Bioinformatics, 36(3), 2020, 948–949
doi: 10.1093/bioinformatics/btz628
Advance Access Publication Date: 8 August 2019
Applications Note

Structural bioinformatics

Protein Homeostasis Database: protein quality control in *E.coli*

Reshmi Ramakrishnan[1,2,†], Bert Houben[1,2,†], Łukasz Kreft[3], Alexander Botzki[3], Joost Schymkowitz[1,2,*] and Frederic Rousseau[1,2,*]

[1]Switch Laboratory, VIB-KU Leuven Center for Brain & Disease Research, VIB and [2]Department of Cellular and Molecular Medicine, KU Leuven, Leuven 3000, Belgium and [3]VIB Bioinformatics Core, VIB, Gent 9052, Belgium

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Alfonso Valencia

Received on April 2, 2019; revised on July 9, 2019; editorial decision on August 4, 2019; accepted on August 6, 2019

Abstract

Motivation: *In vivo* protein folding is governed by molecular chaperones, that escort proteins from their translational birth to their proteolytic degradation. In *E.coli* the main classes of chaperones that interact with the nascent chain are trigger factor, DnaK/J and GroEL/ES and several authors have performed whole-genome experiments to construct exhaustive client lists for each of these.

Results: We constructed a database collecting all publicly available data of experimental chaperone-interaction and -dependency data for the *E.coli* proteome, and enriched it with an extensive set of protein-specific as well as cell context-dependent proteostatic parameters. We made this publicly accessible via a web interface that allows to search for proteins or chaperone client lists, but also to profile user-specified datasets against all the collected parameters. We hope this will accelerate research in this field by quickly identifying differentiating features in datasets.

Availability and implementation: The Protein Homeostasis Database is freely available without any registration requirement at http://PHDB.switchlab.org/.

Contact: joost.schymkowitz@kuleuven.vib.be or frederic.rousseau@kuleuven.vib.be

1 Introduction

In an effort to elucidate which features determine chaperone dependency in *E.coli*, we collected data from all hitherto published large-scale chaperone interaction studies into a meta-dataset (Arifuzzaman *et al.*, 2006; Calloni *et al.*, 2012; Chapman *et al.*, 2006; Deuerling *et al.*, 2003; Fan *et al.*, 2016; Fan *et al.*, 2017; Fujiwara *et al.*, 2010; Houry *et al.*, 1999; Kerner *et al.*, 2005; Martinez-Hackert and Hendrickson, 2009; Mogk *et al.*, 1999; Niwa *et al.*, 2009, 2012). Interestingly, we found overlap between these studies was very poor, and hypothesized therefore that chaperone dependency is not only governed by protein-intrinsic parameters, but also by cellular context, which likely differs between different experimental approaches. Hence, we designed an inclusive classification scheme that takes into account all the studies mentioned above, and complemented this data with a range of experimentally determined proteostatic parameters (abundance, translation rates, solubility, etc.) as well as simple primary-sequence-based calculations (net charge, amino acid composition, etc.), structural features (secondary structure content, contact order, etc.) and bioinformatics predictions (aggregation tendency from TANGO, disorder from IUPred, etc.). We are now making this dataset publicly available through a web interface that not only makes the data readily accessible and easily searchable, but also offers preliminary analysis options, including comparisons with proteome distributions and direct retrieval of significantly distinguishing features between user-defined groups of proteins.

2 Database

The full dataset, constructed as described in the introduction, contains over a hundred proteostatic parameters for 4305 *E.coli* proteins. The data sources used in compiling the database are listed on the website's About page, along with a detailed overview of which study provided which parameter. This information can also be found in our earlier publication describing the original application of our database (Ramakrishnan *et al.*, 2019).

3 Website

In order to provide a user-friendly interface for this complex dataset, we developed a web interface that allows to (i) obtain the client lists of different chaperone fluxes, (ii) inspect features of individual proteins and (iii) perform group analyses on user-defined sets. An overview of the database construction methodology and the analysis functionality offered by the website is shown in Figure 1.

