Universität Potsdam

Doctoral Dissertation

Post-transcriptional Control of Gene Expression

Author:
Celine Sin

Supervisor:
Dr. Angelo Valleriani

A cumulative dissertation submitted in fulfilment of the requirements for the degree of Doctor rerum naturalium

in the

Institute of Biochemistry and Biology

Work completed at the

Max Planck Institute for Colloids and Interfaces
Department of Theory and Biosystems
Stochastic Processes in Complex and Biological Systems

June 2016

Declaration of Authorship

I, Celine Sin, declare that this thesis titled, 'Post-transcriptional Control of Gene Expression' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UNIVERSITÄT POTSDAM

Abstract

Mathematisch-Naturwissenschaftliche Fakultät
Institute of Biochemistry and Biology

Doctor rerum naturalium

Post-transcriptional Control of Gene Expression

by Celine Sin

Gene expression describes the process of making functional gene products (e.g. proteins or special RNAs) from instructions encoded in the genetic information (e.g. DNA). This process is heavily regulated, allowing cells to produce the appropriate gene products necessary for cell survival, adapting production as necessary for different cell environments. Gene expression is subject to regulation at several levels, including transcription, mRNA degradation, translation and protein degradation. When intact, this system maintains cell homeostasis, keeping the cell alive and adaptable to different environments. Malfunction in the system can result in disease states and cell death. In this dissertation, we explore several aspects of gene expression control by analyzing data from biological experiments. Most of the work following uses a common mathematical model framework based on Markov chain models to test hypotheses, predict system dynamics or elucidate network topology. Our work lies in the intersection between mathematics and biology and showcases the power of statistical data analysis and math modeling for validation and discovery of biological phenomena.

allgemeinverstaendlichen Zusammenfassung - English

UNIVERSITÄT POTSDAM

allgemeinverstaendlichen Zusammenfassung (English)

Mathematisch-Naturwissenschaftliche Fakultät
Institute of Biochemistry and Biology

Doctor rerum naturalium

Post-transcriptional Control of Gene Expression

by Celine Sin

The central dogma of molecular biology proposes that the flow of genetic information starts with DNA, which is copied into RNA and then translated into proteins (Crick 1970). This system of information transfer provides for two very natural checkpoints where gene expression can be adjusted – either at the level of mRNA (i.e. controlling transcription or mRNA degradation processes) or the level of proteins (i.e. controlling translation or protein degradation processes). Within each checkpoint, there are a multitude of processes simultaneously active, adjusting and fine-tuning the concentrations of mRNAs and proteins. In concert, all of these processes contribute to maintain a stable internal environment in the cell. Malfunction in the system can result in disease states and cell death. In this dissertation, we explore several aspects of gene expression control by analyzing data from biological experiments. Our work lies in the intersection between mathematical modeling and biology, showcasing the power of statistical data analysis and modeling for validation and discovery of biological phenomena.

allgemeinverstaendlichen Zusammenfassung - German

UNIVERSITÄT POTSDAM

allgemeinverstaendlichen Zusammenfassung

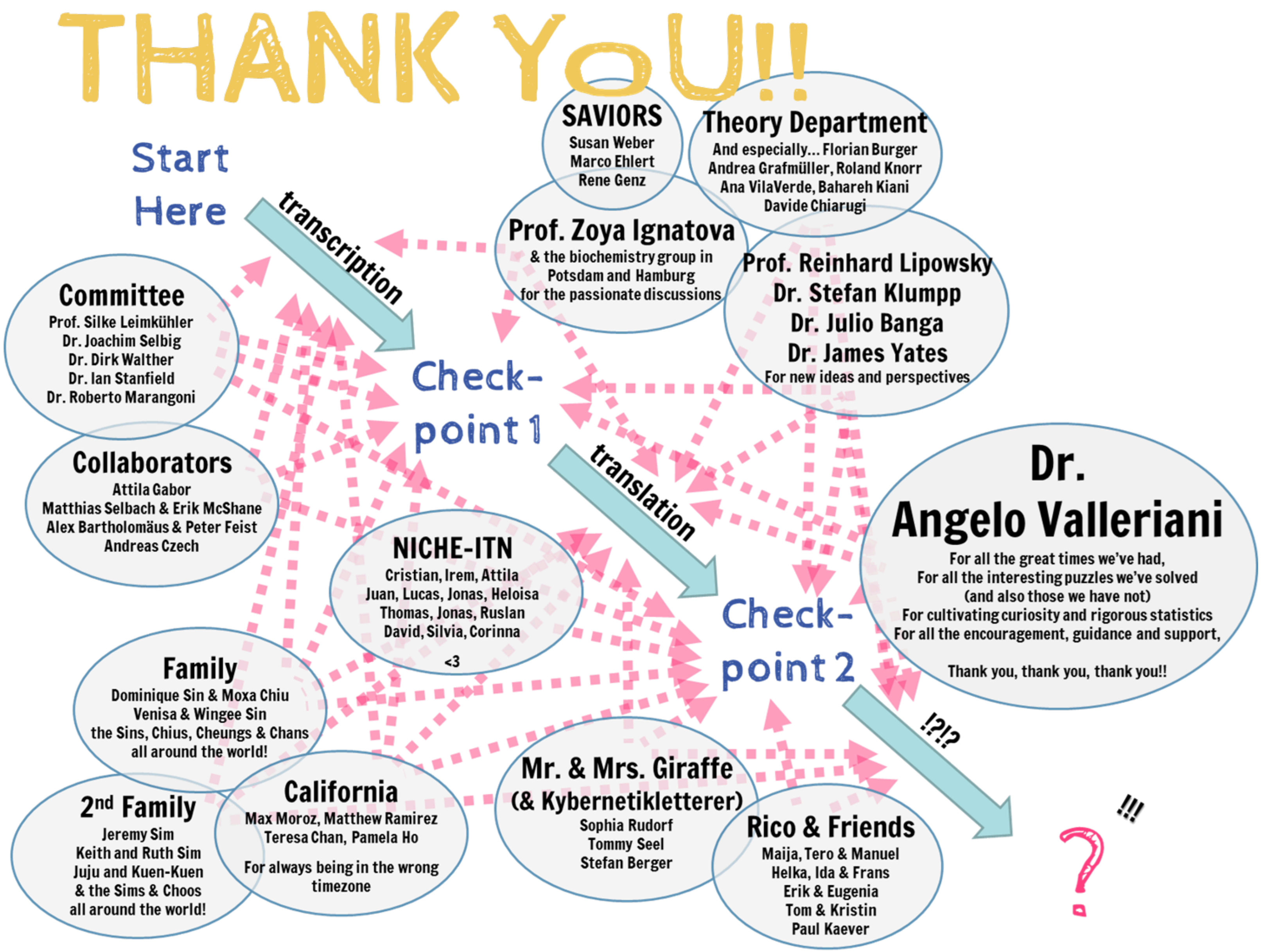
Mathematisch-Naturwissenschaftliche Fakultät
Institute of Biochemistry and Biology

Doctor rerum naturalium

Post-transcriptional Control of Gene Expression

by Celine Sin

Das „zentrale Dogma der Molekularbiologie" besagt, dass der Fluss genetischer Information mit der DNS startet, die dann auf die RNS kopiert und in Proteine übersetzt wird (Crick 1970). Dieses System der Informationsübertragung bietet zwei natürliche Eingriffspunkte, an denen Genausprägungen manipuliert werden können – entweder auf dem Level der mRNS (z.B. durch Kontrolle der Transkriptions- oder mRNS- Degradationsprozesse) oder auf dem Level des Proteins (z.B. durch Kontrolle der Translations- oder Proteindegradationsprozesse). An jedem Eingriffspunkt sind eine Vielzahl unterschiedlicher Prozesse zeitgleich aktiv, um die Konzentrationen von mRNS und Proteinen präzise einzustellen. All diese Prozesse tragen dazu bei, die Zelle intern im stationäzen Zustand zu halten, denn eine Fehlfunktion im System kann zu Krankheitszuständen oder zum Zelltot führen. In dieser Arbeit untersuchen wir verschiedene Aspekte der Kontrolle der Genausprägungs, indem wir Daten biologischer Experimente analysieren. Unsere Arbeit liegt hierbei zwischen den Bereichen der mathematischer Modellierung und der Biologie und zeigt den immensen Nutzen von statistischen Analysemethoden und mathematischer Modellbildung zur Validierung und Neuentdeckung biologischer Phänomene auf.



THANK YoU!!

Start Here

transcription

Check-point 1

translation

Check-point 2

!?!?

!!!

?

SAVIORS
Susan Weber
Marco Ehlert
Rene Genz

Theory Department
And especially... Florian Burger
Andrea Grafmüller, Roland Knorr
Ana VilaVerde, Bahareh Kiani
Davide Chiarugi

Prof. Zoya Ignatova
& the biochemistry group in
Potsdam and Hamburg
for the passionate discussions

Prof. Reinhard Lipowsky
Dr. Stefan Klumpp
Dr. Julio Banga
Dr. James Yates
For new ideas and perspectives

Committee
Prof. Silke Leimkühler
Dr. Joachim Selbig
Dr. Dirk Walther
Dr. Ian Stanfield
Dr. Roberto Marangoni

Collaborators
Attila Gabor
Matthias Selbach & Erik McShane
Alex Bartholomäus & Peter Feist
Andreas Czech

NICHE-ITN
Cristian, Irem, Attila
Juan, Lucas, Jonas, Heloisa
Thomas, Jonas, Ruslan
David, Silvia, Corinna
<3

Dr.
Angelo Valleriani
For all the great times we've had,
For all the interesting puzzles we've solved
(and also those we have not)
For cultivating curiosity and rigorous statistics
For all the encouragement, guidance and support,

Thank you, thank you, thank you!!

Family
Dominique Sin & Moxa Chiu
Venisa & Wingee Sin
the Sins, Chius, Cheungs & Chans
all around the world!

2nd Family
Jeremy Sim
Keith and Ruth Sim
Juju and Kuen-Kuen
& the Sims & Choos
all around the world!

California
Max Moroz, Matthew Ramirez
Teresa Chan, Pamela Ho

For always being in the wrong
timezone

Mr. & Mrs. Giraffe
(& Kybernetikletterer)
Sophia Rudorf
Tommy Seel
Stefan Berger

Rico & Friends
Maija, Tero & Manuel
Helka, Ida & Frans
Erik & Eugenia
Tom & Kristin
Paul Kaever

Figure 1: Model of Celine expression control. Very complicated system with many inputs and parameters. Yet, relatively robust! May we never lose our curiosity, nor our drive for understanding. Keep a song in our hearts and a smile on our faces... and always ask questions!

List of Contributions

In this cumulative dissertation, I have had the fortune to make many collaborations with other scientists. The following outlines my personal contributions to the projects. All work was supervised by Dr. Angelo Valleriani in the department of Theory and Biosystems at the Max Planck Institute for Colloids and Interfaces.

1. Celine Sin, Davide Chiarugi, and Angelo Valleriani. Single-molecule modeling of mRNA degradation by miRNA: Lessons from data. *Bmc Systems Biology*, 9:S2, June 2015

   All authors took part in conceiving the project, developing the model and writing the manuscript. I extracted the data from Braun et. al., 2011, preformed the data analysis and parameter estimation.

2. Alexander Bartholomaus, Ivan Fedyunin, Peter Feist, Celine Sin, Gong Zhang, Angelo Valleriani, and Zoya Ignatova. Bacteria differently regulate mRNA abundance to specifically respond to various stresses. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2063), March 2016

   Large collaboration with many members. I designed many of the metrics. I made the initial calculations on copy number, ribosome density and translational burden. Our main contribution to the manuscript was the fold change analysis. I also made the calculations to test for reproducibility and statistical significance. Additionally (but not included in the manuscript), I preformed additional analyses on codon usage, ribosome speed and trends by decile. The structure used to manipulate the data is documented in Appendix C.

3. Celine Sin, Davide Chiarugi, and Angelo Valleriani. Quantitative assessment of ribosome drop-off in E. coli. *Nucleic Acids Research*, 44(6):2528–2537, April 2016

   This project was the evolution of [2]. I discovered that there was a length dependent relationship to ribosome density, through the decile analysis. Furthermore, I showed that the relationship between length and ribosome density followed a power law. I concluded that this could be due to the phenomenon of ribosome drop-off, and we designed an analysis to test it. DC preformed all the bioinformatics and coded the scripts for analysis. I used the mapped reads from DC to confirm the results in parallel using my own scripts.

4. Erik McShane, Celine Sin, Henrik Zauber, Jonathan N. Wells, Neysan Donnelly, Xi Wang, Jingyi Hou, Wei Chen, Zuzana Storchova, Marsh Joseph A., Angelo Valleriani, and Matthias Selbach. Global quantification of cellular protein degradation kinetics. *Submitted to Cell, currently in revisions after first referee reports.*, 2016

   Another large collaboration with many members. I was involved very early on with the idea of age-dependent degradation. I designed the normalization schemes. I implemented all the normalization schemes and parameter estimation machinery. I made the calculations to describe the lifetime properties of each protein (e.g. average lifetime, half-time, steady state distributions, etc.) I also made many of the calculations to evaluate reproducibility, sensitivity and statistical significance. The structure used to manipulate the data is documented in Appendix D.

5. Celine Sin, Davide Chiarugi, and Angelo Valleriani. Degradation parameters from pulse-chase experiments. *Plos One*, 11(5):e0155028, May 2016

   All authors took part in conceiving the project, developing the experiment and writing the manuscript. I extracted the data from Wheatley, et. al., 1980. I preformed the data analysis, parameter estimation and sensitivity analysis on the system.

List of Publications

1. Celine Sin, Davide Chiarugi, and Angelo Valleriani. Single-molecule modeling of mRNA degradation by miRNA: Lessons from data. *Bmc Systems Biology*, 9:S2, June 2015

2. Alexander Bartholomaus, Ivan Fedyunin, Peter Feist, Celine Sin, Gong Zhang, Angelo Valleriani, and Zoya Ignatova. Bacteria differently regulate mRNA abundance to specifically respond to various stresses. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2063), March 2016

3. Celine Sin, Davide Chiarugi, and Angelo Valleriani. Quantitative assessment of ribosome drop-off in E. coli. *Nucleic Acids Research*, 44(6):2528–2537, April 2016

4. Erik McShane, Celine Sin, Henrik Zauber, Jonathan N. Wells, Neysan Donnelly, Xi Wang, Jingyi Hou, Wei Chen, Zuzana Storchova, Marsh Joseph A., Angelo Valleriani, and Matthias Selbach. Global quantification of cellular protein degradation kinetics. *Submitted to Cell, currently in revisions after first referee reports.*, 2016

5. Celine Sin, Davide Chiarugi, and Angelo Valleriani. Degradation parameters from pulse-chase experiments. *Plos One*, 11(5):e0155028, May 2016

Contents

List of Figures

0 INTRODUCTION

1 mRNA degradation through the miRISC pathway

2 Bacteria differently regulate mRNA abundance to specifically respond to various stresses

3 Quantitative assessment of ribosome drop-off in *E. coli*

4 Global quantification of cellular protein degradation kinetics

5 Degradation Parameters from Pulse-Chase Experiments

6 Discussion and Summary

Supplementary 1 Single-Molecule Modeling of mRNA Degradation by miRNA: Lessons from Data

Supplementary 2 Bacteria differently regulate mRNA abundance to specifically respond to various stresses

Supplementary 3 Quantitative assessment of ribosome drop-off in *E. coli*

Supplementary 4 Global quantification of cellular protein degradation kinetics

List of Tables

3 Quantitative assessment of ribosome drop-off in *E. coli*

5 Degradation Parameters from Pulse-Chase Experiments

Supplementary 1 Single-Molecule Modeling of mRNA Degradation by miRNA: Lessons from Data

Supplementary 3 Quantitative assessment of ribosome drop-off in *E. coli*

Supplementary 5 Degradation Parameters from Pulse-Chase Experiments

Dedicated to my dear grandpa

冼金泉

He taught me
... to fix almost anything,
... to build all that I can imagine,
and
... to shape my world into the future that I envision.

-Valentine's day, 2013

Part I

The Papers

0

INTRODUCTION

In this cumulative dissertation, we explore several aspects of gene expression control by analyzing data from biological experiments. Chapter 1-4 are ordered temporally by the point in the process which the mechanism(s) studied might occur. Chapter 1 explores the biochemical network for degradation of mRNA through the miRISC pathway [1]. Chapter 2 presents a global picture of the transcriptional and translational state of *E. coli* under several stress environments [2]. Chapter 3 quantifies the effect of ribosome drop-off [3]. Chapter 4 is a proteome wide study of degradation [4]. Chapter 5 presents the mathematical modeling framework used in most of our work and derivation of the experimentally measurable quantities. Furthermore, the implications of pulse time on the measurable dynamics and considerations for optimal experimental design are discussed [5].

Abstract

This section begins with an outline of the gene expression process with emphasis on possible points of control. Next, we provide a layperson's overview of the experimental techniques utilized to capture the data involved in our analysis. Finally, we briefly describe the mathematical modeling framework to translate single molecule predictions into population averages.

0.1 Overview of the Gene Expression Process

The central dogma of molecular biology proposes that the flow of genetic information starts at the level of DNA which is copied into RNA and then translated into proteins [9]. This system of information transfer provides for two very natural checkpoints where gene expression can be adjusted – either at the level of mRNA (i.e. controlling transcription or mRNA degradation processes) or the level of proteins (i.e. controlling translation or protein degradation processes). Within each checkpoint, there are a multitude of processes simultaneously active, adjusting and fine-tuning the concentrations of mRNAs and proteins.

DNA
transcription
mRNA
translation
proteins

FIGURE 1: The Central Dogma of Molecular Biology [9]

0.1.1 Gene Expression Control through mRNA synthesis

Genetic information is stored as a nucleotide sequence in the DNA [9]. DNA is transcribed into RNA by RNA polymerase [11, 12]. The first level of control involves regulating the amount of RNA available, by modulating the synthesis and degradation rates or deactivating RNA products. Increased transcription potentially increases the gene expression rate by increasing the amount of mRNA available for translation. Conversely, decreased transcription rate potentially decreases the gene expression rate [13]. RNA products can also be temporarily deactivated, or sequestered away from the translation machinery, thus decreasing the gene expression level [14–16].

RNA is synthesized by RNA polymerase. In *E. coli*, there are over 4000 genes but only 2,000 RNA polymerase molecules in normal growing conditions [17]. Thus, RNA polymerase somehow must know which genes to transcribe – this process is heavily regulated by transcription factors. Transcription factors are proteins that bind directly to DNA and can enhance transcription (activators), or block transcription (repressors) [18, 19].

In prokaryotes, one of the main mechanisms to control the initiation of RNA polymerase on specific genes is provided through the σ subunit [17]. RNA polymerase is a multiunit enzyme composed of two α units, one β unit, one β' unit and one σ unit (also known as σ factor) [20]. Before initiation, RNA polymerase scans the DNA, looking for promoter sequences located upstream of genes which signal the start of genes [21]. The σ factor recognizes specific promoter sequences and binds to them. Binding causes the DNA to unravel, exposing the bases for RNA synthesis and initiating transcription [22]. While it was once believed that σ factor dissociates from the complex and activates other RNA polymerases once transcription is initiated (σ cycling), leaving the $\alpha^2\beta\beta'$ complex to complete RNA transcription by itself [23], new single molecule measurements seem to suggest otherwise [24].

There are seven types of σ factors, each recognizing a different promoter sequence [17]. Many genes share the same promoter sequence. By controlling the type of σ factor associated with the RNA polymerases, different subsets of genes can be expressed [19, 25]. In bacteria, the σ^{70} family is often active, recognizing promoter regions of genes mostly used for general housekeeping in normal cell growing conditions [26]. Other σ factors are more specialized; for example, σ^{32} is affiliated with promoter regions of genes in response to heat shock [27].

In eukaryotes, the regulation process is more complex, often requiring several transcription factors interacting with the DNA both upstream and downstream of the start site [18]. Furthermore, the amount of mRNA ultimately made available for translation also depends on the rate of post-transcriptional processing steps including 5' capping, splicing, polyadenylation and nuclear export. These topics are outside the scope of this dissertation but a comprehensive review can be found in [28].

0.1.2 Gene Expression Control through mRNA degradation

Once RNA has been synthesized (and for the case of eukaryotes: processed and exported), degradation is the final process that affects the amount of mRNA in the cell. RNA degradation is a crucial process in cell metabolism; it allows for recycling of precursors, removal of errant transcripts and gene expression control [29]. Not surprisingly, the degradation system is very robust and has many redundancies; in *E. coli* there are ~24 RNAases [30] and inactivation or mutation of one or more RNAase species generally does not block RNA degradation [31]. RNA is degraded by a number of different pathways, some specific to distinct populations of transcripts in the cell (i.e. specific genes), and some which are non-specific.

Turnover of mRNA in prokaryotes is relatively rapid, on the order of minutes [32]. While multicellular organisms are able to buffer themselves from external environmental changes, prokaryotes are bathed in their external environment. Thus, sudden changes in the external environment require rapid reprogramming of cell's gene expression. Quick mRNA turnover time enables rapid reprogramming necessary for cell viability; in this way, the cell can adjust the mRNA population to the proteins that are currently needed [2, 29].

Prokaryote mRNA degradation is achieved through a suite of endonucleases and exonucleases. Exonucleases degrade RNA from the 3' or 5' ends while endoncleases cut RNA internally. In general, once a nuclease has attacked an RNA molecule either from internally or from the 3' or 5' ends, the subsequent decay of the RNA quickly follows – RNA decay generally follows 1st order kinetics in an all-or-none pattern and intermediates are rarely observable unless the ribonucleases are inactivated [29].

RNA lifetimes vary greatly and depend on a number of different properties: RNA sequence and secondary structure, subcellular location, translatability of the mRNA and interaction to other RNA or proteins are all possible factors that determine the stability of an RNA [29]. Some degradation pathways are activated on RNA containing specific sequences. Secondary structure on the mRNA can increase or decrease the degradation rate depending on where the secondary structure is found: stem-loops in the 5' UTRs decrease stability by making the mRNA more susceptible to RNAaseE, but stem-loops in the 3' UTR increase stability by blocking 3' to 5' exonuclease action [32]. mRNA that are very actively translated (i.e. being simultaneously translated by many ribosomes) are more protected by nuclease action than naked mRNA [33]. In the same fashion, mRNA interaction with proteins can also protect it from nuclease action [29]. Thus mRNA within a cell can have very different lifetimes. In prokaryotes, the amount of mRNA available for translation is determined solely by the net effects of synthesis and degradation. In Chapter 2, one of the aims of the project is to quantify the level of mRNA expression for *E. coli* in different stress conditions using RNA-seq. A short overview of the experimental procedure can be found in Supplementary section 0.2.2, and the full details can be found in Supplementary Section 2. Measuring the steady state level of over 4000 mRNA, this type of experiment gives us a global view of the transcriptional program. Indeed, we find that the cell adapts different transcriptional programs in response to stress conditions [2].

While many mRNA degradation processes in prokaryotes have homologous counterparts in eukaryotic organisms, eukaryotic organisms have (additionally) a much larger suite of mRNA degradation processes. *S. cerevisiae* turns out to be a convenient organism for studying eukaryotic mRNA degradation processes, as many processes are similar to mammalian cells [13]. Many of these processes are multistep with dynamics that sometimes can be distinctly measured. This allows for interesting possibilities to enhance system sensitivity and/or robustness and, for our intents and purposes, opportunities for modeling

Deadenylation dependent pathways

Degradation of mRNA in mammals can be divided into two categories: deadenylation-dependent or deadenylation-independent. Deadenylation-dependent mRNA decay is one of the major degradation pathways in mammalian cells [13].

Mature mRNA have a polyadenylated (polyA) tail that is added post-transcriptionally during the process of mRNA maturation. The polyA tail is used in the processes of nuclear transport, translation and is a determinant of stability. Mature mRNAs are often found "circularized", where the 3' polyadenylated end is bridged to the 5' end with a chain of proteins. The protein structures protect the 3' and 5' ends from exonucleases, thus increasing the stability of the mRNA. If the polyA tail is shortened (i.e. the mRNA becomes deadenylated), the bridge structure is destabilized, and the 3' and 5' ends become exposed, increasing the probability of degradation.

Deadenylation dependent mRNA decay is a multistep process. In general it is believed that special exonucleases (e.g. PARN, DAN, PAN2/PAN3) begin the process by initiating deadenylation of the polyA tail, and CCR/NOT finish deadenylation via 3' to 5' exonuclease digestion [10, 34]. Without the polyA tail, the interaction of the mRNA to the polyA binding proteins (which stabilize the bridge) cannot occur. Without the protein bridge that circularizes the mRNA, the 5' cap dissociates from the mRNA (through the action of DCP1), leaving both ends subject to exonucleolytic action [35]. While it is currently not possible to measure the intermediate steps of this pathway, experiments with the deadenylases deactivated can be used to understand the dynamics of the system.

miRISC (microRNA Induced Silencing Complex) is one of the deadenylation dependent mRNA degradation pathways. This pathway uses microRNA (miRNA) to selectively target mRNA for degradation. miRNA are short non-coding RNA sequences (approximately 20-30 nucleotides) known to be one of the key regulators of gene expression [15]. They exert their function through complementary base pairing with the target mRNA. Depending on the location of base pairing with an mRNA, miRNA can block ribosome initiation, enhance ribosome initiation, or block ribosome action. In the miRISC pathway, miRNA serves to target the degradation/silencing complex to target mRNAs. The miRISC pathway plays a very prominent role in mRNA degradation [1, 15, 36].

In Chapter 1, we present the process of building a model for the miRISC degradation pathway. The work is made possible by the comprehensive assay of the miRISC degradation pathway by [10]. GW182/AGO is the docking platform for the miRNA and the suite of proteins involved in the miRISC degradation pathway [10, 36]. In [10], it was found that NOT1 and PAN3 make direct interactions with AGO through a coimmunoprecipitation assay. Then, to quantify the contributions of these two protein factors in the degradation pathway, a series of experiments where one or more of the factors (miRNA, NOT1 and/or PAN3) were removed or knocked down were preformed. From this, it was concluded that NOT1 is a more relevant factor than PAN3 in mRNA destabilization. Furthermore, a hypothesis regarding the recruitment of NOT1 and PAN3 in the miRISC degradation pathway was presented.

As a starting point, we used the data from the experiments to build and calibrate a model that represented the preliminary hypothesis proposed in [10]. We translate the pathway proposed in [10] into a Markov Chain model. Markov Chain models describe stochastic processes which move through discrete states connected via a network of transitions. Each experiment corresponds to a different Markov Chain model – deactivating one of the factors is synonymous to deactivation of the corresponding transitions. Thus, starting with the experiment with the most transitions deactivated, we calibrate the reduced model. We continue on, one experiment at a time, until all the transitions are active and we have determined the transition rates between all the states.

Ultimately, we find that the hypothesis presented in [10] is unlikely to be true – if it were true, the data tell us that only a very small fraction of the target mRNA (7%) would pass through the miRISC pathway. We propose an alternative hypothesis where the interaction between GW182/AGO and NOT1/PAN3 occur before interaction with the target mRNA. With this model, we get very good fits for all the data collected.

Deadenylation independent pathways

Alternative to deadenylation-dependant pathways, eukaryotic mRNA degradation can also occur through deadenylation-independent pathways. Endoribonucleolytic pathways, which degrade mRNA internally (instead of from the ends) are less common, but still prevalent. These endonucleases (e.g. PMR1, ERN1/IRE1, Zc3h12a) often target actively translating mRNAs [13].

Surveillance pathways, which degrade mRNA that would potentially make defective proteins, are also independent of deadenylation. There are three types of surveillance pathways: Nonsense Mediated Decay (NMD), NonStop mRNA decay (NSD) and No-Go mRNA decay (NGD). In Nonsense Mediated Decay, mRNA with premature stop codons or introns within the 3'UTR are degraded. This can be caused by mutations in genes, errors in mRNA splicing, or errors in mRNA transcription. Conversely, for mRNA with absent stop codons (i.e. the ribosomes would keep translating all the way to the 3' end), the NonStop mRNA decay pathway is activiated. This could be the result of faulty transcription such that the stop codon is missing, or ribosome frameshifting (the ribosome "reads" in a different frame). Lastly, the No-Go mRNA decay pathway is activated when there are stalled ribosomes along the message. Stalling can be caused by low tRNA availability or strong secondary structure. In these cases, ribosomes cannot properly locate and finish translation of the mRNA. If this continues, many ribosomes will be occupied and unable to make proteins. Ref. [37] provides a very comprehensive review of the various surveillance pathways in eukaryotic cells, including hypotheses of the biochemical pathways as well as common targets for each mechanism.

The last mechanism we review in brief here is the system of stress granules and P-bodies. Stress granules are visible clusters of mRNA and proteins found in the cytoplasm of stressed cells [38]. One of the first images of stress granules were captured by [39] in plant cells under heat shock stress. Since then, the granules have been found in a number of different organisms, including mammalian cells. The process is thought to start by phosphorylation of the initiation factor elF2α [6]. TIA-1 and TIAR, both mRNA binding proteins normally found in the nucleus, move to the cytoplasm and spontaneously aggregate into stress suppressing translation of any mRNA to which they are bound [40]. The mechanism of aggregation is akin to the aggregation of $\alpha\beta$ proteins in Alzheimer's patients. Stress granules are thought to be sites for mRNA sorting, sequestration and storage [41]. mRNA selected for degradation is passed to processing bodies (P-bodies) [41, 42].

0.1.3 Gene Expression Control at the level of translation

After transcription of mRNA by RNA polymerase, the resultant mRNAs are read by ribosomes. In this process, the nucleotide instructions are decoded by ribosomes and transferRNA (tRNA) into amino acid sequences, which are then folded into proteins [9]. The nucleotide instructions are encoded in "genetic code", made of nucleotide triplets, called codons. As there are 4 nucleotides, there are $4^3 = 64$ possible codons [43]. Each codon (apart from the 3 stop codons, UAG, UGA and UAA), signifies one of the 20 amino acids and some of the amino acids are represented multiple times in the code.

tRNAs are the adaptor molecules to bring amino acids to the ribosome: one side carries the amino acid, and the other side has the 3-nucleotide sequence for complementary base pairing with the codon on the mRNA [44]. The tRNA delivers the appropriate amino acid to the ribosome. The ribosome catalyzes the aminotransferase reaction and then moves on to the next codon. When the ribosome reaches a stop codon, the amino acid chain is released and the protein sequence is complete.

The protein synthesis process can be broken up into several steps: initiation, elongation and termination. Each of these steps can potentially contribute to gene expression control. That being said, in general, the amount of protein synthesized is directly correlated to the amount of mRNA present – that is to say, that much of gene expression is controlled at the mRNA level [2, 45]. But there are indeed deviations for some genes, as well as in different conditions [2, 46].

In protein synthesis, initiation is thought to be the rate limiting step. A number of studies have explored the idea of "Ribosome Economics" [47–49]. Ribosomes are extremely energetically expensive to produce and thought to be in limited supply. Thus the idea that cell viability is dependent on appropriate allocation of ribosomal resources depending on the current condition is not too farfetched. Indeed, there are mechanisms to rescue ribosomes inappropriately stuck on non productive transcripts, for example, the three mRNA surveillance pathways [37]. Conversely, there are also mechanisms to sequester ribosomes away. In some conditions, ribosomes were found to "hibernate". By dimerizing the 30S subunits, ribosomes are effectively sequestered away [50].

The ribosome is composed of several subunits which must come together with the help of a number of initiation factors. The 5' cap structure provides a base for the small subunit of the ribosome and initiation factors to assemble. Additionally, a tRNA carrying the first amino acid (always methionine) associates. This complex scans the mRNA to find the start site (AUG), which makes complementary base pairing with met-tRNA. From here, the 60S subunit of the ribosome (large subunit) can associate, and protein synthesis can proceed. Hiding the start site is an effective method of gene expression control [51].

As all coding regions for genes necessarily begin with AUG, all proteins start with a methionine amino acid. This can be exploited to label proteins produced during an experiment. Azidohomoalanine (AHA) is a synthetic amino acid that has a similar shape as methionine. By using a cell culture free from methionine, AHA will be incorporated into newly synthesized proteins. AHA contains an azido group which can be used for click chemistry. In Chapter 4, this strategy is used to label newly synthesized proteins. To separate the existing proteins from the newly synthesized (labeled) proteins, the lysate is run though alkyne agarose beads. The azido group on the AHA-labeled proteins reacts with the alkyne triple bonds in a cycloaddition reaction, attaching themselves to the beads, while the unlabeled proteins wash through. In this way, the newly synthesized proteins are isolated and can be measured. Elongation of the nascent protein chain can begin once the ribosome has been properly assembled on the mRNA. The met-tRNA enters the P-site of the ribosome causing the ribosome to change shape and open the A-site. A new tRNA which matches the current codon on the mRNA enters the A-site. If the matching is correct, the ribosome changes shape again, and catalyzes the peptide bond reaction between the two amino acids (methionine and the new one). Once the bond is formed, the ribosome changes shape again, moves forward along the mRNA chain putting the newly added amino acid in the P-site and opening the A-site again. The "used up" tRNA (formerly met-tRNA in this case) is moved to the E-site. This process continues until the ribosome comes upon a stop codon (UAG, UGA or UAA). A detailed description of the ribosomal conformation changes can be found in [52].

Unsurprisingly, elongation (and thereby protein synthesis) is a much slower process than RNA synthesis. For reference, in *E. coli* elongation occurs at ∼20 amino acids per second [53], whereas RNA is synthesized ∼1000-5000 nucleotides per second [54]. The discrepancy is from the combinatorics of the job at hand and the proofreading steps necessary. There are only 4 nucleotides for RNA synthesis compared to the 20 amino acids for protein synthesis. For each elongation cycle in translation, tRNA diffuse into the A-site, many of which may not match the current codon. These tRNA need to diffuse away before a new tRNA can enter. The waiting time for the correct tRNA to diffuse into the A-site depends on the concentration of the correct tRNA and also the diffusion rates in the system [55]. In this way, the precise codon sequence can influence the rate of protein synthesis; some amino acids are signified by more than one codon, and the tRNA matching these codons have different abundances in the cell [46, 56, 57]. Indeed, much scientific effort has gone into measuring the decoding rates of the 61 sense codons [56, 58–61]. Furthermore, the concentration of specific tRNAs can be different in different conditions, and this in turn can facilitate or impede synthesis of specific proteins [46, 62]. If the waiting time is too long, mRNA degradation pathways can be activated to "rescue" the ribosomes. Alternatively, the ribosome may be blocked by 3-dimensional objects in the path of translation, also increasing the waiting time. For example, large secondary structures, protein mRNA interactions, other ribosomes or RNA polymerase (in the case of prokaryotes, where transcription and translation are sometimes coupled) can all impeded ribosome progress [57]. In these cases, the ribosome can drop off and thus would not finish translation of the protein. Furthermore, ribosomes can also drop off in the normal elongation process. Each elongation step involves several conformational changes of the ribosome. With each conformational change, the ribosome subunits shift and rotate with respect to each other [52]. Again, in these cases, the ribosome can drop off before completing translation of a full protein. The ribosome drop off rate was estimated to be approximately 1 per 10,000 elongation steps in prokaryotes [55, 63, 64].

Despite evidence of ribosomal drop off both in vitro and in vivo [63, 65–67], analysis of ribosomal footprinting experiments failed to find the existence of measurable ribosome drop off [68–70]. Upon closer inspection of the analysis by [68–70], we found that the method normally used to determine drop off was not sensitive enough to detect it. The reason for this is that the sensitivity of these methods depended on the length of the open reading frame (see Section 3.1.3 for more details). In the systems considered, the drop off rate was simply too low with respect to the length of the open reading frame (i.e. low drop off rate and short genes) to be measurable by those methods. In Chapter 3, we use a novel highly sensitive data analysis method to quantify the ribosome drop off rate in previously collected data (including those from [68–70]). We were able to measure a statistically significant drop off rate ranging from $1.4 \cdot 10^{-4}$ to $5.6 \cdot 10^{-4}$ [3]. *E. coli* has an average gene length of 300 codons – if we assume drop off in the range of $4 \cdot 10^{-4}$ events per codon, approximately 10 out of 100 ribosomes will drop off before reaching the stop codon.

Once the ribosome reaches the stop codon, the last step of translation, termination, can begin. The stop codons are recognized by two release factors (RF1 and RF2). These factors trigger hydrolyzation of the ester bond between the last tRNA and its amino acid, releasing the amino acid chain. The amino acid chain may need further folding or post translational modifications before becoming a functional protein. Release factor 3 (RF3) comes to release RF1 or RF2 from the ribosome's A-site, and from there, the ribosome can be disassembled from the mRNA so that it may be reused. In some cases, the ribosome has been shown to be "recycled" to translate the same mRNA.

The processes of initiation, elongation, drop off and termination are all possible processes that could be influenced for the purposes of gene expression control on the level of protein synthesis. Increased rates of initiation, elongation and termination can all potentially increase the rate of protein synthesis. Increased rate of drop off would decrease the rate of protein synthesis. The processes of protein synthesis serve as the penultimate point of gene expression control. Furthermore, protein synthesis serves as a robust target for antibiotics. Because ribosome structure is well conserved within a kingdom, differences in ribosomes between prokaryotic and eukaryotic translation can be exploited to selectively target bacteria [71].

0.1.4 Gene Expression Control at the level of protein degradation

The abundance of protein present in the cell is determined by the net effects of protein synthesis and protein degradation. Proteins are extensively turned over. Similar to mRNA, protein degradation facilitates changes in the gene expression pattern, regulation of cellular processes, recycling of protein precursors and removal of damaged proteins. Protein lifetimes vary for different proteins – for example, in mammalian cells, proteins involved in signaling pathways tend to have short lifetimes, on the order of minutes, while proteins for basic cellular processes such as translation, respiration and metabolism tend to have longer lifetimes, on the order of hours or even days [72].

Proteins are degraded through a number of different pathways. Similar to mRNA degradation, proteins can be degraded through specific as well as non-specific degradation pathways. In eukaryotes, the major protein degradation pathways are through lysosomes (non-specific) and the Ubiquitine-Proteasome system (specific).

Non-specific protein degradation occurs in special organelles called lysosomes. Lysosomes are small, acidic organelles with a suite of more than 50 different enzymes to break down all types of biomolecules, including proteins. Molecules are introduced into the lysosome through autophagocytosis or endocytosis, and the suite of enzymes degrade the contents. In addition to non-specific protein degradation pathways, there are also a number of protein-specific pathways. The most prominent pathway in eukaryotes is the ubiquitine-proteasome system (UPS). In this system, a peptide bond is formed between the target protein and ubiquitine, a small 76 amino acid protein. A single protein can be ubiquinated up to 4 times. Marked proteins are shuttled to proteasomes for degradation.

The UPS pathway is a prime example of a protein degradation which can be described as a multi-step process with several biochemical stages. For these pathways, the widely accepted assumption of exponential decay is often not sufficient to describe the degradation of the target proteins. Furthermore, the erroneous assumption of exponential decay leads to incorrect calculations for average life time and half-lives of the proteins in question.

In Chapter 4 we characterize the degradation curves for over 5000 proteins in the mouse proteome. The data was collected by Erik McShane in the lab of M. Selbach at the Max Delbruck Center in Berlin-Buch using metabolic pulse-chase labeling and quantitative mass spectrometry. By fitting the degradation data to two candidate models and using the Akaike Information Criterion to evaluate the suitability of each model with respect to the available data, we determine that ∼15% of the proteins are degraded in a non-exponential fashion. In all cases, the non-exponentially decaying (NED) proteins were less stable in the first few hours of their life and more stable as they age. Repeating the experiments and analysis in human fibroblast cells, we found that the degradation profiles were mostly conserved. Many non-exponentially degraded proteins are subunits of multiprotein complexes. In all, we show that non-exponential degradation is a non-trivial contribution to protein degradation. It is evolutionarily conserved and it has functional consequences for protein complex formation [4].

0.2 Experimental Techniques

My contribution to the research reported in this thesis is purely focused on mathematical modeling and statistical data analysis. However, a crucial step preceding both of those processes is the collection of experimental data. In this section, I describe some of the experimental techniques for collection of the data used in our studies.

0.2.1 F-luc reporting (Chapter 1)

There are a number of bioluminescent molecules which can be used in biochemical experiments. One such molecule is “Firefly luciferase” (F-luc). This is a class of enzymes produced by fireflies that generate light. The gene to make F-luc can be coupled to a gene of interest. By measuring the amount of luminescence, researchers can determine the expression level of the gene of interest.

In [10], F-luc is coupled to Nerfin, a known target of the microRNA miR-9b.

0.2.2 mRNA Sequencing and Ribosomal Footprinting (Chs. 2 and 3)

RNA sequencing (RNA-seq) is the key technology critical to ribosome profiling experiments. With this technology, we can measure the amount (and identities) of RNA at a given moment in time. mRNA sequencing is often preformed in parallel with Ribo-Seq to serve as a baseline for comparison.

First, the RNA is isolated and digested with nucleases to generate fragments.[1] The fragments are ligated to adapters (the sequence of the adapter is known), and the resulting fragments are amplified by PCR[2]. The adapters associate with the clusters in the

[1] If the experiment will be used as a baseline for a parallel Ribo-Seq experiment, digestion time will be adjusted to produce fragments approximately 30 nucleotides long (the length of the ribosome protected fragments), and further selected by a gel.

[2] Polymerase Chain Reaction: a technology to amplify (make many copies of) a particular RNA or DNA sequence. It is preformed by reacting the template RNA/DNA with RNA/DNA polymerase and the appropriate RNA/DNA bases. The temperature is cycled to promote and dissociate the double strand pairing, thereby copying the template sequence.

sequencing machine, immobilizing the fragment. Then, fluorescently labeled nucleotides (one color for each base) are added, and unincorporated nucleotides are washed away. The fragments grow one nucleotide at a time, and the fluorescence is captured by a camera. The flourescent part is then cleaved away, and the cycle is repeated again. In this way, the sequence is "read out" by sequential photos. Modern machines can read up to 500,000 bases/hour. The runs take 2-11 days to complete, and the output files range from 5-600GB.

Ribosomal footprinting is a variant on RNA-seq. Instead of sequencing all mRNA, the mRNA sequences actively being translated are isolated and sequenced. During translation, a length of the mRNA ∼30nt long is completely inside the ribosome. Conveniently, the ribosome sterically protects the mRNA from nucleases. These "ribosome protected fragments" (RPF) are then isolated and sequenced to determine the locations of the ribosomes.

For sequencing experiments, mapping the fragments to positions on the reference genome is a non-trivial computational task. For this, we use a pipeline of existing software tools. Details can be found in Chapter 3 and Supplementary Section 2.

0.2.3 Pulse Chase (Chs. 4 and 5)

Pulse chase experiments are often used to measure the decay of macromolecules such as proteins or mRNA. The experiment consists of two phases: First in the "pulse" phase, cells are exposed to a labeling compound that is integrated into newly synthesized macromolecules of interest. For example, if one wants to follow the fate of proteins, radio or isotope labeled amino acids can be introduced. If one wants to follow the fate of mRNA, radio or isotope labeled nucleic acids can be introduced. In the second phase, the chase phase, the same compound in the unlabeled form is added in excess, replacing the labeled form. Any macromolecules synthesized during the "chase" phase will not be labeled. At the end of the chase, the macromolecules that were synthesized during the pulse can be tracked.

Following the pulse, the pool of labeled molecules consists of molecules of different ages, with age ranging from a minimum of 0 (newly synthesized) to a maximum of the pulse length. This population of molecules of various ages is collected in the chase phase.

For molecules that decay exponentially, the age of the molecule does not affect the probability of degradation. However, this is not the case for non-exponentially degrading molecules. Thus, pulse-chase experiments with different pulse lengths for non-exponentially decaying molecules result in different degradation patterns, due to different starting age distributions in the pool of measured molecules.

With a mathematical model of the degradation process, one can derive the functions to predict the degradation pattern resulting from a given pulse length. In Chapter 5, we go over the calculation and discuss the implications of the pulse time on the measurable dynamics and considerations for optimal experimental design. The theory is applied to all the calculations in Chapter 1 and 4 as well as the case study example in Chapter 5.

0.2.4 Quantitative Mass Spectrometry (Chapter 4)

Mass spectrometry is a method to help identify molecules in a sample. The sample is fragmented by electron beams and the fragments are sorted by their mass/charge ratio. The abundances of the fragments can be measured, and the identity of the original molecule can be extrapolated with knowledge of how molecules tend to fragment.

By clever manipulation of the growing conditions in cells which then produce the sample to be measured in the mass spectrometer, Erik McShane and Matthias Selbach at the Max Delbruck Center in Berlin-Buch measure the abundances of over 5000 proteins in the cell over time. To do this, they use a combination of stable isotope (SILAC) and click-chemistry (AHA) labeling (Figure 2).

Cell cultures are cultivated in growth medium containing amino acids labeled with heavy/medium/light isotopes. The three populations (heavy, medium and light) are subject to a labeling pulse where methionine is depleted and replaced by Azidohomoalanine (AHA) for 1 hour. AHA is a non-natural amino acid which incorporates readily in place of methionine and contains an azido group which can be used in click chemistry. Thus in all three populations, the proteins synthesized in the last hour are labeled with AHA.

Measurement of protein abundance over time is achieved by "chasing" the synthesized proteins in each cell population (heavy/medium/light) for a different time. One cell population serves as the "baseline" measurement with chase time = 0, and the other two populations provide data on later times. At the end of the chases, the populations are mixed together and the proteins are isolated. Protein abundance is measured by quantitative mass spectrometry. For each experiment with three isotope populations, we can collect two time points; one isotope population is used as a baseline representative of the proteins present at t = 0.

The peaks from the differently weighted populations are compared to each other to obtain the relative ratio difference between the time points. In other words, the information we get is that the amount of protein at t = currentTime is X times less then the amount of protein at t = 0.

Each protein fragments in a predictable way, resulting in a characteristic pattern of peaks. The ratio calculation of any one protein and time is aggregated from the data of several peaks. Peaks are analyzed with the MaxQuant software package [73]. The number of peaks evaluated for each calculation is known as the quantification "counts".

In general, proteins of low abundance are not as easy to measure as proteins of high abundance. Furthermore, smaller proteins are not as easy to measure as large proteins. Low-abundance and small proteins will tend to have lower quantification counts. We omit data where the number of quantification counts does not pass a minimum threshold. Lastly, based on the assumption of log-normal errors, all further analysis is with the logged data.

0.3 Mathematical Modeling framework

Models are representations of complex systems or processes. Mathematical models use mathematical concepts and symbolic language to describe the connections and interactions within a system. Indispensible for systems biology, models can be used to better understand biological systems, study the effects of different conditions and make predictions about the system's response.

In Chapter 1, we use a mathematical model to explore the biochemical network for mRNA degradation through the miRISC pathway. In Chapter 3, we model ribosome density on mRNA length as a result of drop-off and use this model to fit the data from RiboSeq experiments. In Chapter 4, we evaluate the protein degradation patterns in the mouse proteome for signs of non-exponential decay. To this aim, we use the data to calibrate two simple models – exponential decay and a 2-stage model – and find model better describes each protein.

Mathematical models can be developed by translating physical phenomena into mathematical equations. For example, a biochemical degradation process which occurs in several steps can be conveniently modeled as a Markov chain. In this model, each Markov state represents a biochemical state of the degradation process. The rate of

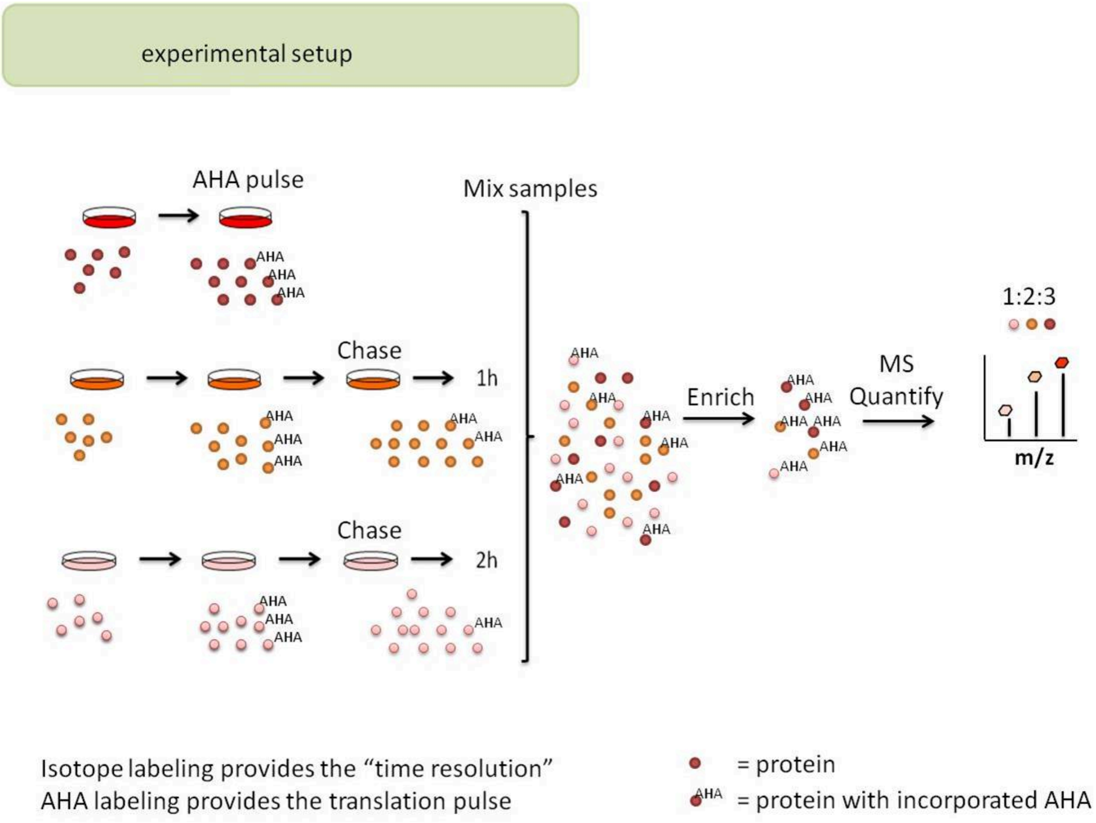

experimental setup
AHA pulse
Mix samples
Chase
1h
Chase
2h
AHA
Enrich
MS
Quantify
1:2:3
m/z
Isotope labeling provides the "time resolution"
AHA labeling provides the translation pulse
= protein
= protein with incorporated AHA

Figure 2: Experiment scheme for measuring protein degradation. Cell cultures are cultivated in growth medium containing amino acids labeled with heavy/medium/light isotopes, red/orange/pink, respectively. The three populations (heavy, medium and light) are subject to a labeling pulse where methionine is depleted and replaced by Azidohomoalanine (AHA) for 1 hour. Depiction of the process by Erik McShane.

reaction or transitions between the states are the fluxes between the Markov states. Thanks to this modeling framework, we are able to create a theoretical bridge between the single molecule perspective of the biochemical process of degradation and the cell culture measurements, as explained in Chapter 5.

In summary, the Markov chain gives us a description of a probabilistic stochastic process over the states. From this, we can solve the equations to find the lifetime distributions of the molecules, and then transform them into predictions of the population average [74]. While single molecules are difficult to measure in the lab, population averages are much more easily measured. These measurements can be used to establish the structure of the underlying biochemical network and solve for the transition rates in the model.

0.4 Closing remarks

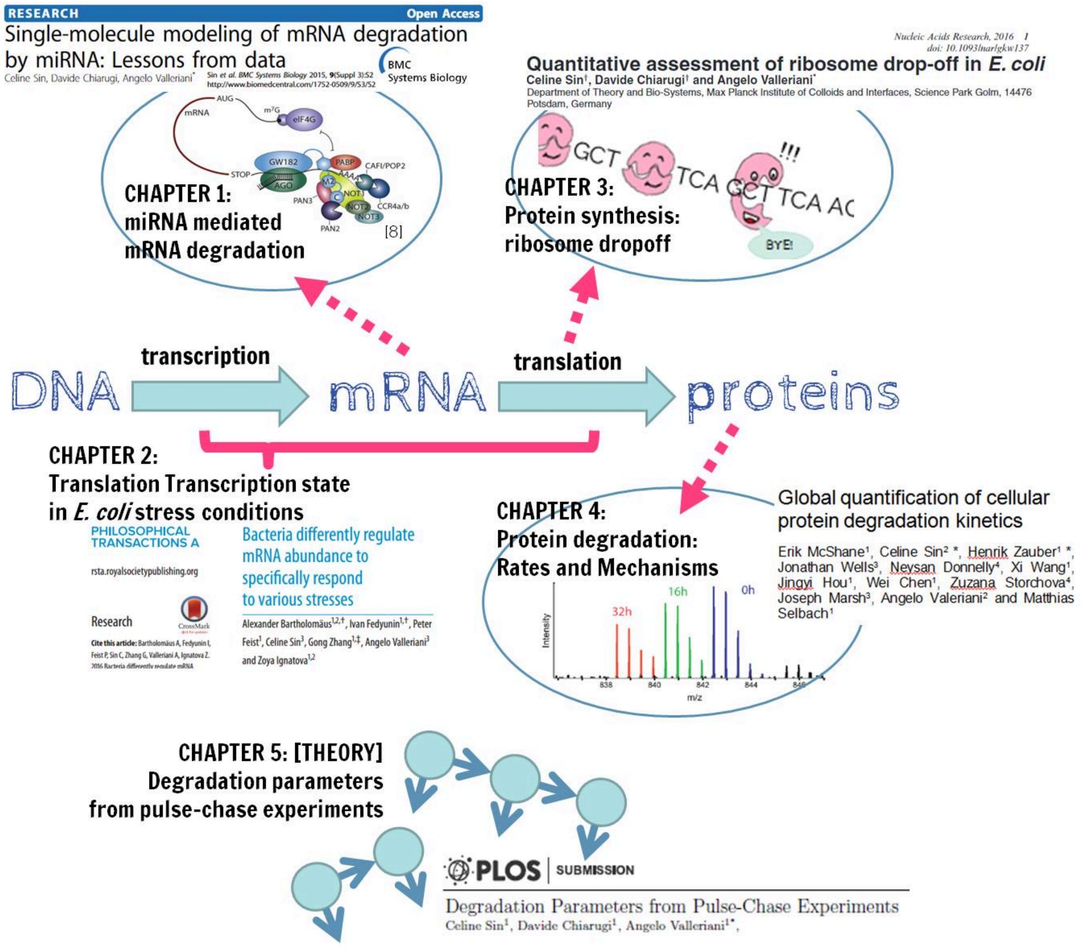
In the process of gene expression, there are a number of points where gene expression control can occur. I have arranged the Chapters 1-4 in order of when the mechanism(s) studied might occur.

The first checkpoint is to control the amount of mRNA available for translation. In Chapter 1, we explore the biochemical network for one of the major pathways for mRNA degradation - the miRISC pathway. The second checkpoint is to control the amount of proteins. The processes in Chapter 2 encompasses both checkpoints in *E. coli* under stress conditions. With mRNA-seq, we get a picture of the transcriptional state representative of the net effects of transcription and mRNA degradation. With Ribo-seq, we get a picture of the translational state, representative of the effects of initiation and elongation[3]. Chapter 3 quantifies the effect of ribosome drop-off. Changes in ribosome drop-off rate affect the protein synthesis rate (second checkpoint). And lastly, we have a proteome wide study of protein degradation in Chapter 4.

Chapter 5 presents the mathematical modeling framework used in most of our work and derivation of the experimentally measurable quantities. Furthermore, the implications of pulse time on the measurable dynamics and considerations for optimal experimental design are discussed.

This dissertation is organized as a "cumulative dissertation". Part I contains the papers as well as a discussion of the work completed. Chapter 1, 2, 3 and 5 have been published, and Chapter 4 is under review. The supplementary sections to each of the papers are found in Part II. Other work not included in the papers are found in the appendices, as well as documentation for the tools developed in-house for analysis.

[3] but not complete synthesis nor protein degradation



RESEARCH Open Access

Single-molecule modeling of mRNA degradation by miRNA: Lessons from data

Celine Sin, Davide Chiarugi, Angelo Valleriani*

Sin *et al. BMC Systems Biology* 2015, **9**(Suppl 3):S2
http://www.biomedcentral.com/1752-0509/9/S3/S2

BMC Systems Biology

CHAPTER 1:
miRNA mediated
mRNA degradation

Nucleic Acids Research, 2016 1
doi: 10.1093/nar/gkw137

Quantitative assessment of ribosome drop-off in *E. coli*

Celine Sin†, Davide Chiarugi† and Angelo Valleriani*

Department of Theory and Bio-Systems, Max Planck Institute of Colloids and Interfaces, Science Park Golm, 14476 Potsdam, Germany

CHAPTER 3:
Protein synthesis:
ribosome dropoff

CHAPTER 2:
Translation Transcription state
in *E. coli* stress conditions

PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Research

Cite this article: Bartholomäus A, Fedyunin I, Feist P, Sin C, Zhang G, Valleriani A, Ignatova Z. 2016 Bacteria differently regulate mRNA

Bacteria differently regulate mRNA abundance to specifically respond to various stresses

Alexander Bartholomäus[1,2,†], Ivan Fedyunin[1,†], Peter Feist[1], Celine Sin[3], Gong Zhang[1,‡], Angelo Valleriani[3] and Zoya Ignatova[1,2]

CHAPTER 4:
Protein degradation:
Rates and Mechanisms

Global quantification of cellular protein degradation kinetics

Erik McShane[1], Celine Sin[2] *, Henrik Zauber[1] *, Jonathan Wells[3], Neysan Donnelly[4], Xi Wang[1], Jingyi Hou[1], Wei Chen[1], Zuzana Storchova[4], Joseph Marsh[3], Angelo Valeriani[2] and Matthias Selbach[1]

CHAPTER 5: [THEORY]
Degradation parameters
from pulse-chase experiments

PLOS | SUBMISSION

Degradation Parameters from Pulse-Chase Experiments

Celine Sin[1], Davide Chiarugi[1], Angelo Valleriani[1*],

Figure 3: Overview of the dissertation
miRISC cartoon from [10]

0.5 Author contributions

In this cumulative dissertation, I have had the fortune to make many collaborations with other scientists. The following outlines my personal contributions to the projects. All work was supervised by Dr. Angelo Valleriani in the department of Theory and Biosystems at the Max Planck Institute for Colloids and Interfaces.

Chapter 1: Single-molecule modeling of mRNA degradation by miRNA: Lessons from data [1]

> All authors took part in conceiving the project, developing the model and writing the manuscript. I extracted the data from Braun et. al., 2011, preformed the data analysis and parameter estimation.

Chapter 2: Bacteria differently regulate mRNA abundance to specifically respond to various stresses [2]

> Large collaboration with many members. I designed many of the metrics. I made the initial calculations on copy number, ribosome density and translational burden. Our main contribution to the manuscript was the fold change analysis. I also made the calculations to test for reproducibility and statistical significance. Additionally (but not included in the manuscript), I preformed additional analyses on codon usage, ribosome speed and trends by decile. The structure used to manipulate the data is documented in Appendix C.

Chapter 3: Quantitative assessment of ribosome drop-off in *E. coli* [3]

> This project was the evolution of [2]. I discovered that there was a length dependent relationship to ribosome density, through the decile analysis. Furthermore, it appeared that the relationship between length and ribosome density followed a power law. Together, we concluded that this could be due to the phenomenon of ribosome drop-off, and we designed an analysis to test it. DC preformed all the bioinformatics and coded the scripts for analysis. I used the mapped reads from DC to confirm the results in parallel using my own scripts.

Chapter 4: Global quantification of cellular protein degradation kinetics [4]

Another large collaboration with many members. I was involved very early on with the idea of age-dependent degradation. I designed the normalization schemes. I implemented all the normalization schemes and parameter estimation machinery. I made the calculations to describe the lifetime properties of each protein (e.g. average lifetime, half-time, steady state distributions, etc.) I also made many of the calculations to evaluate reproducibility, sensitivity and statistical significance. The structure used to manipulate the data is documented in Appendix D.

Chapter 5: Degradation Parameters from Pulse-Chase Experiments [5]

All authors took part in conceiving the project, developing the experiment and writing the manuscript. I extracted the data from Wheatley, et. al., 1980. I preformed the data analysis, parameter estimation and sensitivity analysis on the system.

1

mRNA degradation through the miRISC pathway

A version of this chapter has been published as:
Sin C, Chiarugi D, Valleriani A (2015) Single-molecule modeling of mRNA degradation by miRNA: Lessons from data. BMC Systems Biology 9(Suppl 3):S2.

Online version:
http://www.biomedcentral.com/1752-0509/9/S3/S2

CONTRIBUTIONS: All authors took part in conceiving the project, developing the model and writing the manuscript. I extracted the data from Braun et. al., 2011, preformed the data analysis and parameter estimation.

Abstract

Recent experimental results on the effect of miRNA on the decay of its target mRNA have been analyzed against a previously hypothesized single molecule degradation pathway. According to that hypothesis, the silencing complex (miRISC) first interacts with its target mRNA and then recruits the protein complexes associated with NOT1 and PAN3 to trigger deadenylation (and subsequent degradation) of the target mRNA. Our analysis of the experimental decay patterns allowed us to refine the structure of the degradation pathways at the single molecule level. Our analyses show that the hypothesized biochemical network must be complemented by additional pathways in order to fit the data. More surprisingly, we found that if the previously hypothesized network was correct, only about 7% of the target mRNA would be regulated by the miRNA mechanism, which is inconsistent with the available knowledge.

