DMP TITLE

Assessing co-translational protein aggregation and the cellular factors that prevent it

ADMIN DETAILS

Project Name: Assessing co-translational protein aggregation and the cellular factors that prevent it
Project Identifier: 12S3722N
Project type: Postdoctoral Fellow - junior
Principal Investigator / Researcher: Bert Houben
Project Data Contact: Béla Z Schmidt
Description: Most proteins need to attain a specific structure, their native fold, to function. Since proteins are largely unstructured upon their genesis, a large portion of the cellular resources is dedicated to helping proteins reach and maintain their native fold. A major threat to this process is protein aggregation. Most proteins contain segments called Aggregation-Prone Regions (APRs) that are normally buried within the folded structure. When APRs are exposed, they tend to engage in intermolecular interactions that prevent their parent protein from folding and lead to the formation of potentially cytotoxic protein aggregates. Indeed, over thirty disorders have been associated with the aggregation of one or several protein species, and this list is expanding. As proteins lack structure during translation, they are likely to expose their APRs. Despite this, the process of co-translational aggregation remains heavily understudied. The proposed research is aimed at studying co-translational aggregation on a proteome-wide scale. I will determine if and where co-translational protein aggregation occurs, and deduce cellular factors specifically aimed at preventing it. Finally, I will assess if and how translation kinetics affect co-translational aggregation. This research will yield unprecedented insight into the process of co-translational aggregation, hitherto unexplored territory in the proteostasis field.

Institutions: KU Leuven

1. GENERAL INFORMATION

Name applicant
Bert Houben

FWO Project Number & Title
Application number: 12S3722N
English Title
Assessing co-translational protein aggregation and the cellular factors that prevent it

Dutch Title
Co-translationele aggregatie en de cellulaire factoren die het voorkomen

Affiliation

- KU Leuven

2. DATA DESCRIPTION

Will you generate/collect new data and/or make use of existing data?

- Generate new data

- Reuse existing data

Describe the origin, type and format of the data (per dataset) and its (estimated) volume, ideally per objective or WP of the project. You might consider using the table in the guidance.

Please see data table in the following pages.

WP	Dataset	Purpose	New/ Existing (source)	Data type	Data subtype	Data format	Size	Unit	Comment
1	HEK293T cells	Polysome purification	Existing data	Experimental_data	Cell_lines	Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C.	3	cryovials	2000000 cells per vial
1	Polysome fractions	Material for ribosome footprinting	New data	Experimental_data	Samples	Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C.	150	eppendorfs	stored frozen
1	Polysome fractionation data in a range of conditions designed to break up native protein-protein interactions, yet leave ribosomes and beta-aggregates intact.	In order to reduce the background in ribosome profiling assays on co-translationally aggregated proteins, conditions need to be determined that reduce co-translational native interactions but leave beta aggregates intact.	New data	Experimental_data	Biophysics_data	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb)	100	Mb	
1	Purified RNA	For ribosome footprinting	New data	Experimental_data	Samples	Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C.	100	eppendorfs	stored at -80
1	Ribosome footprinting data (RNASeq) of co-translationally aggregated polysomes	Detecting sites of cotranslational aggregation in mRNA under physiological conditions and various forms of external	New data	Experimental_data	Omics_data	Nucleotide and protein sequences: raw sequence data trace (.ab1), textbased format (.fasta/.fa) and	2	GB	

		stress (heat, chemical,…)				accompanying QUAL file (.qual), Genbank format (.gb/.gbk);		
1	Mass spectrometry data of co-translationally aggregated polysomes	Detecting sites of cotranslational aggregation in proteins under physiological conditions and various forms of external stress (heat, chemical,…)	New data	Experimental_data	Omics_data	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb); possibly in instrument-specific proprietary format	1	GB
1	Compiled dataset of cotranslational aggregation sites in mRNA	Analysis results of ribosome footprinting data to map cotranslational aggregation sites to human genome and characterize these sites	New data	Derived_and_compiled_data	Research_documentation	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb);	10	MB
1	Compiled dataset of cotranslational aggregation sites in protein sequences	Analysis result of Mass spectrometry data to detect modulators of cotranslational aggregation	New data	Derived_and_compiled_data	Specific	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb);	10	MB
1	Human proteome structures	Cross-referencing structural information with co-translational stalling	Existing data	Canonical_data	Protein_structures	Protein structures: Protein Data Bank format (.pdb / .pdbx);	500	MB
1	Human proteome sequences	Mapping co-translational aggregation data to proteome	Existing data	Canonical_data	Protein_sequences	Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS	20	MB