3.1 Technical implementation

The dataset was imported into a MySQL database, on top of which an interactive frontend was written using AngularJS. The visualizations are dynamically created with the help of D3.js and ECharts. The communication between the frontend and the data model is handled by PHP.

3.2 Chaperone client flux view

The web interface homepage contains an overview of the different chaperone fluxes followed by *E.coli* proteins. Each group title links to a Browse page (see Section 3.3) containing the subset of *E.coli* proteins in the specified chaperone flux, as determined previously (Ramakrishnan *et al.*, 2019).

3.3 Browse

The Browse tab gives access to a page which allows for detailed filtering of the full dataset on any feature or any combination of features. Through the 'select' button, users can select single proteins or protein sets based on filtering options or simply on UniProt accession numbers.

3.4 Protein view

Quick searching using a UniProt accession number from the homepage or selecting an entry from the Browse page leads to a protein view page which shows the values of all the parameters within our database for the selected protein (Fig. 1c). Where possible, these values are plotted either as simple bar plots, or as violin plots depicting the distribution of the entire dataset, and the value of the selected protein. This page provides a convenient way of browsing through protein parameters and comparing protein characteristics with their respective proteome distributions. Upon selecting multiple proteins in the Browse tab, users are given a 'compare' option, which allows for a comparison of the selected proteins. Similar to the single-protein view, values for each element in the group are plotted, alongside a representation of proteome distributions (Fig. 1d). This analysis allows for rapid identification of common features within a group, as well as determination of outliers i.e. proteins that do not follow group patterns.

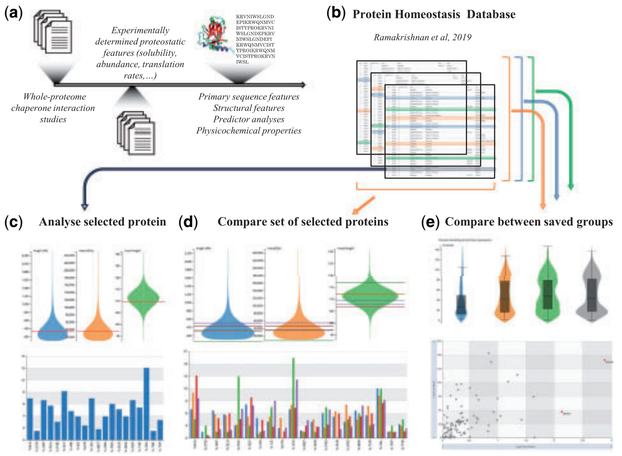
(a)
Experimentally determined proteostatic features (solubility, abundance, translation rates, ...)
KRVNIWSLGND
EPIKRWQNMVC
ISTYPROKRVNI
WSLGNDEPKRV
NIWSLGNDEPI
KRWQNMVCIST
YPROIKRWQNM
VCISTPROKRVN
IWSL
Whole-proteome chaperone interaction studies
Primary sequence features
Structural features
Predictor analyses
Physicochemical properties
(b) Protein Homeostasis Database
Ramakrishnan et al, 2019
(c) Analyse selected protein
(d) Compare set of selected proteins
(e) Compare between saved groups

Fig. 1. Protein Homeostasis Database conception and analysis options. (**a**) The Protein Homeostasis Database was constructed by combining 13 whole-proteome *E.coli* studies on chaperone dependency with large-scale experimental data on proteostatic parameters and a number of calculations and predictions based on primary sequence and structure. This dataset (**b**) is made publicly available through the web interface presented here. Apart from data availability, the web interface offers a range of analysis options (**c**–**e**). (**c**) Users can select a single protein of interest, returning a value for all features in the dataset for that protein, as well as a comparison to the proteome distribution. (**d**) Similarly, users can select a group of proteins, which yields visualizations of group distributions compared to proteome background. (**e**) Finally, multiple groups can be defined and inter-group differences analyzed. This analysis includes preliminary significance testing, hence facilitating the retrieval of distinguishing features between groups of interest