Based on systematic data analysis, we propose the alternative hypothesis that NOT1 interacts with miRISC before binding to the target mRNA. Moreover, we show that when miRISC binds alone to the target mRNA, the mRNA is degraded more slowly, probably through a deadenylation-independent pathway.

The new biochemical pathway we propose both fits the data and paves the way for new experimental work to identify new interactions.

1.1 Introduction

1.1.1 Background

In living cells, the level of protein expression is thoroughly regulated. Many crucial processes for this regulation occur at the post-transcriptional level. In this context, the control mechanisms acting on messenger RNAs (mRNAs) play a pivotal role. Indeed, living cells are endowed with a number of biochemical pathways converging on cytosolic mRNAs, that serve to enhance or repress gene expression. These pathways are known to operate (1) either by enhancement [75, 76] or repression [15, 77] of translation or (2) modulating mRNA lifetimes [15, 77–79]. The global picture emerging from the growing body of experimental evidence, depicts a complex interaction network which affects the mRNAs available for translation. This network is composed of several biochemical pathways, often interwoven and cross-talking [80, 81], and involves mRNA binding proteins as well as non coding RNAs [31, 82, 83]. While there are a number of mechanisms responsible for mRNA degradation in eukaryotic cells [31], the decay of messages mediated by micro-RNAs (miRNAs) plays a prominent role in the control of gene expression [15, 84, 85].

Despite extensive study, the topology and dynamics of the miRNA-mediated mRNA degradation pathway are still unclear. One of the main challenges stems from the fact that intermediate states of the pathway are unknown or difficult to quantify; experimentally, it is only feasible to measure the final state of the pathway (e.g. the decay pattern of the target mRNA). Bridging the gap between observed *decay patterns* and *degradation pathways* is non-trivial [86], since the former refer to a population average and the latter refers to the single-molecule stochastic process of degradation. Here we apply a rigorous strategy to reconstruct the miRNA-mediated degradation pathway, starting from experimentally measured decay patterns. Surprisingly, we find the previously proposed pathway not consistent with the experimental data. We propose an alternative model which both fits the decay pattern and allows us to gain some insight on the network topology.

1.1.2 The experimental data

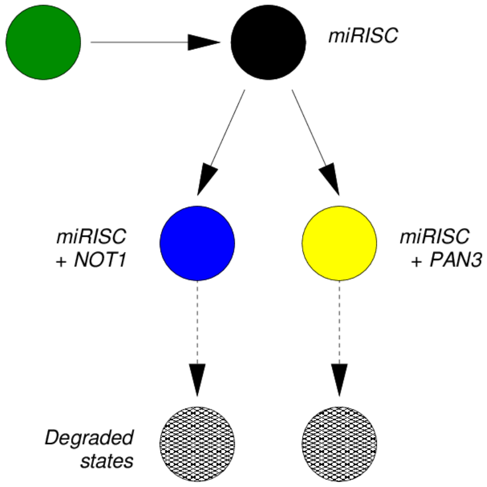


Figure 1.1: The biochemical degradation network hypothesized by Braun *et al.* [10] for the degradation of a target mRNA by miRNA in *D. melanogaster* S2 cells. According to this network, the mRNA in its initial state (green circle) first binds to the miRISC complex thus leading to a new biochemical state (black circle). The miRISC and its target then recruit the proteins NOT1 and/or PAN3, leading to the two states indicated with the blue and yellow circle, respectively. From these two states the mRNA is finally degraded through a complex sequence of events including deadenylation followed by decapping [15]. This unspecified sequence of events is indicated with dotted arrows in all figures of this paper.

To clarify the interplay of the various factors in a degradation pathway involving miRNA, and the protein complexes NOT1 and PAN3, Braun *et al.* [10] performed a series of controlled knock-down experiments in *D. melanogaster* S2 cells containing constructs for the miRNA miR-9b, its target mRNAs, namely the F-Luc-Nerfin mRNAs, and the factors NOT1 and PAN3, which are known to trigger mRNA deadenylation. In each experiment, a subset of NOT1, PAN3 and/or the miRNA miR-9b were selectively knocked down, yielding cell lines expressing different combinations of those factors: a control line without the miRNA, a cell line with miR-9b only, a cell line with NOT1+miR-9b (but not PAN3), a cell line with PAN3+miR-9b (but not NOT1), and finally a cell line with all three factors NOT1+PAN3+miR9-b. After steady state expression of the factors, the transcription of mRNA was blocked and the decay patterns over time were measured and reported (see Figure 1.2).

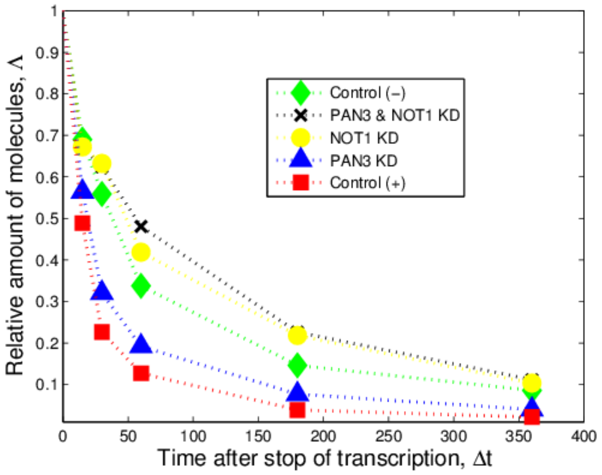


FIGURE 1.2: The experiment performed by Braun *et al.*[10] consists of knocking down several permutations of the target's degradation factors. The "Control (-)" data concern an experimental set-up in which the miR9-b is not expressed; the "PAN3 & NOT1 KD" data concern an experimental set-up in which only the miR9-b is expressed but both NOT1 and PAN3 are knocked down; the "NOT1 KD" data concern a set-up in which miR9-b and PAN3 are expressed while NOT1 is knocked down; the "PAN3 KD" data concern a set-up in which miR9-b and NOT1 are expressed while PAN3 is knocked-down; the "Control (+)" data concern a set-up in which all three factors are expressed. The data have been extracted from figure 7A of [10]. The numerical values are reported in the supplementary materials.

The conclusion of this detailed experimental study is that NOT1 is a more relevant factor than PAN3 in destabilizing the mRNA [10]. When only NOT1 is knocked down, the decay of F-Luc-Nerfin mRNA is significantly slower (yellow curve, Figure 1.2) than the control (red curve, Figure 1.2). In contrast, the effect of PAN3 knock down is less significant (blue curve, Figure 1.2). These findings apparently confirm that the degradation pathway through NOT1 in Figure 1.1 is the most prominent pathway for degradation of the target mRNA. Although this conclusion is relatively robust, the published analyses do not validate the hypothesized biochemical degradation pathway given in Figure 1.1. Indeed, in the negative control (green curve in Figure 1.2), the miRNA is knocked-down so that the formation of a specific silencing complex miRISC is suppressed, yet the target mRNA still decays. Additionally, when only the miRNA is expressed while PAN3 and NOT1 are knocked-down, the target mRNA decays (black curve in Figure 1.2), but is definitively more stable than in the negative control. Both of these cases suggest that the model hypothesized in Figure 1.1 should be expanded to include additional degradation pathways.

An important conceptual consideration is that Figure 1.1 depicts degradation from the single-molecule perspective whereas the curves in Figure 1.2 are averages as a function of time. Therefore, our strategy consists in starting with the network shown in Figure 1.1 and validate it against the experimental decay patterns. At the same time, we will propose alternative parsimonious extensions of the network when the validation fails.

1.2 Methods

As previously mentioned, the relationship between *degradation pathways* (such as the one in Figure 1.1) and *decay patterns* (such as those in Figure 1.2) is not trivial. If the decay pattern was exponential, the halftime of the mRNA population estimated from the decay pattern would be directly related to the rate of decay of single molecules. The decays shown in Figure 1.2 are clearly not exponential; if they were, the traces would appear as straight lines when plotted in a linear-log scale (see Figure 7A in [10]).

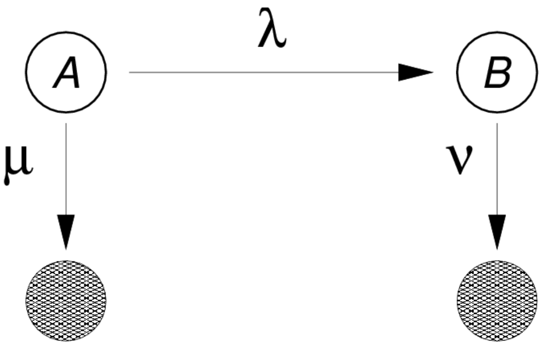


FIGURE 1.3: This two-state model is able to fit all of the decay patterns in figure 1.2, by an appropriate choice of the three fitting parameters, λ, μ, ν. In this context, the mRNA is initially in the state A. From this state it can be degraded directly with a rate μ or change biochemical state, with rate λ, from which it is then degraded with rate ν. Although this network performs a better fit of the data and is preferred to the exponential fitting based on the AIC criterion, the interpretation of the states is unclear. The fitting curves and the parameters are discussed in the supplementary materials.

The issue of relating complex degradation pathways to decay patterns has been tackled in [86] and in [87]. In [86] it was shown that decay patterns similar to those depicted in

Figure 1.2 can be generated by single-molecule networks satisfying certain properties, if one assumes that the transitions between biochemical states can be modeled as first-order chemical reactions. The mathematics supporting this reasoning was presented in [86] and is summarized in the supplementary materials where we show how to derive the necessary mathematical functions using first passage time methods [88–90]. In order to generate decay patterns such as those in Figure 1.2, the corresponding single-molecule degradation pathway must be composed of at least two states from which degradation is possible. Thus, in principle, one can either hypothesize a network of states that represents the biochemical pathway of degradation based on predictions and prior knowledge, or one can use the mathematical relationships mentioned above to find the most parsimonious network that fits that data. The most parsimonious network of states that is able to fit each one of the curves in Figure 1.2 is given by a two-state model depicted in Figure 1.3.

While the network in Figure 1.3 results in a definitively better fit to the data and thus could be used to derive quantities such as the average lifetime and the age dependent degradation rate, it does not address the question as to whether or not the network in Figure 1.1 is a suitable framework for the decay patterns observed in Figure 1.2. To address this question we employ a hierarchical strategy: (i) we start by fitting the negative control decay pattern (the green trace “Control (-)” in Figure 1.2) to the most parsimonious model (Figure 1.3) and thereby fix the corresponding three rates; (ii) we then consider the next decay pattern with one additional decay factor active and enlarge the network of states to accommodate the additional decay factor. We continue until each curve has been evaluated and the corresponding network is built.

1.3 Results

Based on the hierarchical strategy above, we start with the “Control (-)” curve (green decay pattern, Figure 1.2). This curve describes the decay of the mRNA when none of the degradation factors (miRNA, NOT1 and PAN3) are present. In the framework of the hypothesized network in Figure 1.1, this would correspond to all downstream processes inactive thus reducing the network to just the first state (green circle). This is obviously not sufficient to explain the observed decay, since a single state without decay would produce a horizontal line (*i.e.* no decay). This one-state scenario is also not consistent with biological reality. Indeed, even the most stable cellular macromolecule is eventually degraded. In particular, there are many biochemical pathways devoted to mRNA degradation [31].

In the absence of further information, we fit the green ”Control (-)” curve of Figure 1.3 to the most parsimonious (or minimal) network that still captures the dynamics of the data

(Figure 1.4). It turns out that in absence of miRNA-dependent degradation, mRNA molecules can be degraded through two pathways, differing by their kinetic features (see Figure 1.3). The first class of pathways (governed by the rate μ in Figure 1.3) is characterized by a single step and it can be representative of the set of "constitutive" reactions which target mRNAs non-specifically and are catalyzed by enzyme complexes such as the exosome [31]. The second class of pathways (along the path λ and ν in Figure 1.3) exhibits two steps and represents the degradation processes (independent of miRNAs) passing through a control step of a more complex degradation pathway (e.g. the preliminary binding of specific proteins to the target mRNA). One such example independent of miRNA and the NOT1/PAN3 factors is the ARE-mediated degradation pathway [79, 91].

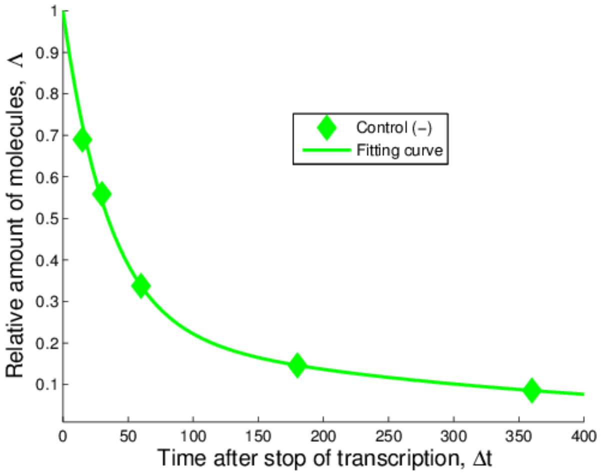


FIGURE 1.4: The negative control decay pattern can be fitted with the single molecule network from figure 1.3. This delivers the rates λ, μ and ν that will be kept constant through any successive enlargement of the network when considering the other decay patterns. The values of the rates are: $\lambda = 0.0008\,\mathrm{min}^{-1}$, $\mu = 0.0276\,\mathrm{min}^{-1}$, $\nu = 0.0028\,\mathrm{min}^{-1}$, with confidence intervals [0.0002, 0.0013], [0.0229, 0.0324], [0.0018, 0.0038], respectively.

1.3.1 The crisis of the original hypothesis

At the next level of our hierarchical approach we consider the decay pattern that results when the miRNA is expressed but NOT1 and PAN3 are knocked-down (black decay pattern in Figure 1.2). When PAN3 and NOT1 are knocked down, the arrows from the black state to the blue and yellow states are absent, resulting in the right pathway having no transition to degradation. However, if we look at the corresponding decay pattern (black line in Figure 1.2), we realize that such a structure is not compatible with the data because a vertex without transition to degradation would imply a flattening of the curve to a steady state amount of mRNA, corresponding to the amount of mRNA

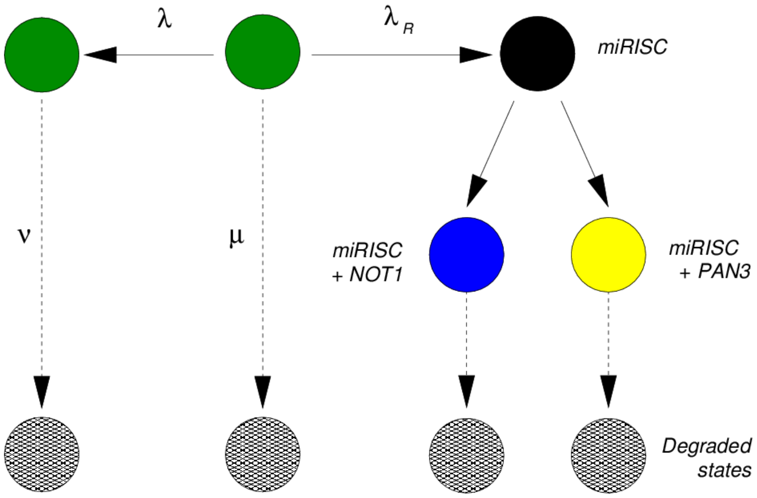


FIGURE 1.5: The initial hypothesis of Braun *et al.* [10] is complemented with an alternative pathway that competes with the miRNA-mediated pathway. Initially, all mRNA start at the state represented by the central green circle and follow one of the two competing paths. The path towards the left (denoted by λ) is exclusive of the miRISC pathway (towards the right, denoted by λ_R), and vice versa. The rates λ, μ and ν have been fixed through the fitting performed in Figure 1.4. The negative control "Control (-)" data concern the decay pattern of the mRNA when only the states represented by the green circles are available.

arrested in this rightmost state (black circle). To model the observed decay of mRNA, we need to postulate an additional transition from the rightmost state (represented by the black circle, after binding with miRISC) to degradation. A possible interpretation is that the additional transition (from the black circle to degradation in Figure 1.6) includes unknown biochemical degradation pathways which are independent of deadenylation. Supporting this hypothesis are the findings reported in [92]: they report that the binding of the miRISC complex to the target mRNA can promote the dissociation of Poli-A-Binding Proteins (PABPs). Indeed, PABPs are known to protect the poli-A tail of the mRNA from being hydrolyzed, thus stabilizing the mRNA. Thus, miRNA-mediated PABP dissociation can trigger NOT1 and PAN3-independent deadenylation, which eventually leads to the degradation of the mRNA [92].

Fitting the data to the network depicted in figure 1.6 reveals several crucial aspects of the hypothesized network of Figure 1.1 and shows the shortcomings of the latter. While the fit of the data using the network in Figure 1.6 works pretty well (see Figure 1.7), it fixes the rate λ_R associated to the binding of miRISC on the mRNA. This rate is therefore independent of the transitions occurring downstream.

The next step in our hierarchical program, however, would be to take the next decay patterns and fit them to the network given in Figure 1.5 activating the appropriate pathway depending on which factor (NOT1 and/or PAN3) is present while keeping λ_R

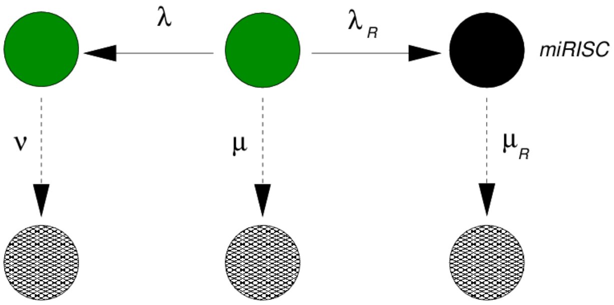


FIGURE 1.6: After fixing the rates μ, λ and ν from the fitting of the negative decay pattern in figure 1.4, we use this network in order to model the decay pattern when miRNA is expressed but NOT1 and PAN3 are knocked down. In order to perform this fit, however, we must add a new transition to degradation after the binding of the miRISC (rightmost transition to degradation). This new transition might contain a complex set of processes which are most likely independent of deadenylation. In the light of recent results, this arrow could include deadenylation-independent decapping [35, 92].

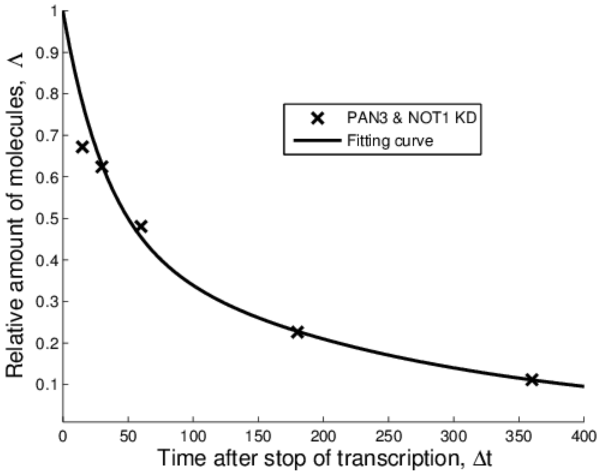


FIGURE 1.7: After fixing the rates μ, λ, ν from the fitting of the negative decay pattern in figure 1.4, we use the network in figure 1.6 to fit the decay pattern when miRNA is expressed but NOT1 and PAN3 are knocked down. The values of the rates are $\lambda_R = 0.0023\,\text{min}^{-1}$ and $\mu_R = 0.0052\,\text{min}^{-1}$ with 95% confidence intervals [0, 0.0052] and [0.0030, 0.0074], respectively.

fixed. Before doing that, however, a simple computation shows that the fraction of mRNA going through the miRISC pathway is given by

$$\sigma_R = \frac{\lambda_R}{\lambda_R + \lambda + \mu} \sim 0.074\,, \tag{1.1}$$

i.e. , about 7% of the whole mRNA binds to miRISC complexes in the absence of NOT1 and or PAN3, based on the network shown in Figure 1.6. This discovery leads to two conclusions. First, the fraction of mRNA that can be manipulated after binding with miRISC is so small that an enlargement of the network by including a separate NOT1 and a separate PAN3 pathway downstream of miRISC binding becomes meaningless. Indeed, attempts to do so lead to very poor fitting of the remaining curves (see supplementary materials). Second, such a small fraction of miRNA-regulated mRNA (about 7%) would indicate that miRNA cannot be considered a strong mechanism of gene regulation, contrary to the experimental evidence that miRNA is a strong regulatory mechanism. Therefore, consistent with the strong role of miRNA in the regulation of mRNA, we are forced to partially reject the hypothesis formulated in Figure 1.1 and revise it in search for other possible interactions between miRISC, PAN3 and NOT1.

Finally, the comparison between the decay pattern fitted in Figure 1.4 and 1.7 shows that binding of miRISC alone does stabilize the mRNA compared to when the miRNA is not expressed. This is a strong indication that miRISC "protects" the target mRNA from the action of alternative, competing degradation pathways.

1.3.2 A new hypothesis arises from the data

Since the initial hypothesis that miRISC binds to the mRNA and then recruits the NOT1 molecule does not result in a reasonable fit, we can hypothesize that miRISC binds to NOT1 *before* recruiting the target mRNA. This hypothesis is formulated in Figure 1.8, which can be used to fit the data where only the PAN3 complex has been knocked down.

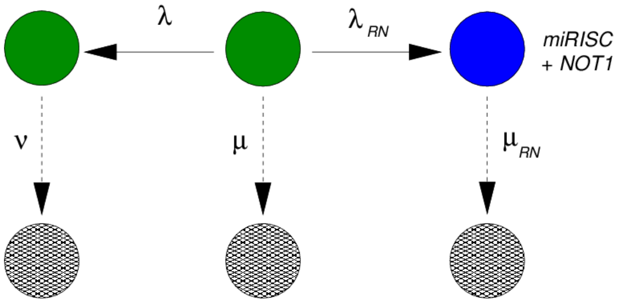


FIGURE 1.8: Biochemical network able to fit the data when PAN3 has been knocked down. After fixing the rates μ, λ and ν from the fitting of the negative decay pattern in figure 1.4, we use this network in order to model the decay pattern when miRNA is expressed but PAN3 is knocked down ("PAN3 KD" data). The transition from the central green state to the state with the mRNA bound to the complex miRISC and NOT1 is ruled by the transition rate λ_{RN}. The downwards transition, ruled by the rate μ_{RN} includes several steps that cannot be specified from these data.

The fit is indeed very good, as seen in Figure 1.9. Based on this result, the sole effect of NOT1 binding to the miRISC leads to a strong increase of the percent of mRNA that are degraded through miRISC activity, given by

$$\sigma_{RN} = \frac{\lambda_{RN}}{\lambda_{RN} + \lambda + \mu} \sim 0.64\,, \tag{1.2}$$

which emphasizes the strong role of NOT1 in the degradation of mRNA.

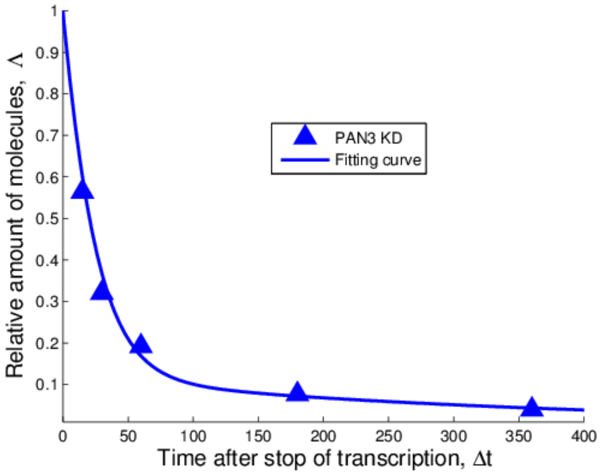


FIGURE 1.9: After fixing the rates μ, λ, ν from the fitting of the negative decay pattern in Figure 1.4, we use the network in Figure 1.8 in order to fit the decay pattern when miRNA is expressed but PAN3 is knocked down. The values of the rates are $\lambda_{RN} = 0.046\,\text{min}^{-1}$ and $\mu_{RN} = 0.0461\,\text{min}^{-1}$ with 95% confidence intervals [0.0305, 0.0687] and [0.0319, 0.0602], respectively. These rates indicate that in the absence of PAN3, about 64 % of the mRNA in the experiment are degraded by the action of miRISC.

The final curve of the experiment in [10] concerns the action of all the factors together. On the basis of the results obtained so far in Figure 1.8 and Figure 1.9 there may be several hypotheses about the possible combined action of PAN3 and NOT1. Since PAN3 alone (yellow curve in the original data shown in Figure 1.2) does not have a significant effect on the decay of the mRNA compared to the action of miRISC alone, we conclude that PAN3 works cooperatively with NOT1 by forming a complex miRISC+NOT1+PAN3 before binding to the target mRNA. This hypothesis is formulated in Figure 1.10.

The data fits the network in Figure 1.10 quite well, as one can see in Figure 1.11.

By using the values λ_{RNP} and μ_{RNP} we can again compute the fraction of target mRNA that is degraded by the action of miRISC+NOT1+PAN3:

$$\sigma_{RNP} = \frac{\lambda_{RNP}}{\lambda_{RNP} + \lambda + \mu} \sim 0.84\,, \tag{1.3}$$

indicating that this model produces the strong regulatory effect of the miRNA on its target as expected.

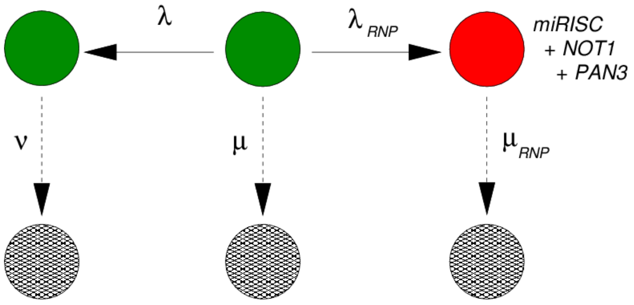


FIGURE 1.10: Biochemical network able to fit the data when all three factors miRISC, NOT1 and PAN3 are expressed. After fixing the rates μ, λ and ν from the fitting of the negative decay pattern in Figure 1.4, we use this network in order to model the decay pattern when all three factors are expressed. The transition from the central state (green) to the state with the mRNA bound to the complex miRISC + NOT1 + PAN3 is ruled by the transition rate λ_{RNP}. The downwards transition, ruled by the rate μ_{RNP} includes several steps that cannot be specified from these data.

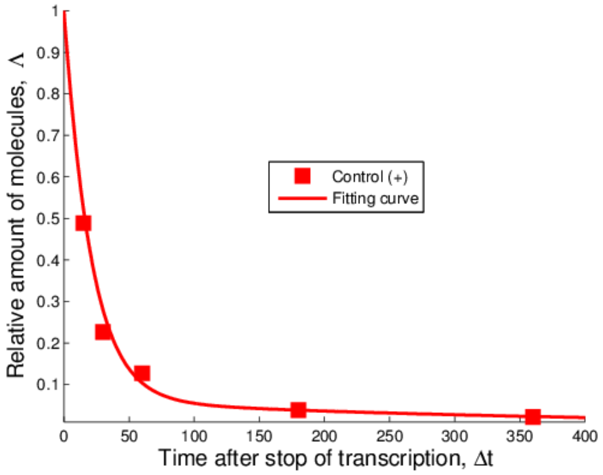


FIGURE 1.11: After fixing the rates μ, λ, ν from the fitting of the negative decay pattern in Figure 1.4, we use the network in Figure 1.10 in order to fit the decay pattern when all three factors, miRNA, NOT1 and PAN3 are expressed. The values of the rates are $\lambda_{RNP} = 0.1501\,\mathrm{min}^{-1}$ and $\mu_{RNP} = 0.0493\,\mathrm{min}^{-1}$ with 95% confidence intervals [0.0908, 0.2094] and [0.0373, 0.0612], respectively. These rates indicate that when both NOT1 and PAN3 are expressed together with the miRNA, about 84 % of the mRNA in the experiment are degraded by the action of miRISC.

1.3.3 The cooperative role of PAN3

We have seen that the expression of PAN3 in a system with miRNA and NOT1 strongly destabilizes the target mRNA and shortens its lifetime. Nevertheless, we can better understand the role played by PAN3 in cooperation with NOT1, when we compare the fraction of target mRNA that are expected to be found in the ”miRISC+NOT1” state in Figure 1.9 (blue circle) with the fraction of target mRNA to be found in the ”miRISC+NOT1+PAN3” state in Figure 1.11 (red circle). This comparison is made in Figure 1.12. There, we find the fraction of mRNAs in each of the three states after denoting state 0 the state in the middle of the network, state 1 the state on its left side and state 2 the state on the right side (blue circle in Figure 1.9 and red circle in Figure 1.11).

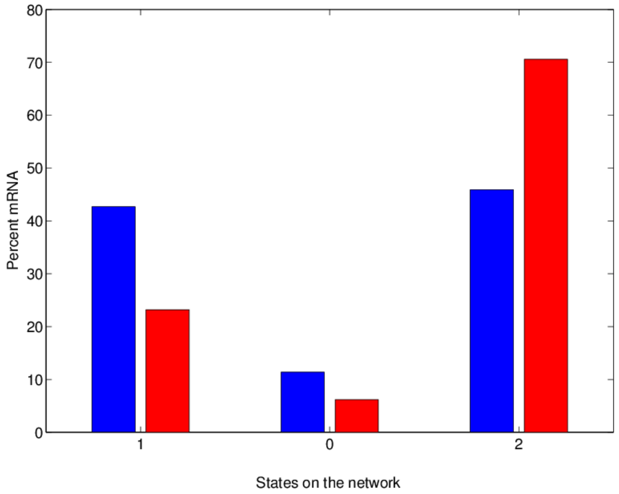


Figure 1.12: Percent of mRNA at steady state expression level in each of the three main states of the networks in Figure 1.9 and 1.11. State 0 represents the mRNAs that are not bound to miRISC and not bound to the competing alternative complexes. The alternative degradation pathways lead to state 1 whereas the miRISC-mediated pathway leads to state 2. The blue histogram refers to the case when only the miRNA and NOT1 are expressed whereas the histogram in red refers to the experiments when all three factors miRNA, NOT1 and PAN3 are expressed. The percent of mRNA in state 2 is strongly increased by the expression of PAN3 at the expenses of the amount of mRNA in state 1 This indicates that the most important role of PAN3 is to shift the balance towards miRNA-mediated decay.

We can see from the bar plot that the major contribution of PAN3 is to shift the balance of forces in favor of the miRNA by subtracting target mRNAs to the alternative pathway. By expressing PAN3, indeed, the amount of mRNA that are found in state 1, corresponding to the mRNA bound to protein complexes competing with the miRISC, decreases by almost 20% of the total mRNA, whereas the amount found in state 0 decreases by only a 5% of the total. This might indicate that the major role played by PAN3 is not to enhance deadenylation but rather to enhance the recruitment of the target mRNA at the expenses of alternative degradation pathways that do not involve miRNA. From the available data it is not possible to establish if the mRNA in these three states are also translational competent or are silenced. From the biochemical point of view, moreover, each of these three states might be a complex of different states sharing the same kinetic characteristics. Nevertheless, experiments designed to estimate the amount of mRNA bound or not bound to miRISC and NOT1 can provide important information to validate this model.

1.4 Summary and Discussion

In this paper, we show that the current hypothesis about the sequence of interactions between miRISC, its target mRNA and the factor NOT1 is not supported by the data. We have shown that the mRNA is also degraded when the miRNA is not expressed, indicating the existance of an alternative pathway, possibly competing with the miRNA pathway.

We also show that when only miRNA is expressed (NOT1 and PAN3 are knocked down), the target mRNA is stabilized, probably because it is protected from the action of an alternative miRNA-independent pathway. We postulate that the binding between miRISC and mRNA is irreversible and leads to the deadenylation independent decay of the target message in agreement with recent experimental studies. However, this assumption is not obligatory. Indeed, one could have hypothesized that binding to miRISC is reversible, and that the presence of miRNA alone just slows down the action of the alternative pathway. With the present data it is not possible to distinguish between these two alternatives.

Finally, our analysis indicates that the miRISC complex and NOT1 interact with each other *before* interacting with the mRNA. We assume that this discovery is not limited to the special miRNA-mRNA pair studied in [10] and is therefore a new general mechanism of mRNA control. Our analyses confirm the conclusions in [10] that PAN3 without NOT1 does not lead to an identifiable destabilization of the mRNA. Nevertheless, we see a strong cooperative effect between PAN3 and NOT1, where PAN3 is able to strongly enhance the binding of the miRISC+NOT1+PAN3 complex to the target mRNA compared to the miRISC+NOT1 complex alone.

Experimentally, one should be able to detect the presence of miRISC+NOT1 complexes in the absence of target mRNA, in order to verify our findings. Moreover, steady state relative amounts of mRNA in the different biochemical states can provide further validation data for our networks and additional information to unveil further details of the miRNA-mediated mRNA degradation.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Conceived the project: AV, CS, DC; Data collection: CS; Model development: AV, CS, DC; Interpretation: CS, AV, DC; Wrote the manuscript: AV, CS, DC.

Acknowledgements

The authors acknowledge support from the ITN Marie-Curie "NICHE".

2

Bacteria differently regulate mRNA abundance to specifically respond to various stresses

A version of this chapter has been published as:
Bartholomaus A, Fedyunin I, Feist P, Sin C, Zhang G, Valleriani A, Ignatova Z (2015) Bacteria differently regulate mRNA abundance to specifically respond to various stresses. Phil. Trans. R. Soc. A 374: 20150069.

Online version:
http://dx.doi.org/10.1098/rsta.2015.0069

CONTRIBUTIONS: Large collaboration with many members. I designed many of the metrics. I made the initial calculations on copy number, ribosome density and translational burden. Our main contribution to the manuscript was the fold change analysis. I also made the calculations to test for reproducibility and statistical significance. Additionally (but not included in the manuscript), I preformed additional analyses on codon usage, ribosome speed and trends by decile.

Abstract

Environmental stress is detrimental to cell viability and requires an adequate reprogramming of cellular activities to maximize cell survival. We present a global analysis of the response of *E. coli* to acute heat and osmotic stress. We combine deep sequencing of total mRNA and ribosome-protected fragments to provide a genome-wide map of the stress response at transcriptional and translational level. For each type of stress, we observe a unique subset of genes that shape the stress-specific response. Upon temperature upshift, mRNAs with reduced folding stability up- and downstream of the start codon, and thus with more accessible initiation regions, are translationally favoured. Conversely, osmotic upshift causes a global reduction of highly-translated transcripts with high copy numbers, allowing reallocation of translation resources to not degraded and newly synthesised mRNAs.

2.1 Introduction

Environmental stress and suboptimal growth conditions reduce cell viability and put cells at risk. Acute stress requires an immediate but specific response in order to maximize cell survival. Multicellular organisms can efficiently buffer external changes using the capacity of internal homeostasis [93]. In contrast, unicellular organisms are directly exposed to sudden alterations of their environment and need to immediately reprogram their cellular activities. By examining the gene expression of cells exposed to different types of stress numerous previous studies suggest that bacteria respond to stress over different time scales [94–97]. Gene expression is subject to extensive regulation at different levels, including transcription, mRNA degradation, translation and protein degradation. It is now clear that in each of these processes a series of regulatory steps is involved [84, 98–100], but little is known about how their combined effect shapes the cellular response to stress in bacteria. Advances in massively parallel nucleotide sequencing platforms and approaches to capture ribosomal position with nucleotide resolution [101] allow precise deconvolution of the control mechanisms of gene expression at translational level. This approach has been pioneered for yeast [101] and applied to bacteria to determine mRNA sequence features leading to non-uniform distribution of the ribosomal reads [69] and to monitor the co-translational binding of chaperones [70]. Here we modified this methodology and combined it with RNA-sequencing (RNA-Seq) to quantitatively assess transcriptional and translational response of *E. coli* to different nutrient conditions and to acute heat and osmotic stress. Our analysis unravelled common and stress-specific response programs which *E. coli* uses to counteract acute stress.

In both type of stress we observed that for a large fraction of genes changes in mRNA levels are co-directional with changes in translation. However, for each stress type, a significant group of genes opposed this trend and showed mainly translational regulation with no changes in the mRNA levels. Analysis of these gene subsets revealed disparate mode of translational regulation at each type of stress.

2.2 Materials and Methods

2.2.1 Media, cultivation conditions and cell quantification

E. coli MC4100 strain was cultured at 37°C to mid-log phase (OD600 $\sim$ 0.3 or 0.4) in LB media or MM medium (12.8 g/l $Na_2HPO_4 \cdot 7H_2O$, 3 g/l KH_2PO_4, 1 g/l NH_4Cl, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, 0.4% glucose). For heat stress, an aliquot of cells grown in LB medium was centrifuged for 4 min at $4000 \cdot g$ at room temperature, resuspended in LB medium preheated to 47°C and further incubated for 7 min at 47°C. Osmotic stress was exerted to an aliquot of cells grown in MM medium by adding NaCl to 0.3M (0.6 Osm) and further incubated for 20 min at 37°C. The number of viable cells was determined after staining the cells with BDTMCell Viability Kit with liquid counting beads (BD Bioscience) and analysed by flow cytometry on a FACSCalibur (BD Biosciences) equipped with an argon 488-nm and diode 632-nm lasers. The data was analysed with Cyfoogic 1.2.1 (CyFlo Ltd.).

2.2.2 Cell lysis, polysome isolation and enrichment of RPFs

To isolate mRNA-bound ribosome complexes we used a previously described approach [102] with some modifications. Exponentially growing cells were poured over crushed ice containing 100 μg/ml chloramphenicol and harvested by centrifugation at $5000 \cdot g$ for 5 min at 4°C. All subsequent operations were carried out on ice with pre-chilled buffers. To obtain polysomal profiles, the cell pellet from 100 ml culture was resuspended in 12 ml of ice-cold sucrose-buffer solution (16 mM Tris pH 8.1 supplemented with 0.5 M RNase-free sucrose, 50 mM KCl, 8.75 mM EDTA, 100 μg/ml chloramphenicol, 1.25 mg/ml lysozyme) and gently stirred for 5 min on ice. 0.3 ml of 1M $MgCl_2$ was added to inhibit lysozyme and the suspension was spun at $6000 \cdot g$ for 10 min at 4°C. The pellet was resuspended in 0.7 ml freshly prepared lysis buffer (10 mM Tris pH 7.8 containing 50 mM NH_4Cl, 0.01 M $MgCl_2$, 0.2% triton X-100, 100 μg/ml chloramphenicol and 10 U DNase I (RNase-free, Fermentas)). The lysate was clarified by centrifugation at $10000 \cdot g$ for 10min at 4°C and frozen at -80°C if not analysed immediately. Up to 0.4 ml of clarified lysate was layered onto 15 to 50% (w/v) sucrose gradient (buffered in 10 mM Tris pH 7.8 supplemented with 50 mM NH_4Cl, 10 mM $MgCl_2$, 0.2% triton X-100, 100 μg/ml chloramphenicol) and centrifuged for 1.5 h at 35,000 rpm in SW 55Ti rotor (Beckman) at 4°C. The gradient was slowly pumped out from the bottom of the tubes with a flow rate set to 0.38 ml/min and the absorbance at 254 nm (A254) was recorded via a flow-through UV spectrophotometer cell (Pharmacia LKB-UV-M II). For the isolation of RPFs, an aliquot of 100 A260 units of ribosome-bound mRNA fraction (prior to ultracentrifugation in the sucrose gradients, see above) was subjected to nucleolytic digestion with 10 units/μl micrococcal nuclease (Fermentas) for 10 min at room temperature in buffer with pH 9.2 (10 mM Tris pH 11 containing 50 mM NH_4Cl, 10 mM MgCl2, 0.2% triton X-100, 100 μg/ml chloramphenicol and 20 mM $CaCl_2$). The monosomal fraction was separated by sucrose density gradient (15-50% w/v). The total RNA was isolated from the monosomes using the hot SDS/phenol method. The micrococcal nuclease also cleaved the rRNAs into fragments with a size similar to the RPFs. The sample was enriched predominantly in one rRNA fragment which was removed by subtractive hybridization at 70 0C using a 5'-biotin-5'-GCCTCGTCATCACGCCTCAGCC-3' DNA oligonucleotide along with μMACS Streptavidin Kit (Myltenyi Biotec) to remove the biotin-labeled DNA/rRNA hybrids. 3' adaper was first ligated to the dephosporylated fragments with T4 RNA ligase 2 truncated (New England Biolabs). Following 3' phosphorylation of the 5' adapter with T4 PNK (New England Biolabs), the 5' adapter was ligated to the fragment with T4 RNA ligase 2 truncated (New England Biolabs). The generation of the libraries was performed as described [103].

2.2.3 Random mRNA fragmentation and spike-in with a control RNA

Total RNA was isolated from ∼6 ml exponentially growing bacteria with GeneJET™ RNAPurification Kit (Fermentas). Prokaryotic mRNA is a minor fraction of the total cellular RNA, therefore mRNAs were enriched by removal of 16S and 23S rRNAs [104] with a MICROBExpress™ Bacterial mRNA Enrichment Kit (Ambion). 20 μl of the enriched mRNA was mixed with equal volume of 2x alkaline fragmentation solution (2 mM EDTA and 100 mM Na_2CO_3Na2CO3 pH 9.2) and incubated for 40 min at 95 °C. The reaction was stopped by adding 560 μl 300 mM NaOAc pH 5.5, followed by isopropanol precipitation [105]. For quantification [106], after the random fragmentation the RNA sample was spiked with an external 25-nt RNA standard (5'-AAUGAUAAUUCAAGAAUCAUAACGG-3') which does not align anywhere in the *E. coli* genome. Typically, 10 ng of spike-in RNA were added to 25.5 μg total extracted RNA. The optimal time for fragmentation of the mRNA was determined using GAPDH mRNA (0.25 μg; Fermentas) and the spectra were recorded with BioAnalyzer (Agilent

RNA 6000 Kit). RNA-size selection and generation of the cDNA libraries was performed as described [103].

2.2.4 Mapping of the sequencing reads

Sequencing reads were aligned to *E. coli* K-12 substrain MG1655 genome sequence (GenBank: U00096.2) with FANSe mapping algorithm allowing up to 3 mismatches and enabled indel detection [107]. The reads aligning to rRNA and tRNA genes were subtracted from the data sets. Sequencing reads spanning two (overlapping) ORFs were assigned to one of the two annotated ORFs (NCBI) based on the position of the middle nt[1].

2.2.5 Quantitative RT-PCR

Total mRNA was extracted using the GeneJETTM RNA Purification Kit (Fermentas) and treated with DNase I (Fermentas). The cDNA was synthesised with RevertAid™ H Minus Reverse Transcriptase (Fermentas) and quantitative RT-PCR was performed on a StepOnePlus™ Real-Time PCR system (Applied Biosystems) using template-specific primers. The values were normalised to the amount of the total RNA.

2.2.6 Quantification of the absolute transcript numbers

The copy number of the transcripts (X_i) was quantified using the following method [108].

$$X_i = \frac{C_i}{CL_i}T \tag{2.1}$$

where C_i represents the mapped reads on a transcript i with a length L_i (in nt), $C = \sum_i C_i$ is the sum of all mapped reads within a RNA-Seq data set, whereas T is the length of the transcriptome in base pairs [108]. From the reads and the amount of the added spike-in standard, we determined the number of molecules of spike-in standard in nt. T was calculated by converting the total mRNA mapped reads in each RNA-Seq experiment into molecules of mRNA (in nt) using the number of molecules of spike-in standard (nt) and normalised per cell number quantified by FACS. The length of the transcriptome, T, in base pairs we calculated as follows:

[1] For reads with an even number of nt-s, the nt closer to the 5'-terminus was arbitrarily taken as a middle nt

1. The amount (ng) of added spike-in standard was converted into molecules spike-in in nt using the molecular weight and nucleotide length of the spike-in and the Avogadro number.

2. The total amount of mRNA to which the spike-in was added was converted into nt using the molecules of spike-in (nt) and the sequencing reads mapped to the transcriptome and spike-in standard.

3. The total mRNA length (nt) was divided by the cell number quantified by FACS to calculate the total length (T) of the trasncriptome per cell. $1.81 \cdot 10^8 \pm 2.4 \cdot 10^7$, $7.84 \cdot 10^8 \pm 1.21 \cdot 10^7$, $1.52 \cdot 10^8 \pm 2 \cdot 10^7$ and $4.72 \cdot 10^8 \pm 7.3 \cdot 10^7$ was the total number of cells in LB, MM, under heat or osmotic stress, respectively.

4. Finally, to calculate the copy number of each gene, we multiply the rpkMd values of each gene by T (Equation (2.1)).

2.2.7 Differential gene expression and GO analyses

The number of raw reads unambiguously aligned to ORFs in both RNA-Seq and RPF data sets, for each condition, were normalised by the median of the total number of reads aligned to all ORFs, which gives more stable normalization [109] and is more suitable for differential expression analysis than other normalization approaches using total number of reads aligned to all ORFs (reads per total mapped reads, rpM, or fragment per total mapped reads, FPM) [110]. For comparison between different genes, the number of the aligned reads was normalised by the length of the ORF and the median of the total number of reads aligned to all ORFs (reads per kilobase of ORF per median of the total mapped reads, rpkMd). Furthermore, to avoid false-positives originating from genes with very low read counts, genes with mRNA or RPF read counts <60 were 'upgraded' to 60 in the differential gene expression analysis. The value of minimal 60 reads in both RNA-Seq and RPF-Seq analysis was determined from the analysis of the sampling error between two biological replicates as described [101]. Thereafter, the ribosome density (RD) for each gene was calculated as follows[2]:

$$\mathrm{RD}_i = \frac{\mathrm{RPF}_i[\mathrm{rpMd}]}{\mathrm{mRNA}_i[\mathrm{rpMd}]} \tag{2.2}$$

To assess the biological functions affected by different stress conditions, we identified the GO terms with significant enrichment for each differentially expressed gene set using R (version 2.15.2) and Bioconductor packages ecoli2.db and GO.db. p-values were adjusted

[2] Note, it is similar to the definition of the TE value as described already [101]

for multiple testing using false discovery rate, FDR according to Benjamini-Hochberg [111].

2.2.8 Secondary structure analysis

Structure profiles were computed with RNAfold program with default parameters from the Vienna RNA Package 2.0 [112] using a sliding window with different widths reported in the literature for such analysis. For each nucleotide of a transcript the minimum free energy was calculated moving the window with a single nucleotide at a time. Average profiles for groups of genes were generated by taking the mean of their per-base folding energy contributions. The 23-nt window, with a width equal to the average of our read lengths, revealed very noisy profiles due to the small window size. In contrast, the 101-nt window, albeit informative on long-range interactions, smoothed too many features in the profiles. For our analysis we choose the 39-nt window [113] which gave very similar result to the 51-bp window [114]. The significance of the structural differences between two gene groups was evaluated nucleotide-wise using a two-sample Kolmogorov-Smirnov test. A two-sided test was performed for each base and the corresponding p-values were collected and adjusted for multiple testing using FDR according to Benjamini-Hochberg [111]. To judge the significance of the whole curve, the single p-values were median averaged. We further assessed the uniqueness of the identified groups of genes by comparing the folding energy distributions to those of equally-sized random sets of genes. We analysed 10,000 subsamples with sample size equal to the gene group of interest and calculated a p-value as the number of points crossing the corresponding average folding profile divided by the total number of sampled points. The dispersion of the subsamples between 2.5th and 97.5th percentiles of the distribution of values per nt is shadowed grey (Figure 2.4b and d).

2.3 Results

2.3.1 Difference in gene expression between growth on mineral and complex medium

We first compared the transcriptional and translational activities of *E. coli* (strain MC4100) at rapid growth in complex (LB) medium or at moderate growth rate in medium containing mineral compounds and glucose as the only carbon and energy source (minimal medium, MM). For both conditions we performed deep sequencing of the total mRNA (RNA-Seq) [108] and ribosome-protected fragments (RPF or RPF-Seq

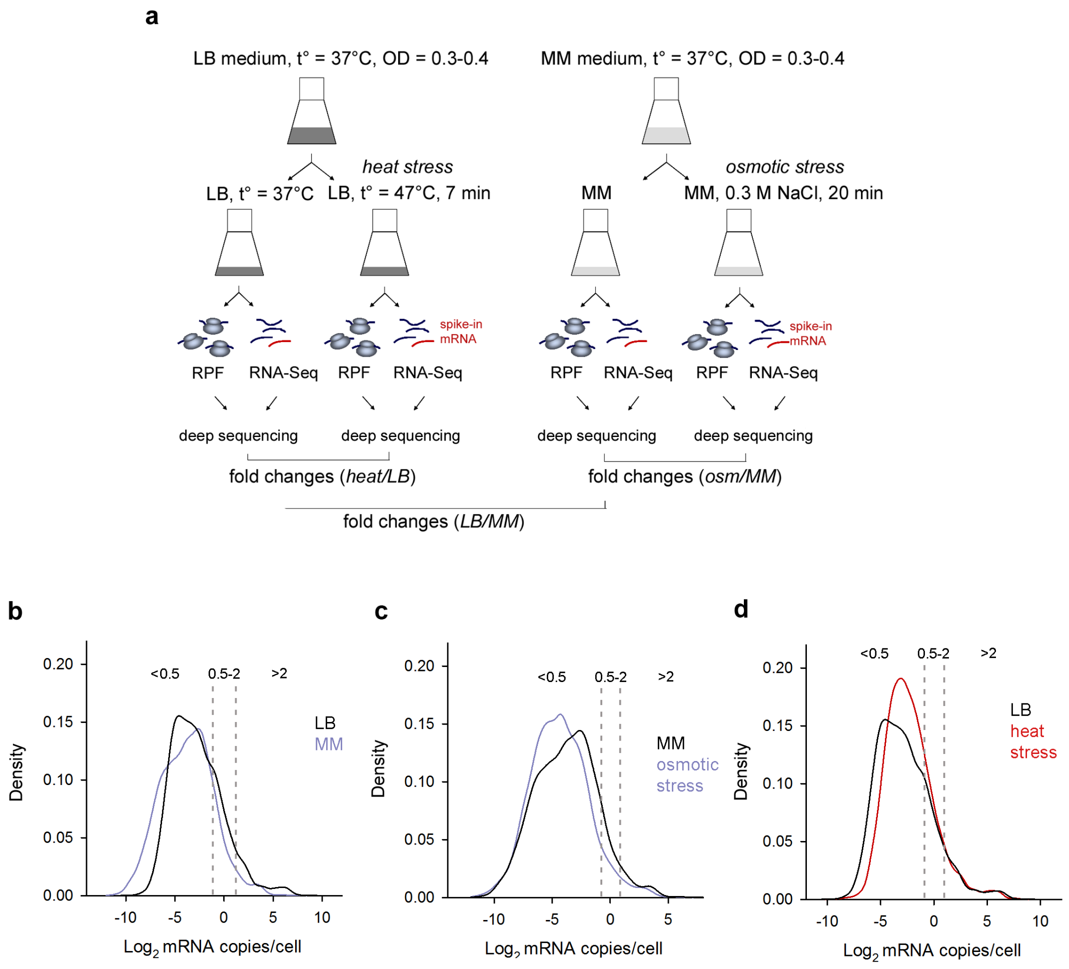


FIGURE 2.1: Changes of the copy numbers upon stress exposure. (a) Schematic of the experiment to analyse differences in expression under different stress conditions. Distribution of the gene copy numbers quantified from the RNA-Seq data sets from cells grown in MM and LB (b), exposed to osmotic (c) or heat stress (d) Dashed vertical lines delimit the three expression groups. The copy numbers between LB and MM are significantly different ($p < 0.01$ according to Kolmogorov-Smirnov test). Note the bimodal distribution evident in the LB density curve corresponding to mRNAs with large copy numbers (> 25).

[103, 105]) (Figure 2.1a). The total mRNA was isolated from each sample and randomly fragmented under alkaline conditions into fragments with similar length to the RPFs (Supplementary Figure 2.1a). We directly ligated 5'- and 3'-adapters to the RPFs and mRNA fragments which enabled us to capture transcripts of high and moderate abundance as well as low-abundance transcripts (Figure 2.1b) with high reproducibility between biological replicates (Supplementary Figure 2.1b, c). For each sequencing set, the unambiguously mapped mRNA and RPF reads were normalised by the median of the total number of reads mapped to all open-reading frames (ORF) and presented as reads per median of the total mapped reads (rpMd) or reads per kilobase of ORF per median of the total mapped reads (rpkMd). Normalization by the median of the total number of reads aligned to all ORFs gives more stable normalization than normalization to reads per total mapped reads, rpM, or fragment per total mapped reads, FPM [109]. We spiked each RNA-Seq experiment (Figure 2.1a) with an external RNA-standard whose sequence does not align anywhere in the *E. coli* genome. Reads mapped to the spike in sequence showed high reproducibility between biological replicates. With this twist, the RNA-Seq data sets, normalised to the living cell counts (determined by FACS) were used to calculate the absolute copy number of each mRNA per cell (Equation (2.1)). From 4325 total mRNA-coding genes of *E. coli* we were able to map reads to 4254 ORFs in LB and 4269 ORFs in MM[3]. Cells grown in LB medium have about three times more mRNA transcripts than cells grown in MM (∼7800 mRNA copies/cell vs. ∼2400 mRNA copies/cell) (Figure 2.1b); correlating to the 4x larger volume of cells cultured in LB. The mean mRNA copy number was 1.79 copies/cell in LB and 0.56 copies/cell in MM. Similar low overall mRNA copy numbers per cell have been theoretically predicted [115] and experimentally determined in ensemble and single-cell measurements for another *E. coli* strain [45]. Relatively low mRNA copies/cell have been reported also for budding [116] and fission [117] yeast; yet the unicellular eukaryotes have on average an order of magnitude higher number of total mRNA molecules/cell which mirrors the larger cell volume and larger size of their genomes compared to *E. coli* . We separated the genes in three groups, similar to the expression zones arbitrarily-defined in unicellular eukaryotic organism, fission yeast [117]: group 1 contains genes with low-abundance mRNA (<0.5 copies/cell); group 2 consists of genes expressed at ∼1 mRNA copy/cell (0.5-2 copies/-cell) and group 3 comprises genes with robust expression at >2 mRNA copies per cell (Figure 2.1b). In general, despite the difference in the total mRNA copy number between LB and MM, similar functional categories (GO terms) were enriched for genes expressed at >2 mRNA copies/cell, whereas there was no significant GO term overlap for genes expressed at ∼1 mRNA copy/cell (Supplementary Figure 2.1d). The high copy number group (>2 copies/cell) in both LB and MM is dominated by mRNAs encoding proteins involved in translation, main metabolism pathways and ATP synthesis (Supplementary Figure 2.1d). Moreover, in LB medium a subset of mRNAs is expressed at comparatively very high copy numbers give a rise to a bimodal distribution of the transcript copy numbers (Supplementary Figure 2.1d); this subset of genes is comprised mostly of such encoding proteins participating in ribosome biogenesis and translation. The mRNAs of the ribosomal proteins were also among the transcripts with the highest mRNA copies in MM, yet their mRNA copy numbers were 5-8 times lower than in LB, correlating with the much lower number of ribosomes per cell for bacteria grown in MM [118]. The second group of genes expressed at ∼1 mRNA copy/cell (0.5-2 mRNA copies/cell) in MM comprises mostly genes involved in the production or conversion of products that are not

[3] In *E. coli*, the total number of genes including rRNA, tRNA and non-coding RNAs is 4496

supplied by the MM growth medium, including amino acids and nucleotides. The group of the genes expressed at a level below one copy per cell on average is the largest group in both media (Figure 2.1b) which could result from a small transcription rate [119] or short lifetime of a transcript [74, 120]. Our results are consistent with recent sequencing experiments from fission yeast [117] or metazoan [119] which also detected a large fraction of mRNAs expressed below 1 mRNA copy/cell. RPF-Seq revealed the position of the translating ribosomes (Figure 2.1a) and enabled quantification of actively translated transcripts. As a quantitative measure of translation of each gene we determined the ribosome density (RD) per gene which is defined as the ratio of the normalised RPF reads to the normalised mRNA reads for each gene (Equation (2.2)). For the majority of translationally-active genes (i.e., for those which RPFs were detected), the mRNA counts are positively correlated with the corresponding RPF counts (Spearman correlation coefficients 0.89 for LB and 0.91 for MM). Next we compared the changes in the gene expression at the level of transcription and translation between LB and MM by assessing the log2-fold changes of the mRNA counts and RD values. This analysis revealed that 605 genes were significantly (>95% confidence) transcriptionally regulated (Supplementary Figure 2.1b and Supplementary Table 1, online[4]). Genes participating in translation, ribosome biogenesis and aerobic respiration were enriched among the transcriptionally controlled genes (Benjamini-Hochberg corrected p-value < 0.001, Supplementary Table 1, online). 239 genes showed significant changes in RD values (95% confidence, Supplementary Figure 2.1c) without any significant changes in the mRNA counts (Supplementary Table 1, online). Interestingly, comparing the cells grown in LB to those cells grown in MM, the mRNA coverage for the majority of factors participating in translation was within the 5th percentile and RD values were not significantly changed. In other words, the RPFs increased proportionally to the mRNA reads keeping the RD values were unchanged (Supplementary Table 1, online), implying similar biosynthetic activities per unit cell volume in both LB and MM.

2.3.2 Similarities and dissimilarities of the response programs to counterbalance various types of acute stress

Next, we evaluated stress-induced reprogramming of gene expression at the level of transcription and translation upon exposure of the cells to two types of acute stress: *E. coli*, grown in LB or MM until the early exponential phase, were subjected to temperature shift or exposed to osmotic upshift by adding NaCl to the medium, respectively (Figure 2.1a). The experimental conditions were chosen to allow for monitoring of rapid

[4] http://rsta.royalsocietypublishing.org/content/roypta/suppl/2016/02/05/rsta.2015.0069.DC1/rsta20150069supp1.pdf

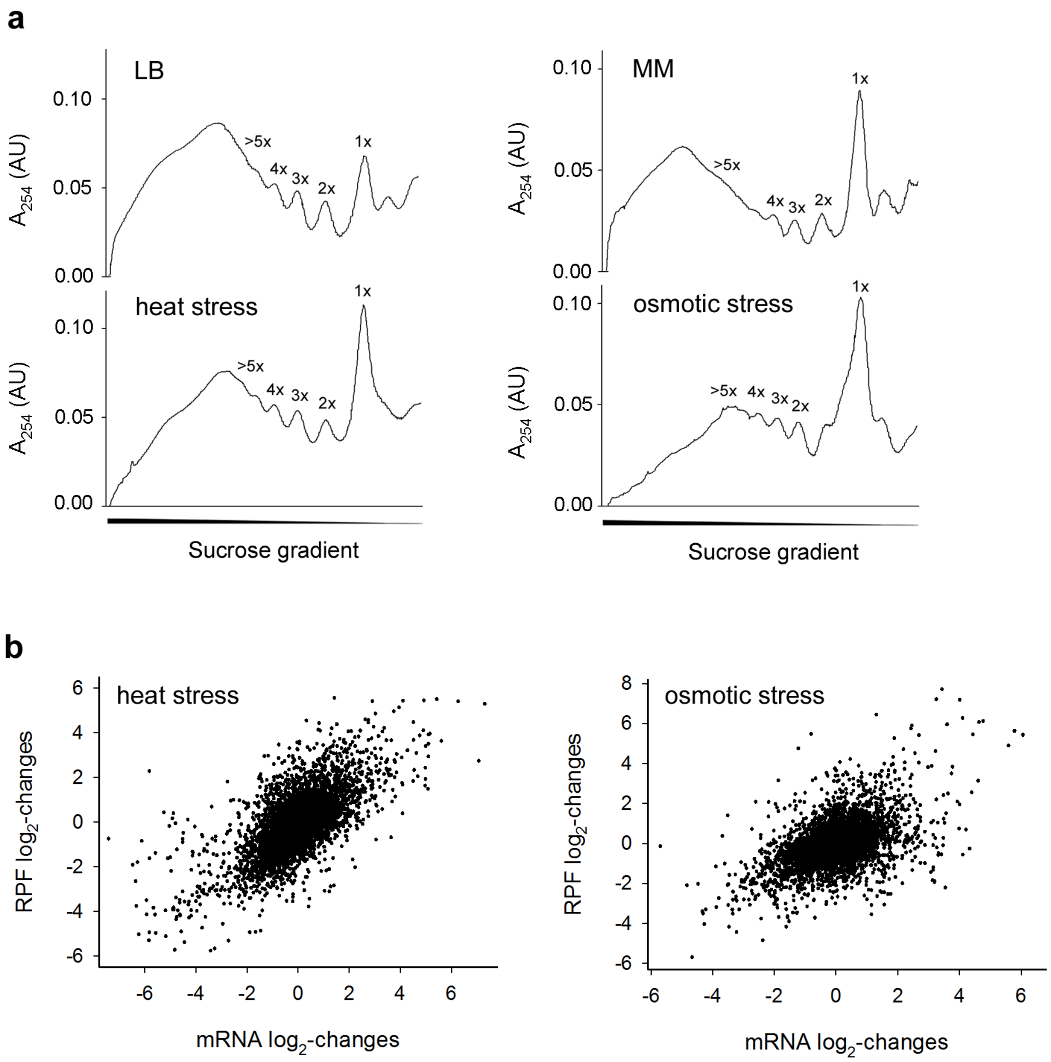


FIGURE 2.2: Transcriptional and translational response to different types of stress. (a) Polysomal profiles of cells grown in different conditions. The profiles changed for cells exposed to stress (compare LB to heat stress and MM to osmotic stress): the peaks corresponding to large polysomes decreases under stress, while the monosomal fraction (1x) increases. The fractions corresponding to mRNA with di-, tri-, tetra- and polysomes are designated as 2x, 3x, 4x and >5x, respectively. (b) Correlation between the log2-fold changes of the normalised RPF and mRNA read counts from cells exposed to heat or osmotic stress. Pearson correlation coefficients are: 0.667 (heat stress) and 0.492 (osmotic stress).

changes in the transcriptional and translational programs during the acute stress response, but not to capture long-term adaptive stress reactions. The different types of stress were applied on different time scales and in different media. Thermal stress was applied for 7 min, the time at which transcription of the heat-shock induced genes is maximal [121]. Osmotic stress was applied in a time-window which is prior to the onset of intrinsic osmolyte synthesis [122] and in MM which was free of any osmoprotective substances that can be taken up to counteract osmotic stress [123]. The mRNA expression of marker genes, known to be specifically upregulated in heat [97] or osmotic stress [124], was verified with quantitative RT-PCR (Supplementary Figure 2.1a,b). Intriguingly, osmotic stress caused a global reduction of transcripts (Figure 2.1c and Supplementary Figure 2.2c), from ~2400 mRNA copies/cell in MM to ~1600 mRNA copies/cell upon osmotic upshift. In particular, mRNAs with higher copies (from the groups of mRNAs with >2 copies/cell) were reduced (Figure 2.1c, Supplementary Figure 2.2e). Conversely, heat stress caused a little global reduction of the transcripts – from ~7800 mRNA copies/cell in LB to ~7200 mRNA copies/cell (Figure 2.1d and Supplementary Figure 2.2d,e).