						Word (.doc/.docx), eXtensible Mark-up Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format;		
1	Human genome cDNA sequences	Mapping co-translational aggregation data to translatome	Existing data	Canonical_data	Nucleic_acid_sequences	Nucleotide and protein sequences: raw sequence data trace (.ab1), textbased format (.fasta/.fa) and accompanying QUAL file (.qual), Genbank format (.gb/.gbk);	1	GB
1	Aggregation-prone regions in human proteome predicted through TANGO	To find predicted aggregation-prone regions	Existing data	Derived_and_compiled_data	Research_documentation	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb);	2	GB
1	Experimentally verified chaperone interactors	To map experimentally verified chaperone-dependent proteins to co-translational aggregation data	Existing data	Derived_and_compiled_data	Research_documentation	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb);	100	MB
1	Master dataset cross-referencing co-translational aggregation sites with structural information,	Analysing co-translational aggregation sites for specific characteristics regarding sequence composition, local	Existing data	Derived_and_compiled_data	Specific	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel	2	GB

	predicted aggregation propensity, experimentally verified chaperone interation sites etc	structure…				(.xls/.xlsx), MS Access (.mdb/.accdb);			
2	Set of trial constructs for mammalian expression	Set of constructs to test optimal configuration of model protein (stalled on the ribosome and exposing an aggregation-prone region). Different combinations of linkers, ribosome-stalling spacers and APRs will be used	New data	Experimental_data	Vectors	Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C.	10	constructs	plasmids stored at -20°C Glycerol stocks in E coli top10 cells stored at -80°C
2	HEK293T cells	To express model constructs	Existing data	Experimental_data	Cell_lines	Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C.	3	cryovials	2000000 cells per vial
2	High-content microscopic analyses of trial constructs	To determine the expression patterns of the set of trial constructs (expression levels, localization, diffuse vs punctate i.e. aggregated, …)	New data	Experimental_data	Digital_images	Digital images in raster formats: uncompressed TIFF (.tif/.tiff), JPEG (.jpg), JPEG 2000 (.jp2), Adobe Portable Document Format (.pdf), bitmap (.bmp), .gif;	5	GB	
2	Quantification of high-content microscopic analyses	To quantify for each of the trial constructs whether they cause observable co-translational aggregation through punctate staining in	New data	Derived_and_compiled_data	Specific	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access	10	MB	

		high-content imaging				(.mdb/.accdb);			
2	lysates of cells expressing trial constructs	To determine whether the trial constructs are actually expressed, whether co-translational stalling works, and whether they in fact cause co-translational aggregation	New data	Experimental_data	Samples	Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C.	50	eppendorfs	
2	Biophysical analyses of trial constructs	To determine whether the trial constructs are actually expressed, whether co-translational stalling works, and whether they in fact cause co-translational aggregation	New data	Experimental_data	Biophysics_data	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb);	10	MB	
2	Immunoprecipitated material	To identify interactors of model construct	New data	Experimental_data	Samples	Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C.	50	eppendorfs	
2	Mass spectrometry data of pulldown of co-translationally exposed aggregation-prone region model protein	Identifying proteins interacting cotranslationally with aggregation-prone regions displayed on the ribosome	New data	Experimental_data	Omics_data	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb); possibly in instrument-specific proprietary format	1	GB	
2	Polysome fractions	Material for ribosome footprinting	Existing data	Experimental_data	Samples	Biological and chemical samples: live animals, frozen	50	eppendorfs	stored frozen

