A Genome-Wide Sequence–Structure Analysis Suggests Aggregation Gatekeepers Constitute an Evolutionary Constrained Functional Class

Greet De Baets [1,2,3], Joost Van Durme [1,2,3], Frederic Rousseau [1,2] and Joost Schymkowitz [1,2]

*1 - **Switch Laboratory,** Flanders Institute for Biotechnology (Vlaams Instituut voor Biotechnologie), 3000 Leuven, Belgium*
*2 - **Switch Laboratory,** Department of Cellular and Molecular Medicine, University of Leuven, Herestraat 49, 3000 Leuven, Belgium*
*3 - **Vrije Universiteit Brussel,** Pleinlaan 2, 1050 Brussels, Belgium*

***Correspondence to Frederic Rousseau and Joost Schymkowitz:** Switch Laboratory, Department of Molecular Cell Biology, University of Leuven, B-3000 Leuven, Belgium. Frederic.Rousseau@switch.vib-kuleuven.be; Joost.Schymkowitz@switch.vib-kuleuven.be*
http://dx.doi.org/10.1016/j.jmb.2014.04.007
Edited by C. L. Brooks

Abstract

Protein aggregation is geared by aggregation-prone regions that self-associate by β-strand interactions. Charged residues and prolines are enriched at the flanks of aggregation-prone regions resulting in decreased aggregation. It is still unclear what drives the overrepresentation of these "aggregation gatekeepers", that is, whether their presence results from structural constraints determining protein stability or whether they constitute a *bona fide* functional class selectively maintained to control protein aggregation. As functional residues are typically conserved regardless of their cost to protein stability, we compared sequence conservation and thermodynamic cost of these residues in 2659 protein families in *Escherichia coli*. Across protein families, we find gatekeepers to be under strong selective conservation while at the same time representing a significant thermodynamic cost to protein structure. This finding supports the notion that aggregation gatekeepers are not structurally determined but evolutionary selected to control protein aggregation.

Introduction

Protein aggregation is mediated by short aggregation-prone regions (APRs) within protein sequences that can self-assemble by intermolecular β-strand interactions to form aggregates of misfolded proteins [1,2]. Whereas APRs are generally buried in the hydrophobic core of native globular proteins thereby precluding aggregation, they are solvent exposed and prone to aggregate in situations of physiological stress or during protein translation, trafficking or degradation requiring tight regulation by molecular chaperones [3].

Proteome-wide analysis of the aggregation propensity of 28 organisms revealed that the flanks of APRs are enriched in charged amino acids and prolines [4]. This observation was further confirmed in other studies [5–10]. These residues, termed aggregation gatekeepers, oppose aggregation by introducing charge repulsion (R, K, D and E), unfavorable entropic contribution by side-chain immobilization (especially K and R), or in the case of proline main-chain entropic destabilization of the β-strand conformation of aggregates [11–13].

In addition to their structural effect as aggregation breakers, gatekeeper residues also appear to contribute to chaperone binding in *Escherichia coli* [4,8,14–17]. Finally, gatekeepers also seem to influence protein abundance in *E. coli* by modifying protein synthesis and degradation rates [18]. Interestingly, the frequency of gatekeeper occurrence is significantly higher in groups of proteins with an essential cellular function [7,19,20], further supporting their role in maintaining proteostasis. In keeping with this notion, mutations that alter gatekeeper residues occur much more frequently in disease-associated mutations than in polymorphisms [8].

Despite the wealth of data showing that aggregation gatekeepers affect protein solubility, chaperone binding and the proteostatic regulation of proteins, it

