses, Bálint; Méhi, Orsolya; Ari, Eszter; Számel, Mónika; Györkei, Ádám; Jangir, Pramod K; Nagy, István; Pál, Ferenc; Fekete, Gergely; Tengölics, Roland et al.
Phylogenetic barriers to horizontal transfer of antimicrobial peptide resistance genes in the human gut microbiota.
NATURE MICROBIOLOGY 4 : 3 pp. 447-458. , 12 p. (2019) MTMT: [30435652] IF: 9.68