						samples in cryovials, samples stored at 4°C.		
2	Immunoprecipitated material of polysome fractions, pulling a specific interactor	To identify binding sites of interactors	New data	Experimental_data	Samples	Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C.	100	eppendorfs
2	Selective Ribosome profiling data (SeRP) - Ribosome profiling on immunoprecipitated materail	Identifying proteome-wide co-translational binding sites of factors identified previously within this work package	New data	Experimental_data	Omics_data	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb); possibly in instrument-specific proprietary format	1	GB
2	Master dataset from WP1	Reference dataset for mapping SeRP data	Existing data	Derived_and_compiled_data	Research_documentation	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb);	2	GB
2	Dataset crossreferencing SeRP data with master dataset from work package 1	Validation of binding of newly identified co-translational modulators of aggregation with experimentally verified sites of co-translational aggregation (WP1)	New data	Derived_and_compiled_data	Research_documentation	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb);	2	GB
3	Translation kinetics data in mammalian cells	To discover patterns of translation rates in and around sites of co-translational	Existing data	Derived_and_compiled_data	Research_documentation	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file	2	GB

		aggregation				(.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb);			
3	Dataset cross-referencing existing ribosome footprinting data (giving an indication of proteome-wide translation kinetics) with information on predicted aggregation propensity, structural information, chaperone interaction sites, and information obtained in WP1 and WP2. This is an expansion of the master dataset from WP2	To perform a proteome-wide analysis of factors that affect translation rates in and around predicted aggregation-prone regions as well as regions experimentally verified to display co-translational aggregation (WP1) and binding sites for co-translationally acting modulators of aggregation (WP2)	New data	Derived_and_compiled_data	Research_documentation	Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb);	5	GB	
3	Set of constructs with differing translation rates around site of co-translational aggregation	Determine whether translation rate around aggregation-prone region affects cotranslational aggregation	New data	Experimental_data	Vectors	Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C.	10	vials	plasmids stored at -20°C Glycerol stocks in E coli top10 cells stored at -80°C
3	In vitro translation samples	See if co-translational aggregation in in vitro translation setup is affected by translation kinetics	New data	Experimental_data	Samples	Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C.	50	eppendorfs	
3	Biophysical analyses of in vitro translation	See if co-translational aggregation in in vitro	New data	Experimental_data	Biophysics_data	Quantitative tabular data: commaseparate	10	MB	

samples	translation setup is affected by translation kinetics	d value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb);

3. LEGAL & ETHICAL ISSUES

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to the file in KU Leuven's Record of Processing Activities. Be aware that registering the fact that you process personal data is a legal obligation.

- No

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, add the reference to the formal approval by the relevant ethical review committee(s)

- No

Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

- Yes

We do not exclude that the proposed work could result in research data with potential for tech transfer and valorisation. VIB and KU Leuven has a policy to actively monitor research data for such potential. If there is substantial potential, the invention will be thoroughly assessed, and in a number of cases the invention will be IP protected (mostly patent protectionor copyright protection). As such the IP protection does not withhold the research data from being made public. In the case a decision is taken to file a patent application it will be planned so that publications need not be delayed. Further research beyond the scope of this project may be necessary for developing a strong IP portfolio.

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?

- No

4. DOCUMENTATION & METADATA

What documentation will be provided to enable reuse of the data collected/generated in this project?
Metadata will be documented by the researcher and technical staff at the time of data collection and analysis, by taking careful notes in the electronic laboratory notebook (E-notebook) and/or in hard copy lab notebooks that refer to specific datasets. All datasets will be accompanied by a README.txt file containing all the associated metadata (see more details below). The data will be generated following standardized protocols. Clear and detailed descriptions of these protocols will be stored in our lab protocol database, and published along with the results.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- The following metadata standards will be used for certain datasets
  - Nucleotide sequence files (vectors and sequencing) : GenBank Sequence Format (https://fairsharing.org/FAIRsharing.rg2vmt)
  - Proteomics data: PRoteomics IDEntifications database (PRIDE, https://www.ebi.ac.uk/pride/)
- For instrument-specific datasets, additional metadata will be associated with the data file as appropriate.