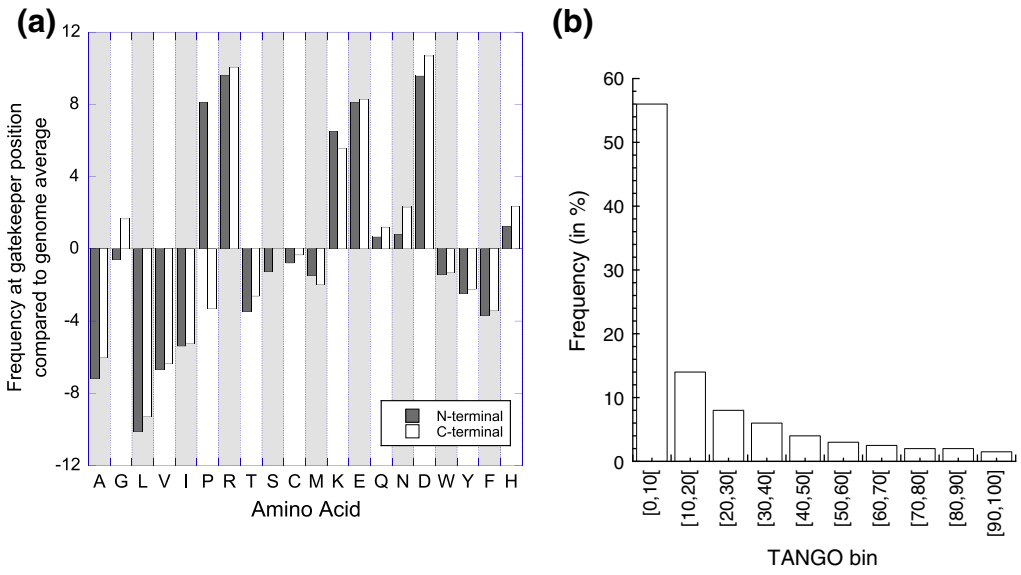

Fig. 1. Aggregation analysis of the reference strain K12 MG1655. (a) Difference plot of amino acid composition at the flank of APRs compared with average proteome composition. Values above or below 0 denote an increase or decrease in frequency, respectively. (b) Histogram showing the TANGO score distribution of APRs in the proteome. A higher TANGO score indicates a higher aggregation propensity.

is still unclear whether the enrichment of gatekeepers at flanks of APR is a selective process resulting from functional constraints or whether gatekeeper enrichment is a secondary effect subordinate to protein structural constraints. It is generally assumed that the specific placement of these aggregation-opposing residues at the flanks of APRs results from the physical impossibility to accommodate charged residues and prolines in the hydrophobic core without disrupting globular structure [10,21] and thus as a corollary that flank positioning of gatekeepers represent thermodynamically neutral or favorable positions compatible with protein structure. On the other hand, *bona fide* functional residues are not optimized for protein stability and are therefore often found to be destabilizing to protein structure despite their high degree of conservation [22–24]. In order to distinguish these two scenarios, here we compared the sequence conservation of gatekeepers with the thermodynamic cost of incorporating these residues in the native protein structure of the *E. coli* genome. For this purpose, we took advantage of a dataset of 34 *E. coli* genomes for which multiple alignments of orthologs from 2659 genes have been generated [25]. A final set of 1955 genes was extracted from this dataset after removal of all transmembrane proteins; thus, the analysis below is performed on cytosolic proteins exclusively.

Results

APRs can be identified using aggregation prediction algorithms such as TANGO [26]. In the reference *E. coli* strain K12 MG1655 [25], we identified 12,755 APRs in 1955 genes. As before, we find that charged residues and prolines are strongly overrepresented (5–10% above genome average) at the flanks of APRs (Fig. 1a) that result in a bias toward low aggregation propensity in APRs (Fig. 1b) suggesting that aggregation is under negative selective pressure [4]. In total, we identified 15,741 charged residues and prolines at the flanks of APRs: 8542 were found at the N-terminal flank of APRs while 7199 were at the C-terminal flanks (5016 APRs are flanked by gatekeepers at both extremities). As previously observed, we find that the enrichment of proline is stereochemically selective as it is only effective at the N-terminal side [18].

Aggregation gatekeepers destabilize protein structure

In order to determine whether the enrichment of gatekeepers is subordinate to protein stability, thus whether gatekeepers are thermodynamically neutral or favorable to protein stability, we extracted from our dataset those sequences that have a crystal structure in the PDB [27] with a resolution better than 4.0 Å, extended with homology models for those sequences that had a homologous structure of the same quality with at least 60% sequence identity. In total, this resulted in a set of 797 out of the 1955 protein families (coverage of 37%), consisting of 436 structures and 361 homology models. These structures were used to investigate the thermodynamic cost of incorporating gatekeeper residues in globular proteins using the FoldX force field [28]. A plot of either side-chain burial (Fig. 2a) or main-chain burial