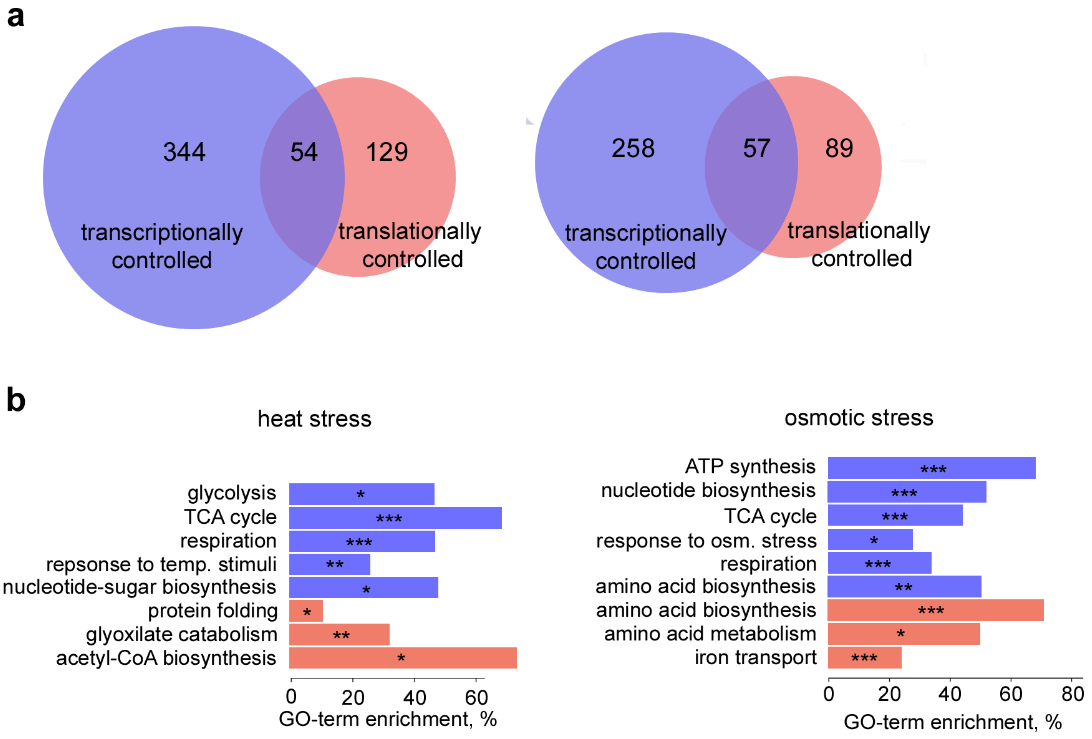


FIGURE 2.3: Fold-change analysis of the cellular response to stress. Venn diagram showing the genes differentially expressed between heat stress vs. LB and osmotic stress vs. MM. Blue, genes with mRNA (read counts) log2-fold changes over the 95th percentile threshold of biological replicates (Supplementary Table 1, online[5]); red, genes with RD log2-fold changes over the 95th percentile threshold of biological replicates (Supplementary Figure 2.1c); purple, genes with significant changes in both, mRNA and RD values. The genes in each diagram correspond to the genes listed in Supplementary Table 1, online. (b) GO terms with significant enrichment under heat stress vs. LB and under osmotic stress vs. MM. The colour code is the same as in Figure 2.3a. ***: $p<0.001$; **: $p<0.01$; *: $p<0.05$

To assess the global effect of stress on translation, we compared the polysome profiles under stress with the corresponding control condition (Figure 2.2a). In both types of stress the amount of the large polysome fraction (>5 ribosomes) decreased in parallel to the increase of the monosome peak (Figure 2.2a) which is likely a result of ribosomal drop-off [122] and implies a general repression of translation. Yet, some translation activity is retained as evidenced by the presence of 2-5-polysomes (Figure 2.2a). We then compared the mRNA and RPF counts for each gene under both stress conditions (Figure 2.2b). Overall, for the majority of translationally-active genes under stress (i.e., for which RPFs were detected), we found that the log-log linear correlation between the mRNA change and RPF change is relatively high and positive for both stress conditions (Figure 2.2b). On a global scale, we observed that changes in mRNA counts are codirectional with changes in the RPF counts, for both induced and repressed genes, in response to heat or osmotic stress, with correlation coefficients of 0.667 or 0.492, respectively (Figure 2.2b). However, for both stress conditions, a sizeable fraction of genes oppose this trend: for these genes, changes in mRNA counts do not correlate with changes of the RPF counts. This raised the question as to whether these gene sets shape the cellular response against each type of stress. To compare expression between different conditions we performed pairwise comparison of RD and mRNA read counts and ranked them according to the log2-fold changes (Figure 2.3a and Supplementary Table 1, online). Comparison of the gene groups with significant changes upon osmotic upshift and heat exposure ($>$ 95% confidence from the biological replicates, Supplementary Figure 2.1b, c) revealed a common pattern in the response to both types of stress: 43.3% of the genes showed similar log2-fold changes at transcriptional level, while only 15.7% displayed similarities of their fold changes at the level of translation, as captured by the RD values Supplementary Table 1, online). When mammalian cells are exposed to thermal stress, RPFs accumulate within the first 30-40 nucleotides at the start of the genes [125, 126]. In bacteria, subunits alternative to the housekeeping σ^{70}(rpoD)-factor of the RNA polymerase orchestrate the expression of stress-related genes to counteract external stress [127]. The mRNA levels of σ^{32} (rpoH, related to heat stress) were upregulated under both osmotic and thermal stress, while σ^{S} (rpoS, related to starvation stress) and σ^{E} (rpoE, related to periplasmic stress) were upregulated only under osmotic upshift (Supplementary Table 1, online). Despite the pervasive upregulation of transcription of rpoH under both osmotic and thermal stress condition, the down-stream genes of the heat-shock regulon (chaperones, proteases of the Clp-family [128]), whose expression is regulated by σ^{32} [128], were upregulated only at elevated temperature with log2-fold changes in the mRNA coverage, ranging from 1.6 to 6.3 (examples: dnaK – 5.6, dnaJ – 5.1, clpB – 6.3, clpX – 1.9, clpP – 1.8) (Fig. 3 and Supplementary Table 1, online). The specific expression of some marker genes, specifically upregulated in heat stress [97], was verified with quantitative RT-PCR (Supplementary Figure 2.2a). Notably, under the heat stress, the translation of the genes from the heat-shock regulon was also enhanced, i.e. the absolute number of RPFs was higher under heat stress, though the overall RD ratio remained similar to LB. Upon exposure to osmotic stress, we did not detect any upregulation of genes from the heat regulon other than clpB (fold-change 1.89), implying the high specificity of σ^{32}-dependent regulation under heat stress.

2.3.3 Genes with lower secondary structure propensity are translated under thermal stress

Bacteria use complex strategies to coordinate temperature-dependent expression, including temperature-sensing RNA sequences – RNA-thermometers – whose structure melts at elevated growth temperature and increases the efficiency of translation initiation [129]. Next, we compared the expression of two established examples of translational regulation in *E. coli*, rpoH and ibpA [129] using our genome-wide analysis under heat stress. For both genes, the absolute number of RPFs increased significantly under heat stress, although the RD for ibpA increased while for rpoH it remained similar to the unstressed condition because of corresponding changes of rpoH mRNAs Supplementary Table 1, online). The cellular concentration of σ^{32} at elevated temperature is comprehensively balanced at different levels, including transcriptional, translational and post-translational protein stabilization [129–131]. Importantly, the RPF reads for both genes did not change under osmotic stress.

In addition to the genes regulated by the σ^{32} factor under heat stress, 94 of the 129 genes exhibited a significant boost in translation (positive fold-change of RD) without any changes in the mRNA read counts (Figure 2.3a and Supplementary Table 1, online)). This group was mostly enriched with genes participating in protein folding ($p < 0.01$) and metabolic activities ($p < 0.01$ and $p < 0.05$; Figure 2.3b). The transcripts of these genes were highly translated, with high RPF coverage along the whole mRNA and no significant accumulation of reads in the 5'-region of the ORFs as in the all-genes group (Figure 2.4a). We hypothesised that this gene set might bear some structural features to facilitate translation and thus we calculated the folding energy in the sequences flanking the initiation start. Typically, the folding profiles of all genes (Figure 2.4b, black line) showed reduced folding stability and fewer paired nucleotides around the initiation start compared to the coding sequence (observed as a peak in the folding energy profile) [113]. The folding energy of the genes translationally upregulated under heat was significantly lower than that of the remaining genes in the genome not only around the initiation site but also over a much broader sequences flanking the initiation (Figure 2.4b). Their folding energy is significantly different than that of randomly sampled genes (gray shadowed area, $p = 0.00204$, Figure 2.4b). 5'-UTRs, which tend to be less structured than the protein coding sequence of the gene [113], were also less structured in the translationally upregulated genes under heat stress (Figure 2.4b). Further downstream of the start codon, along the coding mRNA sequence, the folding energy relaxes to the mean folding profile of all genes (Supplementary Figure 2.3a). Reduced mRNA secondary structure around the start codon facilitates expression [56, 113, 132] and analysis of the mRNA folding stability at the initiation start revealed a positive correlation with the changes in

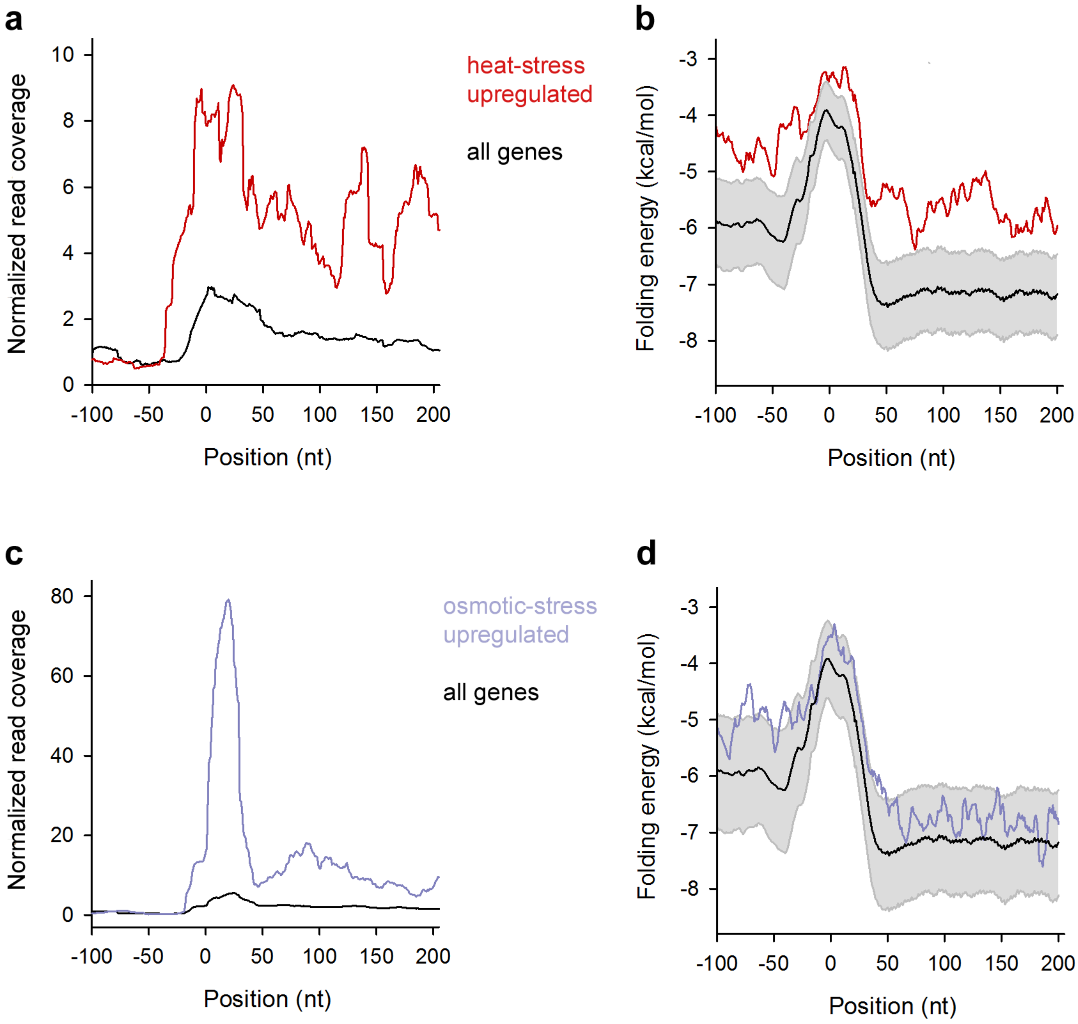


Figure 2.4: Genes translationally upregulated under heat stress have much lower propensity to form secondary structure. Normalised read coverage for upregulated genes under heat (red line) (a) and osmotic (blue line) (c) stress. The RPF reads for all genes in each group were normalised to the median of the corresponding mRNA reads and aligned at the start (zero is the first nucleotide of the start codon in each ORF) and compared to all genes (black lines in a and c). Average folding energy of upregulated genes under heat (red line) (b) and upon exposure to osmotic stress (blue line) (d) compared to all protein-coding genes in E. coli (black lines in b and d). p-values (median averaged from a two-sample Kolmogorov-Smirnov-Test) for the upregulated genes under heat (b) 0.00204 [lower quartile – 0.00027; upper quartile – 0.01434] and osmotic stress (d) 0.58343 [lower quartile – 0.42999; upper quartile – 0.57026]. The area of folding energy distribution sampled from random gene groups of the same size (as the upregulated gene group) is shadowed in gray (b and d). p-value from the sampling for genes under heat (b) and osmotic (d) stress are 0.0014 and 0.26072, respectively.

the RD values under heat stress; lower folding stability correlated with higher changes of the RD values (Supplementary Figure 2.3b). Interestingly, genes whose translation is favoured under heat stress do not originate solely from the subgroup with the least structured initiation area (1-500 gene group, Supplementary Figure 2.3b); rather these upregulated genes contain genes from subgroups with various folding energies, including the group with the most structured initiation region. In *E. coli*, a large fraction (53%) of protein-coding genes is organised in operons as polycistronic mRNAs and is suggested to coordinate the expression of functionally related proteins [133]. Among the genes translationally upregulated under heat stress those residing in operons were significantly less than the genome-wide operon configuration of *E. coli* (45% vs. 53%, $p = 9.36 \cdot 10^{-6}$). Furthermore, even for the few of them encoded within the same polycistronic mRNA we did not detect a coordinated expression. This might be not surprising in light of the fact that translation level of ORFs within the same polycystronic mRNA largely differ as they underlie independent initiation [68] and the strength of the Shine-Dalgarno (SD) sequence differs between the ORFs within one operon [134]. Together, this implies that the response to acute heat stress is at least in part shaped by selection of a subset of genes whose structure up- and downstream of the translation start most likely undergoes significant melting at higher temperatures. Thus, the initiation regions of these genes become more accessible upon heat stress, facilitating ribosome binding and in turn, translation.

2.3.4 Under osmotic upshift cells reallocate translation resources

Exposure of *E. coli* to high osmolarity causes rapid loss of water (plasmolysis), turgor pressure and shrinkage of the cell. A crucial reaction of the cell is the upregulation of osmotic stress-protective genes and genes encoding ATP-driven transporters of ions and small solutes [123, 135]. Their expression was mainly regulated at the level of transcription with significant increase of the mRNA copy numbers (Figure 2.3a and Supplementary Table 1, online)); the RPFs increased proportionally to the mRNA reads while the RD values remained unchanged. The major transporters maintaining the uptake of osmotic substances, proV, proP, proX, proW, otsA, under osmotic upshift were expressed at very high copy numbers (>2 mRNA copies/cell); in contrast their copy numbers were extremely low in MM, <0.5 mRNA mRNA copies/cell. The mRNA expression of those marker genes, known to be specifically upregulated under osmotic stress [124], was verified with quantitative RT-PCR (Supplementary Figure 2.2b). The cell response to osmotic stress at the translational level was clearly distinct from the response to acute heat stress: different sets of genes maintain the cell response at the level of translation, i.e., genes with significant log2-changes of the RD values but invariant mRNA read counts. The group of translationally upregulated genes under osmotic stress was enriched for transcripts of genes participating in amino acid metabolic processes ($p < 0.05$) and iron transport ($p < 0.001$; Figure 2.3b). Although those genes do share some functional similarities they were not enriched in genes encoded by the same polycistronic mRNA ($p = 0.26$ compared to the genome-wide fraction of genes organised in operons).

Next we analysed the cumulative profiles of this gene group (Figure 2.4c). Upregulated genes showed much higher accumulation of RPFs in the first 30 nucleotides and also approximately 10 nt upstream of the gene start (Figure 2.4c). This was clearly not driven by any secondary structural features of the mRNA specific to these genes; their propensity to be involved in secondary structure was not significantly distinguishable from the profiles of the other transcripts (Figure 2.4d). Importantly, translationally upregulated genes under osmotic stress also show higher accumulation of RPFs upstream of the start codon in the vicinity of the SD sequence which raised the question as to whether those genes are enriched in strong SD motifs which will favour their initiation. We performed a search for SD motifs based on the minimum hybridization energy as described in [50], however SD sequences were not significantly enriched in those genes ($p = 0.025$). Osmotic stress, but not thermal stress, resulted in a 35% reduction of the mRNA transcripts globally, whereby transcripts with high copy numbers in normal conditions (>2 mRNA copies/cell) were the most affected (Figure 2.1c and Supplementary Figure 2.2c). Over 25% of the genes in the high copy number group had significantly reduced transcripts (Supplementary Figure 2.2e). The subset of the genes with >2 mRNA copies/cell which underwent reduction in their copy numbers upon osmotic stress were also highly translated (Supplementary Figure 2.1f) implying that the reduction of these abundant, highly translated transcripts will release a large portion of ribosomes.

2.4 Discussion

Here, we quantify mRNA abundance and translational activities of *E. coli* using deep sequencing of ribosome-protected fragments. Even though *E. coli* is an extensively studied organism, the response to different types of acute stress has a complexity not captured by existing studies. A small subset of genes, unique for each type of stress, is regulated only at the level of translation, while the majority of translationally-active genes show concordant co-directional changes between mRNA and translation (RPF). A common feature among the genes translationally upregulated under heat stress is their higher accessibility in the initiation region at higher temperature which in turn facilitates their translation. In general, the genes from this subset are not necessarily genes with the lowest propensity to form secondary structure; some of the genes may undergo the most significant melting at elevated temperatures. In contrast, preferential expression of genes to counterbalance osmotic stress is driven by reduction of highly abundant and highly translated transcripts, thereby making translation resources available to the whole pool of mRNAs which contains mRNAs that remained intact under stress or newly synthesised mRNAs. Under normal conditions more than 90% of the ribosomes are involved in translation [136, 137] thus leaving little capacity to reallocate ribosomes to new mRNAs [138]. Upon exposure to osmotic stress the global transcript level is reduced significantly, by about 33% (from ∼2400 mRNA copies/cell in MM to ∼1600 mRNA copies/cell), thereby preferably highly translated mRNAs with high copy numbers are significantly reduced. Since the number of ribosomes does not change during osmotic stress, as estimated for yeast, the hypothesis that transcript reduction is linked to redistribution of translational capacity is supported by a theoretical study suggesting that in *E. coli* continuous growth rate and nutrient quality are balanced by the ribosome allocation [139]. Moreover, transcript reduction was also observed by exposure of yeast to osmotic upshift [140] implying conserved features of response across the species. The effect of transcript reduction is highly specific for osmotic stress: the transcript reduction upon heat exposure is 7.6% (from ∼7800 mRNA copies/cell in LB to ∼7200 mRNA copies/cell). The five major transporters maintaining the uptake of osmotic substances, proV, proP, proX, proW, otsA were expressed at normal conditions at very low copy number <0.5 copies/cell and upregulated to >2 copies/cell upon osmotic upshift. Other osmotic stress-related genes are upregulated from ∼1 mRNA copy/cell in MM to >2 copies/cell upon osmotic stress. Under normal conditions their translation will compete with genes involved in key physiological processes, translation and energy metabolism (expressed at >2 copies/cell). Thus, the reduction of the total amount of mRNA serves to replenish the pool of free ribosomes, which in turn boosts the translation on the remaining pool of mRNAs. Ultimately, proteins mediate stress response and their levels have to be rapidly adjusted to ensure cell adaptability and survival under stress. Our observations are in a qualitative agreement with an earlier observation suggesting that the average copy number correlates well with the average protein concentration [45]. Previous analysis comparing transcript and protein abundance in *E. coli* concluded a relatively good correlation on a global, population level, while on a single cell level, the mRNA and protein levels correlate poorly. This poor correlation is often attributed to different time scales of transcript and protein turnover [45]. Although each RPF read on an mRNA is producing a protein [68], the RPF coverage does not give an information on protein amounts, thus we cannot judge the contribution of protein degradation. In summary, our data reveals a multifaceted stress-specific response in bacteria towards acute thermal and osmotic stress, enabled by a versatile induction of transcriptional and translational programs. Unicellular organisms lack the internal homeostatic buffering capacity of multicellular species, and for survival under acute stress they need a quick reprogramming of their cellular activities. For the majority of genes, changes in the mRNA level correspond to the translation level for both types of stress. Outside of these genes, a unique fraction of genes are regulated only at the level of translation in response to the two stress conditions we studied. Indeed, reprogramming of cellular activities by translating existing mRNAs is more efficient than the response shaped by de novo transcription followed by translation enabling the cell to respond faster to changing environments.

2.5 Additional Information

Acknowledgments We thank Wei Chen and Mirjam Feldkamp (Max Delbrück Center, Berlin) for assistance with the sequencing.

Funding Statement This work was supported by the Deutsche Forschungsgemeinschaft (ZI) and European Union (grant NICHE ITN) (ZI, AV).

Data Accessibility The sequencing data have been submitted to Gene Express Omnibus (GEO; GSE68762) database.

Competing Interests We have no competing interests.

Authors' Contributions IF performed all ribosome profiling and deep-sequencing experiments. AB and PF analysed the sequencing data and contributed to data interpretation. GZ mapped the sequencing data; AV and CS contributed to the fold-change analysis. IF and ZI conceived concepts, planned and designed the experiments and ZI and AB wrote the manuscript.

3

Quantitative assessment of ribosome drop-off in *E. coli*

A version of this chapter has been published as:
Sin C, Chiarugi D, Valleriani A (2016) Quantitative assessment of ribosome drop-off in *E. coli*. Nucleic Acids Research 44(6): 2528-2537.

Online version:
http://nar.oxfordjournals.org/content/44/6/2528.full.pdf

CONTRIBUTIONS: This project was the evolution of [2]. I discovered that there was a length dependent relationship to ribosome density, through the decile analysis. Furthermore, I showed that the relationship between length and ribosome density followed a power law. I concluded that this could be due to the phenomenon of ribosome drop-off, and we designed an analysis to test it. DC preformed all the bioinformatics and coded the scripts for analysis. I used the mapped reads from DC to confirm the results in parallel using my own scripts.

Abstract

Premature ribosome drop-off is one of the major errors in translation of mRNA by ribosomes. However, repeated analyses of Ribo-seq data failed to quantify its strength in *E. coli.* Relying on a novel highly sensitive data analysis method we show that a significant rate of ribosome drop-off is measurable and can be quantified also when cells are cultured under non-stressing conditions. Moreover, we find that the drop-off rate is highly variable, depending on multiple factors. In particular, under environmental stress such as amino acid starvation or ethanol intoxication, the drop-off rate markedly increases.

3.1 Introduction

Translating messenger RNA (mRNA) into proteins is a complex polymerization process that lies at the heart of protein synthesis. Ribosomes play a pivotal role in this process, decoding of the genetic information contained in the mRNA into amino acid sequences [141].

Given their crucial role, ribosomes are designed to be accurate and robust processive machines. Nevertheless, inherent to all biological processes, errors can occur during protein synthesis. One of the possible errors is premature termination of the translation process; here the ribosome fails to complete the synthesis of a full-length protein.

Various mechanisms are known to mediate translation abortion. Some of them are believed to be relevant mainly when the cell faces stressing conditions that hamper mRNA translation, e.g. amino acid starvation. In bacteria, at least four abortion-mediating factors, namely the tmRNA-SmpB complex [142, 143], RF3 [144], ArfA [145] and ArfB [146] are known to help rescue stalling ribosomes through processes that eventually lead to premature termination of protein synthesis. More surprisingly, translation abandonment can be also part of a proof reading mechanism that interrupts the synthesis of miscoded polypeptides [147]. Besides these factor-mediated pathways, unspecific events, often referred to as nonsense errors [148] or processivity errors [55, 65, 149, 150] can interrupt the elongation of the nascent peptide. Some examples of these errors include false termination due to a false stop codon resulting from frameshift and accidental peptidyl tRNA dissociation from the translation complex [151, 152]. Also, local depletion of ternary complexes can provoke longer pausing events, which may trigger the drop-off of the ribosome [153]. Both factor-mediated translation abandonment and processivity errors prevent the ribosome from reaching the final stop codon. Hence, irrespectively of the mechanism involved, we will use the term "*ribosome drop-off*" to denote all the events that entail the premature detachment of the ribosomes from the mRNA template.

Ribosome drop-off is not limited to stress conditions; it occurs even when the cell is in a non stressing environment [55, 63, 65–67, 149]. In these conditions, the frequency of drop-off events in not affected by external stress and, thus, it is expected to assume a "basal" value. In addition to the seminal works of Kurland and co-authors – reviewed in [149] – other proposals have explored the magnitude of the ribosome drop-off "basal rate" or the dynamics of the phenomenon.

In Refs. [65–67] ribosome drop-off was clearly detected and estimated for the β-galactosidase gene through different *in vitro* approaches. In Ref. [63] an *in vivo* experiment estimated the drop-off rate for *E. coli* to be $4 \cdot 10^{-4}$ events per codon. In Ref. [154], theoretical arguments demonstrate that the presence of a basal drop-off rate leads necessarily to an exponential distribution of ribosomes along the mRNA, while in Ref. [148] a model-based approach elucidates the impact of ribosome drop-off on protein synthesis. Thus, a well assessed quantitative estimate of the rate of ribosome drop-off could have a strong impact on modeling of ribosomal traffic and protein synthesis [155, 156] as well as provide further hints to understand the relationship between gene length and protein abundance [64, 157].

In spite of these well-assessed findings, so far the analysis of data from Ribosome-profiling (Ribo-seq) experiments failed to find the existence of measurable ribosome drop-off frequency in non stressing conditions [68–70].

The ribosome profiling technique [158] begins with drug-mediated interruption of the cellular translation process, followed by the hydrolysis of the mRNA regions that are not covered (protected) by the ribosomes. The residual mRNA oligomers (known as ribosome protected fragments, RPF, because they are the mRNA fragments that were protected by the ribosome) are deep sequenced. Then, the positions of the ribosomes are determined by mapping the sequences to the reference genome.

In this setting, the relative abundance of RPFs that map to different parts of the single genes, usually evaluated in terms of ribosome density (RD) and measured in number of RPFs per codon, is typically used to estimate the protein synthesis rate for each gene. The distribution of RPFs along the genes can also provide information about the possible presence of ribosome drop-off: an average decrease of the RD from the 5' end to the 3' end of each open reading frame (ORF) reveals that a significant number of ribosomes fail to reach the 3' end.

In this work we reevaluate the analysis of Ribo-seq data with the goal of quantifying weak signals of ribosome drop-off. The methods used so far were not able to detect

Dataset **#**	**Series** (GSE ID)	**Organism** (E. coli)	**Samples** (GSM ID)	**Ref.**
1	**68762**	MC4100	1680885 - 1680884	[2]
2		MC4100	1680887 - 1680886	
3		MC4100	1680889 - 1680888	
4		MC4100	1680891 - 1680890	
5	**51052**	MG1655	1399615 - 1399616	[159]
6		MG1655	1399610 - 1399611	
7		BW25113	1399617 - 1399618	
8		BW25113	1399612 - 1399613	
9	**56372**	MG1655	1360042 - 1360030	[160]
10		MG1655	*1360043 - 1360031*	
11		MG1655	1360044 - 1360032	
12		MG1655	*1360045 - 1360033*	
13		MG1655	1360046 - 1360034	
14		MG1655	*1360045 - 1360035*	
15	**58637**	MG1655	1415871 - 1415869	[161]
16		MG1655	1415872 - 1415870	
17	**53767**	MG1655	1300279 - 300282	[68]

TABLE 3.1: Coordinates of the datasets analyzed in this paper. Column 1: Samples ID (referenced throughout the paper); Column 2: GEO Series ID; Column 3: Organisms used for the sequencing; Column 4: GEO Samples ID (left: Ribo-seq sample; right: RNA-seq sample). Column 4: Publication of reference. The entries reported in italic are the technical replicates of the entries reported above them.

any ribosome drop-off in these data, despite experimental evidence of the phenomena. Through a new way of analyzing the same Ribo-seq data we will show that we find significant evidence of ribosome drop-off and that its rate of occurrence can be determined quantitatively and it is consistent with earlier experimental estimates.

3.2 Materials and Methods

To compute the ribosome drop-off rate in *E. coli* , we analyzed all the related datasets present up to now in the GEO database [162] in which both the Ribo-seq data and the corresponding RNA-seq data were submitted. The GEO coordinates for these datasets are reported in Table 3.1. The experimental datasets analyzed here provide both ribo-seq and RNA-seq data for *E. coli* grown under different normal and stressed conditions. Datasets 1 and 2 refer to *E. coli* grown in LB medium (dataset 1) and then subject to acute heat stress at $47^{\circ} C$ (dataset 2). Datasets 3, and 4 refer to *E. coli* grown in minimal medium (dataset 3) and then subject to acute osmotic stress (dataset 4). Dataset 5 refers to a strain of *E. coli* unable to synthesize Leucine when grown in a medium with Leucine. Dataset 5 will be compared with dataset 6, where the growth medium has no Leucine. Dataset 7 refers to a strain of *E. coli* unable to synthesize Serine when grown in a medium that provides Serine. Dataset 7 will be compared with dataset 8, where the growth medium has no Serine. Datasets 9, 11, 13 refer to *E. coli* grown in LB medium, subject to acute ethanol stress, and subject to chronic ethanol stress, respectively. Datasets 10, 12, 14 are replicas thereof. In datasets 15 and 16, *E. coli* is grown under normal conditions (dataset 15) and then subject to induced high expression of the sigma factor σ^E. Finally, dataset 17 reports on *E. coli* grown under normal conditions (LB medium).

In each case, we started our study from the "raw data" consisting of FASTQ files [163]. These files contain both the sequence of the oligonucleotides (reads) coming from the deep sequencing process (without any *a posteriori* data manipulation) and information about the quality of each read, *i.e.,* the probability that of each nucleotide being correctly sequenced.

Our analysis protocol consists of two subsequent steps. In the "upstream phase", we use existing software tools pipelined together to obtain a reliable mapping of the reads on the reference genome. In the "downstream phase", we performed statistical analysis of the outputs from the upstream step using in-house scripts written in the "R environment" [164].

3.2.1 Upstream analysis

For the upstream analysis of both the Ribo-seq and the related RNA-seq data, we applied the following procedure. The raw data was filtered using CUTADAPT [165] (release 1.8.3) such that only high quality reads were kept (Q-score $\geq$ 40, which corresponds to a sequencing error probability of 0.0001%). This refinement of the read sequences allowed us to reduce the probability of errors in the subsequent mapping phase; the presence of mis-sequenced nucleotides introduces artifacts that can increase the similarities between the query sequences and wrong mapping positions in the reference genome, thus increasing the probability of incorrect mapping. CUTADAPT was then used again to trim the adaptor sequences from the remaining reads.

We then filtered out all the reads that were shorter than 15 nucleotides to reduce the prevalence of multi-mapping errors. Shorter reads have a much higher chance of mapping to multiple places in the reference genome, simply due to combinatorics; thus with short reads, we cannot be confident that the part of the genome that the read mapped to actually reflects the origin of the read.

Afterwards, we mapped the resulting reads against the rRNA sequences of *E. coli* to filter out the reads coming from the sequencing of rRNAs. We used the Bowtie2 aligner [166] (release 2.2.5) setting the running parameters in order to allow a successful mapping only when a high degree of similarity between the query read and the reference sequence occurs. In particular, we set the seed length to 15 (the minimum reads length) and we allowed no mapping mistakes in the seed. In this way, we maximized the probability of ruling out only the rRNA reads.

Finally, the remaining reads were mapped against the whole set of protein coding ORFs in *E. coli* , taken from the EnsemblBacteria database [167]. Among the reads that mapped on the reference ORFs we selected the ones that mapped with the highest score possible for Bowtie2 (MAPQ = 42).

Ultimately, in the upstream phase we aim to minimize the bias that could arise either from the non-optimal quality of the sequenced reads or from the mapping process. For each sample analyzed (Table 3.1, fourth column), we obtained two set of reads (one coming from the raw Ribo-seq data and the other from the corresponding raw RNA-seq data) mapping on the same set of ORFs. These “refined” datasets are used in the subsequent phase of our analysis.

3.2.2 Downstream analysis

In this step of our analysis lies the core of our method. All the procedures described hereafter were implemented through a custom script in the “R environment”. For the sake of readability we report all the details of the downstream analysis in Supplementary Section 3.1.

Computing the average number of RPFs per ORF: a novel binning strategy.

For each dataset reported in Table 3.1, we divided each ORF in bins of ℓ nucleotides and we counted the number of RPFs that mapped in each bin. This results in the RPF matrix composed by cells (i, j) reporting the number of RPFs that, for any given ORF$_i$, map in the corresponding bin j (see Supplementary Figure 3.1).

To normalize the amount of RPFs with the abundance of the corresponding RNA-seq reads, we divided the value in each cell of the RPF matrix by the quantity RNA$_{(i,j)}$

given by

$$\mathrm{RNA}_{(i,j)} = (\text{ total RNA-seq reads for } \mathrm{ORF}_i) \cdot \frac{\ell_{(i,j)}}{L_i}\,, \tag{3.1}$$

where $\ell_{(i,j)}$ is the number of nucleotides of the ORF_i in cell (i, j) and L_i is the length of the ORF_i in number of nucleotides.

In this way, we obtain the matrix NRPF that reports the normalized Number of RPFs per bin in each cell:

$$\mathrm{NRPF}_{(i,j)} \;=\; \frac{\mathrm{RPF}_{(i,j)}}{\mathrm{RNA}_{(i,j)}}\,, \tag{3.2}$$

which is equivalent to assume a uniform coverage of each ORF by the RNA-seq reads mapping on it.

Finally, we computed the average over each column j of the NRPF matrix, obtaining a vector Y that contains the average normalized number of RPFs per bin for the whole set of ORFs. This averaging procedure ensures that sequence-specific features are also averaged out. We then use the vector Y to compute the drop-off rate r as detailed in the next paragraph. Supplementary Figure 3.1 reports a schematic representation of the binning strategy described above.

Estimation of the drop-off rate and its associated error.

To obtain an estimate of the drop-off rate r per codon, we investigated the relationship between the average number of RPFs per bin Y and the bin number X. Inspired by theoretical considerations [154], we studied the dependence of Y from X in the form of the exponential decay

$$Y \;=\; A\,\mathrm{e}^{-RX}\,, \tag{3.3}$$

where $X = 1, 2, \ldots$ is the bin number and A is the intercept, which is of no interest here. The value of R can be referred to as the drop-off rate *per bin* and, widely speaking, indicates the probability per bin that a ribosome prematurely detaches from the mRNA template. The corresponding drop-off rate *per codon* r can be exactly related to R considering that drop-off events can occur anywhere inside each bin. Indeed, if r is the drop-off rate per codon, then the probability that the ribosome does not drop-off within a bin of ℓ_c codons is $(1-r)^{\ell_c}$. Consequently, the probability R that any ribosome drops-off anywhere within the bin is $1-(1-r)^{\ell_c}$ and the drop-off per codon is $r = 1-(1-R)^{1/\ell_c}$.

To obtain a precise estimate of R and its associated error, we relied on a bootstrapping procedure, applied to each column of the NRPF matrix. In this way, for each dataset, we produced 10^5 Y vectors that are independent but statistically equivalent to each other. For each Y vector of any given dataset, we estimated one value of R through the (weighted) linear regression of $\ln Y$ versus X, thus exploiting the relationship described by Equation 3.1. From this procedure we obtained a *normal distribution* of the possible values of R for each original dataset (Supplementary Figure 3.2). The average and the variance of that distribution provide a *first estimate* of the true value of R and its associated variance. We call R_{BS} the estimate of the drop-off rate per bin from the bootstrap. Its standard deviation is called S_{BS}. Both R_{BS} and S_{BS} are specific to each single dataset (Supplementary Table 3.1).

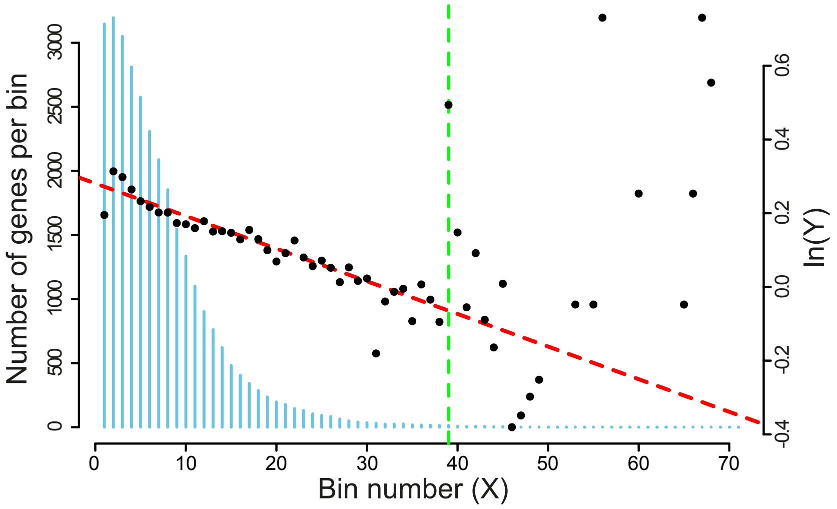


Figure 3.1: Number of elements in each column of the NRPF and BS matrices. The hisogram (blue vertical bars) gives the number of genes contributing in each bin (scale on the left vertical axes). This number decreases from left to right. The plot superimposed to the vertical bars (resulting from the analysis of dataset 17, right vertical axes) shows that the scattering of the plotted values increases with the bin number, indicating an increase of the variance associated to the estimation of average normalized RPF's per bin, Y. The green vertical line represents the cut-off (39 bins of 100 nucleotides) that we chose to obtain the best estimate of the drop-off rate.

We further evaluated our estimate of R for possible systematic errors resulting from the choice of the bin size and the number of bins considered for the regression. The bin size affects the sensitivity of the regression to the drop-off rate. Larger bin sizes result in smoother curves, but lose information to averaging. Furthermore, the number of bins considered in the regression also affects the reliability of the estimation of R. In *E. coli*, the length of ORFs in genes is not uniformly distributed – there are significantly fewer genes with very long ORFs (Figure 3.1). Thus, statistics for later bins are sparse, and the bin average becomes a bad estimator of vector Y.

To evaluate possible systematic errors, for each one of the analyzed datasets we created several simulated datasets. The simulated datasets replicated the original counterparts both in the ORDs lengths and in the number of reads mapping in each ORF. The position of the RPFs, instead, was redistributed along the ORFs according to a nominal drop-off rate. Our aim was to repeat the boostrapping procedure as we did for the real datasets and to measure the systematic deviations from the nominal drop-off rate used to generate the artificial profiles. For each dataset, we repeated the bootstrap process for various combinations of bin sizes and number of bins considered in the regression, looking for the combination that resulted in an R closest to the nominal drop-off rate. From this analysis we found that using a bin size of 100 nucleotides and the first 39 bins for the regression yields an estimate of R which is closest to the nominal value used for all datasets. With these settings, the estimated R was offset from the nominal value by a datasets-specific Δ and an associated standard deviation S_Δ. The results obtained are reported in detail in Supplementary Table 3.1. Thus, for each experimental dataset, the best estimate of the drop-off rate R and its associated standard deviation per bins of 100 nucleotides is given by

$$R = R_{BS} - \Delta S_R = \sqrt{S_{BS}^2 + S_\Delta^2} \ , \tag{3.4}$$

from which we obtain the drop-off rate and standard deviation per codon as $r = 1 - (1 - R)^{3/100}$ and $S_r = 1 - (1 - S_R)^{3/100}$, respectively.

Summing up, our procedure to evaluate R consists of two steps: first, our bootstrap approach allows us to produce a set of "simulated technical replicas". From this, we obtain a provisional estimate of the drop-off rate R_{BS} and its associated standard deviation S_{BS}. Next, we correct for the systematic effects of binning by adding the offset Δ while taking its variance S_Δ into account.

3.3 Results

For each experimental dataset reported in Table 3.1 we applied the analysis protocol described in *Materials & Methods*. The values we obtained for the drop-off rate per codon, r, are reported in Table 3.2 together with the respective standard deviations, S_r, and the 99% confidence intervals (Supplementary Section 3.2).

To check whether the values we obtained for r are significantly different from 0, we performed a Z-test for the mean for each of the r with a significance level of 0.01 (3.2). The results of our tests revealed that in 14 out of 17 cases we measured a drop-off rate significantly larger than 0. In the two datasets 6 and 8, instead, the relationship between X and Y is more complex than the single exponential decay described by Equation 3.1 and, thus, r cannot be evaluated through our method. The corresponding entries in Tables 3.2 are, then, labeled with *n.a.*.

Dataset	r (10^{-4})	S_r (10^{-4})	CI (10^{-4})
1	2.9	0.3	r ±0.8
2	2.2	0.4	r ±1.0
3	2.4	0.2	r ±0.5
4	1.9	0.3	r ±0.8
5	0.7	0.3	r ±0.8
6	*n.a.*	*n.a.*	*n.a.*
7	1.9	0.2	r ±0.5
8	*n.a.*	*n.a.*	*n.a.*
9	2.4	0.5	r ± 1.3
10	2.2	0.7	r ±1.8
11	5.1	0.4	r ±1.0
12	5.6	0.3	r ±0.8
13	2.3	0.3	r ±0.8
14	2.3	0.3	r ±0.8
15	3.0	0.3	r ±0.8
16	0.0	0.4	r ±1.0
17	1.4	0.2	r ±0.5

Table 3.2: Drop off rates detected in the analyzed datasets. Column 1: Datasets ID (see Table 1 for the respective GEO coordinates). Column 2: Drop-off rate per codon. Column 3: Standard deviation associated to r. Column 4: 99% Confidence Interval associated to r.

Further analysis, based on the ANOVA test revealed significant differences among some values of r (see Supplementary Section 3.2.3 for details). Given that the ANOVA test does not tell us *which* of the tested values are significantly different or equal to each other, we performed a series of coupled *post hoc* tests to investigate the sources of the detected variability. In particular, we compared the values we obtained for r within each GEO series. The outcomes of this analysis are reported in the next paragraphs.

3.3.1 Comparing the Drop-off rates in normal and stressing conditions

Datasets 9, 11 and 13: Ethanol-induced stress.

This series refers to a set of experiments performed to elucidate the effect of ethanol intoxication on the translation machinery of *E. coli* MG1655. To this aim, the bacterial cells, first cultured in a minimal medium (M9 minimal medium, supplemented with MgSO4 1 mM, CaCl2 0.1 mM, and glucose 10 g/L) were exposed at $T_0 = 0$ to a toxic concentration of ethanol (40g/l) and sampled after $T_1 = 10$ minutes and $T_2 = 70$ minutes.

Through the Z-test, we compared the drop-off rate that we measured at each time point and, as shown in Table 3.3, it resulted that at T_1 there was a significant increase in the drop-off rate. At T_2, instead, the frequency of drop-off events restored to values similar to the basal rate that we measured at T_0.

Compared samples	**Exp. conditions**	**Z score**	**Sig. level**	Z_B
(Dataset ID)			($\pm Z_{0.0025}$)	
9 *vs.* 11	M9 Minimal Medium *vs.* Ethanol stress - T_1	4.19	± 2.81	± 3.14
9 *vs* 13	M9 Minimal Medium *vs.* Ethanol stress - T_2	0.17	± 2.81	± 3.14
13 *vs* 11	Ethanol stress - T_2 *vs.* Ethanol stress - T_1	6.08	± 2.81	± 3.14

Table 3.3: Results of the Z-tests to compare the drop-off rates of Datasets 9, 11 and 13. Column 1: Dataset ID (see Table 1 for the respective GEO Coordinates). Column 2: Experimental conditions. Column 3: Z-score computed from the comparison of the Drop-off rates. Column 4: Percentiles of the standard normal distribution corresponding to a total rejection area of 0.005. Column 5: Percentiles of the standard normal distribution corresponding to a total rejection area of 0.005, corrected according to the Bonferroni method [168]. This test shows that the drop-off rate under acute ethanol stress is significantly different form the drop-off rate under normal conditions and chronic stress.

The results of our analysis agree with the findings that ethanol alters the structure of ribosomes, inducing an increase of translational errors and stalling events that in turn trigger cellular responses leading to premature translation termination [143]. Figure 3.6, reporting the plots we obtained from our analysis, provides a graphical view of our findings.

Hence, our approach reveals the existence of a basal drop-off rate that can be affected by environmental stress. Moreover, performing our analysis at different time points, we were able to provide some insights on the timing of the stress response: the reliability of the translation process is affected by ethanol only for a limited amount of time.

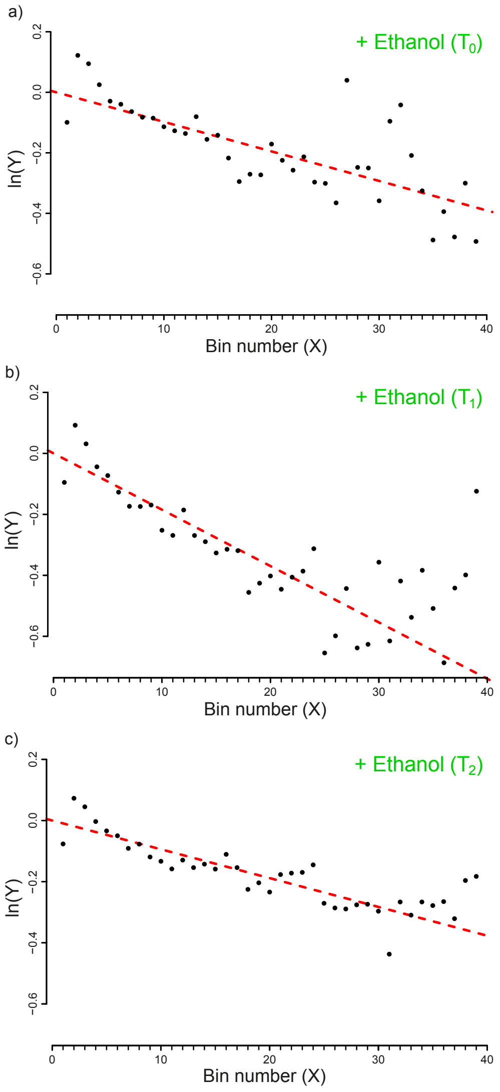


FIGURE 3.2: Plot of the vector Y vs. the number of bins (X). The slopes of the dashed lines correspond to the drop-off rate r reported in Table 3.2. **a)**: Dataset 9 - Control (T_0). **b)**: Dataset 11 - T_1, after 10 minutes of ethanol stress. **c)**: Dataset 13 - T_2, after 70 minutes of ethanol stress. The plots includes only the first 39 bins that we considered in our analysis. To facilitate the comparison with the similar graphs present in the paper we shifted the plots for a distance equal to the intercept of the regression line. The complete plots are reported in Supplementary Figure 3.6.

Datasets 5, 6, 7, 8: Amino acids starvation.

The experiments related to this series report the analysis of *E. coli* MG1655 and *E. coli* BW25113 cells that exhibit auxotrophic phenotypes respectively for the amino acids Serine and Leucine. The auxotrophic strains were grown either in a complete medium containing also the essential amino acids (rich medium - MOPS) or in conditions of starvation of the essential amino acid. Through our analysis, we succeeded in evaluating the basal drop-off rates related to the control experiments (growth in rich medium). More interestingly, the effect of premature ribosomal drop-off in the starvation conditions resulted to be qualitatively different from the normal conditions (Figures 3.3 and 3.4). This finding is consistent with the findings in Ref. [159] where an increase in the drop-off events upon amino-acid starvation was detected. Noticeably, the method used in [159] did not allow the detection of drop-off events in the control conditions. A possible explanation for this discrepancy is that in Ref. [159] a sliding window approach was used to average the ribosome profiles. In our method, the averages are computed in bins of fixed length. Even though the sliding window technique usually enhances the signal to noise ratio, it is also less sensitive to the signal detection with respect to our strategy. Therefore, it is possible that in Ref. [159] low frequencies of drop-off events are not detected because the sensitivity is compromised for the noise dampening. Another likely possibility is that a drop-off rate of 10^{-4} is hardly visible at the scale used in the plots of [159].

Inspecting the plots, the decay of the density profile Y under both starvation conditions seem as it could be better desribed by a two-exponential decay model in which a steeper exponential curve is followed by a less steep one. Thus, the poor fit with a single exponential decay, prevents a computation of r. This finding suggests that a more complex dynamics of the drop-off events is likely to come into play in conditions of heavy cellular stress such as amino acid starvation.

Summing up, our method allowed us to gain preliminary insights on the dynamics of drop-off events measured in different experimental settings. The global increase in the drop-off rate during amino-acid starvation is consistent with the idea that the starvation-induced increase of ribosome stalling events enhances the triggering of the rescue pathways [143] that, eventually, lead to a higher frequency of translation abortions.

Datasets 15 and 16: A novel σ^E-induced sRNA.

The set of experiments related to this dataset were performed to find putative novel targets for the σ^E transcription factor, which is known to play a pivotal role in regulating the homeostasis of the outer membrane [169]. In one of the experiments, σ^E was

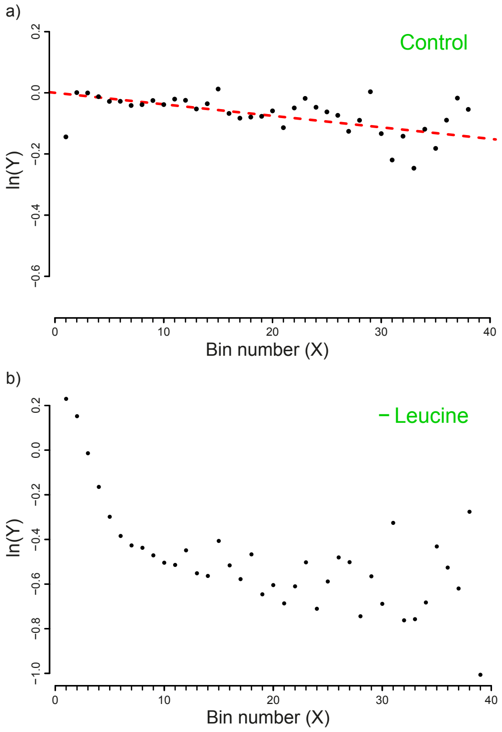


Figure 3.3: Plot of the vector Y vs. the number of bins (X). The slope of the dashed line corresponds to the drop-off rate r reported in Table 3.2. **a)** Dataset 5 - Control (MOPS - Rich medium) **b)** Dataset 6 - Leucine starvation. In this case, due to the poor fit with a single exponential model, we could not compute r. Thus, the regression line is not represented here.The plots include only the first 39 bins that we considered in our analysis. To facilitate the comparison with the similar graphs present in the paper we shifted the plot for a distance equal to the intercept of the regression line. The complete plots are reported in the Supplementary Figure 3.7.

ectopically overexpressed, inducing the overespression plasmid pRpoE through 1mM of IPTG. Two samples were, then, harvested at $T_0 = 0$ and $T_1 = 20$ minutes and analyzed through Ribo-seq and the corresponding RNA-seq.

We analyzed the outcomes of these experiments and measured the drop-off rates at the two time points. Our Z-test reveals a clear difference in the two drop-off rates (see Table 3.4) which is evident also by inspecting the plots reported in Figure 3.5.

In particular, the sample collected at T_1 exhibits a drop-off rate approximately equal to zero, which is the only case we obtained in our analysis. The biological interpretation of this finding is not easy to achieve, due to the scarcity of information regarding the role of σ^E in the regulation of the translation process. Indeed, the transcription factor σ^E is mainly known as a pleiotropic gene expression inducer that promotes the transcription of about 100 genes and three small regulatory RNAs. Our results point towards possible additional roles of σ^E in increasing the reliability of the translation process, at least when it is highly expressed.

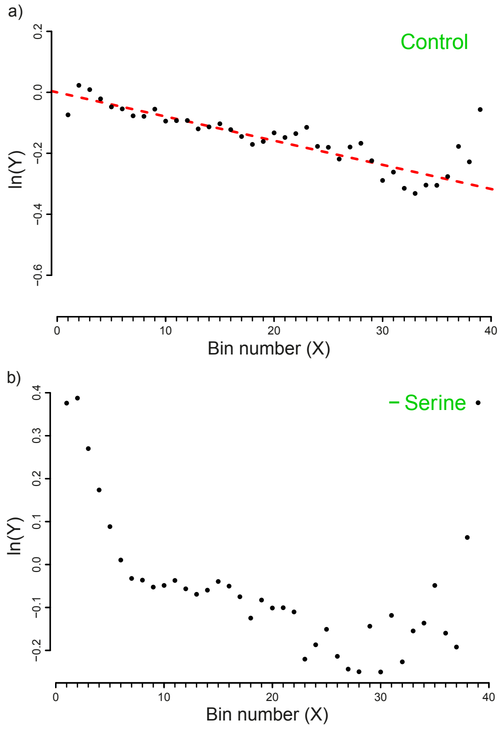


FIGURE 3.4: Plot of the vector Y vs. the number of bins (X). The slope of the dashed line corresponds to the drop-off rate r reported in Table 3.2. **a)** Dataset 7 - Control (MOPS - Rich medium) **b)** Dataset 8 - Serine starvation. In this case, due to the poor fit with a single exponential model, we could not compute r. Thus, the regression line is not represented here. The plots include only the first 39 bins that we considered in our analysis. To facilitate the comparison with the similar graphs present in the paper we shifted the plot for a distance equal to the intercept of the regression line. The complete plots are reported in the Supplementary Figure 3.8.

Compared samples	**Exp. conditions**	**Z score**	**Sig. level**	Z_B
(Dataset ID)		$(\pm Z_{0.0025})$		
15 *vs.* 16	Control - T_0 *vs.* High level σ^E - T_1	6.07	± 2.81	*n.a.*

TABLE 3.4: Results of the Z-tests to compare the drop-off rates of Samples coming from the GEO Series GSE58637. Column 1: Dataset ID (see table 1 for the corresponding GEO coordinates) . Column 2: Experimental conditions. Column 3: Z-score computed from the comparison of the drop-off rates. Column 4: Percentiles of the standard normal distribution corresponding to a total rejection area of 0.005. Column 5: Percentiles of the standard normal distribution corresponding to a total rejection area of 0.005, corrected according to the Bonferroni method [168]. The result of the Z-test confirms a significant difference between the drop-off rates at time $T = 0$ and at time $T = 20$ minutes.

Datasets 1, 2, 3 and 4: Heat and Osmotic Stress.

The data reported in Ref. [2] refer to the analysis of *E. coli* MC4100 cells cultured in the LB medium or in a minimal medium (12.8 g/l Na2HPO4 7H2O, 3 g/l KH2PO4, 1

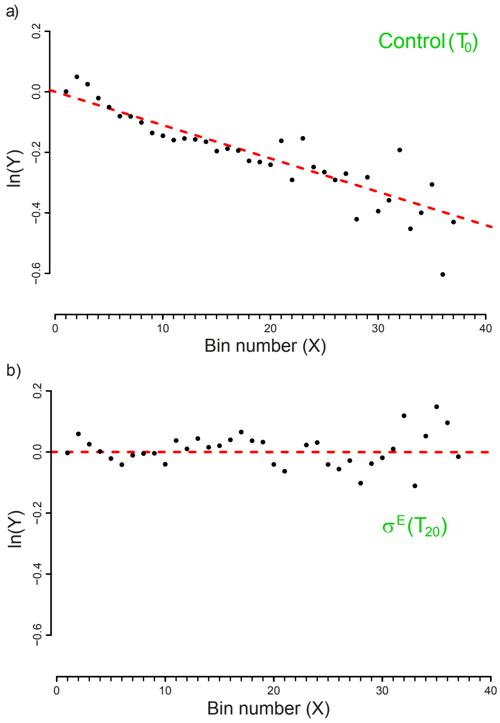


FIGURE 3.5: Plot of the vector Y vs. the number of bins (X). The slope of the dashed lines correspond to the drop-off rates r reported in Table 3.2. **a)**: Dataset 15 - Control (T_0). **b)**: Dataset 16 - T_1, after 20 minutes of σ^E over expression induction. The plots include only the first 39 bins that we considered in our analysis. To facilitate the comparison with the similar graphs present in the paper we shifted the plot for a distance equal to the intercept of the regression line. The complete plots are reported in the Supplementary Figure 3.9.

g/l NH4Cl, 2 mM MgSO4, 0.1 mM CaCl2, 0.4% glucose) and subjected, respectively, to acute heat stress (47°C for 7 minutes) and to acute osmotic stress (NaCl 0.3M for 20 minutes at 37°C). Through our analysis we succeeded in measuring a "basal" rate of drop-off events (see Supplementary Figure 3.10 and 3.11) but we detected no significant differences in the drop-off rates between the control and stress condition. Table 3.5 reports the results of our Z-tests for the mean, showing that the obtained Z-scores (column 3) are in the boundaries of the acceptance area, thus supporting the null hypothesis of equal means.

These results could imply either that the cell-scale translation reprogramming events that are expected to occur in stressing conditions are not strong enough to be detected by our method or that the time scales chosen for harvesting the cells subjected to the stressing conditions were large enough to allow the translation dynamics to be restored to the initial levels. Unfortunately, our method does not allow to discriminate between these two hypotheses.

Compared samples	**Exp. conditions**	**Z-score**	**Sig. level**	Z_B
(Dataset ID)		($\pm Z_{0.0025}$)		
1 *vs.* 2	LB Medium *vs.* Heat Stress	1.61	$\pm$ 2.81	$\pm$ 3.02
3 *vs.* 4	Minimal Medium *vs.* Osmotic Stress	1.21	$\pm$ 2.81	$\pm$ 3.02

Table 3.5: Results of the Z-tests to compare the drop-off rates of samples coming from the GEO Series GSE68762. Column 1: Dataset ID (see table 1 for the corresponding GEO coordinates). Column 2: Experimental conditions. Column 3: Z-score computed from the comparison of the drop-off rates. Column 4: Percentiles of the standard normal distribution corresponding to a total rejection area of 0.005. Column 5: Percentiles of the standard normal distribution corresponding to a total rejection area of 0.005, corrected according to the Bonferroni method [168]. The results of the tests (all the Z scores falls into the acceptance area) show that the drop-off rates measured in normal and stressed conditions do not differ significantly.

3.4 Discussion and Conclusions

Despite clear experimental evidence of ribosome drop-off, past attempts to detect ribosome drop-off in Ribo-Seq data were unsuccessful. In this paper we present a simple data analysis method that is sensitive enough to detect the weak decay of ribosome density over the ORF, which allows us to measure the cell-wide ribosome drop-off rate. With this method, we can even measure the basal rate of ribosome drop-off for cells in non-stressed conditions.