- For other datasets, the metadata will include the following elements:
  - Title: free text
  - Creator: Last name, first name, organization
  - Date and time reference
  - Subject: Choice of keywords and classifications
  - Description: Text explaining the content of the data set and other contextual information needed for the correct interpretation of the data, the software(s) (including version number) used to produce and to read the data, the purpose of the experiment, etc.
  - Format: Details of the file format,
  - Resource Type: data set, image, audio, etc.
  - Identifier: DOI (when applicable)
  - Access rights: closed access, embargoed access, restricted access, open access.

The final dataset will be accompanied by a README.txt document. This file will be located in the top-level directory of the dataset and will also list the contents of the other files and outline the file-naming convention used. This will allow the data to be understood by other members of the laboratory and add contextual value to the dataset for future reuse.

5. DATA STORAGE & BACK UP DURING THE FWO PROJECT

Where will the data be stored?

Digital files will be stored either on KU Leuven servers or in shared laboratory folders of an off-site online backup service. The researchers working on the project will have copies of the data files as well as of the derived and compiled data stored on their personal computers.

The Switch Lab has a professional subscription to an off-site online backup service with unlimited space, version control and roll-back capability, which will be used for storage during the project and after. There is a secondary on-campus physical backup of the online storage which synchronizes with the online content with a one-day delay.

The screening core has a database system in place to handle the data stream from the high content imaging screen, including archiving facilities and will store the data during the project. Representative images and the quantitation of the images will be transferred to the Switch laboratory storage for long term storage.

Vectors: As a general rule at least two independently obtained clones will be preserved for each vector, both under the form of purified DNA (in -20°C freezer) and as a bacteria glycerol stock (-80°C). All published vectors and the associated sequences will be sent to the non-profit plasmid repository Addgene, which will take care of vector storage and shipping upon request.

Other biological and chemical samples: storage at 4°C and/or as frozen samples in cryovials as appropriate.

How is back up of the data provided?
The Switch Lab has a professional subscription to an off-site online backup service with unlimited space, version control and roll-back capability, which will be used for storage during the project and after. There is a secondary on-campus physical backup of the online storage which synchronizes with the online content with a one-day delay.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely. If no or insufficient storage or backup capacities are available then explain how this will be taken care of.

- Yes

The Switch Lab has a professional subscription to an off-site online backup service with unlimited space, which will be used for storage during the project and after.

What are the expected costs for data storage and back up during the project? How will these costs be covered?
Data storage and backup costs are included in general lab costs. The Switch Lab has a yearly subscription to an off-site online backup service paid from the general budget of the laboratory. The yearly cost of the service is 5500 Euros. This cost includes unlimited data storage, not only the data belonging to the present project.
Electricity costs for the -80° and -20° freezers and refrigerators present in the labs as well as the cost of liquid nitrogen cryostorage are included in general lab costs.
Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?
All notebooks and physical data are stored in the labs. Entry to the lab requires ID-card and key. Access to the digital data is u-number and password controlled.

6. DATA PRESERVATION AFTER THE FWO PROJECT

Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).
The minimum preservation term of 5 years after the end of the project will be applied to all datasets.

Where will the data be archived (= stored for the longer term)?
For the datasets that will be made openly accessible, we will use, whenever possible, the existing platforms that support FAIR data sharing (www.fairsharing.org), at the latest at the time of publication.

For all other datasets, long term storage will be ensured as follows: -Digital datasets will be stored on storage space of an online data-backup service. -Vectors: As a general rule at least two independently obtained clones will be preserved for each vector, both under the form of purified DNA (in -20°C freezer) and as a bacterial glycerol stock (-80°C). -Other biological and chemical samples: storage at 4°C and/or as frozen samples in cryovials as appropriate.