3.5 Comparing saved groups

Finally, users have the option of saving selected groups with user-defined names. Through the 'compare' button, saved groups can then be compared with each other. This yields combined violin- and boxplots showing the distributions of each feature for the selected groups, as well as for the proteome background. To readily identify interesting features, a volcano plot is also generated, depicting the most extreme fold change between all groups versus the negative logarithm of Kruskal-Wallis *P*-value (Fig. 1e, lower panel). In doing so, this plot readily offers information on how strongly the selected groups differ, as well as the statistical significance of these differences. Features above specific thresholds are indicated in red and hovering over the data points in the volcano plot shows the correlated feature, which allows for convenient identification of significantly distinguishing characteristics.

Acknowledgements

We thank the following researchers for feedback on methods and calculations: Hideki Taguchi (Tokyo, Japan), Tamir Tuller (Tel Aviv, Israel), Wim Vranken (Brussels, Belgium), James McInerney (Manchester, UK) and Geert Molenberghs (K.U. Leuven, Belgium).

Funding

The Switch Laboratory was supported by grants from the European Research Council under the European Union's Horizon 2020 Framework Programme ERC Grant agreement 647458 (MANGO) to JS, the Flanders institute for biotechnology (VIB), the University of Leuven ('Industrieel Onderzoeksfonds'), the Funds for Scientific Research Flanders (FWO), the Flanders Agency for innovation by Science and Technology (IWT, SBO grant 60839) and the Federal Office for Scientific Affairs of Belgium (Belspo), IUAP, grant number P7/16. R.R. was supported by an Erasmus Mundus fellowship.

Conflict of Interest: none declared.

References

Arifuzzaman,M. *et al.* (2006) Large-scale identification of protein-protein interaction of *Escherichia coli* K-12. *Genome Res.*, **16**, 686–691.

Calloni,G. *et al.* (2012) DnaK functions as a central hub in the *E. coli* chaperone network. *Cell Rep.*, **1**, 251–264.

Chapman,E. *et al.* (2006) Global aggregation of newly translated proteins in an *Escherichia coli* strain deficient of the chaperonin GroEL. *Proc. Natl. Acad. Sci. USA*, **103**, 15800–15805.

Deuerling,E. *et al.* (2003) Trigger factor and DnaK possess overlapping substrate pools and binding specificities. *Mol. Microbiol.*, **47**, 1317–1328.

Fan,D.J. *et al.* (2016) Large-scale gene expression profiling reveals physiological response to deletion of chaperone dnaKJ in *Escherichia coli*. *Microbiol Res.*, **186**, 27–36.

Fan,D.J. *et al.* (2017) Global analysis of the impact of deleting trigger factor on the transcriptome profile of *Escherichia coli*. *J. Cell. Biochem.*, **118**, 141–153.

Fujiwara,K. *et al.* (2010) A systematic survey of in vivo obligate chaperonin-dependent substrates. *EMBO J.*, **29**, 1552–1564.

Houry,W.A. *et al.* (1999) Identification of in vivo substrates of the chaperonin GroEL. *Nature*, **402**, 147–154.

Kerner,M.J. *et al.* (2005) Proteome-wide analysis of chaperonin-dependent protein folding in *Escherichia coli*. *Cell*, **122**, 209–220.

Martinez-Hackert,E. and Hendrickson,W.A. (2009) Promiscuous substrate recognition in folding and assembly activities of the trigger factor chaperone. *Cell*, **138**, 923–934.

Mogk,A. *et al.* (1999) Identification of thermolabile *Escherichia coli* proteins: prevention and reversion of aggregation by DnaK and ClpB. *EMBO J*, **18**, 6934–6949.

Niwa,T. *et al.* (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proc. Natl. Acad. Sci. USA*, **106**, 4201–4206.

Niwa,T. *et al.* (2012) Global analysis of chaperone effects using a reconstituted cell-free translation system. *Proc. Natl. Acad. Sci. USA*, **109**, 8937–8942.

Ramakrishnan,R. *et al.* (2019) Differential proteostatic regulation of insoluble and abundant proteins. *Bioinformatics*, **35**, 4098–4107.