The other analytical approaches reported in literature so far were unsuccessful because the proposed binning strategy was not sensitive enough. These approaches typically divide each ORF into two halves and compare the number of reads that map on each half (see Ref. [68] and Supplementary Section 3.1.3). A significant reduction of reads in the second half would reveal that a certain number of ribosomes have not successfully completed translation. These results are typically illustrated by means of scatterplots where the ribosome density in the first half are plotted against the ribosome density in the second half (Supplementary Figure 3.3 and 3.5). When there is no significant difference between the densities in the two halves, the plotted points will cluster around a straight line with slope = 1. For an ORF where the density of the second half is significantly lower than the first half, the corresponding point would fall below the straight line. At least in principle, this method is mathematically sound. However, it has a major drawback: the sensitivity of this approach depends critically on the ORF length. When the frequency of drop-off events is not large enough with respect to length of the ORFs, the difference in ribosome density between the two halves of the ORF are too small to be detected in a log-log scatterplot. As a consequence, if the genome of interest prevalently contains short genes, the scatterplot-method is not sensitive enough to detect the drop-off. This would lead to the wrong conclusion that, at the genome scale, the ribosome drop-off rate is not measurable.

Fortunately, our analysis technique is not affected by the length of the ORFs. We used our method to analyze various datasets referring to the bacterium *E. coli* cultured in different experimental conditions. The values we obtained for the drop-off rate ranged from a minimum of $1.4 \cdot 10^{-4}$ to a maximum of $5.6 \cdot 10^{-4}$ events per codon. These values make ribosome drop-off not negligible at the cellular level. Indeed, if we consider a drop-off rate of $4 \cdot 10^{-4}$ per codon and an ORF length of 300 codons (approximately the average ORF length for *E. coli*), it turns out that on average, 10 out of every 100 ribosomes will fail to complete the translation of the messenger.

Furthermore, taking into account the speed of ribosomes and the number of ribosomes actively involved in translation [53], and assuming a drop-off rate of $4 \cdot 10^{-4}$ per codon in all growth conditions, the number of premature ribosome drop-off ranges from 1400 per minute per cell at slow growth conditions to 29000 per minute per cell at fast growth conditions. Even considering the lifetime of a cell [136], the total number of drop-off events in a slowly growing population is about $14 \cdot 10^4$ events per cell cycle at slow growth (doubling time 100 minutes) and $75 \cdot 10^4$ at fast growth conditions (doubling time 25 minutes).

Furthermore, we come to a more general result relating drop-off rate and the length of genes. We found that, for a given drop-off rate, there is a limiting gene length above which the translation process becomes ineffective due to the high number of expected drop-off events. In the case of low drop-off rate, this threshold length is usually higher than the maximum gene length of *E. coli.* However, in those cases where ribosomes drop off with a higher frequency, the completion of translation is only reliable for shorter mRNAs. This suggests that when living organisms face conditions leading to increased drop-off rate (e.g. amino acids starvation) only a subset of genes can be effectively expressed. Since the probability of a ribosome to complete translation decreases exponentially with the ORF length, the magnitude of ribosome drop-off becomes an important evolutionary constraint of ORF length. If the genome of an organism is composed of ORFs that are too long relative to the drop-off rate, the reliability of translation may not support cell viability.

Our result is related also to the ongoing discussion concerning the existence of a high density of reads at the beginning of the ORF, a phenomenon sometimes referred to as the “ramp” [170]. With the exception of the data referring to the acute amino acids starvation, the analysis of our density profiles shows that in the samples considered here there is no phenomenological cross-over between the beginning of the ORF and the more downstream bins. Qualitatively, this means that our results would not change after eliminating the two first upstream bins. This indicates that in *E. coli* there is only one mechanism, namely ribosome drop-off, responsible for the decrease of the reads density in the whole ORF.

Contrary to previous Ribo-seq analysis results, we have shown that the magnitude of ribosome drop-off is highly variable and dependent on case-specific factors, including experimental conditions and the protocol used to collect Ribo-Seq data. Since the estimation of translation rates from Ribo-seq data assumes negligible ribosome drop-off, these estimations should be reevaluated to correct for possible biases due to drop-off events. In fact, we speculate that ribosome drop-off could be a possible explanation for the ubiquitous negative correlation between gene length and protein synthesis rate.

3.5 Acknowledgements

The authors would like to thank Z. Ignatova, A. Bartholomaeus, L. Calviello for useful discussions and help at an early stage of this work.

Conflict of interest statement.

None declared.

4

Global quantification of cellular protein degradation kinetics

A version of this chapter has been submitted to Cell as:
McShane E, Sin C, Zauber H, Wells JN, Donnelly N, Wang X, Hou J, Chen W, Storchova Z, Marsh JA, Valleriani A, Selbach M. Global quantification of cellular protein degradation kinetics.

CONTRIBUTIONS: Another large collaboration with many members. I was involved very early on with the idea of age-dependent degradation. I designed the normalization schemes. I implemented all the normalization schemes and parameter estimation machinery. I made the calculations to describe the lifetime properties of each protein (e.g. average lifetime, half-time, steady state distributions, etc.) I also made many of the calculations to evaluate reproducibility, sensitivity and statistical significance. The structure used to manipulate the data is documented in Appendix D.

Abstract

Do young and old protein molecules have the same probability to be degraded? To answer this question we used metabolic pulse-chase labeling and quantitative mass spectrometry to obtain degradation profiles for thousands of proteins. We find that more than 10% of proteins are degraded non-exponentially in mouse fibroblasts. In all of these cases, proteins are less stable in the first few hours of their life and become more stable as they age. We find that degradation profiles are conserved between mouse and human and are similar in two different cell types. Many non-exponentially degraded (NED) proteins are subunits of multiprotein complexes that are produced in super-stoichiometric amounts relative to their exponentially degraded (ED) counterparts. Within complexes, NED proteins have larger interaction interfaces and assemble earlier than ED subunits. Amplifying genes encoding NED proteins increases their initial degradation. Consistently, our decay profiles can help to predict how DNA copy-number alterations affect protein levels. Together, our data show that non-exponential degradation is common, evolutionarily conserved and has important functional consequences for protein complex formation and aneuploidy.

4.1 Introduction

Pioneering experiments by Rudolph Schoenheimer established that proteins are in a dynamic state of synthesis and degradation [171]. The subsequent discovery of lysosomes and the ubiquitin-proteasome system (UPS) provided detailed insights into the molecular mechanisms of cellular protein homeostasis [172, 173]. It is now well-established that proteins are extensively turned over, that this process is specific, and that the stability of individual proteins can vary under different physiological conditions [174].

Despite these mechanistic insights, the kinetics of cellular protein degradation are still not well understood. Early analyses indicated that intracellular protein degradation follows first order kinetics [175, 176]. Accordingly, protein degradation is thought to be an exponential decay process in which young and old proteins have the same degradation probability per unit time (i.e., degradation rate) (Figure 4.1A). However, there is substantial evidence that protein degradation does not always follow first-order kinetics. Pulse-chase experiments by Wheatley et al. indicated that a substantial fraction of proteins are degraded within the first 2 h after synthesis [7]. In addition, the newly synthesized immature forms of proteins, like the cystic fibrosis transmembrane conductance regulator (CFTR) and basigin (CD147), were found to be rapidly degraded while

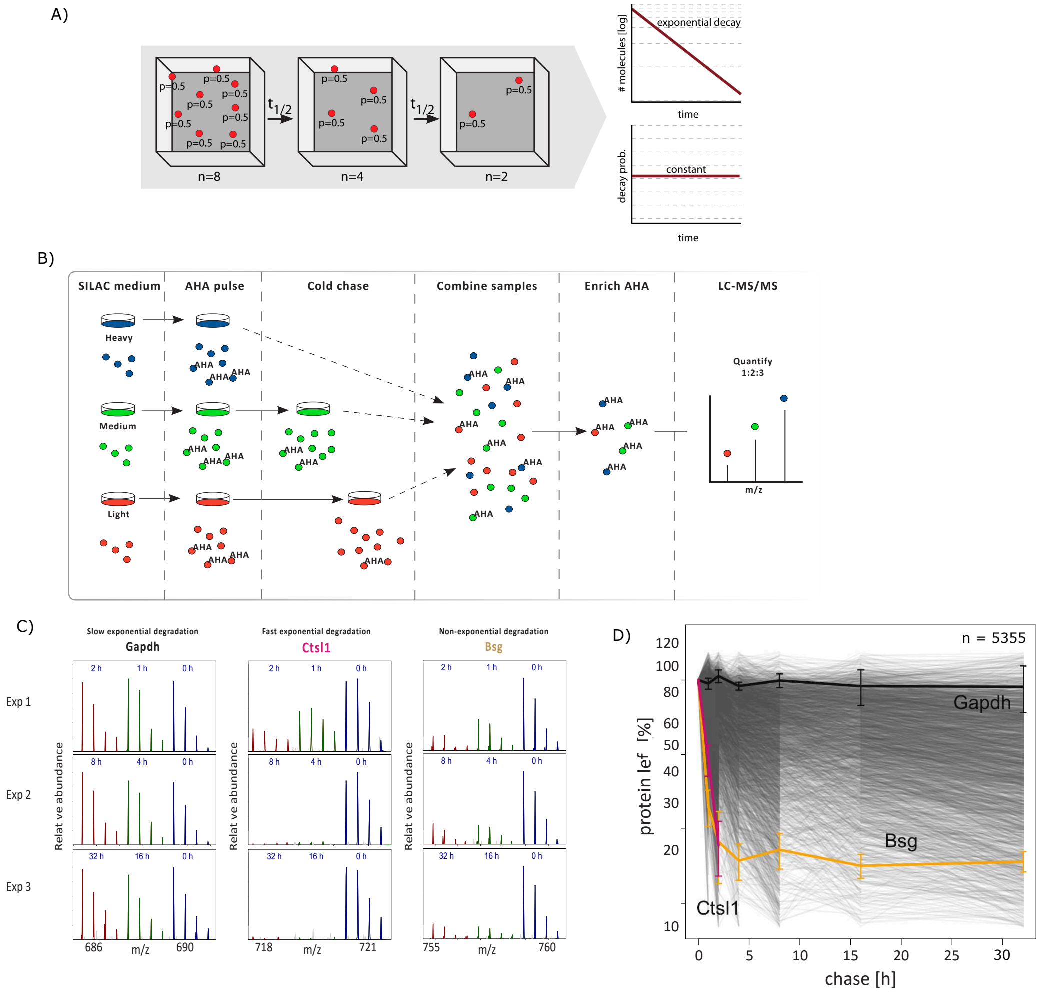


FIGURE 4.1: A) Exponential decay implies that 50% of protein molecules are degraded within one half-life time period. Plotting the remaining number of molecules over time results in a straight line (in a semi log plot). The degradation probability within one half-life time period is 0.5 and constant, i.e. young and old molecules have the same degradation probability. B) Experimental setup for global pulse-chase experiments. Fully SILAC labelled NIH 3T3 cells are pulse-labelled with azidohomoalanine (AHA) and either directly harvested (time point 0 h) or chased in medium containing methionine (cold chase). Samples are combined and AHA containing proteins are enriched by click chemistry. Proteins are digested "on-bead" and peptides are analysed by LC-MS/MS. C) Measured MS spectra for three peptides representing the major types of decay profiles detected. Filamin alpha (Flna, ASGPGLNTTGVPASLPVEFTIDAK) shows slow exponential degradation, Cathepsin L1 (Ctsl1, NLDHGVLLVGYGYEGTDSNK) shows fast exponential degradation, Basigin (Bsg, VLQEDTLPDLHTK) shows non-exponential degradation. Three experiments with different chase times were combined using Heavy samples (t = 0h) as a common reference. D) Decay profiles of individual proteins from one of three biological replicates (grey traces). Highlighted profiles depict proteins shown in C) and are based on all three replicates ($\mu \pm \sigma$). Outliers (> 130% protein left, < 1% of data points) were removed.

the “older” mature forms were stable [177, 178]. It was also observed that proteins can be ubiquitinated co-translationally [179] and that most ubiquitinated proteins in a cell are relatively young [180]. Finally, it has been estimated that around 30% of newly synthesized proteins are quickly degraded [181], although this number was questioned in later studies [182]. Collectively, these studies suggest that decay probabilities of proteins can vary as a function of their molecular age. However, to the best of our knowledge, the degradation kinetics of individual proteins has not yet been investigated on a proteome-wide scale.

To systematically assess cellular protein degradation kinetics we sought to perform pulse-chase experiments on a proteome-wide scale. The general idea is to metabolically label a population of proteins with a short pulse and then to quantify how much of this population is left after different lengths of chase. Traditionally, such experiments are carried out using radioactive amino acids followed by SDS-PAGE and autoradiography [175, 176]. However, this set-up either does not reveal the identity of the labeled proteins or, when coupled to immunoprecipitation, is not scalable and is thus not suitable to study degradation kinetics on a proteome-wide scale. We and others have previously used stable isotope labeling by amino acids in cell culture (SILAC) to study protein synthesis and turnover [72, 183–189]. However, in order to achieve good labeling efficiencies, these approaches require labeling times of several hours, which limits their temporal resolution. Recently, metabolic labeling with bioorthogonal amino acids has emerged as an attractive alternative [190]. Cells can incorporate the artificial amino acid azidohomoalanine (AHA) into newly synthesized proteins instead of methionine [191]. AHA contains an azide group enabling capture of proteins via click chemistry [190]. Combining AHA with SILAC enables relatively short pulse times [192, 193].

Here, we combined AHA and SILAC labeling to obtain the first global survey of protein degradation kinetics. We find that a sizable fraction of proteins are degraded non-exponentially. Degradation profiles were similar in mouse fibroblasts and human epithelial cells. Many non-exponentially degraded (NED) proteins are members of heteromeric protein complexes. These NED proteins are typically over-produced relative to other members of the same complex. Thus, in contrast to recent findings in bacteria [68], disproportional protein synthesis appears to be a common and evolutionarily conserved process in mammalian cells. Our data allowed us to predict how protein levels change in response to gene copy number alterations in aneuploid cells. Therefore, our kinetic profiles are also relevant for gene expression control. Our first global quantification of protein degradation kinetics reveals an unexpected layer of posttranslational regulation with important functional implications.

4.2 Results

4.2.1 Combining metabolic pulse labeling and click-chemistry for global pulse-chase experiments

To perform proteome-wide pulse chase experiments we combined AHA and SILAC labeling (Figure 4.1B). First, cells are fully labeled heavy, medium-heavy or light using SILAC. Second, all three cell populations are pulse labeled with AHA for 1 h. This relatively long pulse labeling time was chosen to allow sufficient label incorporation. Heavy cells are harvested immediately after the pulse while medium-heavy and light cells are chased in AHA-free medium for different lengths of time. All three cell populations are then combined and lysed. Mixing cell populations at this early stage minimizes the impact of sample processing steps on quantification and thus increases robustness. AHA-containing proteins are then purified from the mixed lysate. After digestion into peptides, SILAC-based quantification reveals how much of the pulse labeled fraction remains at different time points.

To establish AHA pulse chase we first analyzed click chemistry-based capturing of heavy AHA labeled proteins after mixing them with unlabeled cells (Supplementary Figure 4.1A). After enrichment, more than 80% of identified proteins were exclusively detected in the heavy form and the rest had a median enrichment ratio higher than 10. Thus, AHA enrichment is highly efficient. We next checked if AHA itself affects protein degradation kinetics. To this end, we pulse labeled cells with both AHA and radioactive cysteine (Supplementary Figure 4.2A). Chasing for six and 24 h did not reveal any apparent impact of AHA on protein degradation compared to control experiments using methionine (Supplementary Figure 4.2B-D). Also, peptides derived from both the N-terminal and C-terminal halves of AHA-labeled proteins had similar intensities, suggesting that AHA does not induce premature termination (Supplementary Figure 4.3). Collectively, these data indicate that AHA pulse-chase (AHA p-c) enables efficient enrichment of newly synthesized proteins with no apparent impact on protein stability, consistent with previous reports [190, 194–196]

For global quantification of protein degradation kinetics we performed three parallel triple-SILAC experiments with different chase times (1, 2, 4, 8, 16 and 32 h) in mouse fibroblasts (NIH 3T3). Heavy cells were always harvested immediately after the pulse and served as a common reference point. Exemplary mass spectra for a stable protein (filamin A, Flna) and an unstable protein (cathepsin L, Ctsl1) show expected slow and fast degradation, respectively (Figure 4.1C). Interestingly, the levels of basigin (Bsg) quickly decreased after the chase but then stabilized after about 4 h. This is consistent with the observation that most of the newly synthesized immature basigin is degraded while the mature form of the protein is stable [177]. We combined the data from the three triple-SILAC experiments to obtain kinetic profiles with seven time points. To compensate for differences in cell numbers we normalized the data using a selected set of very stable proteins (Supplementary Figure 4.4, Materials and Methods). We also subtracted background signals which could otherwise give rise to erroneous degradation profiles (see Materials and Methods for details). The entire large scale experiment was carried out three times, thus yielding data from three independent biological replicates. In total, this analysis yielded profiles for 5,247 proteins (Figure 4.1D, Supplemental Table S1). Overall, reproducibility was very good with coefficients of variation (CVs, computed in log space) of less than 10% at time point 32 h for about 90% of the proteins (Supplementary Figure 4.5).

4.2.2 Stochastic modeling reveals extensive non-exponential protein degradation

To model our experimental protein degradation profiles we adapted a Markov chain-based approach previously used to study mRNA decay (Figure 4.2A) [86]. We considered two different models. In the first model proteins only exist in a single state “A” that is characterized by a constant decay probability. This “1-state model” therefore describes exponential degradation (ED). The second model has an additional state: Newly synthesized proteins first populate state A from where they can either be degraded or transit to state B, which is characterized by a different decay probability. This “2-state model” thus describes non-exponential degradation (NED). To distinguish between both scenarios we compared the relative quality of both models for each protein degradation profile using the Akaike information criterion (AIC) [197]. This approach considers the trade-off between the goodness of fit and model complexity (that is, the number of parameters). Hence, the 2-state-model is only preferred when the improved fit outweighs the increased complexity. The AIC thus provides a conservative estimate of the fraction of non-exponentially degrading proteins. Degradation profiles of Ctsl1 and Flna were better explained by the 1-state model (Figure 4.2B). In contrast, the degradation profile of basigin was better explained by the 2-state model.

Overall, we found that the profiles of 510 proteins are better explained by the 2-state model (Figure 4.2C, AIC probability > 0.8). This corresponds to about 14% of the 3,642 proteins that passed our quality criteria (see Supplementary Methods). We conclude that a sizable number of proteins show a tendency towards NED. Hence, the implicit assumption that protein decay generally follows first order kinetics does not hold for one out of seven proteins in our dataset.

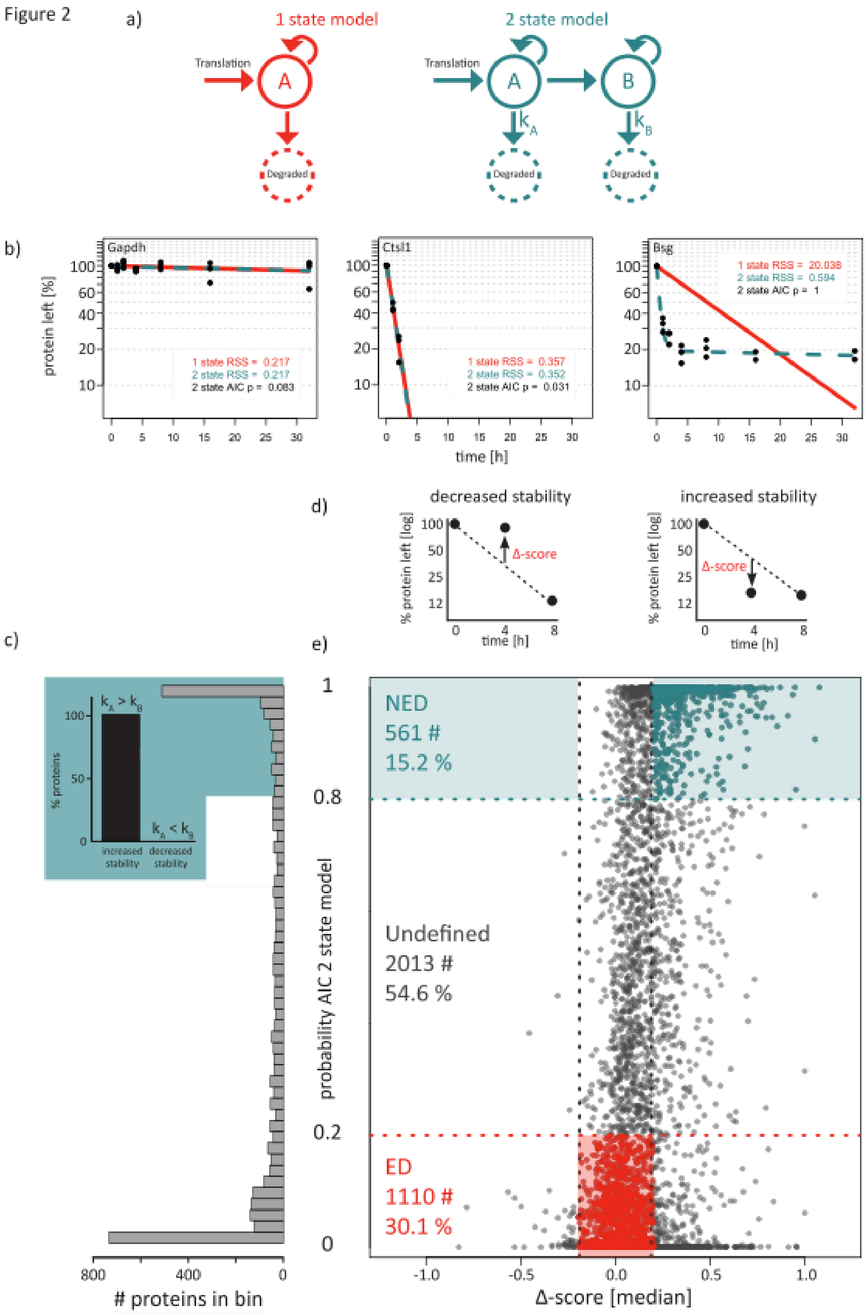


Figure 4.2: A) Graphical representation of the two Markov models applied. The 1-state model and 2-state model reflect exponential degradation and non-exponential degradation, respectively. B) Fitting both models to the exemplary profiles from Figure 4.1D. For Flna and Ctsl1 both models have residual sum of squares (RSS) of similar size. The Akaike information criterion (AIC) therefore recommends the simpler 1-state model. In contrast, the profile of Bsg is better explained by the 2-state model. C) Histogram of all probabilities for the 2-state model for all proteins that passed our quality criteria. D) All proteins with a 2-state probability > 0.8 had a larger initial degradation rate (kA) than later state degradation rate (kB). E) The delta score (Δ-score) as a simple measure for the extent of non-exponential degradation. For each profile, a straight line is drawn between the 0 and 8 h time point (in semi log plot). The Δ-score corresponds to the distance of the measurement at 4 h from this line. F) Plotting the Δ-scores against the 2-state model probability similar to a volcano plot. Non-exponentially degraded (NED) proteins were selected based on a Δ-score > 0.15 and 2-state probability > 0.8. Exponentially degraded (ED) proteins were required to have absolute Δ-scores < 0.15 and 2-state probabilities < 0.2. All other proteins were classified as undefined (UN).

In principle, the 2-state model can describe two different scenarios: When the degradation rate in state A is higher than in state B ($k_A > k_B$) the model describes proteins that become more stable as they age. Alternatively, when the degradation rate in state B is higher than in state A ($k_B > k_A$) the model describes destabilization of older proteins. Intriguingly, we only observed age-dependent stabilization (Figure 4.2D). This is surprising, as we originally expected proteins to become more unstable over time due to the cumulative effects of age-related damage. However, in our cell line model and within the time period monitored (32 h) we did not find any evidence for this. Our finding is consistent with previous evidence that newly synthesized proteins are preferentially degraded [7, 179–181].

While our AIC-based p-value describes the relative quality of both models, it does not provide information about the extent of NED for individual proteins. We therefore also measured the distance of intermediate data points from the linear fit in the semi log plot (Figure 4.2E). This Δ-score is thus a simple measure for the extent of non-exponentiality of individual degradation profiles (see Materials and Methods for details). We then used the Δ-scores and the AIC probabilities to plot the data similar to a volcano plot (Figure 4.2F). Based on this figure, we selected a high confidence set of NED proteins by requiring a 2-state AIC probability of at least 0.8 and a minimal Δ-score of 0.15. Conversely, proteins with absolute Δ-scores < 0.15 and 2-state probabilities < 0.2 were classified as ED. With these cut-offs, about 10% (n=333) were classified as NED and 49.3% as ED (n=1624). The remaining proteins were classified as undefined. For all subsequent analyses we relied on this classification. In summary, our first global quantification of cellular protein degradation kinetics reveals that many proteins become more stable once they have survived the first few hours of their existence.

4.2.3 Validation

Although AHA labeling does not appear to globally affect protein degradation (Supplementary Figure 4.2), it is still possible that labeling with an artificial amino acid introduces systematic biases. We therefore wanted to validate our data with independent methods. First, we compared protein degradation rates in the present study with previously published dynamic SILAC data from the same cell line [72]. First, we assumed exponential degradation and calculated protein half-lives from the degradation profiles by simply taking the time point at which 50% of the protein is left ("exponential half-life"). For ED proteins this resulted in good agreement with the dynamic SILAC data (Figure 4.3A). This is reassuring, especially since dynamic SILAC does not involve artificial amino acids and the way in which half-lives are computed is different. In contrast, NED proteins had overall shorter half-lives in AHA p-c data when we assumed

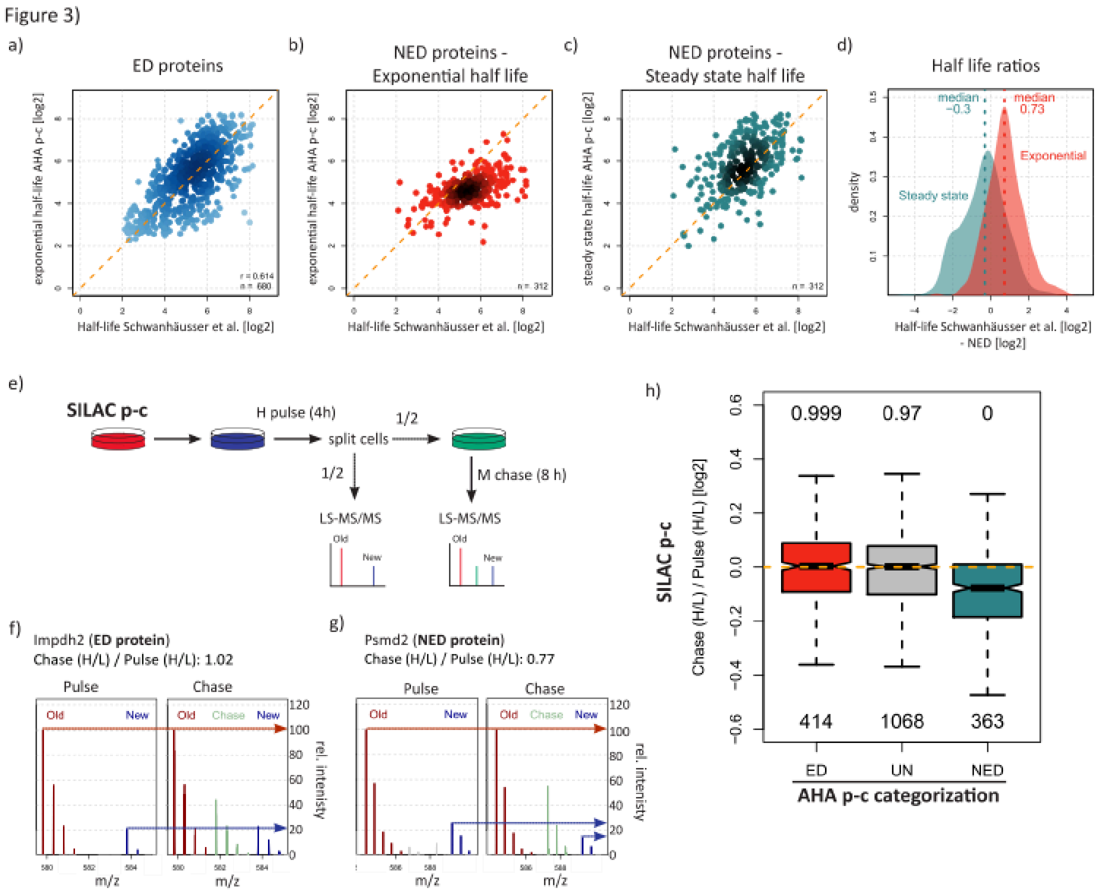


FIGURE 4.3: A-D) Comparison of half-lives measured by AHA p-c to published half-lives measured by dynamic SILAC [72]. A) Half-lives of ED proteins are overall similar in both datasets. B) NED proteins have shorter half-lives in AHA p-c when an exponential fit is used. C-D) Computing "Steady state" half-lives based on the 2-state model improves the agreement with dynamic SILAC data. Proteins with half-lives $>$ 300 h were excluded in A-D because they cannot be accurately quantified. E) Experimental design for direct validation of non-exponential decay. Light (L) cells are pulsed with Heavy (H) SILAC medium for 4 h and split into two populations. The first population is harvested immediately while the second is chased for 8 h in Medium-heavy SILAC medium (M chase). If new proteins are less stable than old proteins their H/L ratio is expected to decrease during the chase. Two example spectra from an ED (F) and an NED protein (G) confirm this expectation. H) All proteins identified in the SILAC p-c experiment were divided into the three populations as defined by AHA p-c. NED proteins show significantly reduced H/L ratios after the chase compared to ED and UN proteins. This holds true independent of the order of labels and length of chase (Supplementary Figure 4.6). P-values from a one sided Kolmogorov-Smirnov test are displayed on top for significantly different distributions ($\alpha = 0.05$).

exponential degradation (Figure 4.3B). This is expected since the exponential fit does not take into account that these proteins become stabilized at later time points. We therefore used our Markov chain-based model to compute "steady state half-lives" [74]. This half-life corresponds to the time it takes for half of the protein molecules present at steady state to be degraded (see Materials and Methods). Importantly, this led to better overall agreement with the dynamic SILAC data (Figure 4.3C and D). Collectively, these findings indicate that the AHA p-c data is reliable and that our modeling approach is meaningful.

Next, we wanted to directly validate our classification of proteins as ED or NED. To this end, we designed a novel SILAC-based strategy (Figure 4.3E and 4.6). First, we pulse labeled light cells with heavy SILAC medium for 4 hours. At this time point, the proteome consists of two populations of protein molecules: Light protein molecules, which are older than 4 hours, and heavy proteins with an age between 0 and 4 hours. We harvested half of the cells to quantify the ratio of both populations at this time point. Our model predicts that NED proteins become more stable when they are older. Consequently, we expect that the heavy to light ratio for these proteins should decrease over time, since the younger heavy proteins should degrade faster than the older light proteins. To test this prediction, we cultivated the other half of the cells on medium-heavy growth medium for 8 additional hours before analyzing them. During this time the heavy and light proteins can only decrease in abundance (barring re-usage of labeled amino acids). When comparing both samples we indeed observed the predicted decrease in the heavy to light ratio for NED but not for ED proteins (Figure 4.3F-H and Supplementary Figure 4.6). Thus, our classification of proteins as ED or NED can be validated by an independent approach that does not depend on non-natural amino acids or protein enrichment.

4.2.4 Non-exponential degradation is mostly mediated by the ubiquitin proteasome system

Having shown that a sizable number of proteins are non-exponentially degraded, we next asked how degradation occurs. The two major cellular protein degradation systems are the proteasome and the lysosome. While the proteasome specifically degrades ubiquitinated proteins, the lysosome can further degrade a wide range of proteins and even entire organelles during autophagy [198]. We therefore used drugs, MG132 or Wortmannin in combination with Bafilomycin A1, to inhibit the proteasome or the lysosome, respectively. To assess non-exponential degradation we performed AHA p-c experiments with 4 and 8 h chase times in the presence of the inhibitor or carrier controls (DMSO). The impact of both inhibitors on NED was quantified by measuring their impact on the Δ-score (Figure 4.4A). Proteasome inhibition reduced NED of most 2-state proteins (Figure 4.4B and Supplementary Figure 4.7A). In contrast, inhibition of the lysosome did not have an observable impact on NED (Figure 4.4C and Supplementary Figure 4.7B). We conclude that the ubiquitin proteasome system is involved in the initial degradation of most NED proteins.

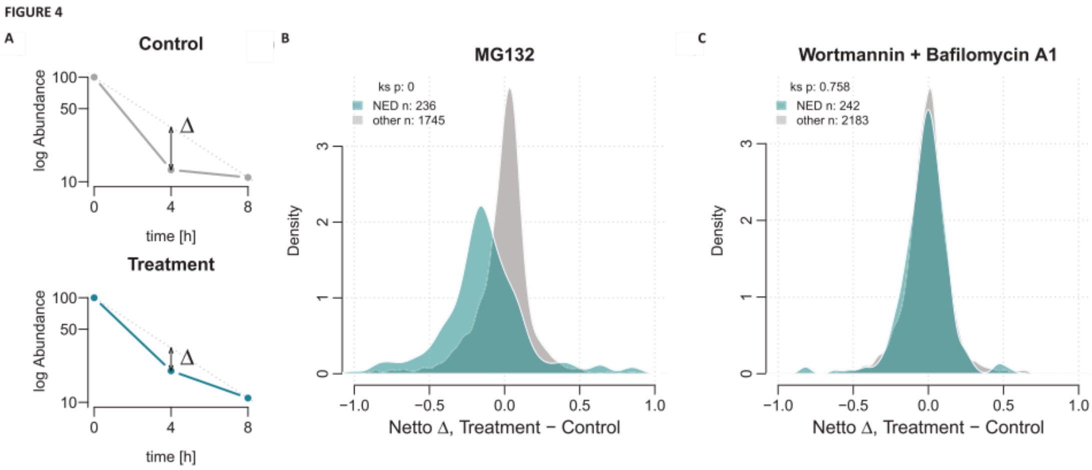


FIGURE 4.4: A) To quantify the impact of inhibitors on non-exponential degradation we compared the Δ-score in treated and control cells. B-C) Distribution of netto Δ-scores for NED proteins (turquoise) and other proteins (grey). The proteasome inhibitor MG132 significantly reduced Δ-scores of NED proteins (B) while the autophagy inhibitors wortmannin and bafilomycin A had no significant impact (C). Controls for inhibitors are displayed in Supplementary Figure 4.7. The numbers of proteins in each group and the p-values derived from a two-sided Kolmogorov-Smirnov test are depicted.

4.2.5 Protein degradation kinetics and complex formation

Next, we sought to characterize features that distinguish ED and NED proteins (Supplementary Figure 4.8, Materials and Methods). This analysis revealed that NED proteins are on average more abundant and have a higher degree of secondary structure. Conversely, ED proteins tend to be more disordered. A particularly interesting finding is that NED proteins are more likely to be members of heteromeric complexes as defined by a recent manually curated database [199] (Supplementary Figure 4.8 and 4.9). To confirm this finding we also mapped our data to published structures of protein complexes (Figure 4.5A, Supplementary Figure 4.10 and 4.11). Again, NED proteins were significantly enriched in complexes relative to ED proteins, with 70% of NED proteins belonging to a heteromeric protein complex.

A long-standing hypothesis is that proteins are stabilized by complex formation [200, 201]. While data for several individual examples supports this idea [202–207] it has, to the best of our knowledge, not yet been investigated systematically. Complex formation could explain non-exponential degradation (Figure 4.5E): If NED proteins were produced in excess relative to ED proteins, the degradation of the non-assembled subunits would lead to non-exponential decay. We therefore analyzed the abundance of newly synthesized proteins directly after the pulse. To compare the relative abundance of proteins in each complex we normalized the data in a complex-centric manner (Figure 4.5F, see Materials and Methods). Indeed, we found that NED proteins are over-synthesized

FIGURE 5

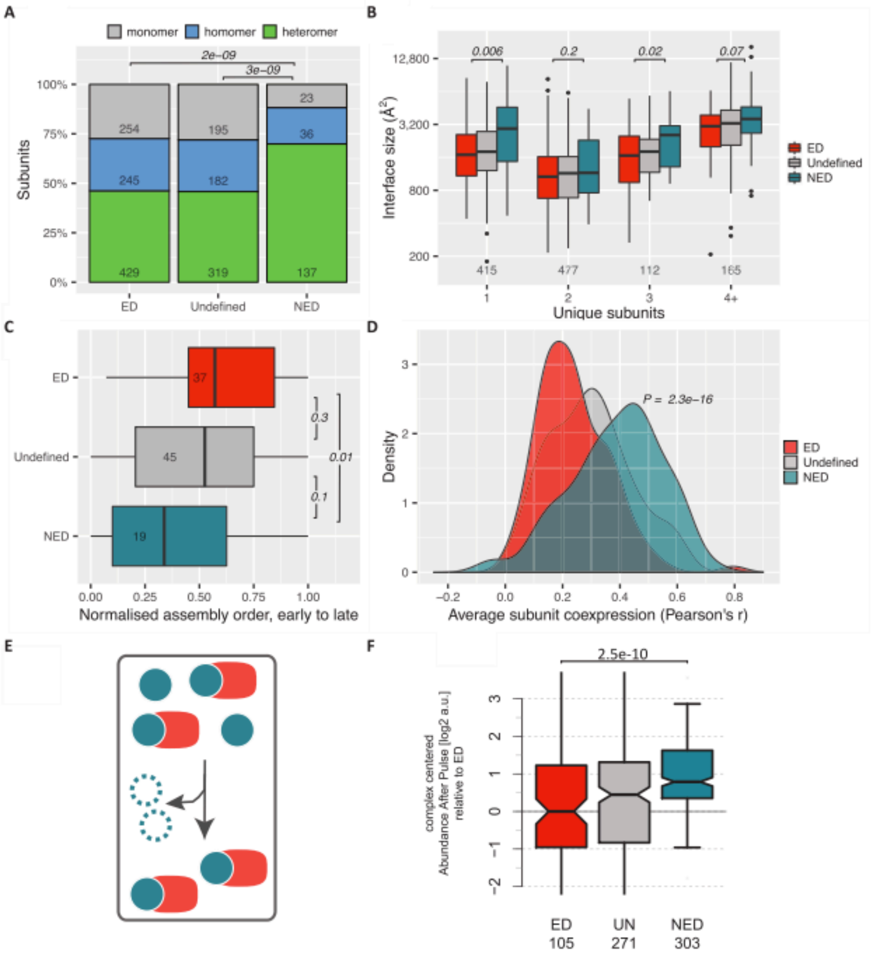


FIGURE 4.5: A) NED proteins are significantly overrepresented in heteromeric protein complexes. P-values (in italics) were calculated using Fisher's exact test, comparing the number of heteromeric subunits to monomeric or homomeric subunits in each decay class. Numbers within each bar represent raw subunit counts. This trend also holds when ribosomal proteins are excluded (Supplementary Figure 4.10). B) NED proteins tend to form larger interfaces in complexes than ED proteins. To control for the fact that NED proteins are also more likely to be members of larger complexes (Supplementary Figure 4.11), we binned subunits by the number of unique subunits per complex. P-values were calculated using Wilcoxon rank-sum tests, comparing NED to ED. Subunit counts in each bin are given along the bottom of the plot. C) NED subunits of large complexes (¿5 unique subunits) tend to assemble earlier than ED subunits. Normalized assembly scores of 0-to-1 indicate the first-to-last steps of a given (dis)assembly pathway. P-values were calculated using Wilcoxon rank-sum tests and raw subunit counts are given within each box. The difference in normalized assembly order was not significant for smaller complexes, although this is in part limited by the size of the dataset. D) NED proteins show stronger coexpression at the mRNA level. For each subunit, the average coexpression correlation coefficient is calculated with all other subunits within the same complex. P-value is calculated with the Wilcoxon rank-sum test comparing NED to ED. This observation also holds when controlling for complex size (Supplementary Figure 4.12-4.13). E) A simple model could explain non-exponential degradation: NED proteins (turquoise) are synthesized in super-stoichiometric amounts relative to ED proteins (red). Only a fraction of the NED protein molecules is stabilized by complex formation whilst the excess is degraded. F) NED proteins tend to be produced in super-stoichiometric amounts. Protein abundances after the pulse (t = 0h) were normalized in a complex centered manner. Boxplot distributions were normalized relative to the median abundance of ED proteins. The displayed p-value was derived from a two-sided Kolmogorov-Smirnov test (ED vs NED distribution).

relative to other members of the same complex. Moreover, while the initial degradation rates (i.e. the degradation rates of state A) of NED proteins within a complex varied considerably, their second (state B) degradation rates were more similar and close to the degradation rates of the ED subunits (Supplementary Figure 4.14). In summary, these data show that many NED proteins are core components of heteromeric protein complexes that are produced in super-stoichiometric amounts relative to their ED counterparts.

4.2.6 Protein degradation profiles are evolutionarily conserved

All data presented so far are based on the analysis of mouse fibroblasts (NIH 3T3). We therefore asked if our findings are due to specific features of this model system. For example, even though NIH 3T3 cells are derived from primary cells, a recent cytogenetic study revealed a complex rearranged karyotype [208]. It is therefore possible that the super-stoichiometric synthesis of NED proteins is due to genomic amplification of the corresponding genes. In this scenario, protein overproduction would simply be a consequence of the rearranged karyotype. In this case, our data would have little relevance for other model systems.

To test this possibility, we analyzed the diploid human retinal pigmented epithelial cell line RPE-1. Low coverage whole genome sequencing of this cell line confirmed that, with the exception of partial trisomies for chromosome 10 and 12, it is mostly diploid (Supplementary Figure 4.15). We performed three independent large AHA p-c experiments to obtain degradation profiles for 4079 proteins (Supplemental Table 2). 47% and 9.4% of the human proteins that passed the quality filters (n=3133) were classified as ED and NED, respectively (Figure 4.6A, left bar). These fractions are overall similar to the mouse fibroblast data. To compare the RPE-1 and the NIH 3T3 data we grouped the human proteins according to the degradation profiles of their mouse orthologues (Figure 4.6A). Human orthologues of mouse NED proteins were highly enriched in NED proteins. Conversely, human orthologues of mouse ED proteins were highly enriched in ED proteins. In addition, the Δ-scores of the human proteins and their mouse orthologues were well correlated (Figure 4.6B). To illustrate this we depict two complexes colored according to their decay profiles in mouse and human (Figure 4.6C-F).

Together, these data show that non-exponential degradation is not mainly due to "erroneous" protein overproduction caused by genomic rearrangements. Instead, our findings show that (i) protein degradation kinetics are conserved between mouse and human and (ii) at least partially independent of the cell type (fibroblasts and epithelial cells).

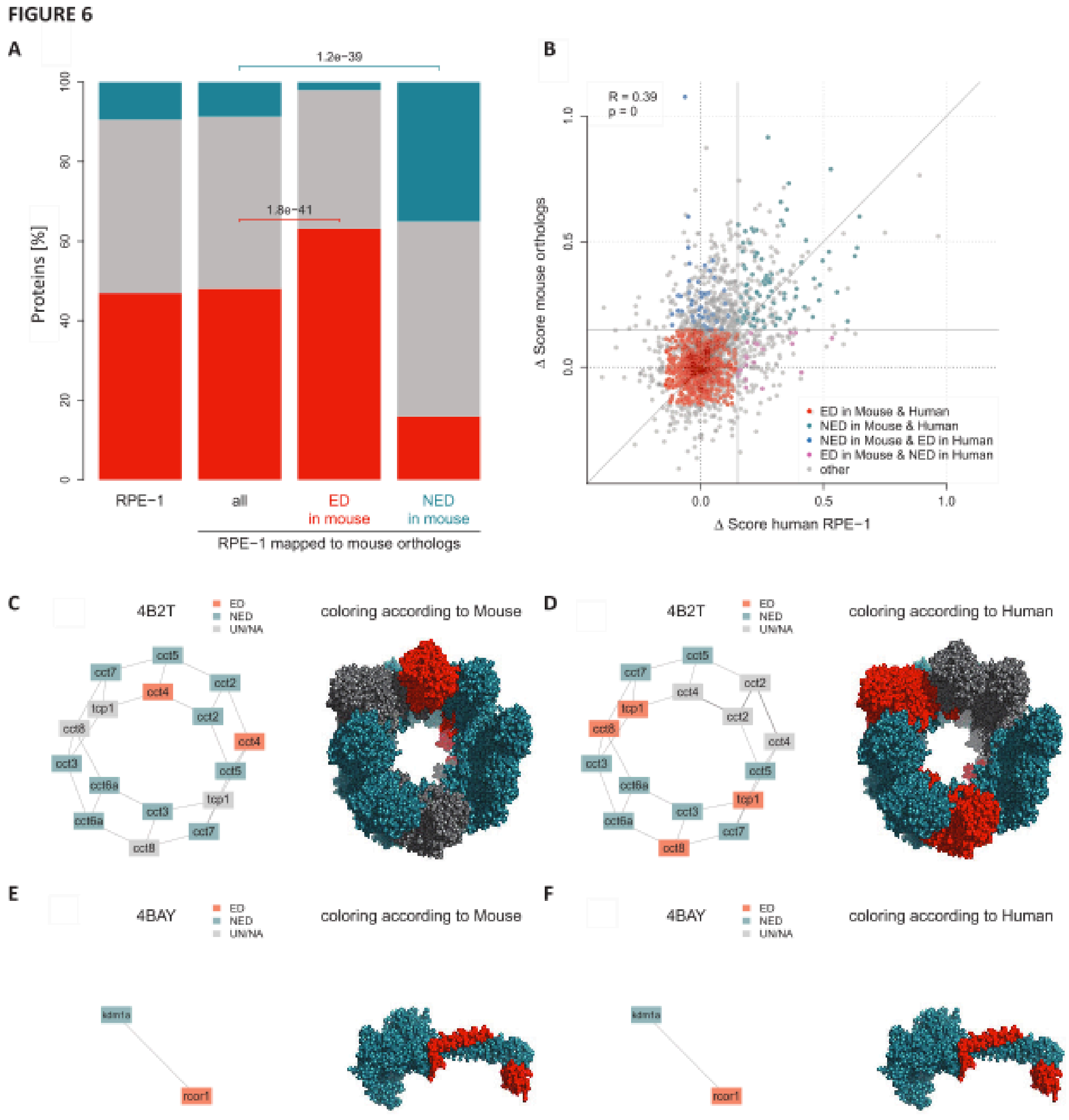


Figure 4.6: A) Relative fractions of NED (turquoise), ED (red) and undefined (grey) proteins in the diploid human epithelial cell line RPE-1. We mapped human proteins to their mouse orthologues and grouped them according to their degradation profile in mouse fibroblasts. Human proteins with ED mouse orthologues are enriched in ED proteins. Similarly, human proteins with NED mouse orthologues are enriched in NED proteins. P-values are based on a hypergeometric test. B) Orthologous human and mouse proteins show highly correlated Δ-scores. Pearson's correlation coefficient (R) is derived from all plotted Δ-score pairs. 3D models of the CCT chaperonin (PDB ID: 4B2T) (C and D) and Kdm1a in a complex with CoREST (PDB ID: 4BAY) (E and F) are shown with coloring based on the degradation profile in mouse (C and E) and in human (D and F).

4.2.7 NED and aneuploidy

Based on our findings we propose a simple model that explains the relationship between protein degradation kinetics and complex formation (Figure 4.7A). According to this model, NED proteins are overproduced relative to the ED proteins in the same complex. Therefore, only a fraction of the overproduced proteins are stabilized by complex formation while the rest are degraded. Importantly, the RPE-1 data show that this overproduction occurs in disomic cells and is thus not generally due to aneuploidy. However, we reasoned that aneuploid cells would allow us to test our model: Genomic amplification of NED proteins should increase their over-production and thus their initial degradation (Figure 4.7A). Consequently, amplification of genes encoding NED proteins should not lead to correspondingly increased protein levels. Thus, NED proteins should be attenuated.

To test this prediction we took advantage of RPE-1 cells that were engineered to carry one additional copy of specific chromosomes [209]. Low coverage genome sequencing verified that these cells are trisomic for chromosome 5 and part of chromosome 11 (Supplementary Figure 4.15). This was further confirmed by chromosome painting (Supplementary Figure 4.16). We therefore went on and performed three independent AHA-pc experiments with these "RPE-1 trisomic" cells (Supplemental Table 3). First, we compared protein production in RPE-1 and RPE-1 trisomic cells using abundance levels after the pulse as a proxy. As expected, proteins encoded by trisomic regions were up-regulated in trisomic cells (Figure 4.7B). Note that chromosome 10 and 12 were also partially trisomic in the parental cell line and therefore excluded from subsequent analyses.

We compared the extent of non-exponential degradation in RPE-1 and RPE-1 trisomic cells using the Δ-score. Proteins encoded in disomic regions of the genome had similar Δ-scores in both cell lines (Figure 4.7C). However, consistent with our prediction, NED proteins in trisomic regions displayed significantly increased non-exponential degradation as measured by the change in their Δ-scores. Importantly, this behavior was specific for NED proteins and not observed for ED proteins. We conclude that increasing the over-production of NED proteins indeed increases their initial degradation.

Aneuploidy has severe developmental effects, it is the leading cause of mental retardation and spontaneous abortions, as well as a hallmark of cancer [210, 211]. However, the functional consequences of aneuploidy are only beginning to emerge. Studies in yeast and mammalian cell lines have shown that mRNA levels generally scale with gene copy numbers. However, protein levels are sometimes attenuated towards the euploid state. It has also been noted that most attenuated proteins are members of multiprotein complexes [209, 212, 213]. Interestingly, however, not all proteins in multiprotein complexes are attenuated. Conversely, not all proteins that are attenuated are part of (known) multiprotein complexes.

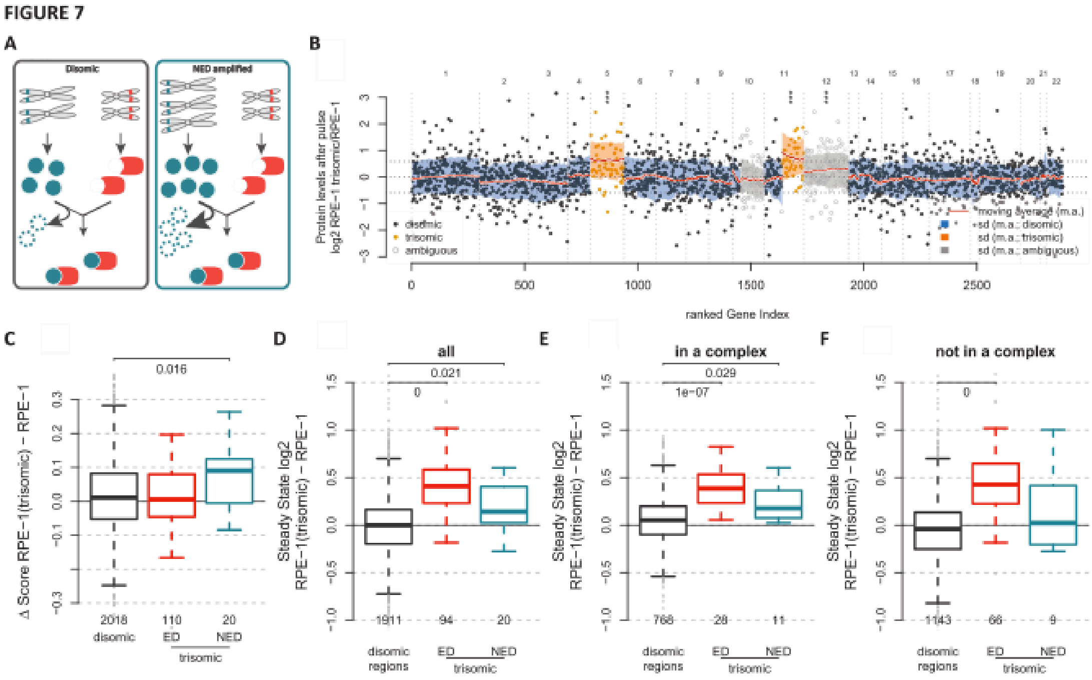


Figure 4.7: A) A model depicting the expected impact of gene amplification on protein synthesis and degradation. In normal (that is, disomic) cells NED proteins are over-synthesized relative to ED proteins in the same complex. Degradation of the excess molecules gives rise to their NED profile. Genomic amplification of NED proteins further increases over-production and thus their initial degradation. B) Log2 fold changes of protein abundances after pulse in RPE-1 cells and RPE-1 cells carrying extra copies of specific chromosomes (sorted by chromosome and genomic position). The data was divided into disomic (black), trisomic (orange) and ambiguous regions (grey) based on genome sequencing data (Supplementary Figure 4.15). Moving averages for each chromosome display the trend of local chromosome expression (red line). Shaded areas show the corresponding moving standard deviations for disomic (blue), trisomic (orange) and ambiguous regions (grey). Chromosomes with significantly different protein abundance in comparison to the diploid chromosome set are marked with asterisks (Kolmogorov-Smirnov test, $\alpha = 0.01$; ***: $p < 0.0001$). C) NED proteins show increased initial degradation (Δ-scores) when the corresponding genes are in trisomic regions. Degradation of ED proteins is not affected. D-F) NED predicts protein level attenuation. We compared steady-state protein levels in RPE-1 and RPE-1 trisomic cells using standard SILAC. Boxplots show log2 fold changes for ED and NED proteins in trisomic regions and compared to proteins from disomic regions. This analysis is shown for all proteins (D), proteins that are part of complexes (E) and proteins that are not part of complexes (F). In all cases, NED proteins show stronger attenuation than ED proteins. The number of analyzable protein pairs is displayed below each boxplot. All distributions were tested for equality against the corresponding disomic protein subset using the Kolmogorov-Smirnov test. P-values are displayed on top for significantly different distributions ($\alpha = 0.05$).

Our model predicts that NED proteins should be attenuated. To test this idea, we quantified relative changes in steady-state protein levels in RPE-1 and RPE-1 trisomic cells using standard SILAC. For proteins encoded in trisomic regions we then compared relative changes in protein levels (Figure 4.7D). Consistent with our model, we found that NED proteins were more attenuated than ED proteins. Importantly, we also observed this effect in the subset of proteins that are part of annotated complexes (Figure 4.7E). Hence, protein degradation kinetics can explain why some proteins in complexes show more attenuation (NED proteins) than others (ED proteins). Moreover, even for the subset of proteins that are not part of an annotated complex, NED proteins showed more attenuation than ED proteins (Figure 4.7F). Collectively, these data show that NED is overall a better predictor of protein level attenuation than complex formation.

4.3 Discussion

Quantifying the kinetics of cellular protein degradation is key to understanding the protein life cycle. So far, available methods only allowed analysis of either individual proteins or proteins in bulk. Here, we developed AHA-pc as a method to quantify protein degradation kinetics globally. We provide the first proteome-wide survey of decay profiles. Using a Markov chain-based modeling approach we find that over 10% of all proteins are most parsimoniously described by a 2-state model, which implies non-exponential degradation. Interestingly, all NED proteins become more stable over time. We find that protein degradation kinetics are similar in two different cell types and conserved between mouse and humans. More than half of the NED proteins are subunits of annotated heteromeric complexes. These subunits tend to be produced in super-stoichiometric amounts relative to other members of the same complex. Based on these findings, we propose a simple model where only a fraction of the newly synthesized NED proteins is stabilized by complex formation while the rest is degraded. Consistent with this model, we observe that up-regulating NED proteins increases their initial degradation. Finally, we show that our data predicts changes in steady-state protein levels during aneuploidy.

While the AHA-pc method has many advantages, it is also important to keep its limitations in mind. For example, due to the technical challenges of AHA-pc our data is less comprehensive and only covers ∼5000 proteins. We do not know how our findings extrapolate to the uncovered part of the proteome. Also, the pulse time of 1 h is too long to capture events on the timescale of minutes. Therefore, our data cannot be used to estimate the extent of co- or peri-translational degradation [7, 181, 214]. Moreover, since our longest chase time is 32 h, we cannot tell what happens to older proteins. It is possible that longer chase times would reveal age-dependent destabilization. Additionally, the seven time points we covered allowed us to distinguish between a 1-state and a 2-state model, more complex multi-stage models would require measurements at more time points. Increasing the number of data points per profile would also decrease the number of proteins we could not classify. Finally, even though we validated our data with independent methods, we cannot rule out that AHA incorporation might affect degradation of specific proteins.

One of our most surprising findings is that mammalian cells seem to overproduce specific subunits of multiprotein complexes. This is in marked contrast to E. coli where subunits were reported to be made in precise proportion to their stoichiometry [68]. Why are mammalian cells producing excess amounts of specific proteins? We would like to discuss four potential answers to this question:

1. It is possible that only a fraction of newly synthesized proteins adopt a functional state while the rest are terminally misfolded and thus degraded [215]. However, since NED proteins tend to be relatively short and well-structured (Supplementary Figure 4.8) this explanation does not seem to generally hold.

2. The observation that that ED proteins tend to be disordered (Supplementary Figure 4.8) suggests an alternative explanation: It has been shown that disordered proteins are often harmful when overexpressed, which is probably due to their tendency to make promiscuous molecular interactions [216]. Overproduction of NED proteins may thus have evolved to ensure that potentially harmful ED proteins are never alone: The super-stoichiometric synthesis of "benign" NED subunits at baseline would be a failsafe mechanism against deleterious effects of unbalanced production of the more harmful ED proteins. Accordingly, NED proteins would have a chaperone-like function towards their cognate ED proteins.

3. The third possibility is that ED proteins evolved as limiting factors to facilitate the coordinated regulation of protein complex abundance: Up-regulating an ED protein would stabilize interacting NED proteins and thus increase the abundance of the entire complex. Conversely, reducing the expression of the limiting ED protein would decrease complex abundance. This explanation resonates well with the finding that mRNAs encoding ED proteins show less co-expression with mRNAs encoding other subunits and have longer 5' and 3' UTRs (Figure 4.5D and data not shown). It also fits to data from yeast that most complexes consist of both constitutive and periodically expressed subunits [217] and to the finding that complex stoichiometry can vary across tissues in mammals [199].

4. Overproduction of NED proteins may be important for ordered complex assembly [218, 219]: Since the formation of protein-protein interactions depends on protein concentration, the relative abundance of multimeric complex may in part determine the assembly order. This explanation is consistent with our observation that NED proteins tend to assemble earlier (Figure 4.5C).

Which (if any) of these four possibilities explains protein overproduction and NED remains to be investigated. It is likely that different reasons are relevant for different proteins. It also important to note that not all NED proteins are components of (known) multiprotein complexes. NED of monomers could be due to many different molecular mechanisms. For example, NED of CFTR and basigin appears to be due to failed protein folding in the ER [177, 178]. Whatever the case may be, our observation that proteins show consistent behavior between mouse fibroblasts and human cells suggests that protein overproduction and NED is a conserved feature of mammalian cells and thus functionally important.

Our results have significant implications for aneuploidy. First, we confirm the previous observation that amplified genes encoding members of multiprotein complexes are often attenuated at the protein level [209, 212, 213]. This finding was interpreted in the light of the longstanding idea that unassembled subunits of complexes are unstable [200, 201]. Accordingly, overproduction caused by gene amplification is attenuated for proteins in complexes. Our findings considerably extend this concept. The difference is that we already observe unbalanced subunit production at baseline (Figure 4.5F). Consequently, our model can explain why only some subunits of complexes show attenuation (Figure 4.7E). More broadly, our data helps to understand how protein levels change in response to altered gene dosage. We expect that this will turn out to be useful for understanding the complex cellular phenotypes of aneuploidy and other types of gene copy number changes. For example, our data might help in understanding the functional significance of somatic copy number alterations in cancer.

4.4 Acknowledgements

We thank Katrin Eichelbaum for help with establishing AHA labeling, Piotr Grabowski for setting up SCX chromatography and Koshi Imami (all MDC) for establishing RP-HPLC on monolithic columns. Yasushi Ishihama (Kyoto University) kindly provided monolithic column (GL Sciences, Japan). We also thank Thomas Sommer (MDC) and his lab members for fruitful discussions. Christian Sommer and Eva Kärgel provided excellent technical assistance. This work was in part funded by a grant of the Helmholtz Association to M.S. and by a grant of the European Union (ITN “NICHE”) to C.S. and A.V.. J.M. is supported by a Medical Research Council Career Development Award (MR/M02122X/1).

5

Degradation Parameters from Pulse-Chase Experiments

A version of this chapter has been published as:
Sin C, Chiarugi D, Valleriani A (2016) Degradation Parameters from Pulse-Chase Experiments. PLoS ONE 11(5): e0155028.

Online version:
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155028

CONTRIBUTIONS: All authors took part in conceiving the project, developing the experiment and writing the manuscript. I extracted the data from Wheatley, et. al., 1980. I preformed the data analysis, parameter estimation and sensitivity analysis on the system.

Abstract

Pulse-chase experiments are often used to study the degradation of macromolecules such as proteins or mRNA. Considerations for the choice of pulse length include the toxicity of the pulse to the cell and maximization of labeling. In the general case of non-exponential decay, varying the length of the pulse results in decay patterns that look different. Analysis of these patterns without consideration to pulse length would yield incorrect degradation parameters. Here we propose a method that constructively includes pulse length in the analysis of decay patterns and extracts the parameters of the underlying degradation process. We also show how to extract decay parameters reliably from measurements taken during the pulse phase.