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?
Electricity costs for the -80° and -20° freezers and refrigerators present in the labs as well as for in liquid nitrogen cryostorage are included in general lab costs. The cost of the laboratory's professional subscription to the online data backup service is 5500 Euros per year (27 500 Euros for 5 years). This cost includes unlimited data storage, not only the data belonging to the present project. Data storage and backup costs are included in general lab costs.

7. DATA SHARING AND REUSE

Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

- No

Which data will be made available after the end of the project?
Participants to the present project are committed to publish research results to communicate them to peers and to a wide audience. All research outputs supporting publications will be made openly accessible. Depending on their nature, some data may be made available prior to publication, either on an individual basis to interested researchers and/or potential new collaborators, or publicly via repositories (e.g. negative data). We aim at communicating our results in top journals that require full disclosure upon publication of all included data, either in the main text, in supplementary material or in a data repository if requested by the journal and following deposit advice given by the journal. Depending on the journal, accessibility restrictions may apply. Physical data (e.g. cell lines) will be distributed to other parties if requested.

Where/how will the data be made available for reuse?

- The data will be shared upon request by mail.
- Possible ways of sharing the generated data:
  nucleic acid sequences: GenBank (https://www.ncbi.nlm.nih.gov/genbank/)
  protein sequences: UniProt KB (https://www.uniprot.org/)
  vectors: AddGene (http://www.addgene.org/depositing/start-deposit/)
  cell lines: direct mailing on dry ice
  microscope images: Image Data Resource (http://idr.openmicroscopy.org/about/)
  proteomics data: PRIDE (https://www.ebi.ac.uk/pride/)
  manuscripts: bioRxiv (https://www.biorxiv.org/)
  other digital data: Zenodo data repository (https://zenodo.org/)

When will the data be made available?

- Upon publication of the research results

Generally, research outputs will be made openly accessible at the latest at the time of publication. No embargo will be foreseen unless imposed e.g. by pending publications, potential IP requirements – note that patent application filing will be planned so that publications need not be delayed - or ongoing projects requiring confidential data. In those cases, datasets will be made publicly available as soon as the embargo date is reached.

Who will be able to access the data and under what conditions?
Whenever possible, datasets and the appropriate metadata will be made publicly available through repositories that support FAIR data sharing. As detailed above, metadata will contain sufficient

information to support data interpretation and reuse and will be conform to community norms. These repositories clearly describe their conditions of use (typically under a Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication, a Creative Commons Attribution (CC-BY) or an ODC Public Domain Dedication and Licence, with a material transfer agreement when applicable). Interested parties will thereby be allowed to access data directly, and they will give credit to the authors for the data used by citing the corresponding DOI. For data shared directly by the PI, a material transfer agreement (and a non-disclosure agreement if applicable) will be concluded with the beneficiaries in order to clearly describe the types of reuse that are permitted.

What are the expected costs for data sharing? How will the costs be covered?
It is the intention to minimize data management costs by implementing standard procedures e.g. for metadata collection and file storage and organization from the start of the project, and by using free-to-use data repositories and dissemination facilities whenever possible. Data management costs will be covered by the laboratory budget.
The receiving party will pay for sharing physical data (e.g. cell lines).

8. RESPONSIBILITIES

Who will be responsible for data documentation & metadata?
Metadata will be documented by the researcher and technical staff at the time of data collection and analysis, by taking careful notes in the electronic laboratory notebook (E-notebook) that refer to specific datasets.

Who will be responsible for data storage & back up during the project?
The research and technical staff will ensure data storage and back up, with support from René Custers and Alexander Botzki for the electronic laboratory notebook (ELN) and from Raf De Coster for the KU Leuven drives.

Who will be responsible for ensuring data preservation and reuse ?
The PI is responsible for data preservation and sharing, with support from the research and technical staff involved in the project, from René Custers and Alexander Botzki for the electronic laboratory notebook (ELN) and from Raf De Coster for the KU Leuven drives.

Who bears the end responsibility for updating & implementing this DMP?
The PI is ultimately responsible for all data management during and after data collection, including implementing and updating the DMP.