5.1 Introduction

The degradation of macromolecules such as mRNA and proteins is a complex biochemical process that is usually carried out by a series of subsequent biochemical reactions. Various experimental techniques can be used to study the decay of macromolecules over time; one of the most widespread methods to measure decay is pulse-chase experiments. A pulse-chase experiment consists of two phases: first, in the pulse phase, cells are exposed to a labeling compound that is integrated into newly synthesized macromolecules of interest. For example, if one wants to follow the fate of proteins or mRNA radio or isotope labeled amino acids[185, 220–223] or nucleotides[224, 225] can be introduced in the culture medium. In the second phase, the chase phase, the same compound in the unlabeled form is added in excess, replacing the labeled form. Any macromolecules synthesized during the chase phase will not be labeled. Throughout the chase, the amount of macromolecules that were synthesized during the pulse diminish due to degradation processes (**Fig 5.1a**). Tracking their amount over time delivers a curve that we call the *decay pattern*. Alternative methods, such as stopping synthesis of the macromolecule and measuring the amount left over time, also provide decay patterns. For example, chloramphenicol is often used to stop protein synthesis in bacterial cells. However, such methods may be very stressful for the cell, potentially resulting in measurements not indicative of normal cell conditions. Ultimately, the process of stopping synthesis corresponds in general to a pulse of infinite duration. From the point of view of data analysis, decay patterns obtained upon stopping the synthesis are a limit case of patterns generated with pulse-chase experiments.

Decay patterns are said to be exponential, if they look like straight lines in a plot linear in time and logarithmic in the (relative) abundance of the molecule. This makes

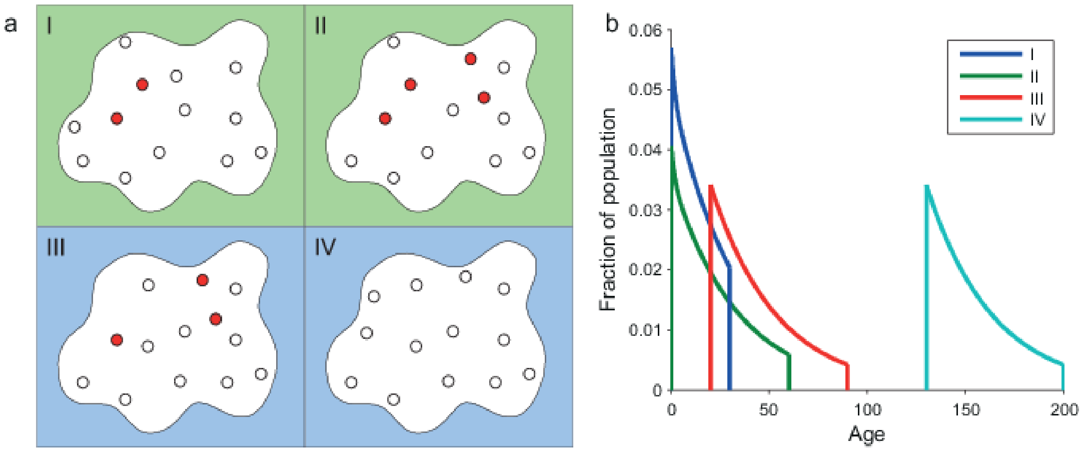


FIGURE 5.1: **Age distribution of molecules during pulse chase experiments** **(a)** Depiction of molecules in a cell during a pulse chase experiment with pulse duration of 70 in arbitrary units (a.u.). For the purpose of illustration we show four snapshots of the experiment. In snapshot I (30 a.u.), pulse has just begun. The white dots depict the population of molecules already existing in the cell before the pulse. All newly synthesized molecules (red dots) are labeled by the pulse and measurable by the experimentalist. As the pulse continues in snapshot II (60 a.u.) we see more labeled molecules appear. Meanwhile, both labeled and unlabeled molecules degrade. In snapshot III (90 a.u.), the pulse has ended since some time. Newly synthesized molecules from this moment on are unlabeled. Again, both labeled and unlabeled molecules degrade. In snapshot IV (200 a.u.), all labeled molecules have degraded. Unlabeled molecules continue to be synthesized and degraded. **(b)** Age distribution of the labeled molecules, each curve corresponds to one phase in panel (a). In snapshot I, the pulse has just begun, and all molecules that are labeled by the pulse are no older than the time elapsed since the pulse has begun. In snapshot II, the pulse has been applied for some time; some labeled molecules may be quite old. In snapshot III the molecules cannot be younger than the time elapsed since pulse has been stopped. By snapshot IV, if there were molecules left, they would have that age distribution.

fitting a single exponential to the decay pattern especially convenient – so much so that it is often used as the default procedure for analysis of decay curves. Unfortunately this procedure is often overused; one can tell when a quick look at the plot clearly indicates that the pattern is not as straight as it ought to be. As a way out of this cumbersome situation, many research studies use the half-time as a measure of the stability of the molecule because it can be estimated even without any fit of the data. For exponentially decaying molecules, there is a simple algebraic relationship between half-time and average lifetime of the molecules, but this relationship does not hold for non-exponentially decaying molecules. Thus, half-time is not a good measure of stability for non-exponential decay[74]. A further complication is that apparently, for some decay patterns, the pulse length affects the half-time [7]. This happens because the half-time is measured by simply extracting it from the data without taking pulse length into account. Thus, the half-time of the resultant population is actually not a quantity representative of the molecule's decay alone, but also of the pulse length. The choice of the function used for the analysis of decay curves makes critical assumptions about the biochemistry of the degradation process at the single molecule level. As we shall see, the choice of the exponential function may lead to achieve incorrect information about the underlying biochemical processes and the stability of the macromolecules.

Another common approach to analyze non-exponential (sometimes called nonlinear) decay patterns is to fit them with a double exponential. Although this certainly improves the fit, it is not yet immediately obvious how the two characteristic timescales from the two exponentials depend on the length of the pulse or on other possible choices of the experimental design.

One fundamental question is: What is the process behind the decay pattern? We know that behind the decay pattern there is a degradation process. When we think about the degradation process, we think at the level of single molecules. Namely, we think at the processes (biochemical networks, perhaps competing against each other) that will eventually lead to the disappearance of the molecule. Taking into account that we are talking about biochemical interactions, the fate of each single molecule can be described quantitatively only in probabilistic terms, with the basic quantity of this description being the probability distribution of the molecule's lifetime. The decay pattern, however, is not a single molecule measurement. Rather, it is the result of billions of molecules, each randomly decaying during and after the pulse.

When pulse-chase experiments are used to follow the macromolecules degradation, a bias may be introduced into the measurements through the pulse length. In this paper we will show that the length of the pulse affects the shape of the decay patterns. If this effect is not taken into consideration, the biochemical information we extract from the decay pattern would be biased. Unfortunately, this effect is usually not considered when choosing the pulse length. The connection between the single molecule perspective and the decay pattern requires an abstract view of the degradation process. Here we present an approach that allows us to correctly bridge the gap between the single molecule perspective and the population level for a wide range of decay patterns. Notably, our method provides a simple recipe that allows us to incorporate the pulse length in the fitting procedure when decay patterns are derived from pulse-chase experiments. The method does not rely on any *a priori* knowledge of the lifetime of the molecule and can be used independently of whether the pulse length is short or long compared to the lifetime of the molecules. First, we recapitulate the general theory providing a new derivation of the required mathematics and then we provide a simple recipe to incorporate the pulse length in the fitting procedure. Our results should be applied to refine the protocols for the design of decay assay experiments.

5.1.1 Aging of molecules

It is rather strange that an apparently simple concept like aging becomes so intriguing when applied to molecules. Molecules, at least complex molecules like proteins and mRNA, age in a fashion which is very similar to what we know based on our daily experience. Damage, shortening or lengthening, the attachment of other molecules (e.g. miRISC complex that attaches to the mRNA[15, 77]) are common phenomena that make the target molecule older. In fact, some processes could even make the target molecule younger, e.g. when damage is repaired or a bound molecule detaches. What does it actually mean when a molecule becomes older, or ages? How can we characterize this phenomenon? Acutally, what best characterizes the effect of aging is not how much time has passed since birth or synthesis but the amount of time left until the end of life.

Let us look at this matter closely. Suppose for a moment that the lifetime distribution of an hypothetical molecule is exponential – as we shall see later, this implies that also the decay pattern is exponential – and we know that in this precise moment the age of the molecule is a, *i.e.* this molecule has been synthesized a time units ago. What is the distribution of its residual lifetime? As a consequence of the assumption of an exponential lifetime distribution, the answer is that the residual lifetime is independent of a and is the same as if the molecule had been synthesized right now. This answer means that the molecule does not age. However, if we know or suspect that molecules like mRNA and proteins must undergo a series of biochemical reactions[226–228] until we consider them as degraded, each step in the biochemical reaction network makes the molecule older, *i.e.* closer to its final end. When the molecule ages in the "complex" way just described, its residual lifetime becomes shorter. Therefore, its lifetime distribution cannot be exponential and so also the decay pattern cannot be an exponential.

Once we accept that molecules age — *i.e.* do not have an exponential lifetime distribution – there is another effect of aging related to the duration of the pulse. Imagine a pulse of a very short duration. The molecules synthesized during the pulse will all have quite the same residual lifetime, since they are likely to be all in exactly the same biochemical state. When the pulse becomes longer, some of the molecules synthesized at the beginning of the pulse may have been already degraded, some are definitively older but still there, and others will be just newly synthesized. In short, we have a mixture of ages in the population, and the composition of the population depends on the pulse length (**Fig 5.1b**). This is an important point for what follows: the initial condition at the time point of chase depends on the mixture of ages in the population of molecules, which depends on the length of the pulse. Therefore, fitting the decay pattern requires a knowledge of the age distribution at the beginning of the chase, which can be computed only if one knows the effect of the pulse on the age distribution. The effect of the pulse on the age distribution can only be known if one has a good model to describe how the molecules age, and has already calibrated the model. A priori, it is not possible to know if a pulse of a given length is long or short compared to the lifetime of the molecules, but as we shall see, this apparent circularity can be solved to provide a unique formula that takes pulse, chase and aging into account.

5.1.2 Degradation processes modeled with Markov chains

Macromolecules are degraded through a number of different pathways. While some can be approximated as exponentially decaying, many of these pathways are multi-step pathways with several complex biochemical stages [185, 226–229]. The biochemical stages are connected to each other to form a network of biochemical states [74, 230–232]. When we describe the dynamics of degradation of a single molecule, we think of this molecule as moving on this network of biochemical states in a stochastic fashion until degradation eventually happens. We will use the term "single step degradation" for pathways with only one measurable rate limiting step (*i.e.* exponentially decaying), and the term "multi-step degradation" for more complex pathways, which necessarily includes the "single step" degradation as a limiting case[74]. The rates of the transitions between the biochemical states would depend on the concentration of the ligands and on their affinity to the target molecule, details rarely available on a large scale. It is therefore convenient to model this process as a Markov chain, which is a simple and mathematically treatable tool to describe stochastic process on discrete states[1, 74, 86, 120, 230, 231]. From this we can model the decay as a single molecule stochastic process, and then solve the equations to derive the lifetime distribution, the age-dependent degradation rate, and the steady state distribution of the fraction of molecules in each state.

Methods

If T is the random variable describing the lifetime of a molecule, we define $f_T(t)$ its probability density function. This definition means that

$$\Pr\{T \leq t\} = \int_0^t f_T(u)\mathrm{d}u \equiv F_T(t)\,, \tag{5.1}$$

with F_T being the probability distribution. For later use, we state here also the equation for the average lifetime

$$\overline{T} = \int_0^\infty u f_T(u)\mathrm{d}u = \int_0^\infty (1 - F_T(u))\,\mathrm{d}u\,. \tag{5.2}$$

The age-dependent degradation rate $\delta(a)$ is defined as the probability that a molecule of age a is degraded in the next unit of time. Formally, it is related to the lifetime probability density f_T by means of a relationship known in the literature as the definition of hazard rate:

$$\delta(a) \;=\; \frac{f_T(a)}{1 - F_T(a)}\,, \tag{5.3}$$

which is a tool commonly employed in survival analysis[233]. When T is exponentially distributed with rate μ, then $\delta(a)$ is a constant equal to μ. When, however, f_T is not exponential then $\delta(a)$ can take many different forms, the most common of which are the monotonously decreasing to a constant larger than zero (molecules stabilize with age) and increasing (molecules destabilize with age). The results presented in this manuscript concern with the derivation of f_T from a model and the fit of this model with the available data, *i.e.* the decay patterns. In some cases, however, it is possible to first formulate a specific form of the age-dependent degradation rate $\delta(a)$, based on some specific model of degradation, and then derive the lifetime density as

$$f_T(t) \;=\; \delta(a)\exp\left(-\int_0^t \delta(a)\mathrm{d}a\right)\,, \tag{5.4}$$

and its probability distribution as

$$F_T(t) \;=\; 1 \;-\; \exp\left(-\int_0^t \delta(a)\mathrm{d}a\right)\,. \tag{5.5}$$

An example of a derivation that starts with the hazard rate was given in Ref. [74].

5.1.3 Decay after a pulse

Consider a system that starts with zero molecules, so that $N(0) = 0$. Through the process of synthesis with constant rate and through a generic complex degradation process, the average number of molecules $N(t)$ will first increase in time and eventually reach a steady state. To derive the full pattern $N(t)$ consider dividing the time t into k small intervals each of duration τ so that $k\tau = t$. Given a constant synthesis rate ω, we have that $N(\tau) = \omega\tau(1 - F_T(\tau))$. At the next time interval, we have that $N(2\tau) = \omega\tau(1-F_T(2\tau))+\omega\tau(1-F_T(\tau))$, where the first term gives the probability that the molecules synthesized in the first interval survive until the end of the second interval. Proceeding in this way one can verify that $N(k\tau) = N(k\tau-\tau)+\omega\tau(1-F_T(k\tau))$. Letting now $k \to \infty$ and $\tau \to 0$ while keeping $k\tau = t$, leads to the differential equation

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} \;=\; \omega\,(1 - F_T(t))\;, \tag{5.6}$$

from which it follows that

$$N(t) \;=\; \omega \int_0^t (1 - F_{\scriptscriptstyle T}(u))\, \mathrm{d}u\,. \tag{5.7}$$

$N(t)$ is thus the amount of molecules present at time t when their origination time was between 0 and t. Therefore, if pulse or synthesis is not stopped, when t goes to infinity the average number of molecules reaches a steady state value where synthesis and degradation balance each other.

At any time, the number of new molecules delivered after an interval Δt of synthesis is thus given by $N(\Delta t)$. Consider now the quantity $N(t + \Delta t) - N(t)$. If synthesis works normally, this quantity must be non negative and has two contributions: the molecules delivered at time t that survived until $t + \Delta t$ plus the newly synthesized molecules that have been delivered after an interval Δt, which is given by $N(\Delta t)$. Therefore, if $N(t)$ is the number of molecules at time t, after an interval Δt the number at $t + \Delta t$ will be $N(t) + (N(t + \Delta t) - N(t)) = N(t + \Delta t)$. If synthesis is stopped after a pulse of length t_p, the number of labeled molecules that can be found at t_p is given by $N(t_p)$. Let us define $N'(t_p + \Delta t)$ as the number of molecules at $t_p + \Delta t$, if we stop synthesis at t_p. This quantity is different from $N(t_p + \Delta t)$, which is the number of molecules one would have had if labeling had not stopped at time t_p. Thus, $N'(t_p + \Delta t)$ is given by

$$N'(t_p + \Delta t) \;=\; N(t_p) + [(N(t_p + \Delta t) - N(t_p)) - N(\Delta t)]\;, \tag{5.8}$$

because the molecules $N(\Delta t)$ that would have been delivered are simply missing. During the chase phase, thus, the relative amount of labeled molecules must decrease according to

$$C(t_p + \Delta t) \;=\; \frac{N'(t_p + \Delta t)}{N(t_p)} \;=\; \frac{N(t_p + \Delta t) - N(\Delta t)}{N(t_p)}. \tag{5.9}$$

If labeling is stopped at steady state, the relative amount of labeled molecules is obtained by Eq. (5.9) by setting $t_p \to \infty$. An explicit formulation of Eq. (5.9) is finally obtained substituting Eq. (5.7)

$$C(\Delta t) \;=\; \frac{1}{N(t_p)} \int_{\Delta t}^{t_p + \Delta t} (1 - F_{\scriptscriptstyle T}(u))\, \mathrm{d}u\,, \tag{5.10}$$

where the denominator just ensures that $C(0) = 1$ and Δt represents the time of measurement after the end of the pulse. If $f_{\scriptscriptstyle T}$ is an exponential probability density, $C(\Delta t)$ decays also exponentially with the same rate independently of the pulse length.

5.1.4 Data analysis

Once F_T is known, the decay pattern during the chase phase is given by $C(\Delta t)$ from Eq. (5.9). In most of the cases, however, only the *experimental* values of $\tilde{C}$ at different time points $\Delta t = \Delta_1, \Delta_2, \ldots, \Delta_n$ are available. Thus one defines a parametric form of F_T and uses the data to fit the parameters that minimize the function

$$M = \sum_{j=1}^{n} \Big(\ln(C(\Delta_j)) - \ln(\tilde{C}(\Delta_j))\Big)^2 , \tag{5.11}$$

the reason for the log being that one can generally assume a lognormal distribution of the measurement error. When data from m experiments on the same molecule using different pulse durations $t_p = t_1, \ldots, t_m$ are available, the function to minimize becomes

$$M = \sum_{p=1}^{m} \sum_{j=1}^{n} \Big(\ln(C(\Delta_j)) - \ln(\tilde{C}(\Delta_j))\Big)^2 , \tag{5.12}$$

where in principle, the Δ_j might also be different from one experiment with one pulse length and another experiment with a different pulse length.

Model calibration (parameter estimation)

We use several functions in MATLAB® to calibrate the models. To minimize the objective function, we use fmincon (bounds: $\kappa_{10}, \kappa_{12}, \kappa_{20} \in [0.000001, 1]$). We also used GlobalSearch (with the default settings) and MultiStart (1000 start points) to better sample the available parameter space.

5.2 Results and Discussion

Most mRNA and protein decay patterns can be fit with one of two simple models: A two-stage model (**Fig 5.2a**) always results in a better fit than a one-stage (exponential decay) model, but many patterns are sufficiently described by the simpler one-stage model. The decision of whether a two-stage model or the exponential fit should be adopted depends on the balance between accuracy of the fit and number of parameters. The Akaike Information Criterion (AIC) is one such measure that can be used to select the better model [6]. In this paper, we concentrate on the use of the two-stage model. The motivation for this is two fold – firstly, for decay patterns from one-stage models (exponential decay), one does not need to take the pulse length into account. The decay curve of the system with exponential decay is independent of the pulse. Secondly, in the examples that we will later present, the one-stage model does not adequately capture the dynamics of the system (**Fig 5.2b**). Through the AIC, we find that one-stage models are too simplistic, and three stage models are not adequately identifiable. For other systems with degradation patterns not well described by a two-stage Markov state model, Eq. (5.10) can be used to derive the appropriate expression. The two-stage model applied to the data of Ref. [7] has an RSS of 0.0011 (**Fig. 5.2b**). To evaluate the goodness of fit, we apply the chi-squared test with three d.o.f. and find that we can not reject the null hypothesis which is that the two-stage model is a good fit (p-value $< 10^{-5}$).

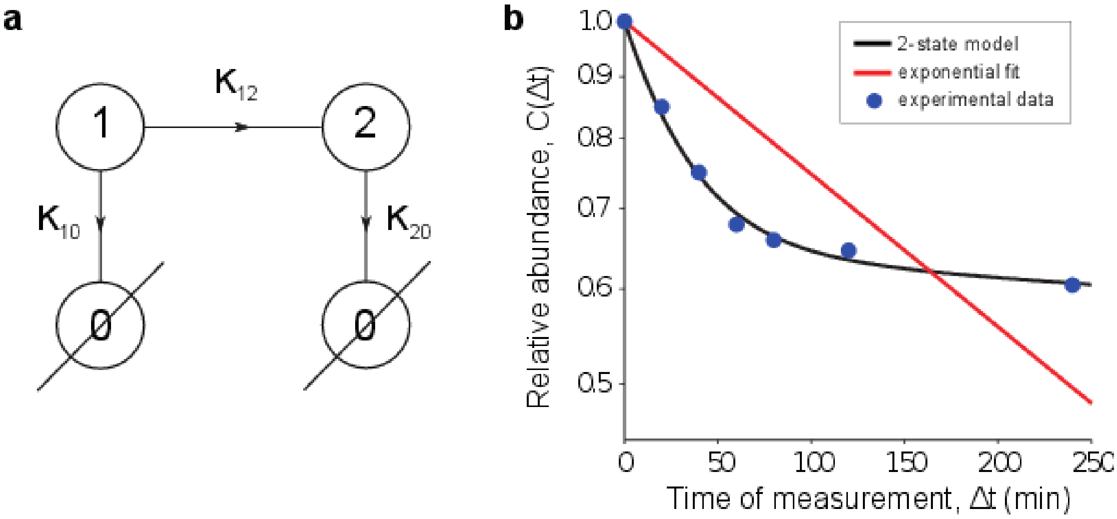


FIGURE 5.2: **Markov chain representation of a Markov process and 2-state model fit to a decay curve (a)** Markov chain representation of a degradation process. Biochemical pathways (such as degradation) can be readily translated into Markov chain models: each biochemical entity is represented as a Markov state (circles) and the reaction speeds are represented as fluxes between the states (arrows). Here we show a possible Markov model of degradation containing two states. Newly synthesized molecules are in state 1. From state 1, there are two possible paths, either to state 2 with rate κ_{12} or degraded with rate κ_{10}. For those molecules that reach state 2, they are degraded with rate κ_{20}. The rates κ_{10} and κ_{20} are necessarily different, otherwise the model collapses into a 1 state model. **(b)** 2-state model fit (black line) and exponential fit (red line) to sample data with 1 minute pulse (data from [7], blue spots). Note that in the log(abundance)-linear(time) scale, the data does not resemble a straight line, thus necessitating a model more complicated than a single exponential. The best fit using Eq. (5.11) with $C(\Delta t)$ from Eq. (5.16) gives the following parameters: $\kappa_{10} = 0.0109$ min^{-1}, $\kappa_{20} = 0.002$ min^{-1}, $\kappa_{12} = 0.0189$ min^{-1}, and pulse $= 1$ min. $\kappa_{exp} = 0.0029$. The decision in favor of the two-stage model is made on the basis of the AIC criterion thanks to its very small RSS.

The general mathematical methodology to describe the decay (see *Methods*) requires a specific form of the lifetime distribution in order to become practically useful. Ideas from biochemical networks indicate that a Markov chain is a useful and flexible framework to develop the basic models of molecular degradation [1, 74]. In terms of a Markov chain the two-stage model consists of two states, say state 1 and state 2, and a degraded state called state 0 connected to each other (**Fig 5.2a**). The rates κ_{12}, κ_{10}, and κ_{20} govern the transitions from state 1 to state 2, from state 1 to state 0, and from state 2 to state 0, respectively. In addition, the rates κ_{10}, and κ_{20} are composed of the degradation rates from each state and the basal dilution rate due to cell division and population growth; the basal dilution rate can have a negligible effect if the timescale of the experiment is shorter than the doubling time of the cell culture.

Using this network we model the life of a single molecule as follows. The molecule is synthesized to be in biochemical state 1 (or it moves very quickly through a series of biochemical steps until it reaches a biochemical state that we call state 1). The molecule dwells for a random amount of time in state 1. The average amount of time spent in state 1 is $\tau_1 = 1/(\kappa_{10} + \kappa_{12})$. The molecule then leaves state 1 and is either degraded, *i.e.* it jumps to state 0 with probability $P_{10} = \tau_1\kappa_{10}$, or it jumps into biochemical state 2 with probability $P_{12} = \tau_1\kappa_{12}$. If the molecule attains state 2, it will dwell in this state for a random amount of time with average $\tau_2 = 1/\kappa_{20}$ until it is eventually degraded (*i.e.* $P_{20} = 1$). The two-stage model does not mean that there are only two biochemical states for a molecule, it says that all other states are visited very quickly and that, at the end, two compound states are sufficient to describe the decay pattern for that molecule.

The lifetime of the molecule is now just the random time required to move from state 1 to state 0, taking into account the possibility to visit state 2 before degradation. Technically, this is the absorption time in 0 starting from state 1 and its probability density is the lifetime probability density $f_T(t)$. The equivalence of lifetime and absorption time is useful because there are plenty of mathematical tools to compute the probability density of the time to absorption. Once the probability density $f_T(t)$ is computed in terms of the (yet unknown) rates κ_{10}, κ_{12} and κ_{20}, and of the pulse length t_p, it can be used to compute the relative abundance $C(\Delta t)$ (Eq. 5.10). Fitting the data will then finally deliver the appropriate values of the rates (**Fig 5.2b**).

The beauty of the two-stage model is that finding the distribution of the lifetime T is simple. If $\rho_1(t)$ and $\rho_2(t)$ are the dwell time distributions on states 1 and 2, respectively, then it results that

$$F_T(t) \;=\; P_{10}\int_0^t \rho_1(\tau)\mathrm{d}\tau + P_{12}\int_0^t \rho_1(\tau)\int_0^{t-\tau} \rho_2(u)\mathrm{d}u\,\mathrm{d}\tau\,, \tag{5.13}$$

with P_{10} and P_{12} defined earlier. In a Markov chain, the distributions are explicitly given by

$$\begin{aligned} \rho_1(t) &= (\kappa_{10} + \kappa_{12})\exp\left(-(\kappa_{10} + \kappa_{12})t\right) \\ \rho_2(t) &= \kappa_{20}\exp\left(-\kappa_{20}t\right)\;. \end{aligned} \tag{5.14}$$

Working through the formulas to obtain $C(\Delta t)$ is now a simple matter of calculus (see Eq. 5.10 and *S1 Supporting Information*). The final result is a formula depending on three unknown parameters (κ_{10}, κ_{12} and κ_{20}) and the known parameter t_p to be used to fit any decay curve.

After defining the two factors

$$\begin{aligned} A_c &= \frac{\kappa_{10}-\kappa_{20}}{\kappa_{10}+\kappa_{12}}\left[1-\exp\left(-(\kappa_{10}+\kappa_{12})t_p\right)\right] \\ B_c &= \frac{\kappa_{12}}{\kappa_{20}}\left[1-\exp\left(-\kappa_{20}t_p\right)\right] , \end{aligned} \tag{5.15}$$

we obtain the relative abundance

$$C(\Delta t) = \frac{A_c}{A_c+B_c}\exp\left(-(\kappa_{10}+\kappa_{12})\Delta t\right)+\frac{B_c}{A_c+B_c}\exp\left(-\kappa_{20}\Delta t\right) , \tag{5.16}$$

which is a single formula to fit the data for any pulse duration t_p. At steady state, when $t_p \to \infty$, the two exponential functions in Eq. (5.15) disappear. For a finite t_p, instead, the two factors A_c and B_c get rescaled in a asymmetric fashion by the effect of the pulse. In the ideal case of a pulse of zero duration, *i.e.* $t_p \to 0$, we would have $C(\Delta t) = 1 - F_T(\Delta t)$. In the limit when decay is exponential, instead, $C(\Delta t)$ is an exponential function independent of t_p.

5.2.1 Examples

With Eq. (5.16) we have been able to determine a single mathematical formula that describes the whole decay pattern. The next challenge is to show that the fitting procedure (*Methods*) works. To show this, we proceed as follows. We have first extracted the decay patterns from Ref. [7], where the decay of proteins content of HeLa cells was monitored after pulse of duration 1, 5, 30, 120, 1200 minutes. We have taken the decay pattern corresponding to a 1 minute pulse and found the best fit for the three rates, $\kappa_{10} = 0.0109\,\text{min}^{-1}$, $\kappa_{20} = 0.0002\,\text{min}^{-1}$ and $\kappa_{12} = 0.0189\,\text{min}^{-1}$ (**Fig 5.2**), which correspond to an average lifetime $\overline{T}$ of more than 53 hours, see Eq. E in S1 Supporting Information. We have then plugged these rates in Eq. (5.16) to fabricate artificial datasets corresponding to 1, 5, 30, 120, 1200 minutes pulse and fit them to prove that the fit gives back the same rates used to fabricate the data. Contrary to the fits to the experimental data 5.2, the result of the fit (**Fig 5.3a**) gives rates that are identical to those used to generate the data (**Table 5.1**). This is good news because we want to be sure that the procedure is self-consistent. We have also added some noise in the form of

$$C^{(\text{noise})}(\Delta t) = C(\Delta t)\exp(\epsilon\zeta) , \tag{5.17}$$

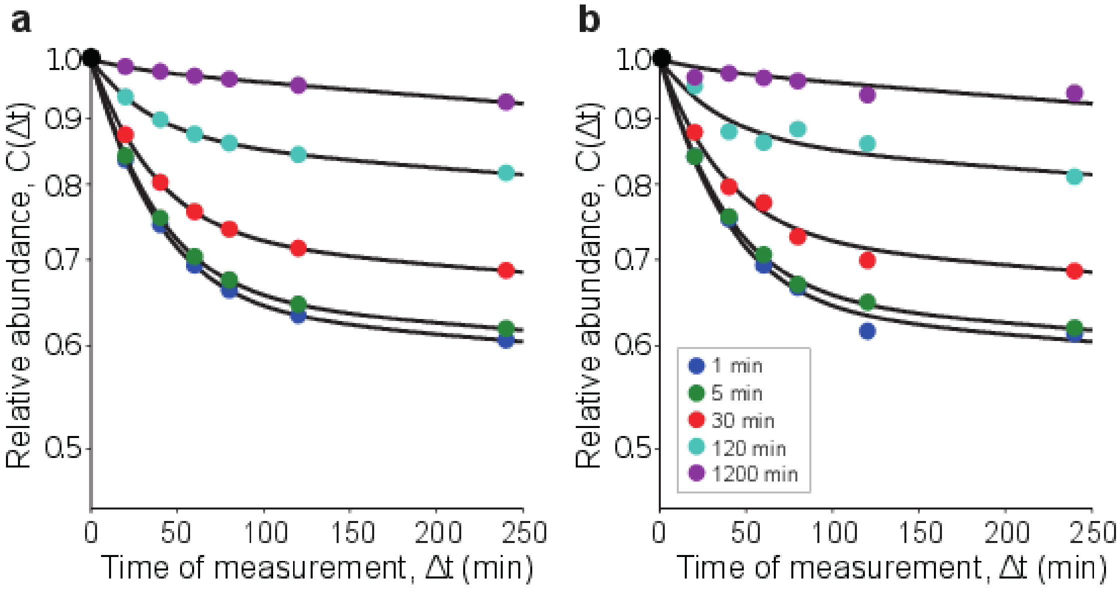


FIGURE 5.3: **Model calibration with fabricated data (a)** Verification of fitting procedure using simulated data separately. Using the parameters obtained from the best fit model to the data from [7] for pulse = 1 min, we fabricate sample data by calculating the abundance over time (dots) for different pulse lengths (1, 5, 30, 120, 1200 minutes) using the function that gives the decay pattern of the relative abundance $C(\Delta t)$, Eq. (5.16), as function of the measurement time Δt. We then fit resultant decay patterns with our fitting routine. We get back the same rates that were used to simulated the data for each experiment (**Table 5.1**). This shows that if the system in the background is unchanging, we can reliably extract the parameters of the system by fitting the decay patterns individually. $\kappa_{10} = 0.0109$ min^{-1}, $\kappa_{20} = 0.002$ min^{-1}, and $\kappa_{12} = 0.0189$ min^{-1}. **(b)** Simultaneous fit of pooled simulated data. Here the simulated data is augmented with a small amount of multiplicative noise, Eq. (5.17). We fit the whole collection of data simultaneously (see *Methods*). Values very close to our original simulation parameters are obtained (**Table 5.1** last row). This shows that under steady experimental conditions, we can reliably extract the parameters of the system by fitting the decay patterns simultaneously.

for $\Delta t > 0$, $\epsilon = 0.01$, and $\zeta \in [0, 1)$ is a uniformly distributed random number which is drawn independently for each Δt. Adding the noise is important to demonstrate the robustness of the fitting procedure and the sensitivity of the fitting function to small perturbations. In fact, the fit of the single fabricated decay patterns deteriorates as the pulse length increases depending on the amount of noise, probably because at long pulses state 2 dominates the behavior of the decay pattern. If the most populated state is state 2, as we have in our example, it becomes more and more difficult to detect events occurring in state 1 from the measurements as the pulse duration increases. This indicates that pulses of short duration should be preferred. Nevertheless, when we search for the three parameters to fit all pulsed curves at once, the procedure becomes quite reliable and the fit quality is very high (**Fig 5.3b** and **Table 5.1** last row).

As a next step, we have fit the individual experimental decay patterns as provided in

pulse	κ_{10} (min^{-1})	κ_{20} (min^{-1})	κ_{12} (min^{-1})
simulation values	0.0109	0.0002	0.0189
1 min	0.0109	0.0002	0.0189
5 min	0.0109	0.0002	0.0189
30 min	0.0109	0.0002	0.0189
120 min	0.0108	0.0002	0.0189
1200 min	0.0139	0.0002	0.0221
whole collection of data + noise	0.0050	0.0003	0.0254

TABLE 5.1: **Model parameters for simulated data with different pulse lengths.** Parameters from the best fit to the data simulated with different pulse lengths and the 2-state model as found by Multistart (MATLAB®) with 1000 start points. Minimization by fmincon (bounds $\kappa_{10}, \kappa_{20}, \kappa_{12} \in [0.000001, 1]$). Notice that the fits yield results identical to those used to generate the data. This proves the self-consistency of the procedure. Conversely in Table 2, the parameter values are not stable; suggesting different degradation system dynamics in each experiment.

Ref. [7] (**Fig 5.4a**) and compared the rates with each other (**Table 5.2**). We find that the rates depart from each other much more than those found with the fabricated data. With the proviso that we have extracted the data from an old low-definition figure in semilogarithmic scale, the differences in the rates definitively increase with the duration of the pulse. Also the search for a unique set of rates that allow fitting the various pulsed curves all together gives a poor result (**Fig 5.4b**). The most likely explanation of the wrong fit is that the labeling has affected the cells and has thus contributed to change its internal environment so that protein degradation after a long pulse is different than protein degradation after a short pulse, an effect that may certainly depend also on the labeling technique.

pulse	κ_{10} (min^{-1})	κ_{20} (min^{-1})	κ_{12} (min^{-1})
1 min	0.0109	0.0002	0.0189
5 min	0.0091	0.0004	0.0294
30 min	0.0139	0.0004	0.0610
120 min	0.0172	0.0003	0.0434
1200 min	0.0759	0.0002	0.0451

TABLE 5.2: **Model parameters for data from Ref. [7].** Parameters from the best fit to the 2-state model as found by Multistart (MATLAB®) with 1000 start points. Minimization by fmincon (bounds $\kappa_{10}, \kappa_{20}, \kappa_{12} \in [0.000001, 1]$ min^{-1}). Notice that each fit yields different parameters. This suggests that the degradation system dynamics in each experiment is different. In contrast Table 1 shows that the fits to simulated data with consistent parameters return the same rates. Data are reported in Table A in S1 Supporting Information

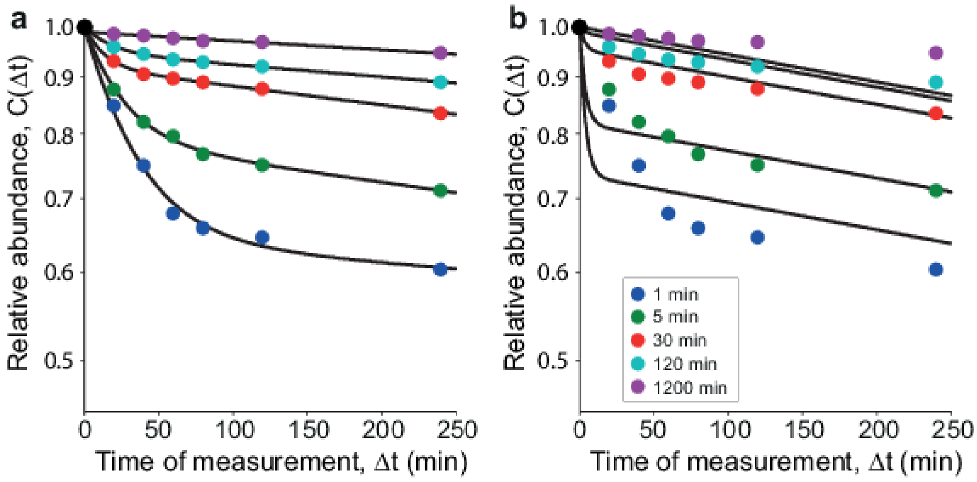


FIGURE 5.4: **Model calibration with data from Ref.** [7] **(a)** Decay patterns from Ref. [7] fit individually. Each decay pattern from Ref. [7] is fit with the 2-state model using Eq. (5.11) with $C(\Delta t)$ from Eq. (5.16). We find that for some of the decay patterns, the parameters obtained from the fitting are different from the others (**Table 5.2**). This implies that the underlying system has changed in the different experiments. Possibly the labeling procedure has affected the cells and contributed to a change in the internal environment. **(b)** Simultaneous fit of pooled real data with Eq. (5.12). Here we pool the decay patterns and fit them simultaneously with the 2-state model. No good fits were found, despite using global and multistart techniques in the parameter search process. This implies that the underlying systems across the experiments can not be described by one unified model, at least not the two state model that we have considered. Possibly the labeling procedure has affected the cells and contributed to a change in the internal environment.

5.2.2 Just pulse, no chase

So far, we have focused our derivations on experiments where marked molecules are synthesized during a pulse of synthesis, and measurements of the marked molecules are taken after the chasing period. Another experimental set up is to have the pulse period up until the time of measurement, without "chase" [234, 235]. The steady state value, $N^{(\mathrm{st})}$, is a convenient value to use for normalizing the measurements. Thus, the ratio

$$P(\Delta t) = \frac{N(t)}{N^{(\mathrm{st})}}\,, \tag{5.18}$$

is a convenient quantity for the relative abundance of labeled molecules. Here, Δt is the time of measurement from the start of labeling. Using the same framework, from Eqs. (5.7) and Eq. D in S1 Supporting Information we obtain

$$P(\Delta t) = \frac{A_p}{A_p + B_p}\left(1 - \exp(-(\kappa_{10} + \kappa_{12})\Delta t)\right) + \frac{B_p}{A_p + B_p}\left(1 - \exp(-\kappa_{20}\Delta t)\right)\,, \tag{5.19}$$

where the two factors A_p and B_p are given by

$$\begin{aligned} A_p &= \kappa_{20}(\kappa_{10} - \kappa_{20}) \\ B_p &= \kappa_{12}(\kappa_{10} + \kappa_{12}), \end{aligned} \tag{5.20}$$

as the expression for P according to a two-stage model.

If the steady state measurement of molecules is not available, any other time point can also be used. With other time points used as normalization, the expression for P is more complicated, but still solvable by hand for the two-stage model.

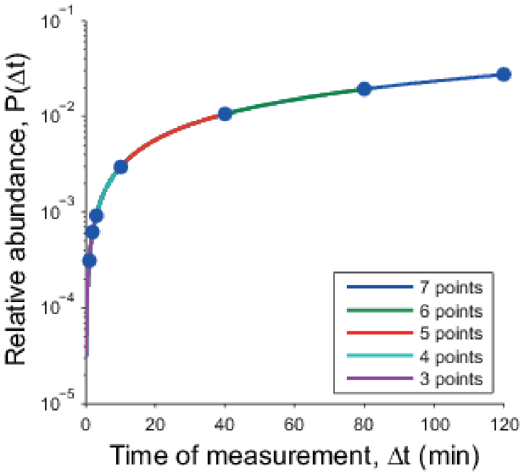


FIGURE 5.5: **Simulation of pulse no chase experiments** Simulation of pulse no chase experiments, where the pulse is applied up until the time of measurement. Data is produced with rates $\kappa_{10} = 0.0109\,\text{min}^{-1}$, $\kappa_{20} = 0.0002\,\text{min}^{-1}$ and $\kappa_{12} = 0.0189\,\text{min}^{-1}$ using (5.19). Traces show the results of several fits, each fitting taking into consideration one additional data point (**Table 5.1**).

	κ_{10} (min^{-1})	κ_{20} (min^{-1})	κ_{12} (min^{-1})
simulation values	0.0109	0.0002	0.0189
3 points	0.0104	0.0002	0.0240
4 points	0.0119	0.0002	0.0402
5 points	0.0109	0.0002	0.0189
6 points	0.0109	0.0002	0.0190
7 points	0.0109	0.0002	0.0189

TABLE 5.3: **Model parameters for simulated pulse no chase data.** Parameters from the best fit to the 2-state model as found by Multistart (MATLAB®) with 100 start points. Minimization by fmincon (bounds $\kappa_{10}, \kappa_{20}, \kappa_{12} \in [0.000001, 1]$ min^{-1}). The first row is the best fit for the data as if the experimentalist only took 3 measurements at $t = 1, 2, 3$ minutes. The 2nd row takes one more measurement ($t = 10$ minutes) into consideration.

One might assume that with more data, we would see an improvement in the determination of the model parameters. Initially we see this is true, but once we have taken data up to t = 20-40 minutes (4-5 points), additional sampling does not significantly improve the fit (**Fig 5.5 and Table 5.1**). This can be explained by examining the sensitivities of the output to the parameters. We define the sensitivity of the output to parameter κ_i as

$$v_{\kappa_i}(\kappa_i, t) = \frac{dP(t)}{d\kappa_i} \tag{5.21}$$

This equation describes the amount that the output should change if the values of κ_i and time change. The expressions are quite lengthy, but we can write them for each of the parameters in the system in closed form. By plugging in the "true" values for $\kappa_{j \neq i}$ we can plot the sensitivities as a function of time and κ_i.

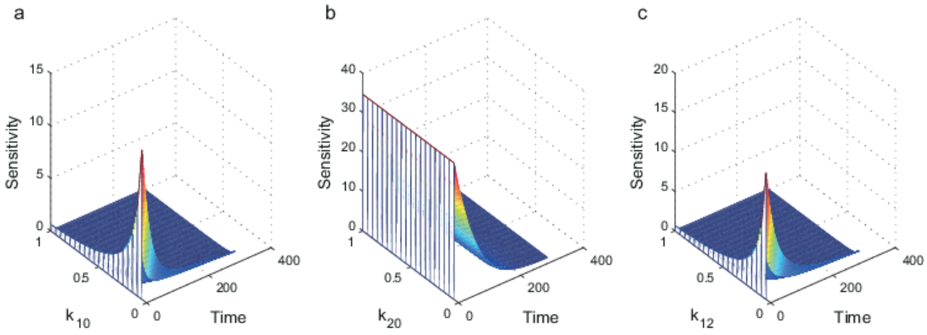


FIGURE 5.6: **Sensitivity of $P(\Delta t)$ to the parameters κ_{10}, κ_{20} and κ_{12}** We plot the sensitivities of the measurements in a pulse no chase experiment assuming a 2-state model with $\kappa_{10} = 0.0109$ min^{-1}, $\kappa_{20} = 0.0002$ min^{-1} and $\kappa_{12} = 0.0189$ min^{-1}. **(a)** Output is only sensitive to small values of κ_{10}. **(b)** Output is sensitive to a range of κ_{20}. **(c)** Output is only sensitive to small values of κ_{12}. For all parameters, taking more measurements at later timepoints does not help the parameter estimation because the output is not sensitive to deviations in the parameters at late times.

We discover that the output at late time points are insensitive to the parameters. Furthermore, the output is only sensitive to small values of κ_{10} and κ_{12} (**Fig 5.6**). Note that here we have calculated the local sensitivities – the result is likely to differ for different ranges of κ's.

5.2.3 Dynamical properties

The half-time $t_{1/2}$ is traditionally used as a quantity to measure the stability of the molecules. Unfortunately, this quantity is meaningful only when the decay is exponential because in that case there is a universal relationship between the half-time, the rate of degradation and the average lifetime. In a non-exponential decay, the universality of the relationship is lost. For these systems, the average lifetime $\overline{T}$ is a more meaningful quantity. Indeed, the steady state quantity of molecules (assuming stationary growth conditions and non-synchronized cells) is given by

$$N^{(\mathrm{st})} = \omega \overline{T}\,, \tag{5.22}$$

(*Methods*) thus allowing, at least in principle, to estimate the synthesis rate ω once the steady state amount of molecules per cell has been determined.

Once the rates κ_{ij} have been estimated, a number of other properties can be easily computed. The average time spent on state 1 is given by $\nu_1 = 1/(\kappa_{10}+\kappa_{12})$. The fraction of molecules able to pass from state 1 to state 2 is given by $\kappa_{12}\nu_1$. The molecules able to reach state 2 will occupy this state for an average amount of time given by $\nu_2 = 1/\kappa_{20}$. Finally, the fractions π_1 and π_2 of molecules found in states 1 and 2 is given by

$$\pi_1 = \frac{\kappa_{20}}{\kappa_{12}+\kappa_{20}} \qquad \text{and} \qquad \pi_2 = 1 - \pi_1\,. \tag{5.23}$$

Furthermore, if for any reason the half-time $t_{1/2}$ is sought, it can be computed numerically by inverting the equation

$$C_\infty(t_{1/2}) = \frac{1}{2}\,, \tag{5.24}$$

where C_∞ is the expression given in Eq. (5.16) when $t_p \to \infty$. Notice that the use of Eq. (5.24) for any finite $t_p < \infty$ would lead to an apparent the half-time different from the true value $t_{1/2}$ thus making the apparent half-time dependent on the length of the pulse.

5.3 Conclusion

Even if steady state expression levels of molecules depend only on their average lifetime, non-exponential decay has an effect concerning timing and dynamics of cellular response[120, 232, 236], and cell-to-cell variation in cellular content when cultures are subject to stochastic effects [237]. For this reason, it is an aspect of regulation that has to be taken into account if one wants to understand the reaction of cells to stress or to environmental changes. Furthermore, disentangling various hypotheses concerning the nature and the structure of biochemical pathways responsible for degradation[1, 15, 77] is possible only when models take the complexity of the pathways into account. Nevertheless, the derivation of the average lifetime required to compute the steady state properties (and the synthesis rate, if an independent measurement of the steady state abundance is available) requires the correct mathematics, which in most of the cases is not given by the exponential fit.

In this manuscript we focused on the two-stage model as a good approximation that describes most of the non-exponential decay patterns. This generalization is based on extensive experience from fitting thousands of mRNA and protein decay patterns. Nevertheless, there is no reason in principle to restrict the number of states to two, since the number of biochemical steps related to the destabilization of molecules is conceivably much larger. While it is not difficult to draw and mathematically describe larger networks with more states and alternative pathways, there is rarely enough data available to fix all the parameters [6]. A lucky exception is provided by sets of experiments where measurements are taken with different parts of the biochemical network deactivated [1, 15].

Pulse-chase experimental techniques offer the advantage of a low impact on the metabolism and well-being of the cells. Yet, little attention has been devoted to the fact that the pulse has an effect on the decay pattern if the decay pattern is not exponential. By working through the mathematics of single molecule decay, we derive and demonstrate in a novel approach a series of equations that anyone can use to fit complex decay patterns. The values of the rates would then finally provide a valuable tool to compute other dynamic quantities.

When assessing the nature of the decay pattern we recommend to compare the fit of the two-stage model with the simpler exponential model to ensure that the data supports the more complex model. As a guideline, long pulses (and thus steady state measurements) tend to obfuscate short term dynamics and thus appear as exponential decay, as this catches the long term dynamics of the degradation process. We recommend the use of the shortest pulse as possible in order to detect short time effects. More pulses of different lengths increase the robustness and our confidence in the model and its calibration. As a byproduct, fitting curves generated from pulses of different lengths may allow discovery of perturbing effects from labeling, when the fits of the individual curves reveal very different rates beyond what one would expect from noise. One caveat to the current approach outlined here is the assumption that the quantities measured are reflective of synthesis and degradation only. However, in in vivo systems, each cell division results in a dilution effect of the molecules. Thus for such experiments it is imperative to choose experimental times well within the cell division time, or take cell division into account.

5.4 Acknowledgments

CS and AV thank M. Selbach and E. McShane for useful discussions.

5.5 Supporting Information

S1 Supporting Information. Supplementary note. The file contains some supplementary calculations and the table with the experimental data extracted from Ref. [9].

6

Discussion and Summary

The central dogma of molecular biology proposes that the flow of genetic information starts at the level of DNA which is copied into RNA and then translated into proteins [9]. This system of information transfer provides for two very natural checkpoints where gene expression can be adjusted – either at the level of mRNA (i.e. controlling transcription or mRNA degradation processes) or the level of proteins (i.e. controlling translation or protein degradation processes). Within each checkpoint, there are a multitude of processes simultaneously active, adjusting and fine-tuning the concentrations of mRNAs and proteins. Some of these processes are specific to a particular subset of genes, while others are non-specific. In concert, all these processes contribute in some way to maintain cell homeostasis, keeping the cell alive and adaptable to different environments.

In the previous sections, we have examined several aspects of post-transcriptional gene expression control (Chapter 1-4). Chapter 2 and Chapter 4 are -omics type experiments to study patterns of gene expression across the genome (or proteome, in the case of Chapter 4). In Chapter 2, RNA-Seq coupled with ribosome profiling allows us to evaluate the transcriptional and translational state in *E. coli* under different conditions: LB (normal), minimal media, acute heat stress and acute osmotic stress. In Chapter 4, shotgun proteomics yields protein degradation curves of the mouse proteome. In both of these -omics type studies, identification of gene subgroups is integral reducing the complexity of the big picture to understand trends and patterns in the system.

Chapter 1 and Chapter 3 deal with specific aspects of gene expression control. In general, the abundance of mRNA correlates directly with the level of protein synthesis; more mRNA copies allow more proteins to be made simultaneously. In Chapter 1, we study the biochemical pathway of miRISC mediated degradation, a major pathway of mRNA degradation[1]. Chapter 3 explores the frequency of ribosome dropoff. In the process of protein synthesis, ribosomes move along the mRNA chain to translate the genetic sequences into amino acid sequences. Premature ribosome dropoff results in production of incomplete proteins; thus, elevated rates of ribosome dropoff reduce the rate of protein synthesis[3].

The work completed in this PhD is in the direction of data analysis and mathematical modeling. In Chapter 2, 3 and 4, development of analysis techniques appropriate to the experiments and the scientific question is the main part of our contributions to the work.

Mathematical modeling is a powerful tool for describing, understanding and predicting biological systems. In Chapter 1 and 4, simple Markov Chain models are used to validate hypotheses regarding the biochemical networks underlying the process of degradation. Chapter 5 focuses the analysis of data from pulse-chase experiments, which are widely used to study degradation of macromolecules due to low impact on metabolism and the well-being of cells. Additionally, mathematical models allow us to evaluate systems for parameter sensitivities and make inferences to optimal experimental design. These concepts are touched upon in Chapter 1 and 5 and the supplementary section B.

Our work in the intersection of mathematics and biology showcases the power of statistical data analysis and mathematical modeling for validation and discovery of biological phenomena.

In the following sections, I review the main results of the manuscripts included in the thesis and I show how the research results presented in this dissertation have advanced our understanding of post-transcriptional control of gene expression.

6.1 Overview of main results

6.1.1 mRNA degradation through the miRISC pathway

One major pathway of mRNA degradation is through the action of microRNA and the miRISC complex. While many of the important proteins in the pathway have been identified, the order of protein recruitment and biochemical pathway are unknown. In [10], it was hypothesized that miRISC finds the target mRNA and forms a complex. According to that hypothesis, the protein complexes associated with NOT1 and PAN3 are recruited, triggering deadenylation of the target mRNA and subsequent degradation. This hypothesis can be translated into a network topology (Figure 6.1), which guided the experimental work of [10].

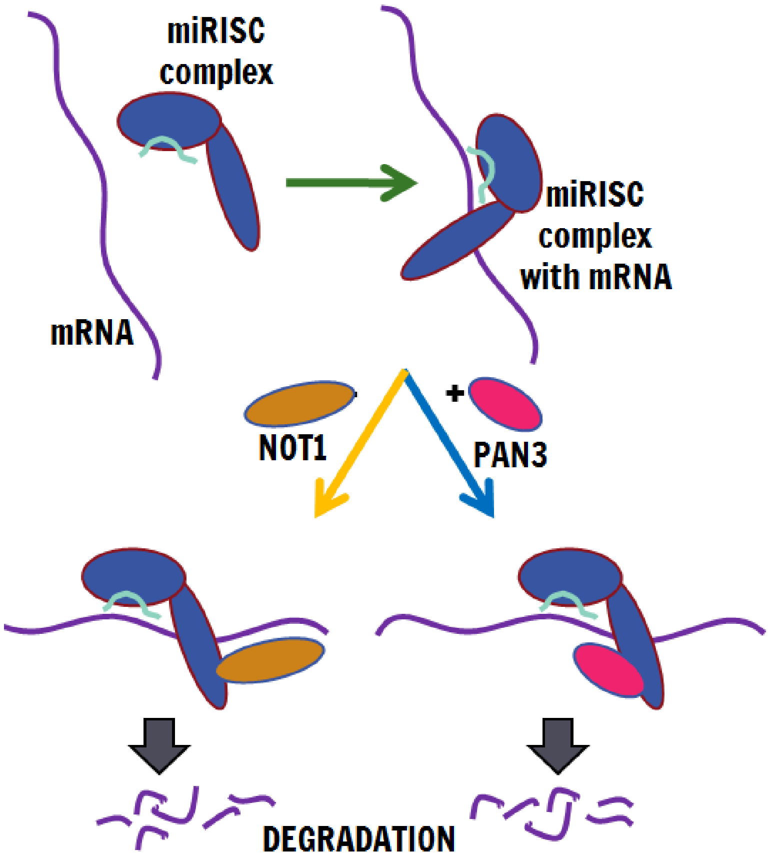


FIGURE 6.1: miRISC hypothesis as presented in [10]

This experimental work provided a measurement of the decay curves of an mRNA targeted by miR-9b with systematic depletion of miR-9b, NOT1 and PAN3. In each experiment, the absence of the factors miR-9b, NOT1 and/or PAN3 corresponds to the network with certain transitions inactivated, thus resulting in a reduced network. Our work consisted of evaluating the biochemical network by means of a Markov chain model. We started with the experiment corresponding to the most reduced network (absence of miR-9b) and found the most parsimonious model that can capture the dynamics in the experiment. We calibrated our model using the data and solved for the unknown transition rates. From here, we immediately found that the hypothesis presented in [10] must be complemented by additional pathways in order to capture the dynamics found in the experimental data.

Next, we considered the experiment corresponding to the next most reduced network. The previously defined transitions were kept constant, and we solved for the newly activated transitions. Following this strategy of network reactivation and model calibration, we reached a point where we are unable to find suitable parameters to capture the dynamics of the corresponding experiment. That is, we found that the data cannot support the hypothesis proposed in [10]. In fact, we find that if the hypothesized network was correct, only $\sim 7\%$ of the target mRNA would be regulated by the miRISC mechanism, contrary to the belief that miRISC is a major contributor to mRNA degradation. Thus we proposed an alternative hypothesis: the complex of miRISC and NOT1/PAN3 occurs before interaction with the target mRNA.

Additionally, we discovered that the data provides some more insights into the dynamics of the network. In the case where NOT1 and PAN3 are absent, the mRNA is more stable, degrading slower. Perhaps the miRNA stabilizes the target mRNA, protecting it from the action of an alternative miRNA-independent pathway. Here we have assumed that the interactions in the system are irreversible. However, this assumption is not obligatory, and with the present data, it is not possible to distinguish between reversible and irreversible interactions.

Our model also supports the conclusions in [10] regarding the contributions of PAN3. PAN3 alone does not enhance the destabalization of the mRNA. However, we saw a strong cooperative effect between PAN3 and NOT1: the degradation system is enhanced when both factors are present. To verify our findings experimentally, one could look for the presence of miRISC+NOT1 complexes in the absence of target mRNA. Moreover, steady state relative amounts of mRNA in the different biochemical states can provide further validation data for our networks and additional information to unveil further details of the miRNA-mediated mRNA degradation. We believe that these aspects are not limited to the special miRNA-mRNA pair studied in [10] and are therefore a new general mechanism of mRNA control.

6.1.2 Bacteria differently regulate mRNA abundance to specifically respond to various stresses

Ribosomal profiling is an experimental technique that produces a high resolution, genome-wide snapshot of translation [101]. When coupled with RNA-seq experiments, a genome-wide description of the translational and transcriptional state is obtained. While *E. coli* has been extensively studied, the translational/transcriptional response to acute stress has not been captured by existing studies. In this paper we study the effects of acute heat stress and acute osmotic stress on the transcription and translational state of the cell.

In agreement with past studies [45], we found that mRNA abundance is highly correlated to the number of ribosomes translating the gene for most genes. Additionally, we found that the response to stress was also well correlated for most genes: that is, increases in mRNA abundance were often paired with increases in ribosome protected fragments (and the inverse). However, for each type of stress, a small subset of genes emerged, seemingly only to be regulated at the level of translation.

In heat stress, the mechanism for translational upregulation may be facilitated by secondary structure (or lack thereof). The subset of translationally upregulated genes in heat stress tended to have secondary structure in the initiation regions that melted at higher temperatures.

In contrast, the translational response to osmotic stress seems to be the result of a more opportunistic mechanism. Under normal conditions more than 90% of the ribosomes are involved in translation [136, 137], leaving little capacity to reallocate ribosomes to new mRNAs [138]. With osmotic stress, there is a 33% reduction of mRNA transcripts (minimal media ∼2400 mRNA copies/cell, osmotic stress ∼1600 mRNA copies/cell). Compare this to heat stress, where there is only a 7.6% reduction of transcripts (LB ∼7800 mRNA copies/cell, Heat ∼7200 mRNA copies/cell). Transcript reduction in response to osmotic stress was also found in yeast [140], implying conserved features of response across the species. Potentially, the massive reduction of mRNA transcripts in response to osmotic stress frees up translational resources to translate mRNA that is newly synthesized or existing mRNA that survive through the stress.

Notably, five major transporters were upregulated in response to osmotic stress. The copy numbers for *proV*, *proP*, *proX*, *proW* and *otsA* increased from < 0.5 copies/cell in normal conditions to > 2 copies/cell upon osmotic stress. Other osmotic stress-related genes were upregulated from ∼1 mRNA copy/cell in minimal media to >2 copies/cell upon osmotic stress. As reference, genes involved in key physiological processes, translation and energy metabolism during normal conditions are expressed at >2 copies/cell.

Ultimately, gene expression must be adjusted to ensure cell adaptability and survival under stress. Unicellular organisms lack the internal homeostatic buffering capacity of multicellular species, and for survival under acute stress they need to quickly reprogram their cellular activities. For the two stress conditions studied, changes in the mRNA level correspond to the translation level for the majority of genes. Outside of these genes, a unique fraction of genes are regulated only at the level of translation. Adjusting the cellular program by translating existing mRNAs should be much faster than a response shaped by *de novo* transcription followed by translation, potentially allowing the cell to respond more rapidly to changing environments.

6.1.3 Quantitative assessment of ribosome drop-off in *E. coli*

The ribosome is the molecular machine for protein synthesis. Ribosomes assemble on mRNA, *initiating* translation. The ribosome then moves along the mRNA reading one codon at a time, catalyzing the aminotransferase reaction *elongating* the growing protein. At each elongation step, there is a chance that the ribosome can *drop off*, resulting in incomplete translation of the gene. If the ribosome successfully completes the amino acid sequence, it must *terminate* the process, releasing the amino acid chain and dissociate from the mRNA. The rates of *initiation*, *elongation*, *drop-off* and *termination* determine the rate of protein synthesis.

The phenomenon of ribosome drop-off has been known since at least the 1970s. Several groups have measured the rate of drop-off both *in vitro* and *in vivo* [63, 65, 66]. In [63], the rate of ribosome drop-off was measured *in vivo* by measuring the accumulation of incomplete protein products in a system where the pathway for incomplete protein degradation was disabled. From these experiments, drop-off in *E. coli* was estimated to be approximately 10^{-4} per elongation step [63].

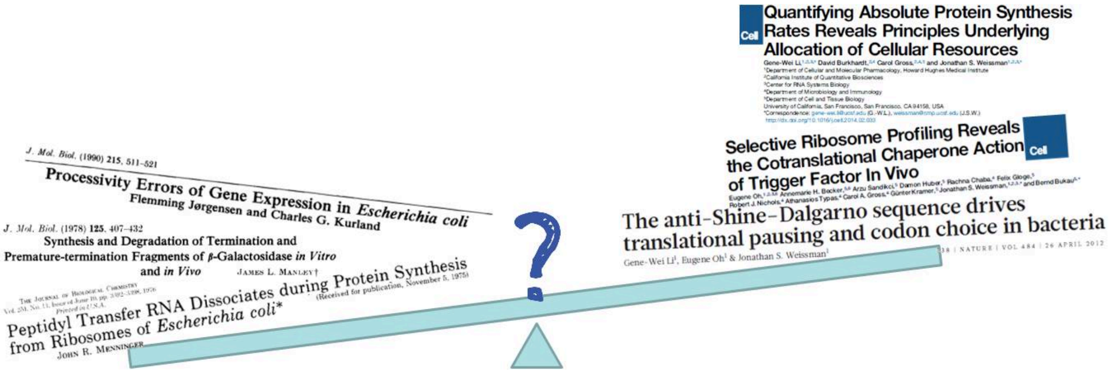

Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources
Cell
Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S. Weissman
Selective Ribosome Profiling Reveals the Cotranslational Chaperone Action of Trigger Factor In Vivo
Cell
The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria
Gene-Wei Li, Eugene Oh & Jonathan S. Weissman
NATURE | VOL 484 | 26 APRIL 2012
J. Mol. Biol. (1990) 215, 511-521
Processivity Errors of Gene Expression in Escherichia coli
Flemming Jørgensen and Charles G. Kurland
J. Mol. Biol. (1978) 125, 407-432
Synthesis and Degradation of Termination and Premature-termination Fragments of β-Galactosidase in Vitro and in Vivo
James L. Manley†
The Journal of Biological Chemistry
Peptidyl Transfer RNA Dissociates during Protein Synthesis from Ribosomes of Escherichia coli*
(Received for publication, November 5, 1975)
John R. Menninger
?

Figure 6.2: Contradictory evidence of ribosome drop-off

Ribosome profiling is a new experimental technique which maps the current locations of the ribosomes on the mRNA. If the ribosomes are indeed dropping off, one would expect that ribosome density at the end of genes should be lower than at the beginning of genes. Surprisingly, ribosome drop-off appeared to be undetectable in Ribo-Seq data [68]. In [3], we presented a method that can detect the drop-off rate from Ribo-Seq data.

We show that analytical approaches reported in literature thus far were unsuccessful because they utilized a binning strategy that obscured the relatively low rate of ribosome drop-off. These approaches typically divide each ORF into two halves and compare the number of reads that map onto each half [68]. If there are ribosomes dropping off before successfully completing translation, there would be a reduction of reads in the second half. This metric can be illustrated in a scatterplot where the ribosome density in the first half is plotted against the ribosome density in the second half (Figure 3.3, Supplementary Section 3.1.3). When there is no significant difference between the densities of the two halves, the points will cluster around $y = x$. For an ORF where the density of the second half is lower than the first half, the corresponding point would be below the line $y = x$. In principle, this approach is mathematically sound. However, the sensitivity of this approach to drop-off depends critically on the ORF length. When the frequency of drop-off events is not large enough with respect to length of the ORFs, the difference in ribosome density between the two halves of the ORF are too small to be detected in a log-log scatterplot. As a consequence, if the genome of interest prevalently contains short genes, the scatterplot-method is not sensitive enough to detect the drop-off, leading to the wrong conclusion that the ribosome drop-off rate is not measurable in the genome scale.

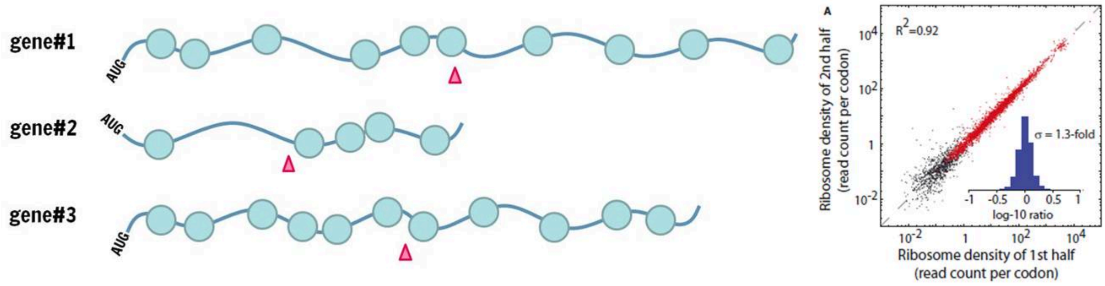


FIGURE 6.3: Preliminary approaches to detect ribosome density from ribosome positions. Scatterplot from [68].

We introduced an approach not affected by the length of the ORFs. The current consensus of ribosome drop-off is that the rate is per elongation step. Thus, instead of binning the ribosome positions by the 1st half vs the 2nd half of the gene, we bin the positions such that the width of each bin is equal (Figure 6.4). That is, we calculate the density of ribosomes per nucleotide. We used our method to analyze various *E. coli* datasets cultured in different experimental conditions, all of which were found in the GEO database. We found drop-off rates ranging from $1.4 \cdot 10^{-4}$ to $5.6 \cdot 10^{-4}$ events per codon. For a drop-off rate of $4 \cdot 10^{-4}$ per codon and an ORF length of 300 codons (approximately the average ORF length for *E. coli*), 10 out of every 100 ribosomes will fail to complete the translation of the messenger on average.

Taking into account the speed of ribosomes and the number of ribosomes actively involved in translation [53], and assuming a drop-off rate of $4 \cdot 10^{-4}$ per codon in all growth conditions, the number of premature ribosome drop-off ranges from 1400 per minute per cell at slow growth conditions to 29000 per minute per cell at fast growth conditions. Considering the lifetime of a cell [53], the total number of drop-off events in a slowly

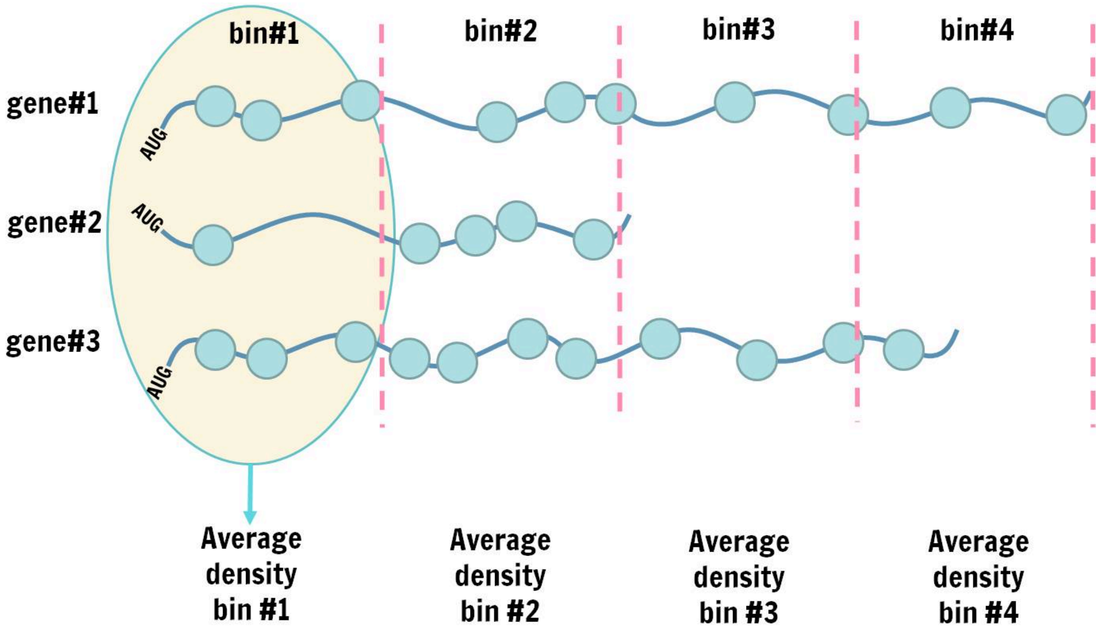


Figure 6.4: Binning strategy to detect ribosome drop-off from riboSeq experiments

growing population is about $14 \cdot 10^4$ events per cell cycle at slow growth (doubling time 100 minutes) and $75 \cdot 10^4$ at fast growth conditions (doubling time 25 minutes).

Furthermore, we come to a more general result relating drop-off rate and the length of genes. For a given drop-off rate, there is a limiting gene length above which the translation process becomes ineffective due to the high number of expected drop-off events. In the case of low drop-off rate, this threshold length is usually higher than the maximum gene length of *E. coli*. However, in those cases where ribosomes drop off with a higher frequency, the completion of translation is only reliable for shorter mRNAs. This suggests that when living organisms face conditions leading to increased drop-off rate (e.g. amino acids starvation) only a subset of genes can be effectively expressed. Since the probability of a ribosome to complete translation decreases exponentially with the ORF length, the magnitude of ribosome drop-off becomes an important evolutionary constraint of ORF length. If the genome of an organism is composed of ORFs that are too long relative to the drop-off rate, the reliability of translation may not support cell viability.

Contrary to previous Ribo-seq analysis results, we have shown that the magnitude of ribosome drop-off is highly variable and dependent on case-specific factors, including experimental conditions and the protocol used to collect Ribo-Seq data. Since the estimation of translation rates from Ribo-seq data assumes negligible ribosome drop-off, these estimations should be reevaluated to correct for possible biases due to drop-off events. In fact, we speculate that ribosome drop-off could be a possible explanation for the ubiquitous negative correlation between gene length and protein synthesis rate.

6.1.4 Global quantification of cellular protein degradation kinetics

Protein degradation is an important part of cell homeostasis. It allows the regulation of cellular processes (through deactivation of enzymes, transcription factors or receptors), recycling of protein precursors (amino acids), and removal of malformed or abnormal proteins.

In general, the literature on the degradation of biological molecules (such as proteins, mRNA) assumes exponential decay. Historically, this may have been the case because many experiments only measured 1-2 data points, and were thus unable to justify the implementation of more complex models. However, it is known that proteins are degraded through a number of different pathways; many of these pathways are multi-step processes with several complex biochemical stages. For these pathways, the widely accepted assumption of exponential decay is often not sufficient to describe the degradation of the target proteins.

Improvements in experimental technique and equipment have enabled experimentalists to generate higher resolution data, both qualitatively and temporally.

Here, the degradation profiles over several time points (t = 1, 2, 4, 8, 16 and 32 hours) for over 5000 proteins were measured by mass spectrometry. One of the early aims of our study was to overturn the long standing assumptions of exponential decay (1-stage model). Using a Markov Chain modeling approach, we found that over 10% of all proteins have degradation dynamics warranting at least a 2-stage model. Biochemically, one possible interpretation on of the 2-state model is that the protein can reach a second configuration (i.e. through folding or post translational modifications) which has a different lifetime expectancy as compared to the first state. All of these proteins were found to become more stable over time, namely their degradation rate becomes smaller with the age of the molecule. The degradation kinetics were found to be similar in two different cell types and conserved between mouse and human.

More than half of the proteins categorized to the 2-stage model are subunits of annotated heteromeric complexes. These subunits tend to be produced in super-stoichiometric amounts relative to the other members of the same complex. Based on these findings, we propose a simple model where only a fraction of the newly synthesized non-exponentially decaying proteins is stabilized by complex formation while the rest is degraded. Consistent with this model, we find that upregulation of non-exponentially decaying proteins increases the initial degradation.

6.1.5 Degradation Parameters from Pulse-Chase Experiments

Pulse chase experiments are often used to measure decay of macromolecules such as proteins or mRNA. However, depending on the dynamics of the target system being measured, the choice of the pulse may affect what dynamics can be derived from the measurements. For some experiments the length of the pulse can be chosen to be sufficiently long to bring the system to steady state, however for others, the toxicity involved with the pulse may limit the time that the pulse can be applied.

For exponentially decaying molecules, there is a simple algebraic relationship between half-time and average lifetime of the molecules, but this relationship does not hold for non-exponentially decaying molecules. Thus, half-time is not a good measure of stability for non-exponential decay [86]. In Chapter 4 we found that at least 10% of the mouse proteome decays in a non-exponentially decaying pattern. In [7], it was observed that the pulse length seems to have an effect on the half-time. In both of these cases, the assumption of exponential decay is inappropriate – the dynamics governing the system being measured can not be described by exponential decay.

Degradation can be modeled as a Markov chain process: each state represents a biochemical state of the degradation process, and the transfer or reaction rates between entities are represented as fluxes between Markov states. Exponentially decaying molecules can be modeled as a 1-state Markov Chain model, while the degradation patterns in Chapter 4 and reference [7] were well described by 2-state models. From this we can model the decay as a single molecule stochastic process and transform them to predictions on the population average (which can be measured with biochemical experiments).

We then use the framework to evaluate parameter sensitivity and optimal experimental design for the two simple degradation models. By working through the mathematics of single molecule decay, we derive and demonstrate in a novel approach a series of equations that anyone can use to fit complex decay patterns. The values of the rates would then finally provide a valuable tool to compute other dynamic quantities. Combined with prior knowledge of the system at hand, this framework can assist in choosing the optimal pulse length for experimentation.

As a guideline, long pulses (and thus steady state measurements) tend to obfuscate short term dynamics and thus appear as exponential decay, as this catches the long term dynamics of the degradation process. We recommend the use of the shortest pulse as possible in order to detect short time effects. More pulses of different lengths increase the robustness and our confidence in the model and its calibration.

6.2 Outlook

This cumulative dissertation probes several aspects of gene expression control, statistical data analysis and mathematical modeling. Thus the future extensions of this work are in several directions.

From a theoretical point of view, I have presented a model building approach (Chapter 1) and a mathematical framework (Chapter 5) that can be applied to study other systems. The model building approach is particularly useful to study biochemical networks where key players (e.g. enzymes or substrates) can be switched off. The mathematical framework around degradation patterns extended to pulse-chase experiments allows analysis of complex degradation systems (i.e. not single exponential). Both studies facilitate further development of optimal experimental design. Together with a hypothesis of the biochemical network, the model building approach in Chapter 1 allows experimentalists to design experiments to test different network topologies. The mathematical framework in Chapter 5 can be used to find the optimal pulse times depending on the degradation network and the parameters to be measured.

From an analysis point of view, the work in Chapter 2 and 3 presents several metrics and methods of analysis for ribosome profiling experiments. Notably Chapter 3 quantifies a basal rate of ribosome drop-off. This rate can be used as a benchmark for ribosome drop-off, thus providing a reference point to evaluate drop-off in specific genes.

6.3 Summary

In this dissertation I have presented several projects to elucidate some part of the gene expression control process or the magnitude of gene expression in different environments. To do this, we have utilized careful statistical data analysis, custom metrics and a robust mathematical framework to study decaying quantities. Chapter 2 and 4 deal with whole genomic / proteomic data sets to study patterns of gene expression in the cell. Chapter 1 and 3 deal with specific aspects of gene expression control. Chapter 1 and 4 utilize simple Markov Chain models to model degradation processes and validate hypotheses about the biochemical networks underlying the degradation. In addition to the predicative and validation power, mathematical modeling can also guide optimal experimental design. This work showcases the power of statistical data analysis and math modeling for validation and discovery of biological phenomena.

There are many more mechanisms and components to the system of gene expression control, all working in concert to adjust and fine-tune the expression level. The number of scientific questions is large, and the level of complexity is near infinite... and so we go on!

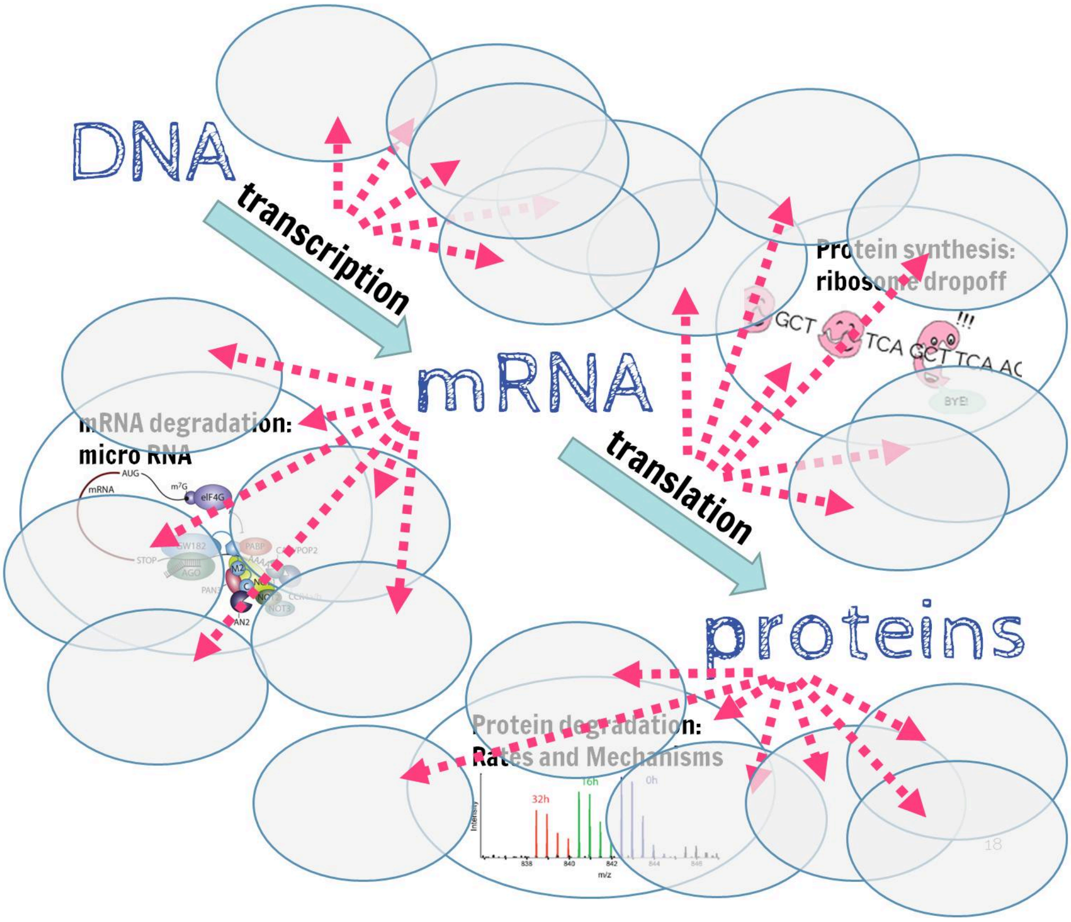


FIGURE 6.5: Future directions

Part II

Supplementary Sections to the Papers

Supplementary 1

Single-Molecule Modeling of mRNA Degradation by miRNA: Lessons from Data

A version of this chapter has been published as:

Sin C, Chiarugi D, Valleriani A (2015) Single-molecule modeling of mRNA degradation by miRNA: Lessons from data. BMC Systems Biology 9(Suppl 3):S2.

Online version:

http://www.biomedcentral.com/1752-0509/9/S3/S2

1.1 From single-molecule pathways to decay patterns

We model the mRNA decay as a single molecule stochastic process. The process starts from an initial state σ_0, which represents the mRNA in its initial condition, and terminates in the degradation state σ_n. The initial and the degradation states are defined here based on the time point from which the mRNA is experimentally detected until it is not detected anymore, respectively, in the context of the experimental technique used in [10]. Each realization of the stochastic process, which starts from σ_0 to the degradation state through the network of states represents the life of a single mRNA molecule. Therefore, the distribution of the random time T elapsing from σ_0 to the degradation state σ_n can be identified with the distribution of the lifetimes of the mRNA molecules. The state space σ is thus made of n transient states and 1 absorbing state σ_n (a generalization to more absorbing states is straightforward, see [90]).

In our modeling framework each state transition describes the occurrence of what we define a *first-order biochemical event*. This definition includes three possible biochemical scenarios:

1. an elementary first-order chemical reaction
2. a pseudo-first order biochemical reaction: the actual reaction order is > 1 but the concentration all of the reactants except for one are at the steady state level. As a result, this reaction behaves as a "true" first order reaction
3. an apparent first-order reaction: a complex chain of reactions with a single rate limiting step, thus exhibiting first-order behavior

In the context of the mRNA degradation pathway we consider the two latter cases only, due to the complexity of the involved biochemical reactions. If we assume that the probability of occurrence of each biochemical event depends only on the current state (Markov property), then we can model the stochastic processes describing the life of mRNA molecules as Markov chains.

Thus, following [86, 88–90], we can write the lifetime probability density ϕ as

$$\phi(t) \;=\; \vec{e}_0 \left[\exp\left(\mathbf{S}_0 t\right) S_A\right] \,, \tag{1.1}$$

where S_0 is a square $n \times n$ matrix. The off-diagonal elements of S_0 are the transition rates between the transient states, and the diagonal elements are composed of the negative sum over all outgoing transition rates from each transient state. The column vector S_A contains all the transition rates regarding transitions from the transient into the absorbing state. The row vector of length n $\vec{e}_0$ has a one in the position corresponding to the initial condition and zeros elsewhere.

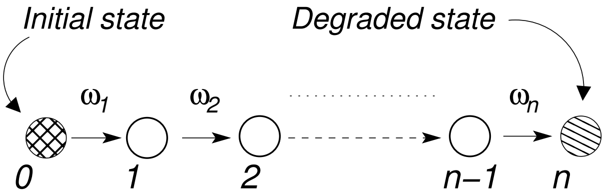


FIGURE 1.1: A generic chain made of n irreversible transitions. The probability density of the random time T from the initial to the degraded state is given by $\phi_n(t)$ derived in Eq. (1.5).

Similar to [86], we must take into consideration that at the beginning of the experiment, the expression of mRNA is at the steady state. Thus, not all mRNA molecules are at the initial state; some of them will be in other states because they are older. If the transcription of mRNAs was ongoing long before the beginning of the decay experiment, one can assume that the age distribution of the population of mRNA is at steady state[1]. In [86] it was shown that when the age distribution is at steady state, the fraction of mRNA remaining Δt time units after the stop of transcription is given by the function

$$\Lambda(\Delta t) \;=\; \frac{1}{\mathrm{E}(T)} \int_{\Delta t}^{\infty} \mathrm{d}t \; (1 - \Phi(t)) \;, \tag{1.2}$$

where $\mathrm{E}(T)$ is the average lifetime, namely the average value of T given by

$$\mathrm{E}(T) \;=\; \int_{0}^{\infty} \mathrm{d}t \; (1 - \Phi(t)) \;, \tag{1.3}$$

and Φ is the probability function of T given by

$$\Phi(t) \;=\; \int_{0}^{t} \mathrm{d}\tau \, \phi(\tau) \, . \tag{1.4}$$

An important technical point for the computation of Λ is that in Equation (1.2), $\mathrm{E}(T)$ is just the normalization factor in order to have $\Lambda(0) = 1$. The networks used in the main article are sufficiently simple, so that the calculation of ϕ can be done easily (see below), without recurring to the relatively complex formulation given in Equation (1.1). Nevertheless, Equation (1.1) is the general form of the function, which depends on the values of the rates in the matrix S_0. When Equation (1.2) has been written in terms of the rates, the rates can be determined by means of a nonlinear fitting of the log of $\Lambda(\Delta t)$ with the log of the data. The use of the logged data is justified by the common assumption that the experimental noise is multiplicative. Equation (1.1) and

[1] Short transcription pulses instead generate an age distribution away from steady state. This interesting extension of the theory will not be considered here because it is not relevant in this context.

(1.2) include the case when there is just one transient state. This is the case when the decay pattern follows an exponential function whose parameter is the rate of degradation, and also the inverse of the average lifetime.

1.1.1 Fitting functions: General aspects

All our fits were performed for two or three parameters and compared, where it made sense, with the exponential fit. Based on the Akaike Information Criterion (AIC) corrected for small sample sizes, we found that the fits with the exponential function did not perform sufficiently well to be considered in this study.

Since all our networks are comprised of irreversible transitions only, the computation of Λ can be split in the sub-tasks of computing the contribution of chains of $1, 2, \ldots, n$ transitions such as the one in Supplementary Figure 1.1. For each of these chains, the probability density ϕ_n for the absorption time in state n can be computed as the convolution of n independent exponential functions, and it is given as

$$\phi_n(t) \;=\; \left(\prod_{k=1}^{n} \omega_k\right) \cdot \left(\sum_{j=1}^{n} \frac{\exp(-\omega_j t)}{\prod_{k\neq j}(\omega_k - \omega_j)}\right)\,, \tag{1.5}$$

if all the ω_j are different from one another. In the networks considered in this work we just need the two cases corresponding to $n = 1$, *i.e.* the exponential function, and $n = 2$. The cumulative functions associated to the ϕ_n are thus given by

$$\Phi_n(t) \;=\; 1 \;-\; \sum_{j=1}^{n} \left(\prod_{k\neq j} \frac{\omega_k}{\omega_k - \omega_j}\right) \exp(-\omega_j t)\,, \tag{1.6}$$

again under the assumption that all rates are different. As we have seen after fitting the data, this assumption is never violated.

Assume now that a given network has, say, three possible paths from the initial state to the absorbing state made of only irreversible transitions. Let us call these three paths a, b, and c and let us assume that these three paths are characterized each by its own set of ω's and that the probability of each path to be taken is p_a, p_b and p_c, respectively. Let ϕ_a, ϕ_b and ϕ_c be the probability densities conditioned on each of the paths separately. Then, the total probability density ϕ is given by

$$\phi \;=\; p_a\phi_a + p_b\phi_b + p_c\phi_c\,, \tag{1.7}$$

and the cumulative probability function is given by

$$\Phi \;=\; p_a\Phi_a + p_b\Phi_b + p_c\Phi_c\,, \tag{1.8}$$

so the decay function Λ from Equation (1.2) reads

$$\Lambda(\Delta t) \;\propto\; \int_{\Delta t}^{\infty} \mathrm{d}t\,\left(1 - \sum_x p_x\Phi_x(t)\right)\,, \tag{1.9}$$

where $x = a, b, c$, and the proportionality constant must be fixed so that $\Lambda(0) = 1$. If we now write the Φ_x as $\Phi_x + 1 - 1$ and we substitute, we obtain

$$\Lambda(\Delta t) \;\propto\; \sum_x p_x\mathcal{O}_x(\Delta t; \vec{\omega}_x)\,, \tag{1.10}$$

where we have defined

$$\mathcal{O}_x(\Delta t; \vec{\omega}_x) \;=\; \int_{\Delta t}^{\infty} \mathrm{d}t\,\left(1 - \Phi_x(t)\right)\,. \tag{1.11}$$

The formulation in Equation (1.10) turns out to be a quite useful, since our Φ_n expressed in Equation (1.6) take an explicit relatively simple form once put in (Equation (1.11)). The vector $\vec{\omega}_x$ is the set of all ω's along the path x. Upon simple explicit integration we have thus, for $n = 1$ and $n = 2$ the two functions

$$\mathcal{O}_1(\Delta t; \omega_1) = \frac{\exp(-\omega_1\Delta t)}{\omega_1}\,, \tag{1.12}$$

and

$$\mathcal{O}_2(\Delta t; \omega_1, \omega_2) = \frac{\omega_2\exp(-\omega_1\Delta t)}{\omega_1(\omega_2 - \omega_1)} + \frac{\omega_1\exp(-\omega_2\Delta t)}{\omega_2(\omega_1 - \omega_2)}\,, \tag{1.13}$$

to be used in the explicit calculations that follow.

1.1.2 Fitting functions: Specific forms

This section delivers the explicit form of the functions used in the fitting of the data.

- The generic two-state network in Figure 3 in the main article has

$$\begin{aligned}\Lambda(\Delta t) \;&\propto\; \frac{\mu}{\mu+\lambda}\mathcal{O}_1(\Delta t; \omega_1 = \lambda + \mu) \\ &+\; \frac{\lambda}{\mu+\lambda}\mathcal{O}_2(\Delta t; \omega_1 = \lambda + \mu, \omega_2 = \nu)\,,\end{aligned} \tag{1.14}$$

where in both cases $\omega_1 = \lambda + \mu$ takes into account that the dwell time on state A is exponential with parameter $\lambda + \mu$. This function was used to fit the data of the negative control (green line) and therefore to fix the rates λ, μ and ν. These three rates are kept fixed in all other networks.

- The networks in Figures 6, 8, 10 in the main article share the same structure and lead to

$$\begin{aligned}\Lambda(\Delta t) \quad &\propto \quad \frac{\lambda}{\lambda+\mu+\lambda'}\mathcal{O}_2(\Delta t; \omega_1 = \lambda+\mu+\lambda', \omega_2 = \nu) \\ &+ \quad \frac{\mu}{\lambda+\mu+\lambda'}\mathcal{O}_1(\Delta t; \omega_1 = \lambda+\mu+\lambda') \\ &+ \quad \frac{\lambda'}{\lambda+\mu+\lambda'}\mathcal{O}_2(\Delta t; \omega_1 = \lambda+\mu+\lambda', \omega_2 = \mu')\,,\end{aligned} \tag{1.15}$$

where $\lambda' = \lambda_R$, λ_{RN}, λ_{RNP} and $\mu' = \mu_R$, μ_{RN}, μ_{RNP} in Figures 6, 8, 10, respectively. To fit the data, as was done in Figures 7, 9, and 11 in the main article, we have kept the three rates λ, μ and ν fixed as they have been determined in the fit of the negative control. This means that the $\Lambda(\Delta t)$ given in Eq. (1.15) was used to fit the two additional rates λ' and μ' for each of the following three data sets.

Δt	(-)	R	RP	RN	RNP
0	1	1	1	1	1
15	0.6899	0.6720	0.6722	0.5641	0.4885
30	0.5588	0.6238	0.6322	0.3206	0.2263
60	0.3375	0.4806	0.4184	0.1919	0.1272
180	0.1462	0.2262	0.2182	0.0766	0.0388
360	0.0857	0.1112	0.1035	0.0397	0.0217

Table 1.1: Experimental data as extracted from the plot published in [10]. The various experimental conditions are given in the first row: (-) = negative control, when the miRNA is not expressed; R = when only miRNA is expressed but both NOT1 and PAN3 are not expressed; RP = data when NOT1 is knocked down; RN = data when PAN3 is knocked down; RNP = all factors are present, no knock downs.

1.1.3 Data used in the study

The experimental data used in this study have been extracted from [10]. We report the numerical values in Supplementary Table 1.1: the first column is the measurement time expressed in minutes; the second column is the data for the negative control, when the miRNA is not expressed, which was fitted with Eq. (1.14); the third column is the data when both NOT1 and PAN3 are knocked down (only miRNA is expressed) and

rate	10^{-3}min^{-1}	10^{-3}min^{-1} 95% CI
λ	0.8	[0.2, 1.3]
μ	27.6	[22.9, 32.4]
ν	2.8	[1.8, 3.8]
λ_R	2.3	[0.0, 5.2]
μ_R	5.2	[3.0, 7.4]
λ_{RN}	46.0	[30.5, 68.7]
μ_{RN}	46.1	[31.9, 60.2]
λ_{RNP}	150.1	[90.8, 209.4]
μ_{RNP}	49.3	[37.3, 61.2]

Table 1.2: Values of the rates as they arise from the fitting of the experimental data. By using our derived functions Λ, Eqs. (1.14) or (1.15), we can perform a non-linear fit of the data listed in table 1.2. The values of the rates are given together with their 95% confidence interval (CI).

was fitted with Equation (1.15); the fourth column is the data when NOT1 is knocked down: since this data is very close to the data in the third column, an additional fit of it was not performed as these two sets of data cannot be distinguished from each other; the fifth column is the data when PAN3 is knocked down and was fitted with Equation (1.15); the last column is the positive control, when all factors are expressed and was fitted with Equation (1.15).

1.1.4 Fitting procedure and algorithm

Our fitting procedure consisted in finding the set of parameters $\vec{\omega}$ that minimized the square deviation of the logarithm of the data, as

$$\text{RSS} = \sum_{\{\Delta t\}} \left(\log(\text{data}(\Delta t)) - \log(\Lambda(\Delta t))\right)^2 , \tag{1.16}$$

where "data" is any of the columns in the table above and Λ is either Equation (1.14) or Equation (1.15) as explained above. The choice of the logarithm is dictated by the assumption that the noise is multiplicative. The minimization was done by using standard MATLAB routines ("lsqnonlin" and "nlinfit") leading to the same results. By using "nlinfit" we were able to estimate also the 95% confidence interval, reported together with the estimated values in Supplementary Table 1.2:

Since the data concerns averages whereas the biochemical model for the degradation pathway is based on a single molecule perspective, it is to be expected that the confidence intervals are relatively large. In addition, the data used had to be extracted graphically from the plots published in [10] and thus are expected to contain random errors related to reading the data from the plots. Despite these shortcomings, any other of the alternative

condition	E[T] min	% via miRISC	100π_0	100π_1	100π_2
(-)	44.7	0.0	78.8	21.2	0
R	55.6	7.4	58.6	15.7	25.6
RP	52.7	5.5	63.1	16.9	19.9
RN	30.1	63.6	42.7	11.4	45.9
RNP	24.2	84.1	23.2	6.2	70.6

Table 1.3: Steady state properties. For each condition we compute the average lifetime of the mRNA (second column), the percent of molecules that are degraded via miRISC (third column) and the distribution of mRNA among the three main biochemical states of Supplementary Figure 1.2.

models we have tested were not able to fit the data (not shown), thus indicating that the networks proposed in this study are structurally correct.

1.1.5 Steady state properties and statistics

We can associate to each path of the type drawn in Supplementary Figure 1.1 its average absorption time given by

$$\mathrm{E}[T_n] = \sum_{j=1}^{n} \omega_j^{-1}, \tag{1.17}$$

which allows us to compute the average lifetime associated to each of the networks discussed in our paper. Therefore, the average lifetime associated to the decay pattern described by Equation (1.14) is given by

$$\mathrm{E}[T^{(-)}] = \frac{\nu + \lambda}{\nu(\lambda + \mu)}, \tag{1.18}$$

where the upper index (-) indicates that this average lifetime refers only to the negative control decay data. Conversely, the average lifetime associated to the decay patterns described by Equation (1.15) is given by

$$\mathrm{E}[T^{(RX)}] = \frac{\lambda\mu' + \lambda'\nu + \mu'\nu}{\mu'\nu(\lambda + \mu + \lambda')}, \tag{1.19}$$

where the upper index (RX) indicates that this refers to all analyzed networks where miRNA is involved together with none, one or both factors PAN3 and NOT1. When the rates are known one can compute the percent of target mRNA that are involved in miRNA mediated degradation, shown in the third column of Table 1.3. Notice that the halftimes computed in [10] are very similar in value to our average lifetimes. The average lifetimes, however, have a clear meaning here because the steady state amount of mRNA is proportional to the average lifetime, whereas the halftimes only have a meaning in the framework of the exponential decay.

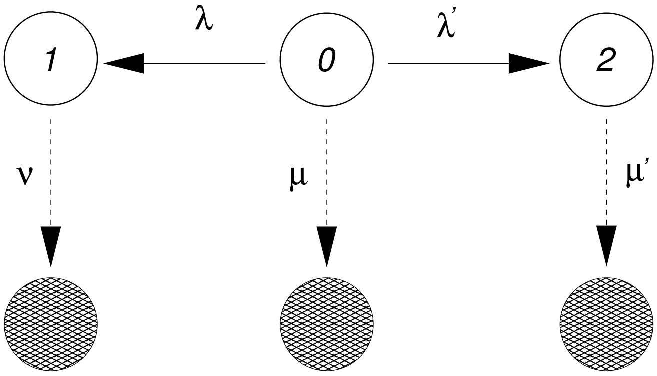


FIGURE 1.2: The network studied in this contribution is made of three states, 0, 1 and 2, whereas state 2 cannot be reached when $\lambda' = 0$ as in the case of the negative control. Using the rates and the Master equation we can compute the stationary distribution of the mRNA's in the three states for all models considered.

Another quantity of biological interest is the percent of mRNA in the different biochemical states of our network at steady state. The three states of the network are drawn in Supplementary Figure 1.2 and the stationary probabilities π_0, π_1 and π_2 can be computed from the Master equation by redirecting the arrows with rates μ, ν and μ' towards the initial state 0. This leads to the stationary distribution

$$\begin{aligned} \pi_0 &= \nu\mu'/(\nu\mu' + \lambda\mu' + \lambda'\nu) \\ \pi_1 &= \lambda\mu'/(\nu\mu' + \lambda\mu' + \lambda'\nu) \\ \pi_2 &= \nu\lambda'/(\nu\mu' + \lambda\mu' + \lambda'\nu) \end{aligned} \tag{1.20}$$

which have a clear limit when $\lambda' \to 0$ for the negative control case. The values of the probability distribution can be used to estimate how many mRNA should be found in the cell in the three different biochemical states, 0, 1 and 2 in Figure 1.2. Notice that the % of mRNA degraded through miRNA targeting can also be computed in terms of normalized fluxes:

$$\%\text{ via miRISC} = \frac{\pi_2\mu'}{\pi_0\mu + \pi_1\nu + \pi_2\mu'}, \tag{1.21}$$

consistent with the values computed earlier, given in the third column of Supplementary Table 1.3

1.2 Why the originally hypothesized pathway is not consistent with the experimental data

There are several reasons why the biochemical network proposed in [10] is not an adequate model for miRNA-mediated mRNA degradation:

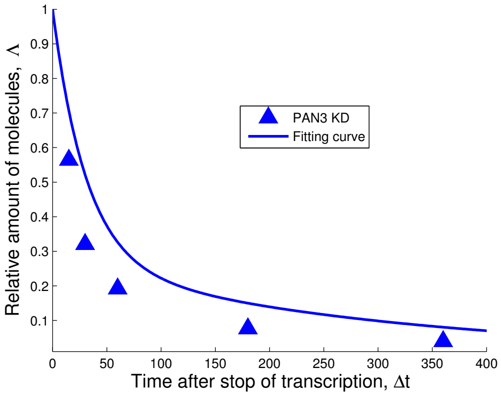


FIGURE 1.3: Fit of the data of the decay pattern when PAN3 has been knocked down. The model used here is the one of [10] complemented with a pathways for alternative routes to degradation. In our interpretation this means that $\lambda' = \lambda_R$ in Supplementary Figure 1.2 is fixed with the data of the experiment with both PAN3 and NOT1 knocked down, while also λ, μ and ν are kept fixed. The only parameter left free to be fixed with this set of data is therefore μ'. However, since too few mRNAs are targeted by miRISC alone, the extra parameter μ' is not sufficient to fit the data.

1. The network proposed in [10] does not contain all the state transitions that are necessary for fitting both the negative control data and the data where the miRNA is knocked-down.

2. Through the action of miRISC alone, only about 7% of the target mRNA are degraded via the action of the miRNA (see second row, third column in Supplementary Table 1.3)

3. The hypothesis formulated by [10] foresees that NOT1 and PAN3 bind to the mRNA after the binding of miRISC. To mimic this scenario, we can thus impose $\lambda' = \lambda_R$ in Supplementary Figure 1.2, thereby implying that only one degree of freedom, namely the parameter μ', remains available for being tuned through the data fitting procedure. As seen in Supplementary Figure 1.3 and 1.4, the model proposed with this assumption does not provide a good fit to the data.

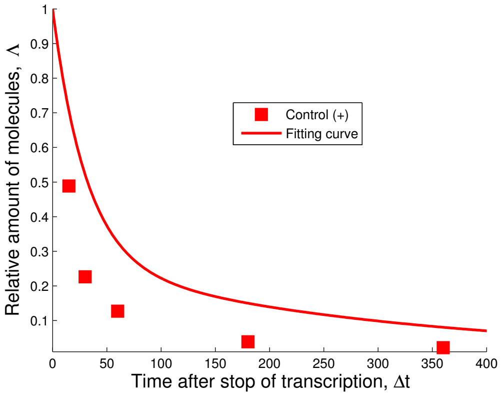


FIGURE 1.4: Fit of the data of the decay pattern when all factors are expressed (no knock downs). The model used here is the one proposed in [10] complemented with a pathway for alternative routes to degradation. In our interpretation this means that $\lambda' = \lambda_R$ in Supplementary Figure 1.2 is fixed with the data of the experiment with both PAN3 and NOT1 knocked down, while also λ, μ and ν are kept fixed. The only parameter left free to be fixed with this new set of data is therefore μ'. However, since too few mRNAs are targeted by miRISC alone, the extra parameter μ' turns out to be not sufficient to fit the data.

Supplementary 2

Bacteria differently regulate mRNA abundance to specifically respond to various stresses

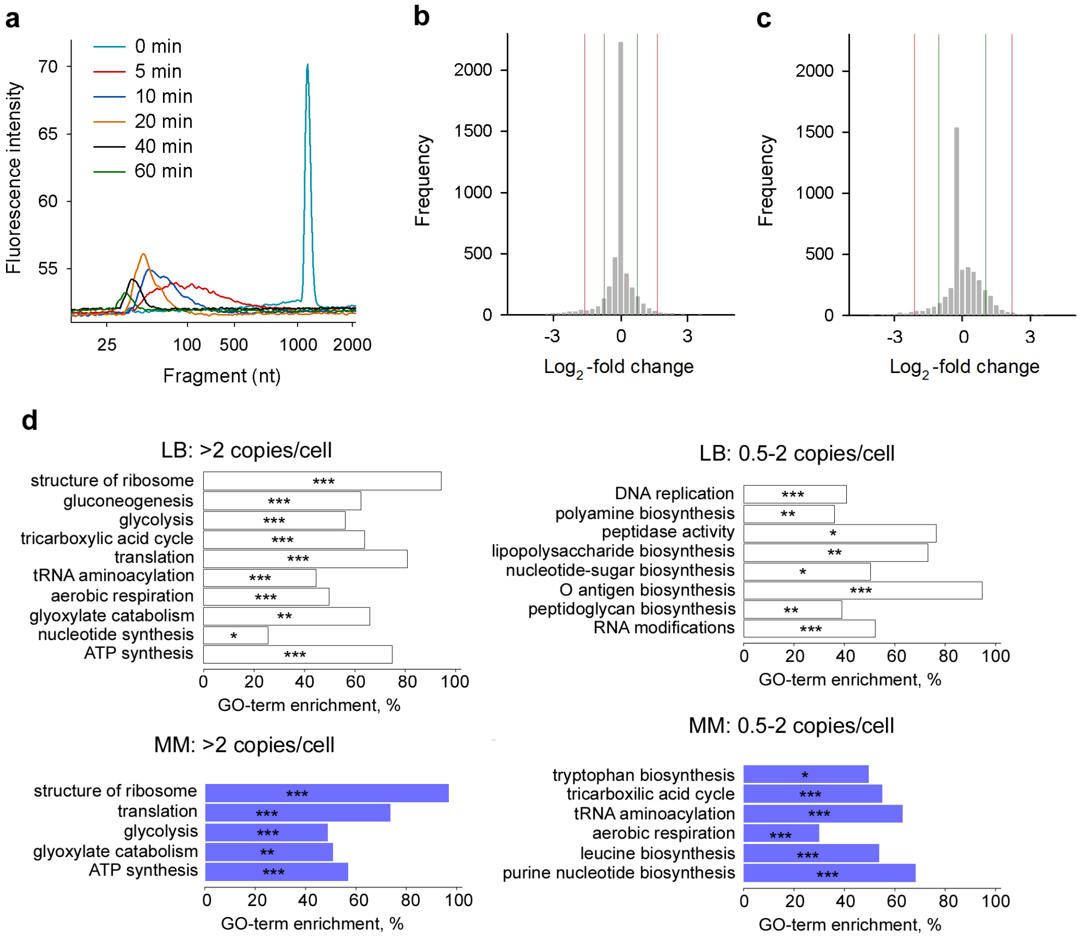
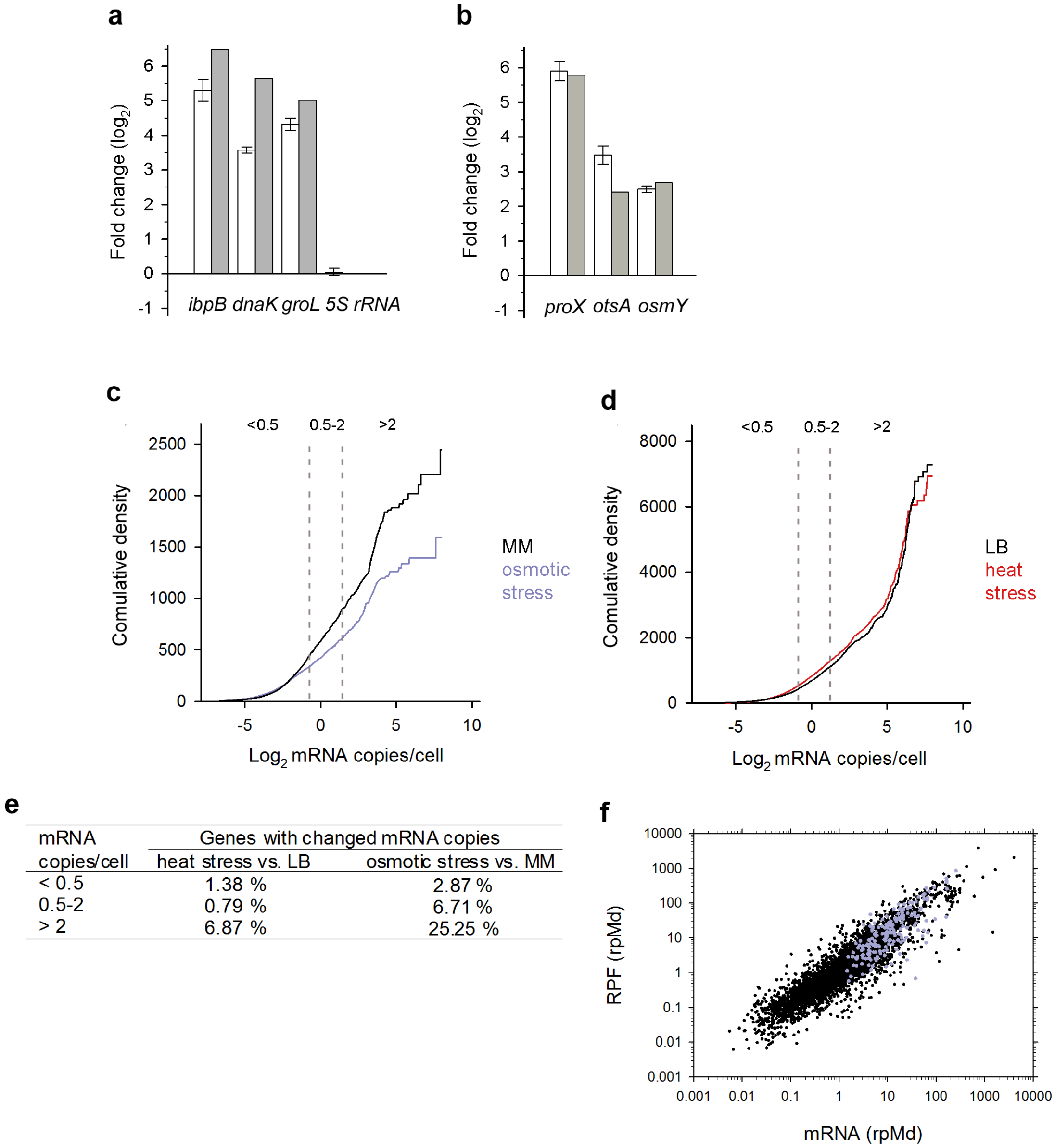


FIGURE 2.1: Optimization of the conditions for data generation, reproducibility and GO analysis. (a) Optimization of the time for mRNA random fragmentation using GAPDH mRNA. The optimal time for alkaline fragmentation is 40 min; a symmetric peak is formed indicative of a relatively uniform distribution of the digested mRNA with fragment maximum length of around 30 nt; this length corresponds to the length of RPFs. Upon longer incubation the intensity of the peak decreased and the maximum shifted to smaller lengths of the fragments. (b, c) Reproducibility of randomly fragmented mRNAs (b) and RD (c) of two biological replicates of *E. coli* grown in LB medium. The histogram of log2-change of each gene between the two replicates is plotted. Red line delineates 99th percentile and green line 95th percentile overlap between the log2-distibutions of the two biological replicates. 95th percentile was used as a statistical threshold for the log2-fold change analysis between different conditions. The log2-fold change at 95% for mRNA (b) equals to 1.6928 and for RD (c) 2.0653. For the replicates the Spearman correlation coefficients of the normalized read counts (rpkMd) are: 0.85 (b) and 0.9 (c). The Spearman correlation coefficients indicate that the ribosomal profiling and RNA-Seq analyses are reproducible. (d) GO terms with significant enrichment among the genes with high copy numbers (>2 copies/cell) and mRNA with ~1 copy/cell of bacteria grown in LB (white) and MM (blue).
***, p<0.001, **, p<0.01, *, p<0.05.



mRNA	Genes with changed mRNA copies	
copies/cell	heat stress vs. LB	osmotic stress vs. MM
< 0.5	1.38 %	2.87 %
0.5-2	0.79 %	6.71 %
> 2	6.87 %	25.25 %

FIGURE 2.2: Differences in the expression under stress. The mRNA levels of specific stress-related genes were quantified under thermal (a) and osmotic (b) stress by real-time qRT-PCR (white bars) and compared to the RNA-Seq data set (gray bars). The values are presented as a fold change (log2; mean ± SEM) compared to the control cells grown in the corresponding control medium (i.e., heat stress vs. LB (a) and osmotic stress vs. MM (b)). (c,d) Cumulative plots of the copy number changes upon osmotic stress exposure compared to growth in MM (c) and upon heat shock compared to LB (d). Dashed vertical lines correspond to the gene groups defined in Figure 2.1c,d, respectively. (e) Percentage of genes with mRNA copy number changes within each group (the three expression groups are defined as in Figure 2.1b). Cells exposed to heat stress were compared to growth in LB and cells subjected to osmotic upshift were compared to growth in MM. (f) Scatter plot of the normalized RPFs of each gene with the sequencing reads from the normalized reads of the randomly fragmented mRNAs from cells grown in MM (as in Figure 2.2b). Spearman correlation coefficient is 0.91. Blue dots depict the genes in the expression group of >2 copies/cell with significant mRNA copy number change between MM and osmotic upshift.

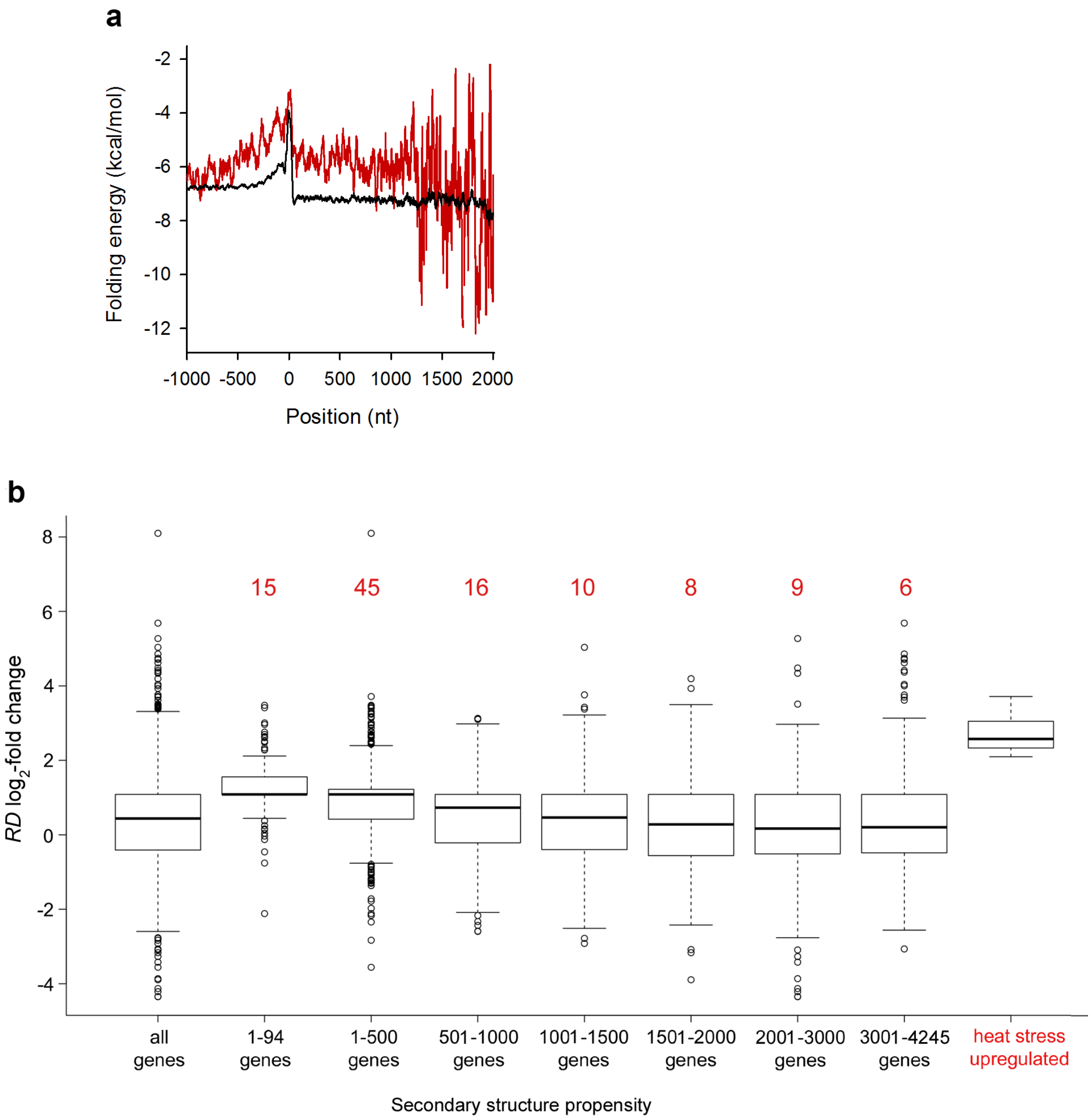
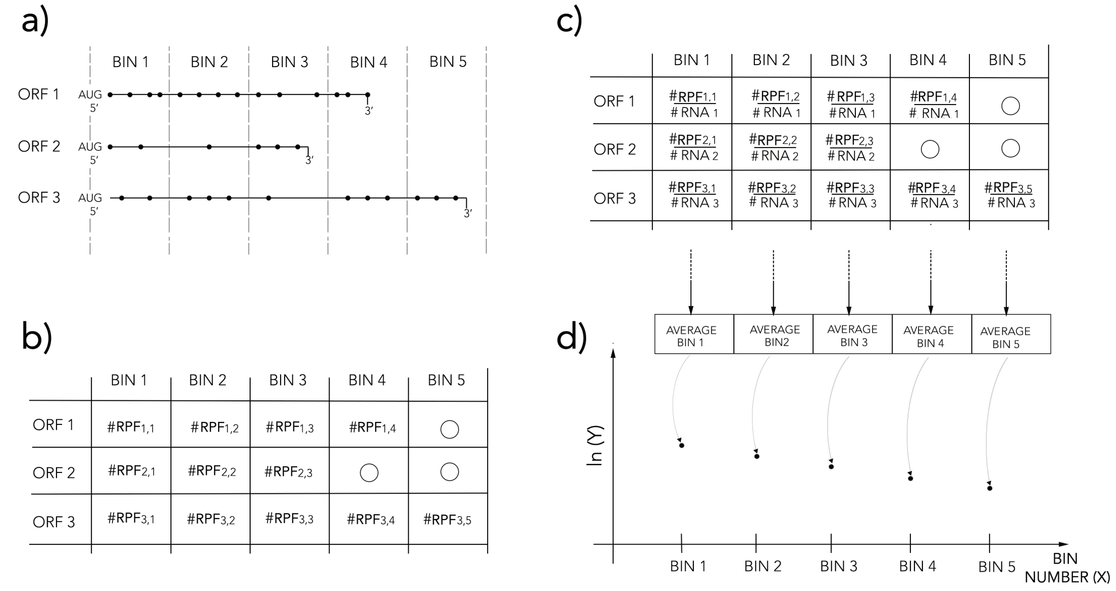


Figure 2.3: Secondary structure analysis. (a) Average folding energy of translationally upregulated genes under heat (red) compared to all protein-coding genes in *E. coli* (black). Zero is the first nucleotide of the start codon in each ORF. This figure is similar to Figure 2.4b; however it shows the folding energy for the whole gene length. The highly scattered area at sequence positions >1300 nt is due to the lower number of genes with length larger than 1300 nt. (b) Boxplot analysis of the log2-changes of the RD values from LB to heat stress of different protein-coding gene groups starting with the genes with the lowest propensity to form secondary structure at the initiation start (1-500 genes) to the gene group with the highest propensity to form secondary structure (3001-4245 genes). "All genes" comprises of all the protein-coding genes in *E. coli*. The group of the translationally upregulated genes under heat stress (red) comprises 94 genes. For comparison, a gene group of the same size, the 94 least structured genes (1-94 genes), is included in the boxplot. The distribution of RD values of the heat upregulated group differs significantly from that of all genes (Kolmogorov-Smirnov-test: $p < 10^{-15}$).

Supplementary 3

Quantitative assessment of ribosome drop-off in *E. coli*

3.1 Downstream Analysis



	BIN 1	BIN 2	BIN 3	BIN 4	BIN 5
ORF 1	$\#RPF_{1,1}$	$\#RPF_{1,2}$	$\#RPF_{1,3}$	$\#RPF_{1,4}$	○
ORF 2	$\#RPF_{2,1}$	$\#RPF_{2,2}$	$\#RPF_{2,3}$	○	○
ORF 3	$\#RPF_{3,1}$	$\#RPF_{3,2}$	$\#RPF_{3,3}$	$\#RPF_{3,4}$	$\#RPF_{3,5}$

	BIN 1	BIN 2	BIN 3	BIN 4	BIN 5
ORF 1	$\frac{\#RPF_{1,1}}{\#RNA_1}$	$\frac{\#RPF_{1,2}}{\#RNA_1}$	$\frac{\#RPF_{1,3}}{\#RNA_1}$	$\frac{\#RPF_{1,4}}{\#RNA_1}$	○
ORF 2	$\frac{\#RPF_{2,1}}{\#RNA_2}$	$\frac{\#RPF_{2,2}}{\#RNA_2}$	$\frac{\#RPF_{2,3}}{\#RNA_2}$	○	○
ORF 3	$\frac{\#RPF_{3,1}}{\#RNA_3}$	$\frac{\#RPF_{3,2}}{\#RNA_3}$	$\frac{\#RPF_{3,3}}{\#RNA_3}$	$\frac{\#RPF_{3,4}}{\#RNA_3}$	$\frac{\#RPF_{3,5}}{\#RNA_3}$

FIGURE 3.1: Illustration of the core steps of our analysis method to measure ribosome drop-off rates in Ribo-seq data. **a)** The ORFs are divided in bins of equal length; **b)** For each bin, the total number of RPFs mapping on it is reported in the RPF matrix; **c)** The elements in each line of the RPF matrix are divided by the average amount of RNA-seq reads mapping in each bin of the corresponding ORF, thus obtaining the NRPF matrix; **d)** The logarithm of the average value of each column of the NRPF matrix, $\ln(Y)$, is plotted against the bin number X.

3.1.1 Bootstrap approach

To obtain an accurate estimate of R_{BS} and of its associated error without making any assumptions on the distribution of the number of RPFs, we implemented a bootstrap approach:

(i) We counted the number of elements E_X in each column of the NRPF matrix, *i.e.* the number of elements contributing to each bin (equal to the number of ORFs with at least X bins). As suggested by Figure 1 of the main text, the number E_X is not constant for all the columns, due to heterogeneous ORF length of the genes in the study.

(ii) From each column X of the NRPF matrix, we sampled a combination of E_X elements randomly with replacement, thus obtaining a matrix (call it BootStrap - BS - Matrix) that has the same dimensions of the NRPF matrix and contains the sampled elements in each column;

(iii) We computed the average for each column of the BS matrix, obtaining the vector Y_i

(iv) Given that the exponential relationship:

$$Y = A\,\mathrm{e}^{-RX} \tag{3.1}$$

holds also for Y_i and X, we computed the rate R_i as the slope of the weighted linear regression of $\ln(Y_i)$ against X.

For each of the studied datasets (listed in Table 1 of the main text), we repeated the sequence of steps described above 10^5 times, thus obtaining 10^5 values for R_i ($R_1, \ldots, R_{10^5}$) for each dataset. Each distribution of the obtained values for R_i seems to follow a Gaussian distribution (Pearson's χ^2 test with P-value < 0.01); therefore, we assumed that the average value of each R_i distribution is representative of R_{BS} for the corresponding dataset. Supplementary Figure 3.2 provides an example of an R_i distribution.

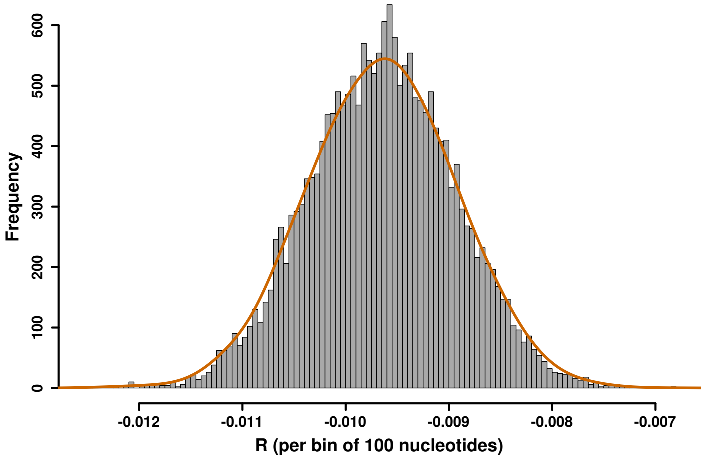


FIGURE 3.2: Sample of a R_i distribution obtained from 10^5 iterations of the bootstrapping process. The superimposed curve represents the best fitting Gaussian function. The data used for this plot come from the analysis of dataset 17 (see Table 1 of the main text).

3.1.2 Evaluation of the error

Here we review the technical details of how we estimate the errors for R. Due to the heterogeneous distribution of *E. coli* ORF lengths, the number of elements contribution to each bin is not constant through the bins; in particular, in both the NRPF and BS matrices, E_X becomes progressively small as X increases (see Figure 3.1, Chapter 3). Thus, the variance associated to the average of the E_X elements becomes progressively large and the bin average becomes a bad estimator of the correspondent element of the vector Y. On one hand, these arguments motivate us to compute the linear regression of $\ln(Y)$ vs. X by weighting each average by its associated variance (weighted linear regression). On the other hand, from the same arguments, it turns out that the reliability of our estimate of R depends both on the number of bins (*i.e.* the length of the Y vector) we consider for the linear regression and on the bin size.

To select the optimal bin size and the optimal Y length needed to obtain the best estimate of R and the associated error, we analyzed a set of simulated data for each of the 17 databases we considered. In each case, we generated a simulated dataset by redistributing the reads of each ORF according to an exponential distribution with parameter equal to a preselected nominal value. This value was chosen equal to the rate R that, for each dataset, can be obtained by applying the bootstrap approach described above. Thus, in the simulations we performed, we preserved some of the features of the original datasets, namely the total number of reads per ORF and the gene lengths.

For each one of the original datasets we considered different bin sizes ranging from 10 to 130 nucleotides and, for each bin size, we generated 1000 simulated sets of ribosome positions. For each simulated data set, we estimated the value of R 47 times, taking into account a different number of bins each time. The minimum number of bins was 2 (the minimum required for linear regression), and the maximum number of bins is 49. In each case, we obtained the best estimate of R for a bin size of 100 nucleotides and a length of the Y vector of 39.

This data, combined with the results of the bootstrap procedure, allowed us to evaluate the systematic error associated to our estimate of R. From these simulations, we conclude that the correct value of R can be obtained by adding an offset Δ from the estimate R_{BS} provided by the bootstrap procedure. The standard deviation associated to the value of R (S_R) can be computed from the square root of the sum of the variance associated to the bootstrap process (S_{BS}) and the variance associated to the offset (S_Δ) by the formula:

$$S_R = \sqrt{S_{BS}^2 + S_\Delta^2} \tag{3.2}$$

The values of Δ and S_Δ we obtained for each dataset, are reported in Supplementary Table 3.1.

As described in the text, the value of R is obtained from the R_{BS} after a correction by Δ, according to the equation:

$$R = R_{BS} + \Delta \tag{3.3}$$

Dataset ID	R_{BS}	S_{BS}	Δ	S_{Δ}
	(10^{-4})	(10^{-4})	(10^{-4})	(10^{-4})
1	112.9	8.7	16.2	3.98
2	88.5	11.7	16.2	3.84
3	96.3	7.3	16.7	3.78
4	80.2	10.0	16.1	3.61
5	39.6	8.0	16.7	3.80
6	*n.a.*	*n.a.*	*n.a.*	*n.a.*
7	79.2	6.9	16.0	3.64
8	*n.a.*	*n.a.*	*n.a.*	*n.a*
9	97.6	17.0	16.7	3.81
10	89.1	22.1	16.7	3.80
11	184.8	12.1	16.9	3.65
12	201.5	10.0	16.9	3.89
13	94.2	7.8	16.2	3.76
14	91.3	8.0	16.5	3.63
15	116.7	9.2	16.8	3.70
16	0.00	12.7	16.2	3.72
17	63.0	7.0	16.5	3.75

Table 3.1: Parameters R_{BS}, S_{BS}, Δ and S_{Δ} used for the evaluation of the error in the estimation of the drop-off rate. Column 1: Dataset ID. Column 2: the value R_{BS} estimated through the bootstrap approach. Column 3: the standard deviation S_{BS} associated to R_{BS}. Columns 4 and 5: results of the simulations performed to evaluate the error Δ and the associated standard deviation S_{Δ}. Column 4: Offset Δ. Column 5: Standard deviation S_{Δ}.

3.1.3 Comparison with other methods

Even though strikingly simple in principle, our binning strategy represents the Columbus' egg that allowed us to detect the signal of ribosome drop-off in Ribo-seq data. The other analytical approaches reported so far in the literature failed to reach this goal, essentially because the proposed binning strategy was not sensitive enough. As outlined in the main text, the usual way of evaluating ribosome drop-off in Ribo-seq data (for an example, see [68]) is by looking for a difference between the number of reads that map to two subsequent halves of each ORF. A significant difference between the two halves (fewer reads in the second half) reveals that a certain number of ribosomes has not successfully completed translation. The results of this analysis are typically illustrated through scatterplots similar to the one reported in Supplementary Figure 3.3, where the number of reads (expressed in terms of ribosome density, i.e. number of reads per nucleotide or per codon) mapping in the first half is plotted against the number of reads mapping in the second half. If there is no significant difference between the quantities reported in the two axes, the plotted points will cluster around a straight line having the slope equal to 1, as it happens in the case reported in Supplementary Figure 3.3. If fewer reads map in the second half with respect to the first half, the point corresponding to that ORF will plot below the line mentioned above.

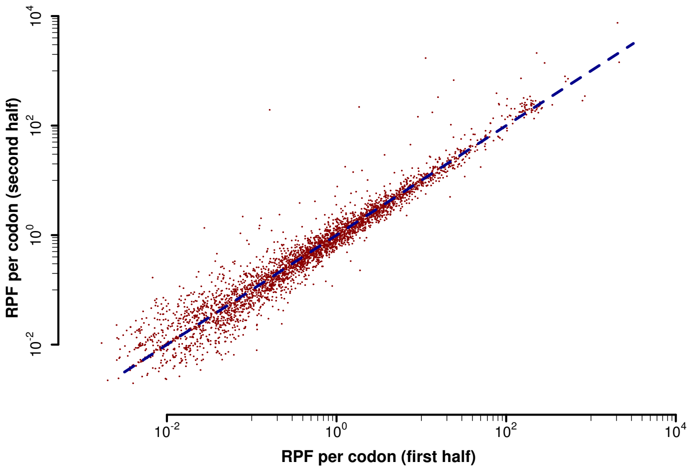


Figure 3.3: Typical scatterplot obtained from the drop-off analysis method proposed in [68]. The ribosome density (number of RPFs per codon) of the first half of each gene is plotted against the ribosome density of the second half. The clustering of the plotted points along the dashed line indicates that the ribosome drop-off rate is negligible. Data taken from [68].

While this method is mathematically sound, it has a major drawback – the sensitivity of this approach depends critically on the ORF length. When the frequency of drop-off events is not large enough with respect to length of the ORFs, the difference in ribosome density between the two halves of the ORF are too small to be detected by eye and cannot influence the correlation coefficient in a log-log scatterplot. Supplementary Figure 3.4 provides an illustration of this phenomenon.

As a consequence, if the genome of interest prevalently contains short genes, the method may not be sensitive enough to detect the drop-off in the shorter genes. Then, the

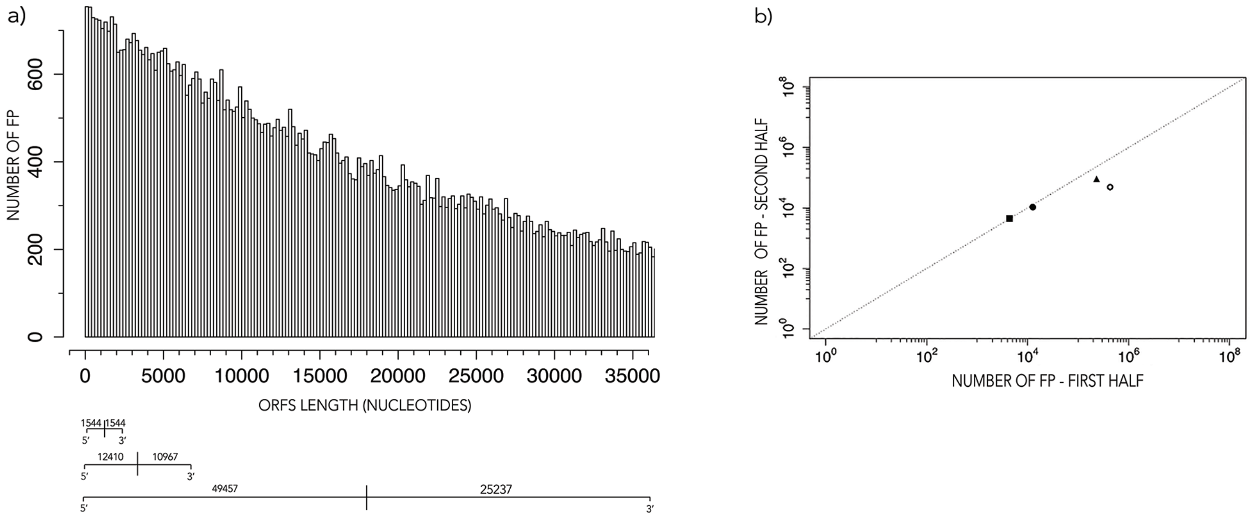
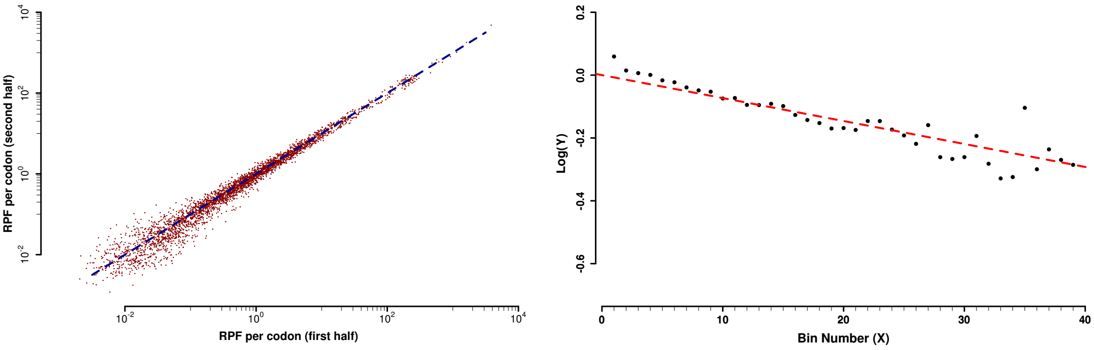


Figure 3.4: Simulation for illustrating how a relatively low drop-off rate is not detected by the analysis method proposed in [68]. **(a)**: histogram describing the simulated decrease in the number of RPFs mapping on the ORFs. This decrease is generated as an exponential decay with rate 4 × 10^{-4} per codon, corresponding to the *E. coli* drop-off rate estimated in [63]. The drawing below the histogram depicts three ORFs of different lengths: 1000 nucleotides (close to the average ORF length in E. *E. coli*), 7077 nucleotides (the maximum gene length in *E. coli*) and 3500 nucleotides, a length about 5 times longer than the maximum ORF length in *E. coli.* The numbers above the three depicted ORFs report the number of RPFs mapping on the two halves according to the distribution above. **(b)**: Scatterplot obtained plotting the the number of RPFs mapping on the two halves of the sample ORFs depicted in Figure (a); square: ORF length = 1000; solid circle: ORF length = 7077; triangle ORF length = 35000; empty circle: ORF length = 100000 (not depicted in (a)). Note that a significant deviation from the dashed line is obtained only when non-biologically possible ORFs lengths are considered .

conclusion would necessarily be that, at the genome scale, the ribosome drop-off rate is not measurable. It turns out that the sensitivity of this method is too low to measure ribosome drop off on a global level if the drop off is occurring in a biologically viable cell; we discuss the specifics of this argument later.

Conversely, our analysis is not affected by the length of the ORFs. We can illustrate this scenario at the genome-wide level: Supplementary Figure 3.5 reports the results of a simulation where we spatially redistributed the RPFs associated to each ORF in the dataset of Ref. [68]. The RPFs are distributed according to an exponential distribution with parameter r equal to 1.40 × 10^{-4}, which corresponds to the drop-off rate per codon that we estimated for this dataset. In this way, we generated an artificial dataset very similar to the one from [68] except we have tailored the distribution of the RPFs on the ORFs to explicitly mimic the presence of ribosome drop-off.

We then used this dataset as a benchmark to test the capabilities of our method and the one proposed in [68] to detect the drop-off. As shown in Supplementary Figure 3.5a, the scatterplot reporting on the differences of the number of RPFs in the two halves of each ORFs shows a clustering around the straight line with slope equal to 1, which would deceptively suggest that no drop-off events occurred. The plot resulting from our analysis (Supplementary Figure 3.5b) is correctly consistent with the presence of drop-off, even when it occurs at a low rate. Interestingly, if we repeat the simulation described above in the case of an hypothetical genome with genes whose length is markedly longer than those in *E. coli*, the method proposed in [68] successfully detects the signals of drop off.



(A) (B)

Figure 3.5: Results of a simulation in which the distribution of the RPFs on the *E. coli* ORFs was artificially set with a drop-off rate of 1.40×10^{-4} per codon. **(a)** using the method proposed in [68] the drop-off is not detected (the plotted points cluster along the dashed line); **(b)** our method allows the measurement of the drop-off rate corresponding to the slope of the dashed line, obtained through the approach described in the text. The ORFs length is measured in number of bins of 100 nucleotides. The plot includes only the first 39 bins that we considered in our analysis. To facilitate the comparison with the similar graphs present in the paper we shifted the plot so that the y-intercepts of all plots will match.

Thus, it turns out that the sensitivity of the method proposed in [68] suffices only when the average gene length goes beyond a biologically reasonable threshold. Indeed it is important to notice that for a given drop-off rate, the translation process remains reliable only if the gene lengths remain bounded within certain limits, roughly by $1/r$. If we assume a drop-off rate (or, analogously, a drop-off probability) which is constant along the whole length of the various mRNAs, the distribution of the RPFs density will decrease along the messengers according to an exponential distribution. In this case, the probability P_S that a ribosome will reach the stop codon located L codons away from the start codon (technically the survival probability) is:

$$P_\text{S} = (1-r)^L \sim \exp(-rL)\,, \tag{3.4}$$

where r is the drop-off rate per codon, with $r \ll 1$. Thus, the probability for a ribosome to successfully complete the translation process sharply decreases with the gene length. If we set r at the value of $4 \cdot 10^{-4}$ per codon, as estimated in [63], this probability falls under 50% for genes longer than 1700 codons, meaning that, on average one ribosome over two will drop-off the mRNA. In other words, the magnitude of the drop-off rate represents an important constraint for the possible genes lengths in living organisms, because exceptionally long genes would never be reliably translated. This scenario is consistent with the experimental results presented in [66], where the translation efficiency was observed to markedly decrease when progressively longer β-galactosidase gene constructs were assayed to detect ribosome drop-off.

3.2 Statistical tests

3.2.1 Computing the Confidence Interval

The 99.9% confidence interval (CI) was computed through the equation:

$$\mathrm{CI} = r \pm \mathrm{Z}_{0.005} \cdot S_r \tag{3.5}$$

where S_r is the standard deviation associated to the estimation of r and $\mathrm{Z}_{0.005}$ is the Z-score corresponding to $1 - \frac{0.999}{2}$. The values we obtained for the confidence intervals provide us with two important clues about the accuracy of our estimate and the features of r. First, the relatively small value of $\mathrm{Z}_{0.005} \times S_r$, often referred to as the margin of error, indicates that our approach yields accurate estimate of r. Moreover, the definition of CI tells us that the "true value" of r (i.e. the average of the population of all possible r) lies between the boundaries of the CI with a probability of 99.9 %; there is only a 0.1 % probability that the values outside of the CI are a reliable estimate of r. The 4^{th} column of Table 2 of the main text shows that in all the cases but one (dataset 16) the values close to 0 are not in the range of the CI. This suggests that, in these cases, the drop-off rate r is significantly different from 0.

3.2.2 Z-test for the mean

To check whether the values we obtained for the drop-off rate per codon (r) were significantly different from 0, we performed a series of Z-tests for the mean. In particular, we evaluated the probability that r belongs to the normal distribution H_0 having the average $\mu_0 = 0$ and the same standard deviation S_r associated to r. In other words, we verified whether we can reject the null hypothesis ($\mathrm{H}_0 : r = 0$), which would indicate that the alternative hypothesis ($\mathrm{H}_1 : r \neq 0$) is more likely to be true. To do this, we computed the Z-score for each r hypothesizing that it belongs to H_0 ($\mathrm{Z}_{r|\mathrm{H}_0}$) through the equation:

$$\mathrm{Z}_{r|\mathrm{H}_0} = \frac{r-0}{S_r} \tag{3.6}$$

and we compared it to $Z_{0.01}$ which is the Z-score corresponding to the significance level 0.01 (probability of type 1 error or false positive rate). To check whether the significance level we chosen was meaningful for our purposes, we evaluated the type 2 error (false negative rate) and the power of the test, considering the alternative hypothesis in which r belongs to a normal distribution with average $\mu_a = r$ and S_r. The detailed results of these tests are reported in Supplementary Table 3.2.

3.2.3 The ANOVA test

To detect any possible significant difference between the values of r we measured, we considered all the values of r significantly different from 0 and, through the ANOVA test, we checked the null hypothesis that all of them are approximately equal, against the alternative hypothesis that there are at least two values of r that are significantly different. For the ANOVA test, we considered 13 groups (one for each estimated r significantly different from 0) each composed by 10^5 elements, i.e. , the number of elements composing each r_i distribution. Thus, the degrees of freedom "between" turned out to be 12 while the degrees of freedom "within" resulted to be 12999987. We set the significance level to 0.001 and the test indicates that we should reject the null hypothesis.

3.3 Drop-off rate and growth medium

According to [149], the kinetic properties of ribosomes are influenced by the growth medium. In particular, cells cultured in the same media should be characterized by ribosomes with very similar features. To check whether this holds for ribosome drop-off rate, we used a two-tailed Z-test to compare the values of r we obtained from samples coming from different series (*i.e.* from different laboratories) referring to experiments in which the bacteria were grown in the same medium in non-stressed conditions (control cultures). For the Z tests, we choose a significance level of 0.005. Supplementary Table 3.3 reports the results of the comparisons between samples referring to the rich medium (MOPS) culturing conditions.

By comparing columns 2, 3 and 4, in two of three cases the Z-score falls into the rejection area (*i.e.* the Z-score is out of the boundaries delimited by the Bonferroni-corrected significance levels). Hence, in spite of the fact that the cell cultures are grown in the

Dataset ID (Ref. Table 1)	**Drop-off rate** Per codon ($\times 10^{-4}$)	$Z_{0.01}$	$Z_{r\|H_0}$	**Power** π
1	2.89	2.33	10.2	0.97
2	2.16	2.33	6.12	0.96
3	2.40	2.33	9.80	0.98
4	1.91	2.33	5.95	0.93
5	0.69	2.33	2.61	0.83
6	*n.a.*	2.33	*n.a.*	*n.a.*
7	1.88	2.33	8.00	0.90
8	*n.a.*	2.33	*n.a*	*n.a.*
9	2.43	2.33	4.65	0.96
10	2.18	2.33	3.24	0.91
11	5.05	2.33	13.3	0.98
12	5.56	2.33	17.4	0.99
13	2.33	2.33	8.93	0.98
14	2.24	2.33	8.48	0.96
15	3.01	2.33	10.2	0.94
16	0.00	2.33	1.22	0.82
17	1.40	2.33	5.88	0.98

Table 3.2: Results of the (right tail) Z-tests to assess whether the drop-off rates are significantly different from 0. Columns 1 and 2: GEO coordinates of the datasets (Series, Ribo-seq sample, RNA sample). Column 3: Drop-off rate per codon. Column 4: percentile of the standard normal distribution corresponding to a rejection area (right) of 0.01. Column 5: Z-score associated to the comparison between the distribution with average r and standard deviation S_r versus the null distribution with average 0 and the same standard deviation S_r. Column 6: power of the corresponding Z-test.

same medium, the basal drop-off rate turns out to be different sometimes, implying that the experimental variability or some differences in the experimental protocols may have affected the value of the ribosome drop-off rate, at least in some cases.

In Supplementary Table 3.4 we report the results of the comparison between samples referring to bacteria grown in the minimal medium.

In this case the Z-score is consistent with the hypothesis of significantly equal drop-off rates.

Compared samples	**Z score**	**Significance level**	Z_B
(Sample ID)		($\pm Z_{0.0025}$)	
15 vs. 17	4.24	± 2.81	± 3.14
5 vs. 17	2.00	± 2.81	± 3.14
5 vs.15	5.85	± 2.81	± 3.14

TABLE 3.3: Results of the Z-tests for comparing the drop-off rates of samples coming from different GEO Series (different laboratories), referring to cultures in Rich (MOPS) Medium. Column 1: Samples ID, referring to Table 1 of the main text. Column 2: Z-score computed from the comparison of the drop-off rates. Column 3: percentiles of the standard normal distribution corresponding to a total rejection area of 0.005. Column 4: percentiles of the standard normal distribution corresponding to a total rejection area of 0.005, corrected according to the Bonferroni method.

Compared samples	**Z score**	**Significance level**	Z_B
(Sample ID)		($\pm Z_{0.0025}$)	
9 vs. 3	0.05	± 2.81	*n.a.*

TABLE 3.4: Results of the Z-tests for comparing the drop-off rates of samples coming from different GEO Series, referring to cultures in Minimal Medium. Column 1: Samples ID, referring to Table 1 of the main text. Column 2: Z-score computed from the comparison of the drop-off rates. Column 3: percentiles of the standard normal distribution corresponding to a total rejection area of 0.005. Column 4: percentiles of the standard normal distribution corresponding to a total rejection area of 0.005, corrected according to the Bonferroni method.

Summing up, the outcomes of our tests do not provide us a clear response about possible correlations between the ribosome drop-off rate and the growth medium. Nevertheless, given the variability we observed, our analysis reveals an important information: even though cells are cultured on the same medium, the data coming from different laboratories might unpredictably be significantly different in terms of ribosome drop-off rates, possibly due to differences in the experimental protocols.

3.4 Complete plots

In this Section we report all the plots referring to the databases we analyzed. For the sake of readability, the plots reported in the main text are cut at the 39th bin and shifted vertically by a value corresponding to the intercept of the fitting line. These modifications are not present in the plots reported hereafter.

3.4.1 Datasets 9, 11 and 13 : Ethanol-induced stress

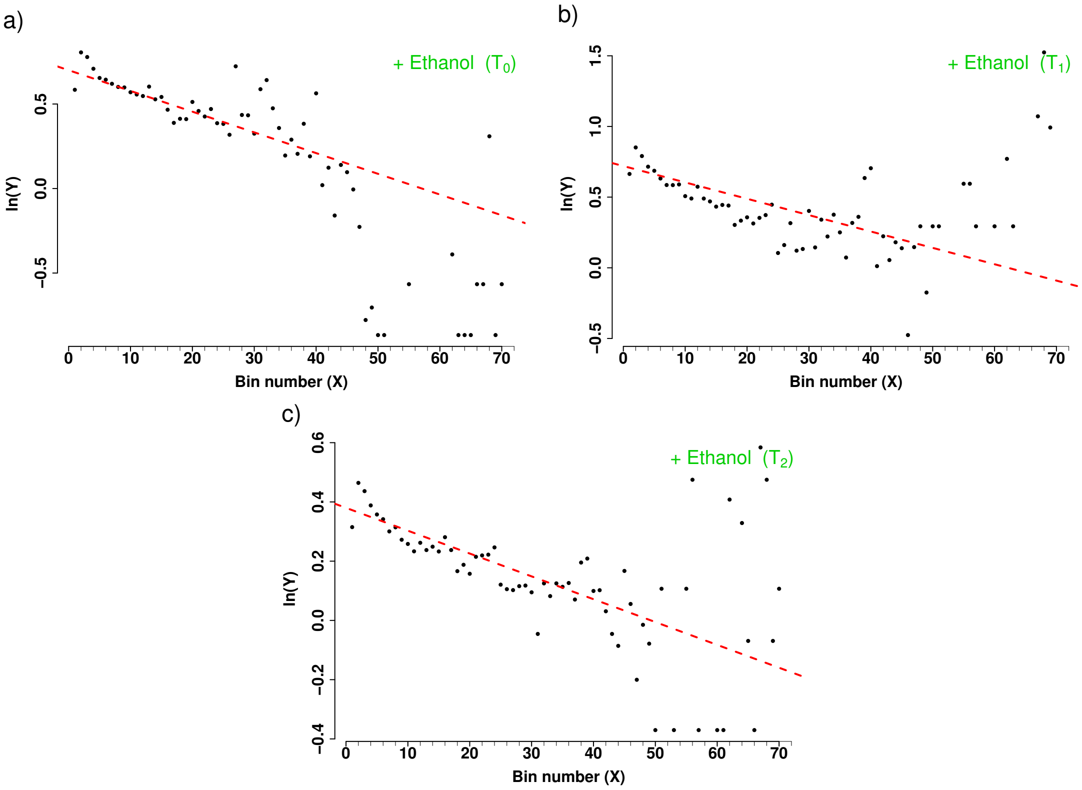


FIGURE 3.6: Plot of the vector Y vs. the number of bins (X). The slopes of the dashed lines correspond to the drop-off rate r reported in Table 1 of the main text. **a)**: Dataset 9 - Control (T_0). **b)**: Dataset 11 - T_1, after 10' of ethanol stress. **c)**: Dataset 13 - T_2, after 70' of ethanol stress.

3.4.2 Datasets 5, 6, 7, 8: Amino acids starvation

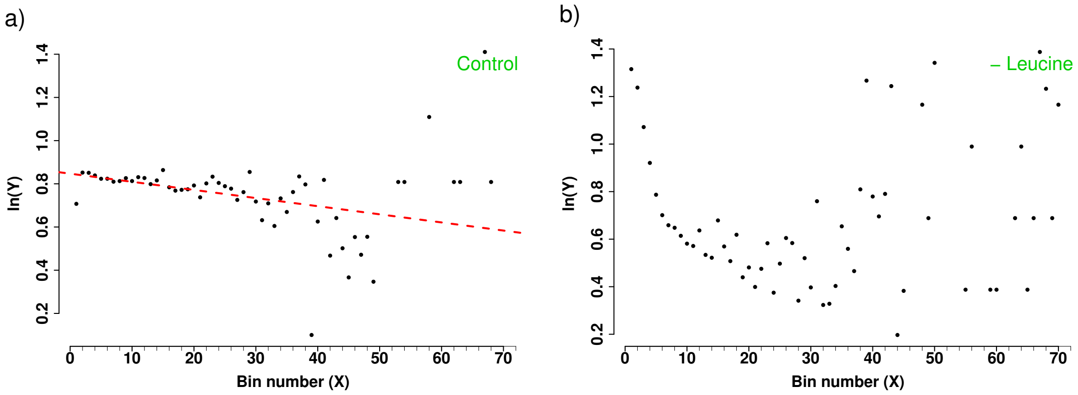


FIGURE 3.7: Plot of the vector Y vs. the number of bins (X). The slope of the dashed line corresponds to the drop-off rate r reported in Table 1 of the main text. **a)** Dataset 5 - Control (MOPS - Rich medium) **b)** Dataset 6 - Leucine starvation. In this case, due to the poor fit with a single exponential model, we could not compute r. Thus, the regression line is not represented here.

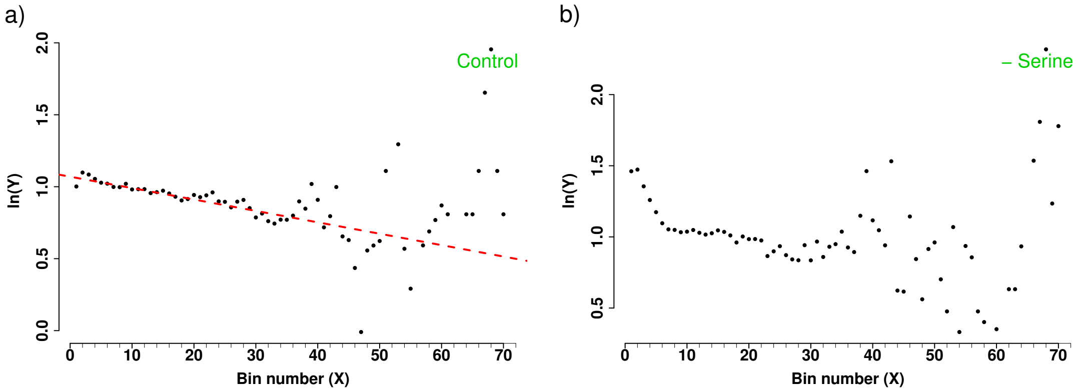


FIGURE 3.8: Plot of the vector Y vs. the number of bins (X). The slope of the dashed line corresponds to the drop-off rate r reported in Table 1 of the main text. **a)** Dataset 7 - Control (MOPS - Rich medium) **b)** Dataset 8 - Serine starvation. In this case, due to the poor fit with a single exponential model, we could not compute r. Thus, the regression line is not represented here.

3.4.3 Datasets 15 and 16: a novel σE -induced sRNA

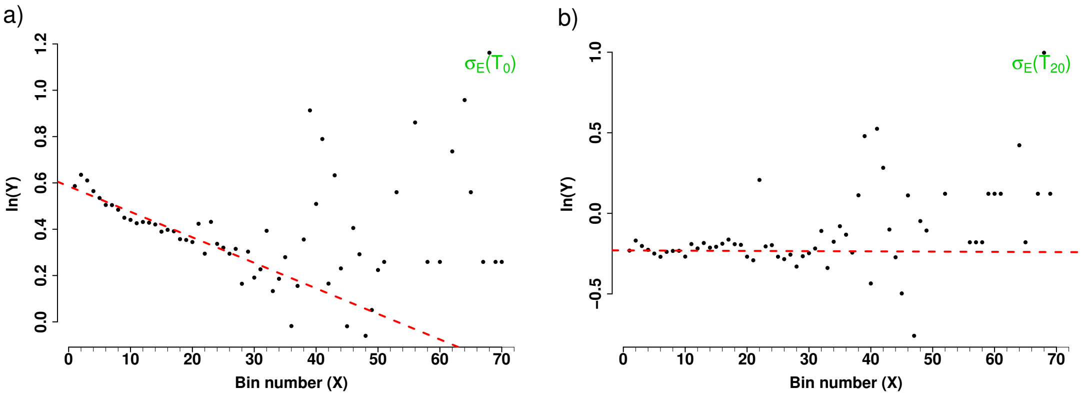


FIGURE 3.9: Plot of the vector Y vs. the number of bins (X). The slope of the dashed lines correspond to the drop-off rates r reported in TTable 1 of the main text. **a)**: Dataset 15 - Control (T_0). **b)**: Dataset 16 - T_1, after 20 minutes of σ^E over expression induction.

3.4.4 Datasets 1, 2, 3 and 4: Heat and Osmotic stress.

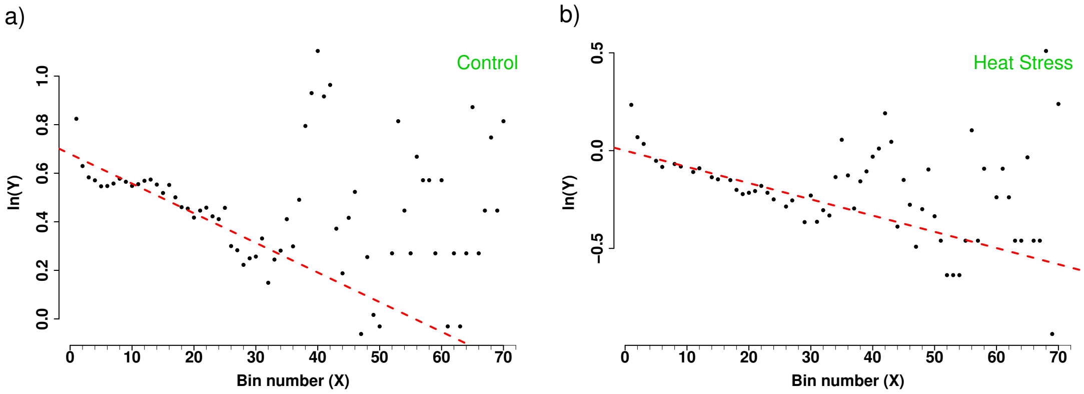


FIGURE 3.10: Plot of the vector Y vs. the number of bins (X). The slope of the dashed line corresponds to the drop-off rate r reported in Table 1 of the main text. **a)** Dataset 1 - Control (MOPS - Rich medium) **b)** Dataset 2 - Acute heat stress (47°C for 7')

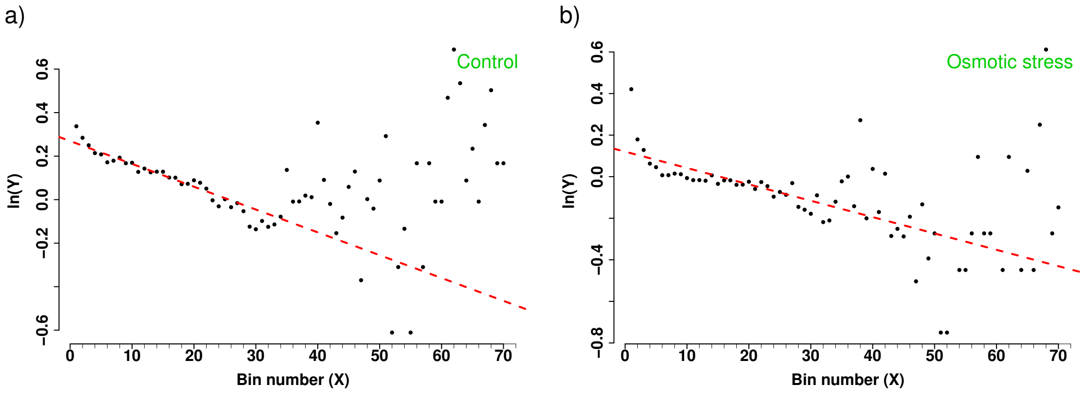


FIGURE 3.11: Plot of the vector Y vs. the number of bins (X). The slope of the dashed line corresponds to the drop-off rate r reported in Table 1 of the main text. **a)** Dataset 3 - Control (Minimal medium) **b)** Dataset 4 - Osmotic stress (NaCl 0.3M for 20' at 37°C).

Supplementary 4

Global quantification of cellular protein degradation kinetics

A version of this chapter has been submitted to Cell as:
McShane E, Sin C, Zauber H, Wells JN, Donnelly N, Wang X, Hou J, Chen W, Storchova Z, Marsh JA, Valleriani A, Selbach M. Global quantification of cellular protein degradation kinetics.

4.1 ^{35}S cysteine pulse-chase in combination with AHA or methionine

NIH 3T3 mouse fibroblast cells (ATCC) were grown in SILAC DMEM (life technologies) complemented with glutamine (Glutamax, life technologies) and 1% Penicillin and Streptomycin (life technologies), 10% dialyzed fetal calf serum (Pan-Biotech), light L-arginine (Arg0, Sigma-Aldrich) and light L-lysine (Lys0, Sigma-Aldrich) referred to as Light SILAC DMEM [238]. Confluent cells in 6-well plate wells were washed in prewarmed PBS before being starved of methionine and cysteine for 45 min in DMEM (Invitrogen) prepared as above. Cells were then pulsed for 1h with 80 μCi final concentration of 35S-Cysteine (Perkin Elmer) in combination with either 1mM AHA or 1mM methionine (Dieterich et al., 2006). Cells were then washed twice in SILAC DMEM before either being directly lyzed or chased for 6 or 24 h in cold medium with either 50 μM cycloheximide (Sigma Aldrich) or 10 fold cysteine (Sigma Aldrich) added to prevent re-incorporation of the radiolabeled amino acids. After chase, cells were scraped and lyzed in modified radio-immunoprecipitation buffer (50 mM Tris HCl (pH 7.4), 1mM EDTA, 150mM NaCl, 1 % Nonidet P-40, 0.25% Na-deoxycholate and 0.1% SDS) containing 2 fold protease inhibitor cocktail (Roche). All samples were frozen at -80 °C before being thawed on ice for 30 min in the presence of an endonuclease (Benzonase, Merck). Samples were spun down to clear cell debris. The resulting supernatant was diluted in LDS sample buffer (Invitrogen) complemented with DL-Dithiothreitol (DTT, Sigma Aldrich) before being boiled at 95°C for 5 min. Proteins were resolved by SDS-PAGE using a 10% polyacrylamide gel. Proteins were fixed on the gel by 5% Acetic acid and 50% methanol and then stained by Commassie (colloid blue stain kit, Novex). The gel was vacuum dried for 2 h using a gel drying system (Bio Rad) at 75°C. Vacuum dried gels were scanned using a scanner (Cannon) for quantification of total loaded protein amount. The radioisotope signal was measured by exposing the gel to a magnetic photostimulable phosphor plate overnight. The plate was then scanned on a phosphorimager (Typhoon FLA 9500, GE Healthcare). Radioactive and Coommassie images were quantified using the ImageQuant software (GE Healthcare). Each lane was quantified separately and background signal was estimated by marking a lane with no proteins loaded. The measured background was subtracted from the signal. The radioactive signal was then further normalized to total protein input estimated by the Coommassie staining. The experiment was repeated three times with technical triplicates for each biological replicate. Statistics and plotting was performed in Excel (Microsoft).

4.2 Enrichment efficacy of AHA labeled proteins

NIH 3T3 mouse fibroblasts were cultured in a 10 cm plate in heavy SILAC DMEM (life technologies) prepared as above but with heavy L-arginine (Arg10, Sigma-Aldrich) and heavy L-lysine (Lys8, Sigma-Aldrich) instead of their light counterparts. Cells were washed twice in pre-warmed heavy SILAC DMEM depleted of methionine (Biosera) before being starved of methionine in the same for 1h. The starvation was followed by a 3 h incubation with 1mM AHA. The AHA-labeled heavy cells were washed in ice cold PBS and then scraped in ice cold Urea lysis buffer (Click-iT protein enrichment-kit, Invitrogen), and two fold protease inhibitors (Roche) and mixed 1:1 with cells grown in Light SILAC DMEM. The lysate was treated with benzonase on ice for 20 min before being sonicated in an ice bath. Samples were spun down at 20,000 rcf and supernatant was transferred to a new tube containing alkyne agarose beads and the click reaction was performed overnight following the "Click-iT protein enrichment kit"-protocol (Invitrogen) as described earlier [239]. Proteins were reduced by heating to 70°C in the presence of 10mM DTT in SDS buffer and later alkylated by the addition of 40 mM iodoacetamide (Sigma) final concentration. Beads were sequentially washed in SDS buffer, 8 M Urea in 100mM Tris (pH8) and 80% acetonitrile by centrifugation and decanting supernatant. Proteins were digested "on bead" in 5% acetonitrile in ABC buffer first 3 h by LysC and then over night with trypsin. The peptide solution was acidified by addition of TFA before peptides were desalted and stored on StageTips as described in [240]. In short, C18 material (3M) was put in 200 μL pipette tips and activated by methanol. Organic solvents were washed away by Buffer A (5% acetonitrile and 0.1% formic acid) before peptides were loaded onto the stageTip. Salts were washed away by washing the retained peptides in Buffer A. Peptides were eluted using Buffer B (80% Acetonitrile and 0.1% formic acids) and organic solvent was evaporated using a speedvac (Eppendorf). Samples were diluted in Buffer A (5% acetonitrile and 0.1% formic acid) before peptides were separated on a 150 mm long column with an inner diameter of 75 μm packed in house with ReproSil-Pur 120 C18-AQ 3μm resin (Dr. Maisch GmbH) (referred to as "15 cm column" from now) using a 4 h linear gradient with 250 nl/min flow rate of increasing Buffer B concentration on a High Pressure Liquid Chromatography (HPLC) system (ThermoScientific). Peptides were ionized using an electrospray ionization (ESI) source (ThermoScientific) and analyzed on a Q Exactive mass spectrometer (ThermoScientific). The mass spectrometer was run in data dependent mode selecting the top 10 most intense ions in the MS full scans (Orbitrap resolution: 70,000; target value: 3,000,000 ions; maximum injection time of 20 ms) for higher energy collision induced dissociation. The resulting MS/MS spectra from the Orbitrap had a resolution of 17,500 after a maximum ion collection time of 60 ms with a target of reaching 1,000,000 ions. For one of the three replicates the peptides were pre-fractionated into 4 fractions on IPG strips using iso-electric focusing as previously described [241]. These four fractions were analyzed using 2 h HPLC-gradients.

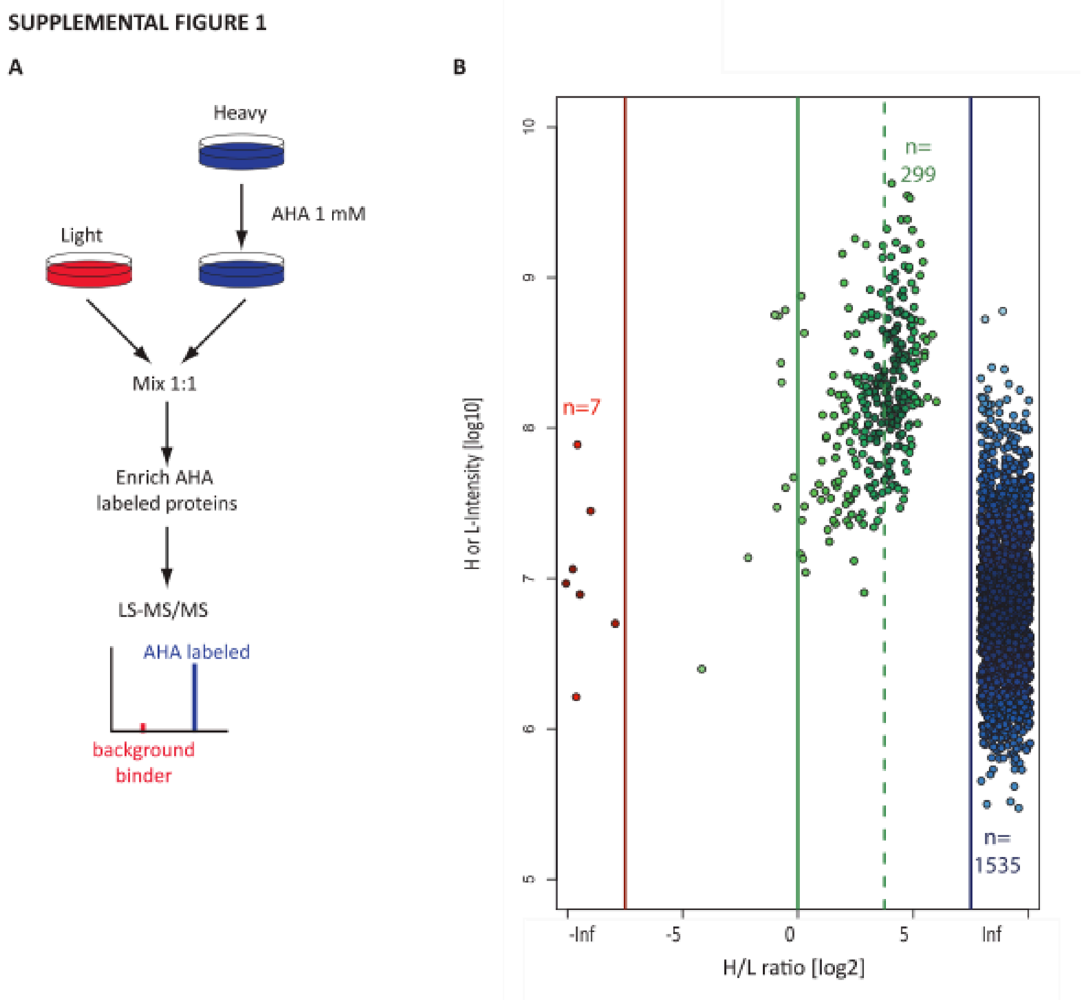


FIGURE 4.1: A) SILAC labeled heavy mouse fibroblasts were labeled for 3 h with AHA. The cells were lyzed and mixed 1:1 with Light unlabeled proteins. The AHA labeled proteins were enriched and the samples were analyzed using LC-MS/MS. In theory, only heavy labeled proteins should be enriched. B) The vast majority of proteins show up only in the heavy form (infinite ratio (Inf.), n=1535) with very few exceptions that were exclusively identified in the light form (-Inf., n=7). Light proteins included proteins with very few peptides identified and extra cellular proteins (e.g. Keratin) probably representing contaminants. A number of highly abundant proteins (n=299) showed up in both channels with a mean ratio 1:14 (green dotted line). H-intensity only was used when plotting proteins which showed up in both channels not to exaggerate their relative abundance. Shown is one representative experiment of three.

The resulting raw files were analyzed using MaxQuant software version 1.4.1.2 [73]. Default settings were kept except that ‘reQuantify‘ was deactivated and ‘match between runs‘ was turned on. Lys8 and Arg10 were set as labels and oxidation of methionines, n-terminal acetylation and deamidation of aspargine and glutamine residues were defined as variable modifications. Carbamidomethyl of c-termini was set as fixed modification.

The in silico digests of the mouse Uniprot database (2013-01) and a database containing common contaminants were done with Trypsin/P. The false discovery rate was set to 1% at both the peptide and protein level and was assessed by in parallel searching a data base containing the reversed sequences from the Uniprot database. Plotting was done using R version 2.15.1 (R Foundation for Statistical Computing, Vienna, Austria) and figures were modified in Illustrator (Adobe).

4.3 AHA pulse-chase of SILAC labelled NIH 3T3 mouse fibroblasts

Fully triple SILAC (Light: Lys0, Arg0; Medium: Lys4, Arg6; Heavy: Lys8, Arg10) labelled mouse fibroblast were grown as in the "enrichment efficacy experiment". Experiments were performed when cells reached ~25% cell density so that full confluency would not be reached during the 32 hours of chase time. For the first two replicates two 10cm plates were used per time point and for the third replicate two 15cm plates were used to increase the starting material. Cells were labeled with 1mM AHA after 1h of methionine starvation. After the 1h pulse Medium and Light cells were washed in first PBS then SILAC DMEM before being chased in the same medium. The Heavy cells that were used for time point 0 h were instead washed in ice cold PBS before being scraped in the same, spun down, supernatant was decanted before the cell pellet was frozen. After the chase the medium and light cells were also scraped and frozen. The frozen pellets were thawed and lyzed as described for "enrichment efficacy experiment". Also, the click reaction and washing of beads, denaturation, alkylation and digestion was performed as above. In one, out of the three experiments, the peptides were pre-fractionated by IEF into 12 fractions as described above. In two experiments the peptides were separated using strong anion exchange (SAX). The SAX protocol was performed as in [242]. In short SAX material (3M) was put in 200 μL pipette tips and activated by methanol. The SAX tip was then washed by high pH buffer (20 mM Acetic acid, 20 mM phosphoric acid, 20mM boric acid, pH was adjusted to 11 by titrating in 1M sodium hydroxide) before peptides were loaded onto the tip. The peptides were then eluted stepwise by decreasing the pH of the buffer in discrete steps (pH of 11 (flow though), 8, 5 and 3 all prepares as above with the addition of 0.25 M NaCl to the pH 3 buffer). The eluted peptides were stored on stageTips. IEF and SAX fractionated peptides were separated on a HPLC system as described above by either 4 or 2 h gradients with a 250 nl/min flow rate on a 15 cm column packed in house with C18 material. Peptides were ionized using an ESI source and analyzed on a Q-exactive with the above described settings. The acquired raw-files were analyzed using MaxQuant version 1.4.1.2 [73]. Default settings were kept except that 'reQuantify' was deactivated and 'match between runs' was turned on. Lys8 and Arg10 were set as heavy labels and Lys4 and Arg6 as medium heavy labels. Oxidation of methionines, n-terminal acetylation and deamidation of aspargine and glutamine residues were defined as variable modifications. Carbamidomethyl of c-termini was set as a fixed modification. The in silico digests of the mouse Uniprot database (2014-10) and a database containing common contaminants were done with Trypsin/P. The false discovery rate was set to 1% at both the peptide and protein level and was assessed by in parallel searching of a data base containing the reversed sequences from the Uniprot database. For all downstream analysis we used non-normalized SILAC ratios (see below for normalization procedure) with a minimum of 2 SILAC counts. Reverse database hits, potential contaminants and proteins only identified by site were all excluded. In addition, based on the enrichment efficacy experiments described above we applied a stringent cut-off excluding all data points smaller than 10% protein remaining.

4.4 Data normalization

Normalization is a common challenge for experiments measuring abundances – differences in starting material, labeling efficiency, instrument sensitivity, etc. are all contributors to deviations in the scale of measurements. A common normalization strategy is to normalize the data to the median value in each experiment or replicate, assuming that the median values should not change through the series of experiments. However, for pulse chase experiments, we expect the measured quantities (and thus also the median) to decay over time, thus this strategy cannot be used. In our data, we expect each time point to have an unknown and potentially different multiplicative factor which affects all measurements at that time point. Thus, we aim to estimate the average value of the multiplicative factor that affects the real data at each time point without assuming any protein degradation rates a priori while being robust to experimental errors. Our normalization scheme is based on the assumption that there are stable proteins within the pool of proteins measured, whose amounts decay very little during the time course of the experiment (Figure 4.4) [72, 207]. Without noise, the signals corresponding to these proteins would remain unchanged and equal to 100% left throughout the experiment. With noise, we can identify these very stable proteins because the Medium/Heavy and Light/Heavy SILAC ratios of these proteins should be among the highest. Using this method, we identify the most stable proteins, and then calculate the multiplication factor necessary to normalize the data for each time point such that the geometric mean of the measurements of these very stable proteins will have a signal of 100%. To find the most stable proteins, we consider proteins with data at all time points in all replica – one reason for this is that it ensures that all the potential candidates are able to contribute to the normalization factor. Furthermore, being in this subset suggests that

SUPPLEMENTAL FIGURE 2

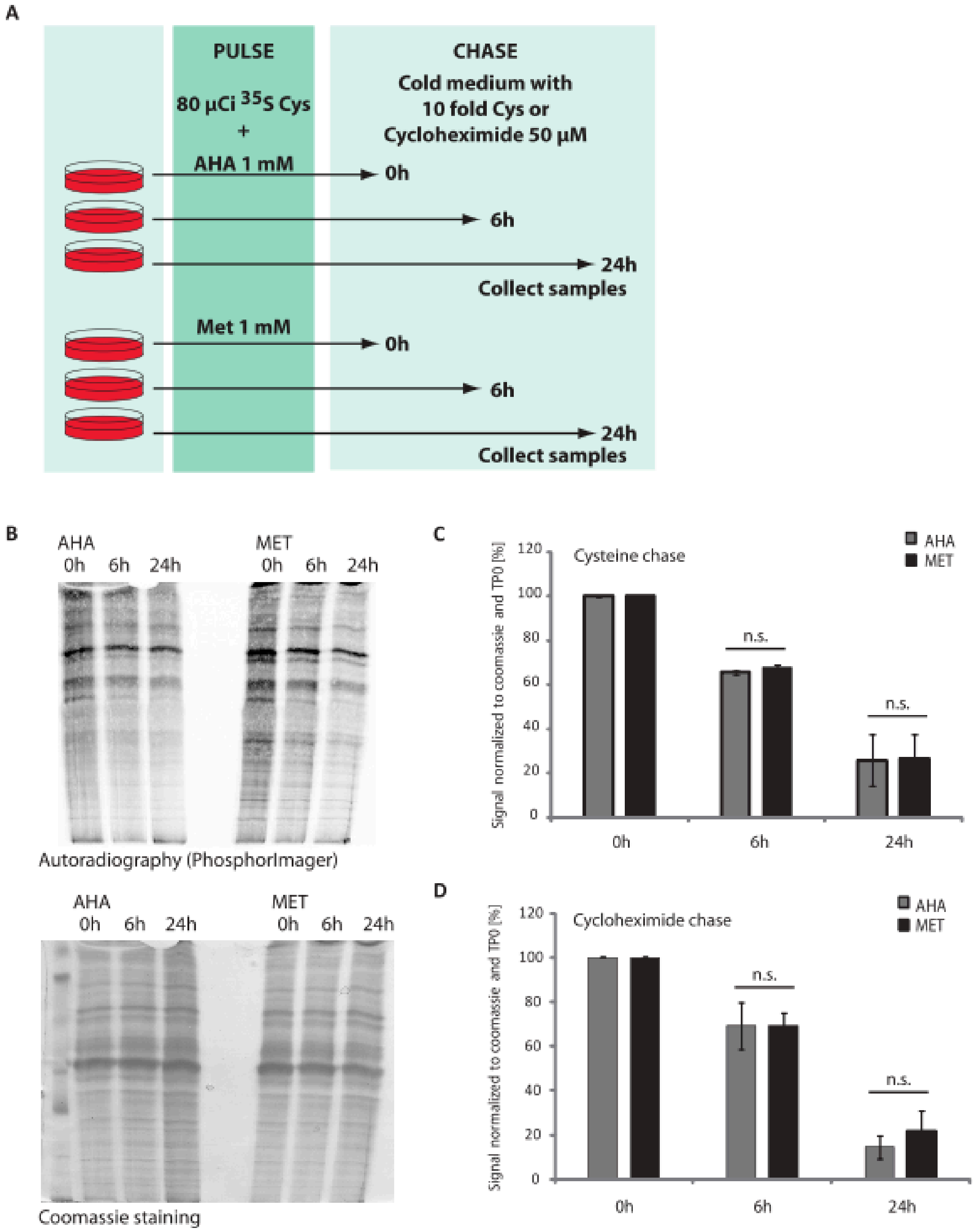


FIGURE 4.2: A) NIH 3T3 mouse fibroblasts were starved for methionine and cysteine for 45 min before being pulsed with 35S Cysteine in combination with either Methionine or Azidohomoalanine for 1h. Cells were then washed before being chased in either excess of Cysteine (10x) or cycloheximide (50 μM). Proteins were resolved on a SDS-PAGE gel before being fixed and the gel dried. B) Radioactivity was measured by a phosphorImager and total protein was assessed by coomassie blue staining. Quantification of protein degradation using C) Cysteine chase (n=3) and D) Cycloheximide (n=3). Error bars: SE, n.s.: not significant two sided t-test.

SUPPLEMENTAL FIGURE 3

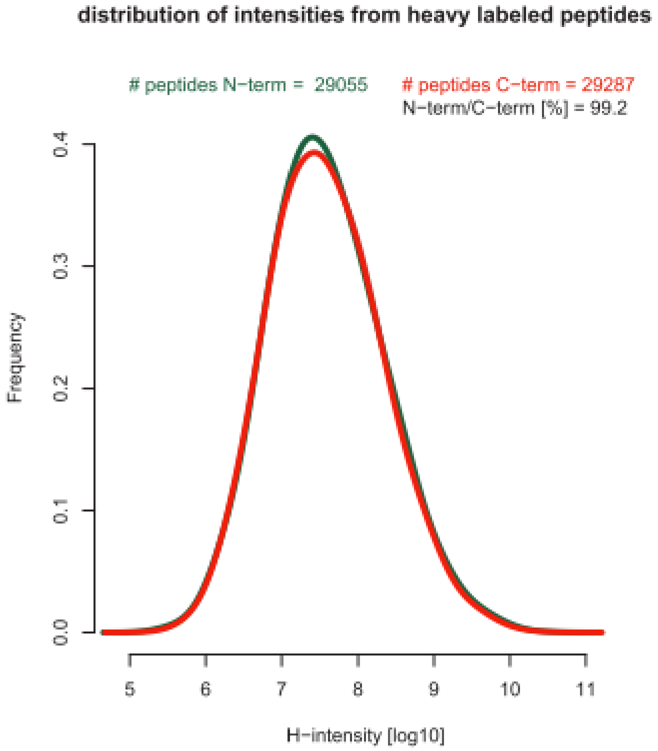


Figure 4.3: All proteins in the mouse Uniprot database were split in half resulting in a C-terminal and N-terminal part of each protein. All peptides detected in one AHA p-c experiment were matched back to the new database and sorted in the two categories N or C-terminal (peptides matching to the middle were excluded). Plotted is the distribution of heavy intensity (∼abundance after the pulse) for the detected peptides. If AHA labeling would induce premature Ribosomal fall off we expected to see a high N to C-terminal ratio in the number of detected peptides and a higher abundance of N-terminal peptides.

these proteins are reliably measurable. For each of these proteins, we assign a score, defined as

$$score_i = \sum_{t \in \{8,16,32\}} \text{Percentile Rank}_i(t) \tag{4.1}$$

Where the index i denotes the protein and PercentileRank$_i(t)$ maps the rank of each protein's signal strength (from smallest to largest, at time t) to the interval $(0, 1)$. Proteins with higher signals at each time point will have higher scores. Thus for a protein who has the highest signal at all time points would have a score equal to the number of time points, which we call maxScore (i.e. the range of scores is $(0, maxScore]$). Each protein has up to 3 scores, one from each replica. From the three scores, we calculate the deviation of the score from the maximum score:

$$dev_i = \sum_{j \in \#\text{replicas}} (\text{maxScore} - \text{score}_{i,j})^2 \tag{4.2}$$

Candidates for normalization are those proteins with the lowest deviations. This normalization scheme is based on 4 key assumptions:

1. All groups of cells (heavy/medium/light) produce and degrade proteins equally.
2. Proteins degrade at different rates, which can be differentiated in the time scale of our experiments)
3. Proteins degrading the slowest have the highest Medium and Light to Heavy ratios (and thus lowest deviations)
4. The slowest degrading proteins do not degrade at all (in the time scale of our experiment) From the data, we find the population of proteins with the lowest score deviations ($n = 200, < 5\%$ of total population) and deem these to be the stable proteins. This set is chosen intentionally large in order to mitigate the effects of outlying data points.

4.5 Parameter fitting

In this study, we consider two simple models: a 1-state model (exponential decay) and a 2-state model (non-exponential decay). In the 1-state model, proteins are in state A just after synthesis. From state A, they are degraded at the rate k_A. The system is memoryless, meaning that the life expectancy for any single protein molecule does not change as the molecules age. For true exponential decay, the data should resemble a straight line when plotted in a log-linear plot. While the one-state model is a good approximation for some decay patterns, other decay patterns have dynamics that are not well described by a one-state model [86].

In the two-state model, proteins are in state A after synthesis. From state A, the molecule can immediately degrade at the rate of k_A, or it can transition to state B with the rate k_{AB}. Molecules that reach state B are degraded with rate k_B. From the analysis point of view, one important distinction between the 1-state and 2-state models is that we lose the property of memorylessness; for 2-state models, the history of a specific molecule (which determines whether the molecule is in state A or state B) changes the life expectancy of the molecule. In short, the life expectancy of the molecule depends on the age of the molecule. In pulse-chase experiments, the duration of the pulse affects the composition of molecule ages at the beginning of the chase – for a very short pulse, the molecules synthesized in the pulse are likely to have the same age. However, for a longer pulse, molecules synthesized at the beginning of the pulse are "older") while there are some molecules which are just newly synthesized. In short, the length of the pulse must be taken into account for the calculations to accurately uncover the dynamics of degradation.

The derivation of the mathematical description consists of two steps: First is to translate the single molecule dynamics model (e.g. the one-state or two-state model) into the degradation from steady state at the level of population averages. The translation of single molecule dynamics to population averages has been covered in [86]. The second step takes the pulse into account and returns the degradation curve of the population averages (e.g. the measurements from the experiments). Calculation of the response of the system resulting from a pulse has been covered in [5].

In our formalism, the function $\Lambda(t)$ defines the theoretical decay pattern, namely the fraction of molecules left after a decay of t time units. This function is expressed in terms of parameters defined through the underlying degradation model. In our degradation model, we have assumed that the proteins follow either a 1-state model, in which there is only one degradation parameter, or a 2-state model, where there are three parameters. The equations used for fitting are as follows:

$$\Lambda(t) = e^{-k_A t} \qquad \text{for the 1-state model} \tag{4.3}$$

$$\Lambda(t_p + t) = \frac{G(t) - G(t_p + t)}{G(0) - G(t_p)} \qquad \text{for the 2-state model} \tag{4.4}$$

where$G(t) = k_{AB}(k_{AB} + k_A)e^{-k_{20}t} + k_B(k_A - k_B)e^{(-k_{AB}+k_A)t}$ and t is the measurement time after the end of the pulse.

Parameter estimation is performed by MATLAB through nonlinear fitting by minimizing the square deviation from the logarithm of the experimental data and the logarithm of the theoretical function. The routine employed for the nonlinear fit is fmincon.

After parameter fitting we applied two quality criteria for selection of proteins for downstream analysis. First, only proteins which had measurements for more than three data points were kept. Second, profiles with RSS > 0.05 were not considered for downstream analysis.

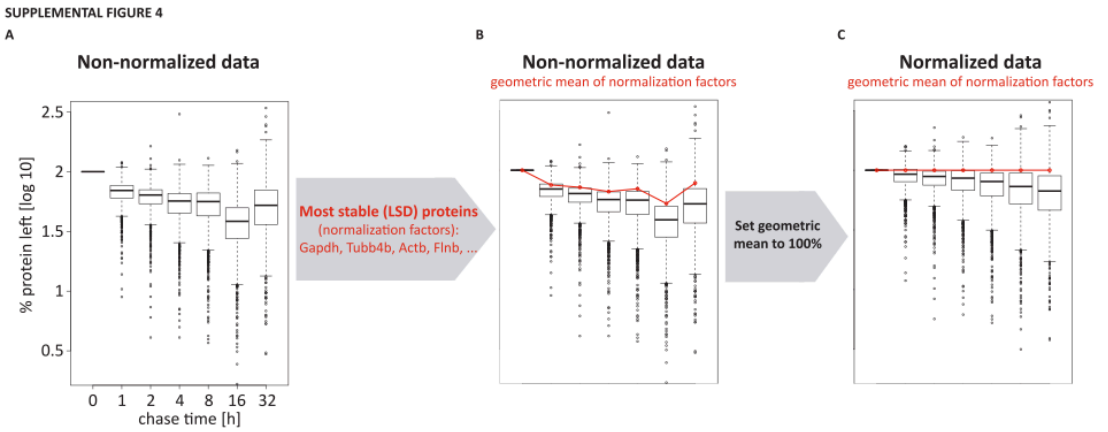


FIGURE 4.4: The raw data required normalization due to input differences (due to pipetting errors, cell density variation, and so forth). Plot A) displays the non-normalized raw data from one biological replicate. For normalization an assumption of non-changing data points over time or conditions are required (e.g. median normalization). In our case we look for very stable proteins. We defined the most stabile proteins over all biological replicates as proteins with the consistently lowest (H/L or H/M) ratios at all time points (LSD proteins, see Material and Methods). Example LSD proteins for mouse are shown in the left grey arrow. Each time point was then independently normalized by setting the geometric means for the LSD proteins (red points in B)) to 100% (i.e. no degradation). The resulting shift in the data can be seen in C) with new median values for LSD protein as red dots.

4.6 Model selection by the Akaike Information Criterion [6]

The Akaike Information Criterion indicates the quality of the model for a given set of data. Based on information theory, the AIC aims to find the model with minimal Kullback-Leibler distance between the proposed model and the "true" model (as assessed from the data). Models with more parameters have more degrees of freedom during the parameter estimation process, and can often deliver a more accurate fit to the data. However, there is the danger of over-fitting – that is the model could be approximating not only the system's dynamics, but also quantities not related to the degradation, such as measurement noise. To decide which model we should adapt for each protein, we calculate the AIC for each model. The model resulting in the lowest AIC is the preferred model. We use the AIC with correction for small sample sizes to evaluate each of the two models fitted to each protein degradation pattern:

$$AIC = 2k + n \ln \frac{RSS}{n} + \frac{2k(k+1)}{n-k-1} \tag{4.5}$$

where n is the number of data points, k is the number of parameters, and RSS is the residual sum of squares. The AIC penalizes models with more parameters, worse fits,

SUPPLEMENTAL FIGURE 5

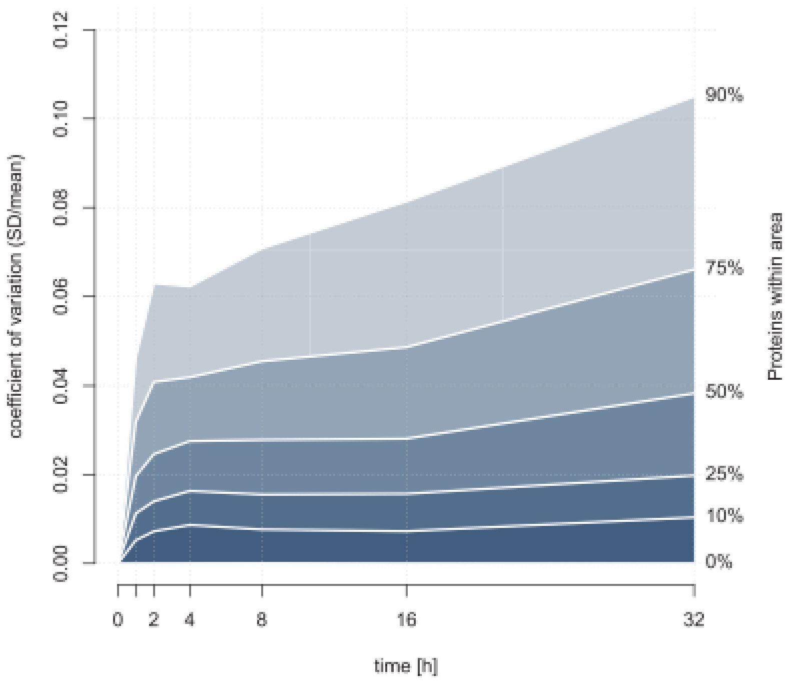


Figure 4.5: Plot of the coefficient of variation (CV), i.e. the standard deviation (SD) divided by the mean in log space, over all time points. The CV was calculated for each time point and for each decay profile individually (see Figure 4.1D for examples of the multiple means and SDs per profile). Within the outermost grey area in the plot 90% of the decay profiles resides. The second outermost area contains 70% of decay profiles and so forth. The CV increases with the chase time indicating the increased spread in data points at later time points. Overall 90% of the profiles were within a CV of 11% after 32h of chase.

and less data. That is, the AIC quantifies the tradeoff between fit accuracy and model complexity.

Furthermore, we can calculate the probability that a particular model is the preferred one (relative to the other models we consider) by:

$$\Pi_{AIC_i} = \frac{exp(\frac{AIC_{min}-AIC_i}{2})}{\sum_j exp(\frac{AIC_{min}-AIC_j}{2})} \tag{4.6}$$

4.7 Δ-score calculations

If a protein is exponential degradation one can draw a straight line in a semi log plot between time point 0 h (100% protein left) and the measured value for another time point and derive the relative protein abundance at any other time point. However, if a measurement for another time point is not on the line, allowing for measurement and quantification errors, the protein is non-exponentially degraded. We used this fact to estimate the size and direction (increased or decreased stability with age of the molecule) of the non-exponentiality of degraded proteins. We used the median log "percent protein left" at time point 8 h (tp8) after chase to calculate the expected relative protein abundance at time point 4h assuming exponential degradation. For this we solved the straight line equation (y = mx + c) for x = 4h, where the intercept c is log(100%), and the slope m is calculated using the value at tp8:

$$y(4h) = -\frac{\log(100\%) - tp8}{8h} 4h + \log(100\%) \tag{4.7}$$

Finally, we calculated the distance from the measured median log "percent protein left" at time point 4 h (tp4), to the expected value, y(4h):

$$\Delta - \text{score} = y(4h) - tp4 \tag{4.8}$$

This calculation was repeated for all proteins. The time points 4 and 8h were selected because of the observation that most of the initial degradation of NED proteins had happened by 4 h chase. Thereby we expected to be able to catch age-dependent stabilization (or destabilization) by comparing these two time points. Also, few proteins (see Supplemental Table 1) had a half-life shorter than 2.5h and could thereby theoretically not be detected at the 8 h time point. In addition, these were almost exclusively exponentially degraded according to the AIC call.

4.8 SILAC p-c (confirmation experiment)

To exclude issues related to using non-natural amino acids and to the enrichment process (e.g. background binders) we performed a pulse-chase experiment using only Stable isotopes labeled amino acids. Mouse fibroblasts were grown to 80% confluency in 15cm plates in Light SILAC DMEM. Cells were washed three times in PBS before being pulsed in Heavy SILAC DMEM for 4h (or as annotated in Figure 4.7). Cells were then washed in PBS before being trypsinated for 2 min at 37°C. Cells were resuspended in PBS before half of the cells being transferred to a 10cm plate containing Medium SILAC DMEM and the other half spun down and pellet then frozen. After the Medium chase (see Figure 4.4 and Supplementary Figure 4.7, for different chase length) cells were spun down and frozen. In addition, "label-swap" experiments were also performed in this fashion. However, in the label-swap experiments the cells were pulsed with Medium and chased in Heavy amino acids. Cell pellets were lyzed and proteins denatured in 0.2%

SDS, 0.1M DTT and 50mM ABC (pH8) by boiling for 10 min at 95°C. After cooling, Benzonase was added for 10min before cell lysates were spun down and supernatants were transferred to fresh tubes. Proteins were alkylated by adding iodoacteamide to a 0.25 M final concentration, in the dark, for 20 min. Proteins were precipitated by Wessel-Flügge precipitation as previously described [243]. In short proteins were precipitated, to get rid of SDS, by sequentially adding, methanol, chloroform and finally water before spinning down the samples at 10,000 rcf [244]. The retrieved protein pellet was resuspended in 6 M Urea, 2 M Thiourea in 10 mM Hepes (pH8). Proteins were digested with LysC before being diluted in ABC buffer and trypsinated overnight. The resulting peptide solution was desalted on StageTips before being eluted in buffer B as described above. The peptides were resolved on a 4 m long monolithic column using a 12h gradient of increasing buffer B concentration and a flow rate of 500 nl/min. Peptides were ionized by ESI and analyzed on a Q-exactive orbitrap. Resulting rawfiles were analyzed on MaxQuant with the same parameter settings as above. Plotting and statistics were performed using R and figures were modified in Illustrator.

4.9 Inhibitor treatments + controls

Inhibitor treatment experiments were performed as the AHA p-c experiments but only with three time points (0, 4 and 8h). In addition to pulsing the cells with 1mM AHA different inhibitors or vector control dimethyl sulfoxide (DMSO, Biomol) were added. Proteasomes were blocked using 20 μM MG132 (Cayman chemical) and a robust inhibition of autophagy was secured by a combination of 250 nM Bafilomycin A1 (Invivogen) and 500 nM wortmannin (Calbiochem) both treatments were added only during the chase. Inhibition of autophagy by Bafilomycin/Wortmannin was monitored by in parallel taking samples for western blotting as previously described [245]. In short, scraped cells were spun down and directly lyzed in LDS sample buffer supplemented with DTT. Samples were run on 4-12% Bis-Tris gradient gels (NuPAGE, Invitrogen) before being blotted onto PVDF membrane (Immobilion-P, Millipore) using a wet blotting contraption (Invitrogen). The Autophagy blocked cells were probed against LC3 (Novus NB100-2220) and afterwards the membrane was stripped at 37°C for 15min in stripping buffer (2% SDS (Roth), 2% β-mercaptoethanol in 65mM Tris Base (pH6,7, Roth)) before being re-blotted using an anti-β-actin antibody (Sigma-Aldrich A5441). Treated cells for mass spec analysis were scraped, lyzed, and had their AHA labeled proteins clicked to alkyne-agarose beads as described above. Proteins were reduced with DTT and alkylated before beads were washed all as in the main AHA p-c experiment. Proteins were digested "on bead" by LysC and then trypsinated overnight. Peptide solution were put on 4mm/1ml C18 columns (Empore, 3M) and washed in buffer A. Peptides were eluted in buffer B

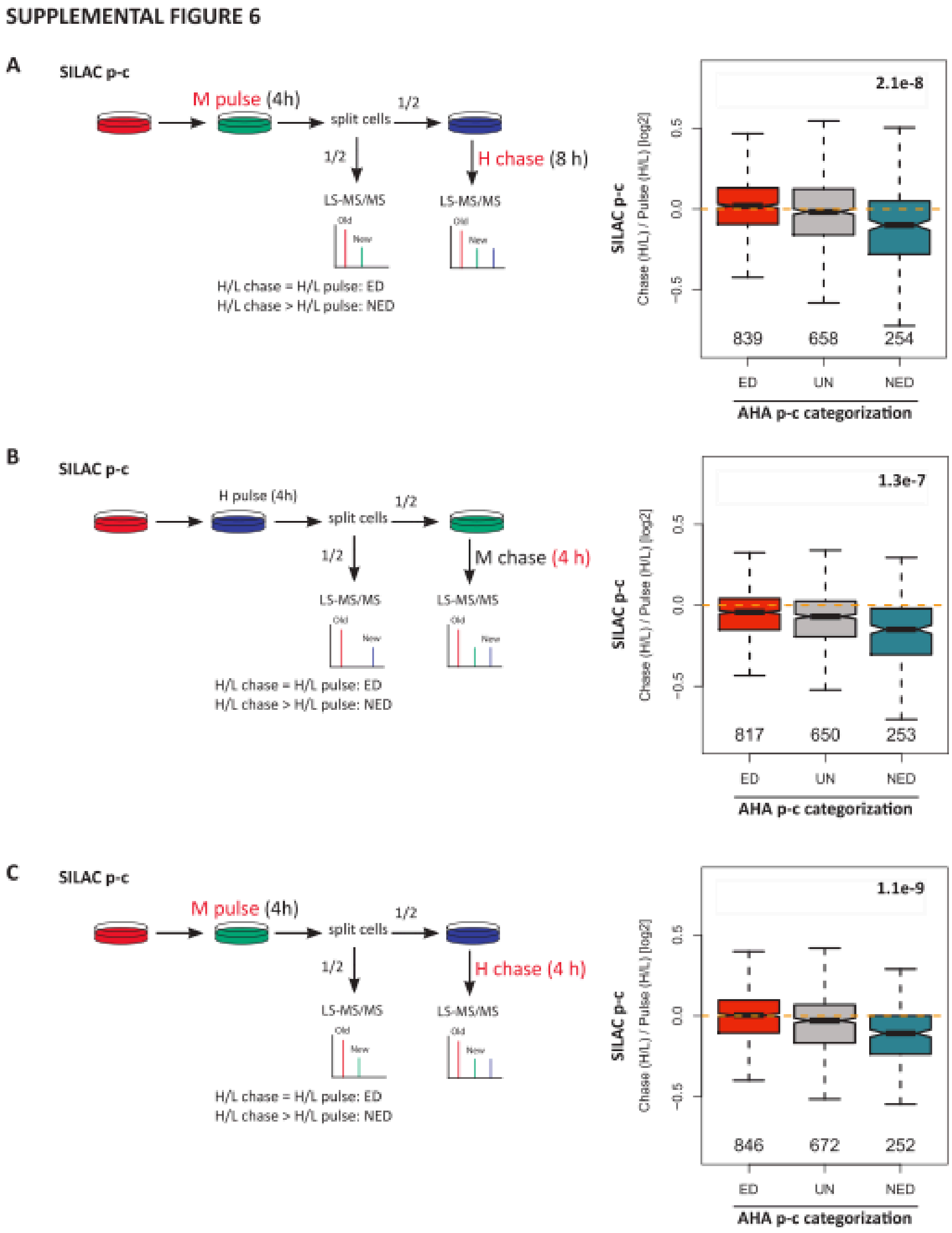


Figure 4.6: Confirmation experiments as shown in Figure 3. The differences from the experimental procedure used for the main figure is highlighted in red before each corresponding plot. A) A label swap experiment in where the Heavy medium used for the pulse was swapped with the Medium-heavy used for the pulse. B) The chase time was limited to 4h. C) Also with 4 h chase but with the same label-swap as in A). P-values from a one sided Kolmogorov-Smirnov test are displayed on top for significantly different distributions ($\alpha = 0.05$).

and vacuum dried. MG132 treated samples were separated using an online SCX/WAX approach. Samples were loaded on a column packed first with C18 material "trap" and then with a 2:1 mixture of WAX 3μm beads (PolyLC Inc. PolyWAX LP) and 3μm SCX beads (PolyLC Inc. PolyWAX LP) [246]. The peptides were subsequently eluted with increasing salt concentration (ammonium acetate in 4, 8, 16, 32, 64 and 500mM steps) onto the C18 trap part of the pre-column. Each fraction eluted from the SAX/SCX material where then separated as normal on a 15cm C18 column with 2h gradients of increasing buffer B concentration with a 250nl/min flow rate. Wortmannin/Bafilomycin A1 treated samples were put on SCX tips and washed in no salt buffer, as described above, to minimize polymer contamination. Samples were then eluted with 500mM ammonium acetate before being desalted on stageTips. Samples were elute from stageTips by Buffer B, vacuum dried, re-suspended in Buffer A and then separated on a HPLC system using an 8h gradient as previously described [239]. In brief, increasing buffer B concentration on a 2,000 mm monolithic column with a 100-μm inner diameter filled with C18 material that were kindly provided by Yasushi Ishihama (Kyoto University) (from now on referred to as "2 m monolithic column") with a flow rate of 300nl/min. Peptides were ionized using an ESI source and analyzed on a Q-exactive with the above described settings. ESI and mass spectrometer setting were, for all samples, as described above. The resulting raw files were analyzed as the standard AHA p-c experiment with MaxQuant. Timelines were reassembled from non-normalized protein ratios, resulting into three time points (0, 4 and 8 h) for each inhibitor treatment and the corresponding DMSO control. Proteins were filtered for being represented by at least two peptide identification events. Each time point was normalized to the geometric mean of the identified intersection of the LSD proteins that were identified in the mouse dataset used for the mathematical modeling. From this normalized dataset Δ-Scores were calculated for each, treatment and control, as described above. The difference of the Delta Scores between treatment and DMSO control was compared for the two protein subsets identified as NED and identified as not NED (others). Both distributions were tested for equality using the Kolmogorov-Smirnov test. The corresponding p value is reported in the figure legend for each treatment.

4.10 Degradation profile prediction from different protein features

The following features were selected to test each for prediction power of protein degradation profiles. The "Part of a Complex" feature distinguished proteins that are part of a complex from proteins that are not part of a complex [199]. Proteins were defined as being part of a complex, if they are listed in a published manually curated protein

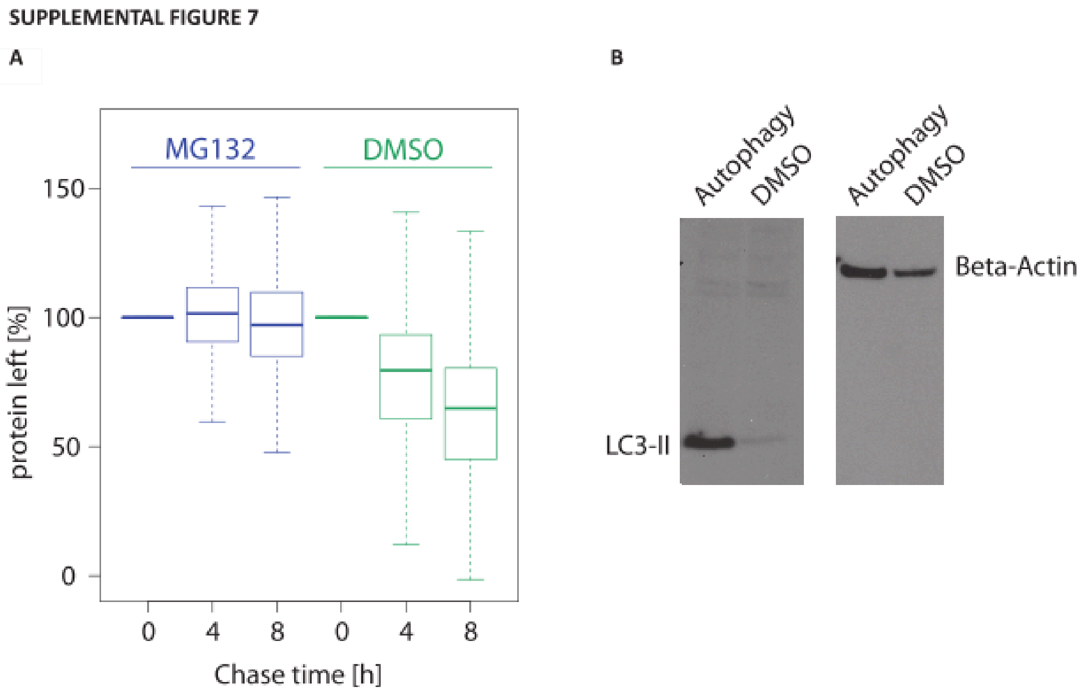


Figure 4.7: A) We estimated the effect of MG132 on protein degradation by plotting "% protein left" for all proteins with or without treatment. MG132 clearly stabilized the majority of the measured proteins. B) To control that autophagy was indeed inhibited, cells were treated with 250 nM bafilomycin A1 and 500 nM Wortmannin ("Autophagy" in figure) in parallel with the experiment for LC-MS/MS analysis. After 8 h chase, cells were lyzed and the samples were run on a SDS-PAGE. Samples were subsequently western blotted before being probed against the autophagy marker LC3 (Novus NB100-2220) and after stripping re-blotted using anti-Beta-actin (Sigma-Aldrich A5441).

complex database (unfiltered version; [199]). Protein Length refers to the protein sequence length and was taken from the Uniprot fasta table (version 10.2011). "Protein abundances in steady state" refer to average protein copy numbers per cell [72] mapped by Uniprot accessions and gene names if the Uniprot accession was not mapped. The feature "Low Complexity Region" were obtained from the "mmusculus_gene_ensembl" dataset from the biomart database (status 14.10.2015). Listed lengths of Low Complexity regions were summed up per protein. Disordered, Helix and Beta Sheet fractions per protein were taken from secondary structure predictions using the s2d method [247]. All structural features ("Low complexity", "Disorder", "Helix" and "Beta Sheet") were normalized to protein length. For each feature a ROC-curve was generated and the area under the curve calculated using the pracma R-package. The robustness of the calculated AUCs was tested by running 200 bootstrap repetitions. The 90% confidence intervals of the resulting AUCs are shown as error bars in the corresponding bar plot. Each feature was additionally randomized resulting in a real AUC and a random AUC population. Each feature prediction was tested for being absent or present by reversing the sorting vector. In each case the positive AUC was reported and labeled with

SUPPLEMENTAL FIGURE 8

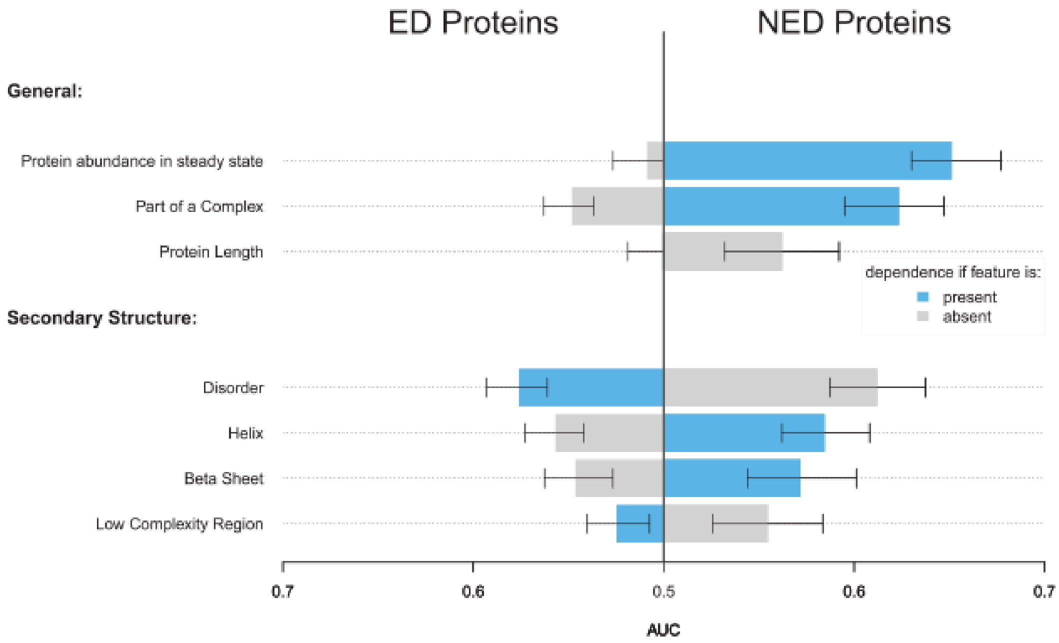


Figure 4.8: Protein features were tested to predict either exponential (ED) or non-exponential degradation (NED) in mouse by calculating the area under the curve (AUC) of a receiver operating characteristic (ROC) curve. Error bars were derived from 200 bootstrap repetitions and indicate the 90% confidence interval of the corresponding calculated AUCs. Features were additionally categorized in being present (blue) or absent (grey) in the tested protein fraction. For example, the bar for “protein length” is colored grey for NED proteins, indicating that NED proteins tend to be shorter than ED and undefined proteins.

“absence” or “presence” of the corresponding feature.

4.11 Protein structural dataset

Starting from the entire set of protein structures in the Protein Data Bank on 2016-02-24, we searched for all polypeptide chains with >70% sequence identity to a human or mouse gene. For genes that map to multiple chains, we selecting a single chain sorting by sequence identity, then number of unique subunits in the complex, and then the number of atoms present in the chain. Pairwise interfaces were calculated between all pairs of subunits using AREAIMOL [248]. The normalized assembly order was calculated for all complexes, excluding those containing nucleic acid chains, by first predicting the (dis)assembly pathway as previously described using all the pairwise interfaces from each heteromeric complex [218] and implemented in the assembly-prediction package [249]. For subunits with multiple copies within a single complex, the average assembly order of each subunit type was considered. The normalized assembly order was defined so that

SUPPLEMENTAL FIGURE 9

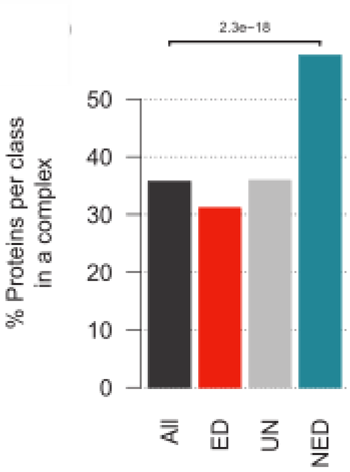


FIGURE 4.9: Fraction of proteins listed in a manually curated protein complex database (Ori et al 2016). Fractions are shown for all, exponentially degraded (ED), non-exponentially degraded (NED) and undefined (UN) proteins. The p-value is derived from a hypergeometric test testing for enrichment against all proteins.

SUPPLEMENTAL FIGURE 10

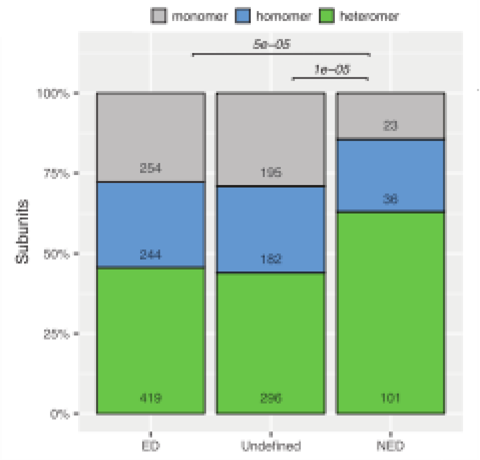


FIGURE 4.10: Due to the very large size of ribosomes, they can significantly skew small datasets. Here we show that the observed tendency of NED proteins to form heteromeric complexes is independent of their enrichment in ribosomes (equivalent to Fig. 5A).

the first subunit to assemble has a value of 0, the last has a value of 1, and the average value for all unique subunits in a complex is equal to 0.5.

4.12 Non-structural dataset

To complement the analysis of protein complexes of known structure, we also performed analyses on the perform coexpression analyses on the non-redundant "core" set of mammalian complexes was downloaded from CORUM [250] (downloaded 2015-10-20). As CORUM preferentially uses human complexes in its non-redundant set, homologous mouse versions of each complex were generated by replacing each subunit/gene with its mouse counterpart, provided sequence identity was at least 70%. Sequence identities were calculated by collecting all mouse sequences for which NED/ED classifications were available and running BLAST on these against all genes in the CORUM core set. In cases where the identity of a subunit was ambiguous (as defined by CORUM), the first possible subunit for which homology data was available was selected.

SUPPLEMENTAL FIGURE 11

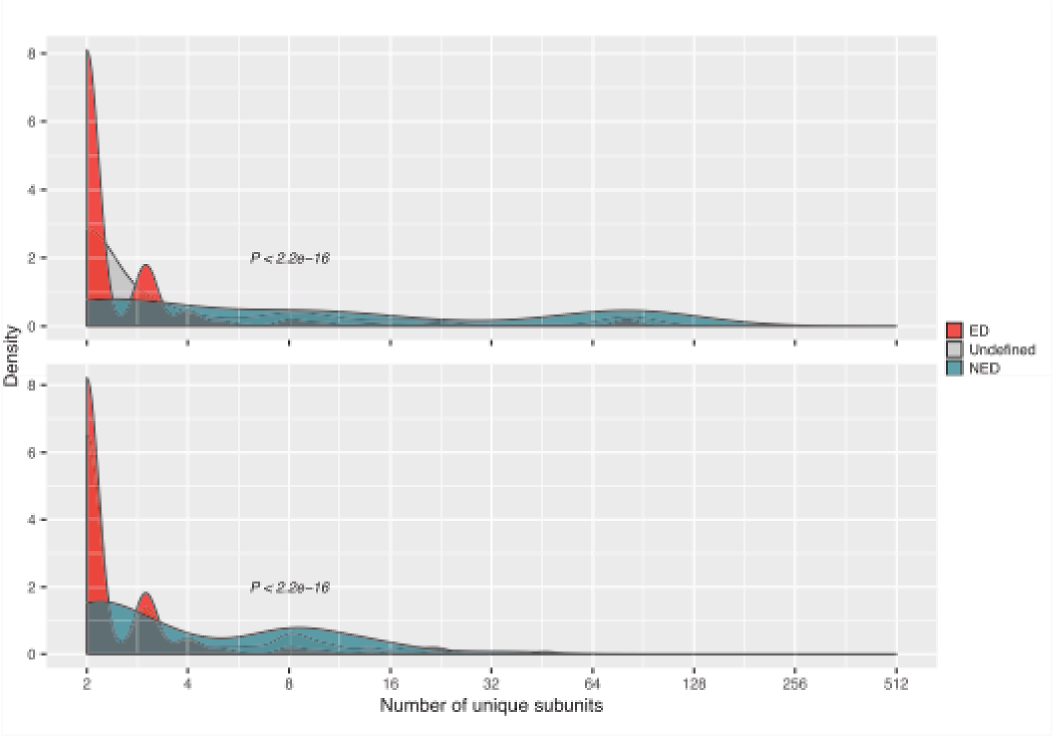


Figure 4.11: The distribution of subunits from heteromeric complexes of different sizes shows that NED proteins are significantly enriched in large complexes. P-values are comparing NED vs. ED subunit distributions, and were calculated using a modified Kolmogorov-Smirnov test to account for the discrete distribution of subunit counts (Arnold, 2011) (see R-package "dgof"). The top panel includes ribosomal subunits and the bottom panel excludes them.

SUPPLEMENTAL FIGURE 12

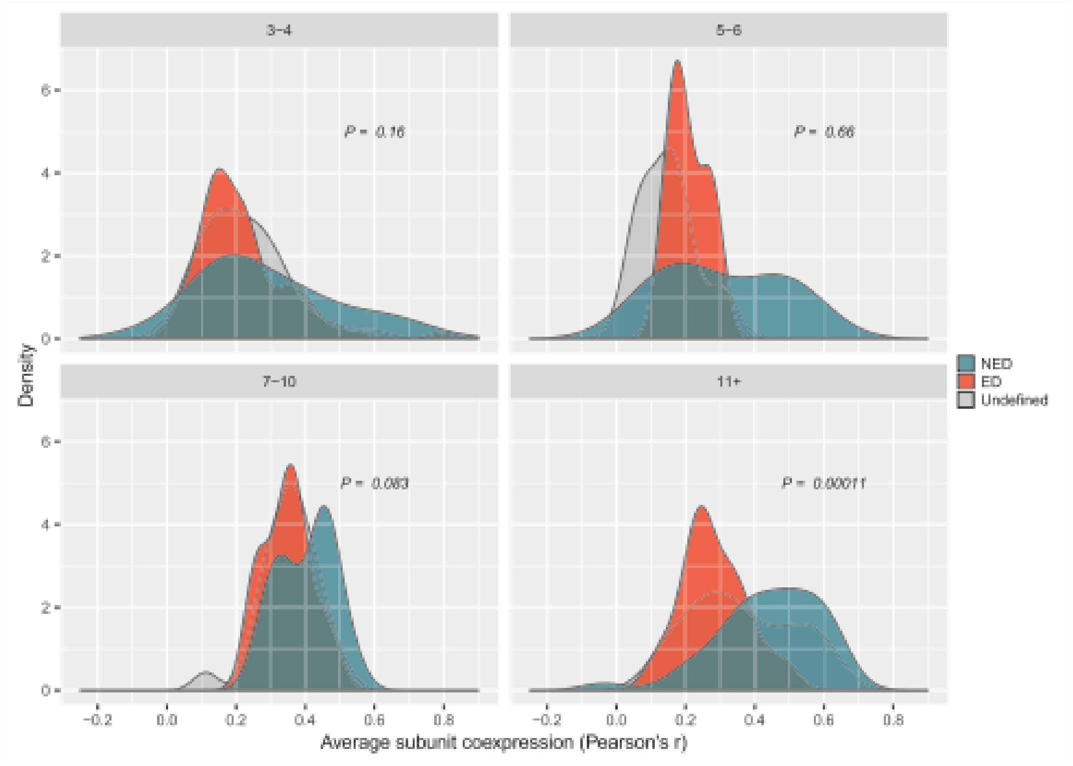


FIGURE 4.12: To control for the possibility that the observation in Fig. 5D is due to the fact that NED subunits typically form larger complexes, we binned subunits by the size of the complex from which their coexpression was calculated. Controlling for this, we find that the tendency for NED proteins to be more highly coexpressed still holds. Numbers in the top row of each facet describe the size range of complexes in each bin, measured by number of unique subunits, e.g "3-4" includes all subunits from complexes with between 3 and 4 unique subunits.

4.13 Coexpression analyses

Coexpression data was downloaded from CoexpressDB [251] (mouse dataset: Mmu.v13-01.G20959-S31479; human dataset: Hsa.v13-01.G20280-S73083). For each complex, the mean coexpression of each available subunit was calculated, using all other subunits in the complex. Cases where fewer than three subunits were present in the complex were discarded, due to calculations of average coexpression being superficially identical.

SUPPLEMENTAL FIGURE 13

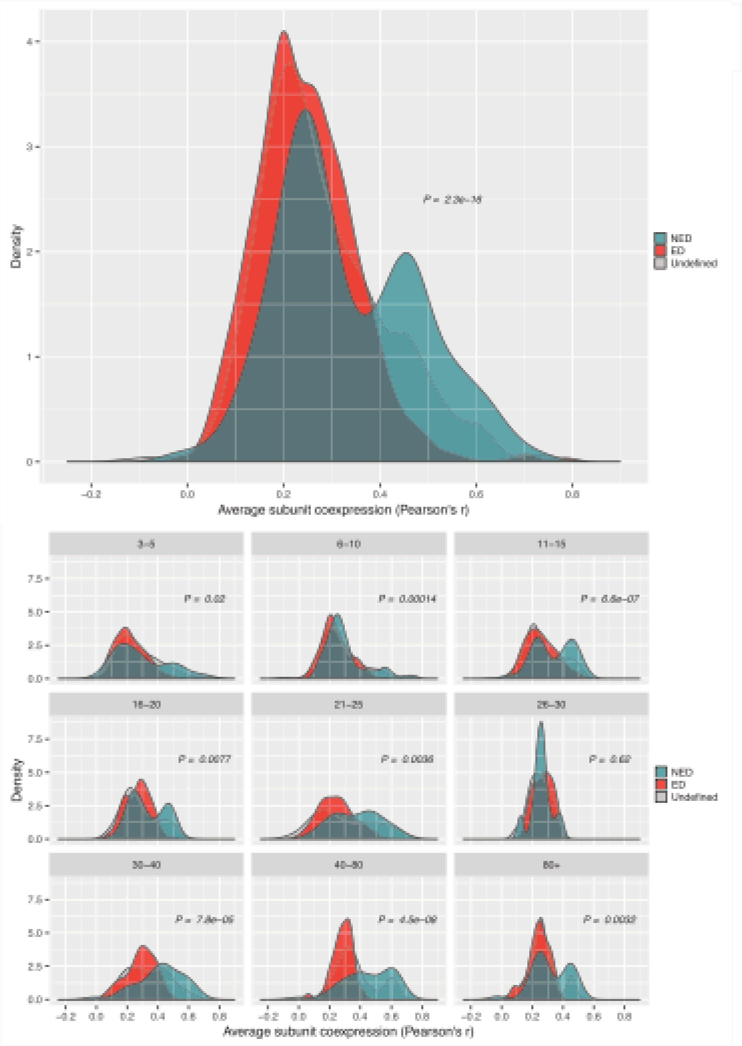


FIGURE 4.13: In order to confirm that the observations displayed in Fig. 5 are not an artifact of the structural data, we replicated the procedure using data from the "core" (i.e. non-redundant) set of CORUM complexes (Ruep et al. 2009). To maximize available data, non-mouse protein complex subunits were mapped to mouse genes wherever sequence identity was at least 70% (see Methods). Mean, per-complex subunit coexpression was then calculated using as before for each subunit in the inferred mouse homologs. The upper panel display coexpression data from the full set of CORUM complexes, whilst lower panels are binned by the size of the complex, with facet titles representing the size range of each bin (see Supplementary Fig. 12).

4.14 Estimation of relative protein abundance after pulse (iBAQ)

To estimate the protein abundance after the pulse (i.e. the relative amount of newly synthesized proteins) we used intensity based absolute quantification (iBAQ, [72]. First, all the intensities reported directly after the pulse, i.e. the H-intensities, for each protein group were divided by the number of observable peptides to correct for observability biases. Second, all the corrected H-intensities were normalized by using the median H-intensities for the LSD-proteins (see Section 4.4). This allowed the combination of experiments. Finally, we reported the median H-Intensity from all experiments as the relative abundance. The median was used to avoid counting highly abundant proteins which show up in all replicates multiple times. For a complex centered analysis of the relative protein abundances after pulse, identified proteins from the mouse dataset were mapped to a filtered version of PDB (see previous section) using gene names. Protein abundances were normalized in a complex centered manner: First all proteins that mapped to a complex were extracted. Second, stoichiometric differences between subunits within a complex were normalized out by dividing each protein abundance against the listed stoichiometry in the PDB database. Third, abundances of all proteins of a complex were normalized to the average abundance of each complex. The resulting complex centered abundances were compared between the protein subsets ED, NED and UN. Only proteins from complexes with at least one ED and one NED subunit were considered for the analysis. The distributions between the ED and NED fraction were tested for equality applying a Kolmogorov-Smirnov test as is implemented in R.

4.15 Human cell lines

The cell line RPE-1 hTERT (referred to as RPE-1) was a kind gift from Stephen Taylor (University of Manchester, UK). The derivative cell line RPE-1 5/3 11/3 12/3 (referred to as RPE-1 trisomic) was generated using microcell-mediated chromosome transfer as described below. The donor mouse A9 cell line was purchased from the Health Science Research Resources Bank (HSRRB), Osaka 590-0535, Japan. Cell lines were maintained at 37°C with 5% CO2 in DMEM GlutaMax (Gibco) containing 10% fetal bovine serum (FBS), 100U penicillin and 100U streptomycin.

4.16 Microcell-mediated chromosome transfer (MMCT)

Microcell fusion was utilized to generate aneuploid RPE-1 from the parental RPE-1 [209]. Briefly, murine donor cells containing an additional human chromosome with a resistance gene were treated for 48h with colchicine (60ng/ml). Cells were then trypsinized and seeded on plastic bullets. The cells were allowed to attach to the surface, and were subsequently centrifuged at 27000rcf for 30min at 30-34°C in DMEM supplemented with 10μg/ml cytochalasin B. Cell pellets were resuspended in serum-free DMEM and filtered to avoid contamination with mouse cells. Filtered microcells were then mixed with phytohemagglutinin (PHA-P) and added to the RPE-1 recipient cell line. The fusion of microcells with the recipient cells was facilitated by polyethylene glycol 1500 (PEG 1500) treatment. Cells containing the additional human chromosome were selected for in medium supplemented with 400μg/ml G418. Chromosomes 11 and 12 carry no genes coding for resistance as the chromosomes were gained spontaneously after the transfer of chromosome 5. Clonal populations arising from a single cell after the MMCT were isolated and further expanded. Subsequently, chromosome spreads combined with chromosome painting were performed (see below).

4.17 Preparation of chromosome spreads and chromosome painting

Cells were grown to 70-80% confluency before treatment with 50 ng/ml colchicine for 3-5h. Subsequently, cells were collected by trypsinization and centrifuged at 250rcf for 10min. Pellets were then resuspended in 75mM KCl and incubated for 10-15min at 37°C. After centrifugation at 150rcf for 10min, cell pellets were resuspended in 3:1 methanol/acetic acid for fixation. Finally, cell pellets were washed several times in 3:1 methanol/acetic acid, spread on a wet glass slide and air-dried at 42°C for 5min. Each sample was labeled with probes for two different chromosomes. Probes (Chrombios GmbH, Raubling, Germany) for chromosome 5, 11 and 12 were tagged with FITC and TAMRA, respectively. The chromosomes were labeled according to the manufacturer's instructions and counterstained with DAPI. Images were obtained by a fully automated Zeiss inverted microscope.

4.18 Genomic DNA sequencing and copy number estimation of RPE-1 and RPE-1 trisomic cells

DNA was isolated using the Blood and Cell Culture DNA kit (Qiagen) according to the manufacturer's recommendations. 1μg genomic DNA was sheared following the protocol from the SureSelectXT Target Enrichment System Kit for Illumina Multiplexed Sequencing (Agilent Technologies). Genomic DNA sequencing library was prepared with 100ng sheared genomic DNA using TruSeq ChIP Library Prep Kit according to the manufacturer's guidance (Illumina). The libraries were sequenced in 1x 100nt manner on HiSeq 2000 platform with a depth of $\sim$30 million reads per library (Illumina). Sequencing reads were aligned to the human reference genome (hg19) using Bowtie (version 2.1.0) with default parameters, and only uniquely mapped reads were kept for downstream analysis. With a sliding window of size 100Kb and a step size of 50-Kb, mapped reads in each window were then counted and used for copy number estimation. With the assumption that most genomic regions for the cells were diploid, we took C_i given by the following formula as the copy number estimates for genomic location at the ith window:

$$C_i = 2 \cdot \frac{R_i}{\text{median}(j \in I) R_j} \tag{4.9}$$

where R_i is the read counts of the ith window. To avoid underestimating copy numbers for regions with multi-aligned reads, we adjusted for mapability based on mfapping of simulated reads with uniform coverage across the genome. The original read counts were divided by the read counts in the same window obtained from the simulation data, and the adjusted read counts were instead used for copy number estimation.

4.19 AHA pulse chase of SILAC labeled RPE-1 and RPE-1 trisomic cells

RPE-1 and RPE-1 trisomic cells were grown, methionine starved, AHA pulsed, chased and lyzed as described for the mouse fibroblast. Experiments were started when cells reach ∼30% confluency and two 15cm plates were used per time point. Click chemistry, denaturation, alkylation, washing and digestion were performed as described for the mouse cells. Peptides were stageTipped on 4mm/1ml C18 columns (3M). Peptides were eluted using 500μl buffer B and speed vacced until a few μl liquid was left. For two of the samples Buffer A was added to 10μl final volume. 5μl of sample was loaded onto a 15cm column and 5μl onto a 2m monolithic column using a HPLC system. The 15cm column and 2m monolithic column samples were analyzed on a Q-Eactive orbitrap system, as described above, deploying 4 and 6h gradient of increasing Buffer B, respectively. For one sample the peptides were further SCX fractionated into 2 fraction 125mM and 500mM ammonium acetate as described above. These samples were analyzed using 4h gradients of increasing Buffer B concentration over a 15cm column. The resulting raw files were analyzed using MaxQuant with the previously described parameter settings with the exception that Andromeda search engine was matching the MS/MS to the human Uniprot database (2014-10). 3 biological replicates were performed per cell line. Normalization, fitting of models, Δ-score and abundance after pulse calculations were performed as for the mouse fibroblasts. For all downstream analysis proteins derived from genes located on autosomes were used except when from chromosome 10 (fully trisomic in both parental and trisomic cell line) and chromosome 12 (clonal expansion of trisomic cells among control cells).

4.20 Relative protein levels at steady state in RPE-1 and RPE-1 trisomic cells

Fully Heavy SILAC labeled RPE-1 and Light labeled RPE-1 trisomic cells were grown to 70% confluency in 10cm plates. Cells were scraped in ice cold PBS before being spun down at 1000rcf and PBS decanted. Cell pellets were lyzed in 1.3% SDS, 0.1M DTT in 50mM ammonium bicarbonate solution. Samples were heated to 95°C for 10min. After cooling the samples, Benzonase was added for another 10min. The samples from the two cell lines were then mixed 1:1 and spun down at 20,000 rcf to clear cell debris. Proteins in supernatant were alkylated by the addition of 0.25M iodoacetamide, final concentration, and left in the dark at room temperature for 20min. After alkylation samples were directly precipitated, to get rid of SDS, by sequentially adding, methanol, chloroform and finally water before spinning down the samples at 10,000 rcf as previously described [243, 244]. The upper water phase was discarded and more methanol was added to the precipitated proteins and the samples were spun down again. Supernatant was discarded and pellet air dried. The pellet was solubilized by shaking the sample in 6M Urea/2M thiourea in 10mM Hepes (ph8). Proteins were digested "on pellet" by Lys-C for 3h at room temperature before the sample was diluted in ABC buffer and Trypsin was added overnight. Peptides were acidified by triflouroacetic acid before being stored on stageTips. Peptides were prepared for HPLC as described above and analyzed using a 6h gradient on a 15cm column packed with C18 material as described above. The Q-exactive was run with standard setting and the raw files were analyzed as described for the AHA p-c experiments. MaxQuant output was filtered as described above but this time normalized SILAC ratios were used for downstream analysis. A label swap experiment was also performed in where RPE-1 cells were grown in light SILAC medium and RPE-1 trisomic cells were grown in heavy SILAC medium.

4.21 Bioinformatics of RPE-1 cells

Delta Scores and abundances after pulse were calculated for the datasets of the RPE-1 and the RPE-1 trisomic cell line as described in the previous sections. For the conservation analysis proteins identified in RPE-1 were mapped to proteins from the mouse data using always the first entry in the gene name column from both protein-groups tables (proteinGroups.txt, as provided by MaxQuant). Mapped proteins from mouse that fell into the categories ED, UN and NED were compared to the fraction of mapped proteins identified in RPE-1: RPE-1 proteins that mapped to the mouse dataset (all, mapped to mouse orthologues), RPE-1 proteins that mapped to the ED subset of mouse

(ED in mouse), and RPE-1 proteins that mapped to the NED subset (NED in mouse). Enrichment of RPE-1 ED or NED definitions in the fraction ED or NED in mouse against the corresponding fraction of all mapped proteins was tested applying a hypergeometric test (phyper function as implemented in R setting lower.tail to False). Delta Scores of mapped RPE-1 and Mouse Genes were compared and Pearson correlation of all available data points was calculated (using cor.test function as implemented in R). The corresponding p-value indicating the significance of the observed correlation coefficient is given in figure legend (Figure 4.5). Complex structures were modeled using the PyMOL Molecular Graphics System, Version 1.7.6.0. Schrödinger, LLC. ED, NED and UN definitions from RPE-1 and mouse were mapped to subunits of each complex using gene names. Protein categories ED, UN and NED were colored red, grey and turquoise respectively for mouse (Figure 4.5C and E) and RPE-1 definitions (Figure 4.5D and F). RPE-1 and RPE-1 trisomic datasets were merged using the leading protein ID in each protein-group. Proteins were linked to chromosome positions using the human reference genome (hg19). Proteins were first mapped to chromosome positions based on Uniprot IDs. Remaining unmapped proteins were mapped using gene name entries as provided in the uniprot fasta file. Proteins were further grouped into the categories disomic, trisomic or ambiguous. Disomic proteins included all proteins that mapped to disomic chromosomes as identified by genome sequencing (see previous section). Trisomic regions included proteins mapping to chromosome 5 and chromosome 11 downstream from position 62 650 000. Chromosome 12 and 10 starting from nucleotide position 62 500 000 were considered to be ambiguous since they show partial aneuploidy already in the RPE-1 cell line. Only autosomes were considered for the subsequent analysis. Distributions of the Delta-Score and Steady-State protein levels were compared between proteins from disomic and trisomic regions and tested for equality applying a two-sided Kolmogorov-Smirnov-test (ks.test as implemented in R). Steady State protein levels were further split into two subsets: proteins in a complex (listed in the manually curated protein complex database from [199] and proteins not in a complex (not listed in the protein complex database).

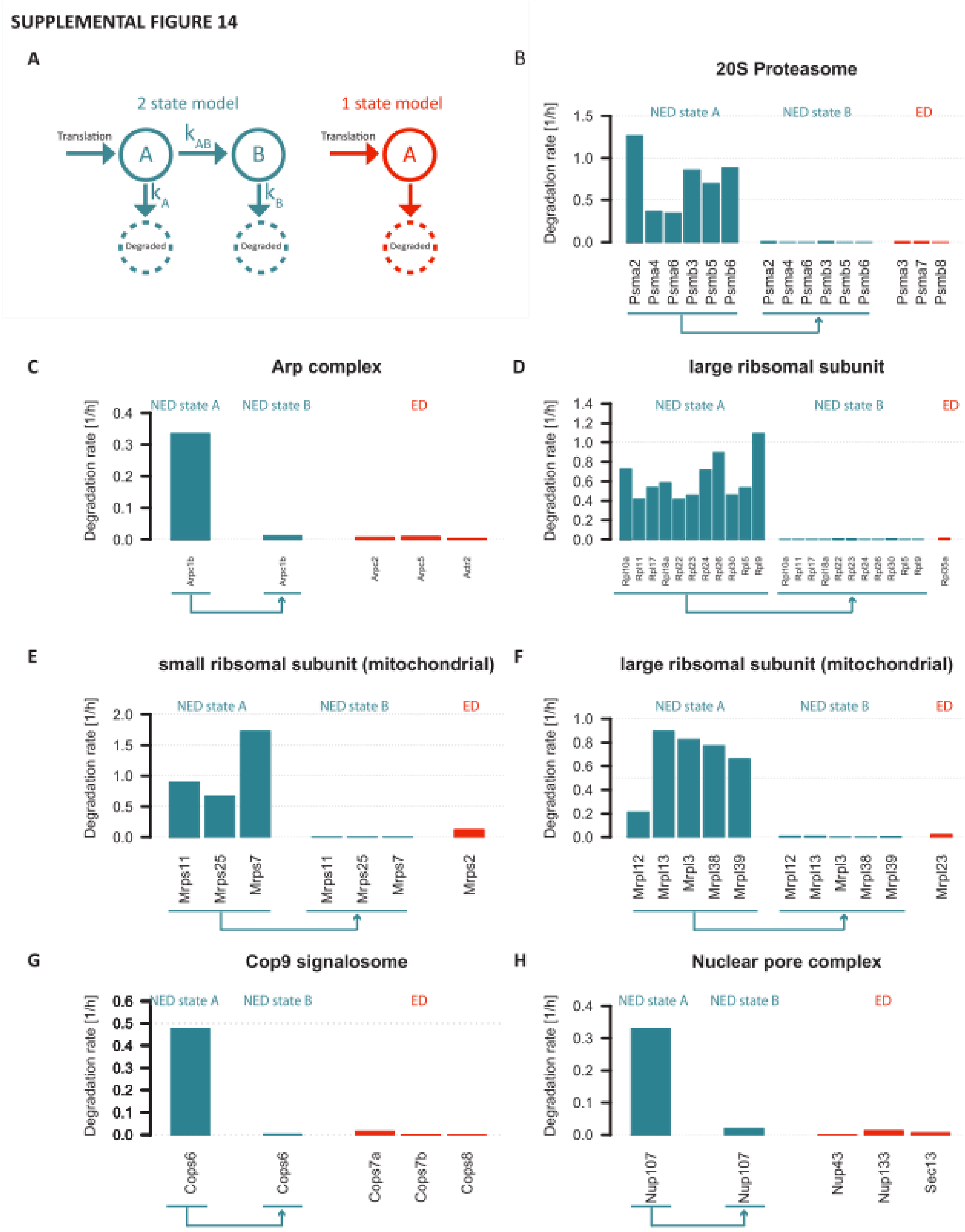


FIGURE 4.14: A) NED proteins have two degradation rates (see Supplemental Table 1). The initial one (kA) is in our model related to the free subunit or subunits in partially assembled protein complexes. The second degradation rate (kB) is associated to the subunits in the holoenzyme. In addition, the ED proteins we hypothesize are quickly incorporated into complexes and thereby only have one degradation rate which is equivalent to the degradation rate of the complex. To see if this is the case we plotted bar plots of the degradation rate A and B and the ED degradation rate (B-H). In all cases probed degradation rate B is more similar to the ED degradation rates of protein in the same complex (lower). Displayed are the degradation rates of all NED and ED proteins belonging to each complex.

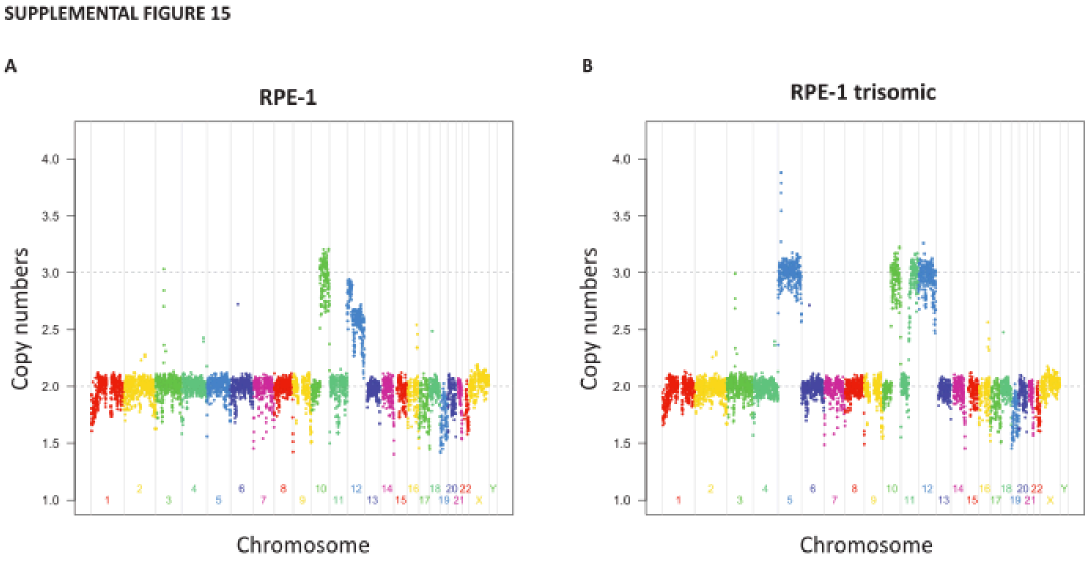


Figure 4.15: Gene copy number estimates based on sequencing of genomic DNA from RPE-1 parental and RPE-1 trisomic cells. Copy number estimates are based on mappability corrected read counts (see Materials and Methods) and ordered over the chromosomes. A) RPE-1 parental cell-line is trisomic for chromosome 10 and displays clonal expansion of a population of cells also trisomic for chromosome 12. B) The trisomic RPE-1 cells are fully trisomic for chromosome 5, 10 and 12 and a region of chromosome 11. For the downstream analysis only chromosome 5 and part of 11 where used while chromosomes 10 and 12 where ignored.

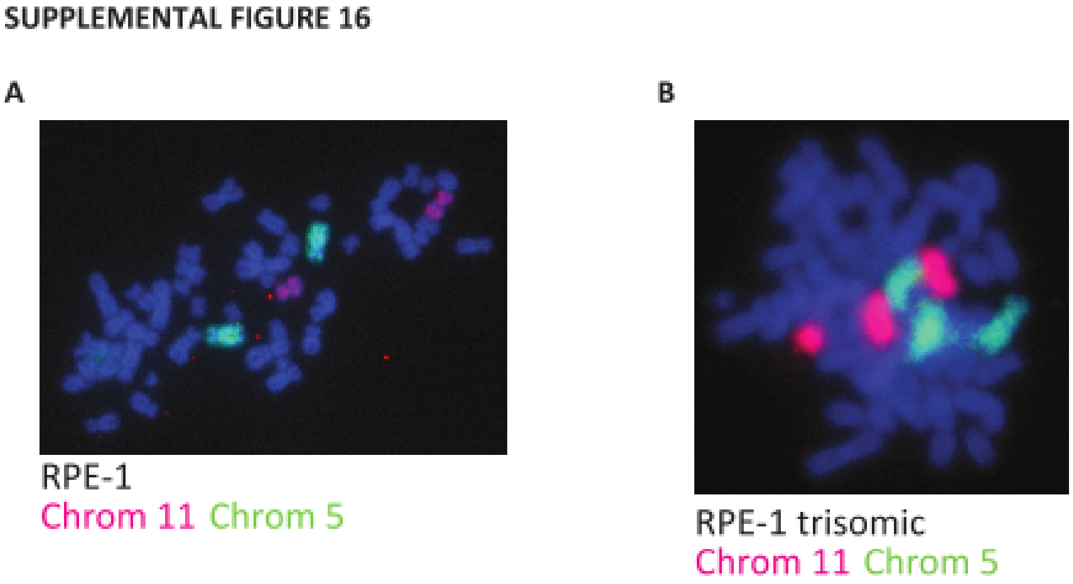


Figure 4.16: Chromosome paintings labeling chromosome 11 (magenta) and 5 (green). A) RPE-1 parental cell line is disomic for chromosome 5 and 11. B) RPE-1 trisomic cells are trisomic for chromosome 5 and part of 11. The trisomic part of chromosome 11 has fused to an unidentified chromosome.

Supplementary 5

Degradation Parameters from Pulse-Chase Experiments

5.1 Integrals

For the two-stage model introduced in section *Results* there are a few minor integrals, that may become interesting in certain applications. Using F_T from Eq. (13) and the $\rho_i(t)$ from Eqs. (14) we obtain

$$F_T(t) = 1 - \frac{(\kappa_{10} - \kappa_{20})e^{-(\kappa_{10}+\kappa_{12})t} + \kappa_{12}e^{-\kappa_{20}t}}{\kappa_{10} + \kappa_{12} - \kappa_{20}}, \tag{5.1}$$

from which the probability density is given by

$$f_T(t) = \frac{(\kappa_{10} - \kappa_{20})(\kappa_{10} + \kappa_{12})e^{-(\kappa_{10}+\kappa_{12})t} + \kappa_{12}\kappa_{20}e^{-\kappa_{20}t}}{\kappa_{10} + \kappa_{12} - \kappa_{20}}, \tag{5.2}$$

and the age-dependent degradation rate is also given by

$$\delta(a) = \frac{(\kappa_{10} - \kappa_{20})(\kappa_{10} + \kappa_{12})e^{-(\kappa_{10}+\kappa_{12})a} + \kappa_{12}\kappa_{20}e^{-\kappa_{20}a}}{(\kappa_{10} - \kappa_{20})e^{-(\kappa_{10}+\kappa_{12})a} + \kappa_{12}e^{-\kappa_{20}a}}. \tag{5.3}$$

Notice that $\delta(a)$ becomes a constant in the limit of an exponential decay, obtained either when $\kappa_{12} \to 0$ or when $\kappa_{10} = \kappa_{20}$. Central to the derivation of $C(\Delta t)$ and $P(\Delta t)$ is the following integral

$$\int_{\Delta t}^{t_p+\Delta t} (1 - F_T(t))\,\mathrm{d}t = \frac{A_c e^{-(\kappa_{10}+\kappa_{12})\Delta t} + B_c e^{-\kappa_{20}\Delta t}}{\kappa_{10} + \kappa_{12} - \kappa_{20}}, \tag{5.4}$$

where A_c and B_c, given in Eqs. (15), contain the dependence on the pulse duration t_p. This integral, allows us to finally derive the average lifetime of the degrading molecules

$$\overline{T} = \frac{\kappa_{12} + \kappa_{20}}{\kappa_{20}(\kappa_{10} + \kappa_{12})}\,. \tag{5.5}$$

5.2 Experimental data from Ref. [7]

Pulse duration / Measurement time	1 min	5 min	30 min	120 min	1200 min
0 min	100	100	100	100	100
20 min	84.9067	87.7513	93.0551	95.8508	98.525
40 min	74.9383	82.043	90.6058	94.4322	98.1468
60 min	67.8501	79.6414	89.7735	93.4221	97.5927
80 min	65.8373	76.7611	89.0894	92.8671	97.0384
120 min	64.5597	75.0701	87.871	91.9931	96.8331
240 min	60.44	71.1759	83.5673	89.0805	94.6325

TABLE 5.1: **Data from Ref.** [7] Each column shows the time course measurements after pulse time of 1, 5, 30, 120, 1200 minutes. Measurements are taken at t = 0, 20, 40, 60, 80, 120 minutes.

Part III

The Appendices

Appendix A

Model Discrimination

In general, models with more parameters will result in a better fit to data, due to the more degrees of freedom available. The important question here is whether or not a more parameterized model (*i.e.* a more complex model) provides a *significantly* better fit to the data and whether or not the available data justifies the inclusion of extra parameters.

A.1 Akaike Information Criterion (AIC) [6]

The Akaike Information Criterion indicates the quality of the model for a given set of data. Based on information theory, the AIC aims to find the model with minimal Kullback-Leibler distance between the proposed model and the "true" model (as assessed from the data). Models with more parameters have more degrees of freedom during the parameter estimation process, and can often deliver a more accurate fit to the data. However, there is the danger of over-fitting – that is the model could be approximating not only the system's dynamics, but also quantities not related to the degradation, such as measurement noise. To decide which model we should adapt for each protein, we calculate the AIC for each model. The model resulting in the lowest AIC is the preferred model. We use the AIC with correction for small sample sizes[1] to evaluate each of the two models fitted to each protein degradation pattern:

$$AIC = 2k + n \ln \frac{RSS}{n} + \frac{2k(k+1)}{n-k-1} \tag{A.1}$$

where n is the number of data points, k is the number of parameters, and RSS is the residual sum of squares. The AIC penalizes models with more parameters, worse fits,

[1] There a few ways to calculate the AIC, see [6] for more information.

and less data. That is, the AIC quantifies the tradeoff between fit accuracy and model complexity.

Furthermore, we can calculate the probability that a particular model is the preferred one (relative to the other models we consider) by:

$$\Pi_{AIC_i} = \frac{exp(\frac{AIC_{min}-AIC_i}{2})}{\sum_j exp(\frac{AIC_{min}-AIC_j}{2})} \tag{A.2}$$

A.2 F-test for nested models [8]

In cases where there are two models under consideration and one of the models is nested within the other, the F-test is another method that can be used to evaluate model suitability. Like the AIC, this method balances the trade off between goodness of fit, the amount of data available and the complexity of the model. We start by calculating the F-statistic:

$$F = \frac{\frac{RSS_1-RSS_2}{p_2-p_1}}{\frac{RSS_2}{n-p_2}} \tag{A.3}$$

where the subscript 1 refers to the more simple nested model, and subscript 2 refers to the more complex model. This tests the null hypothesis that the difference between the simpler and the more complex model is just noise. In particular, the null hypothesis states that

p is the number of parameters

n is the number of data points

RSS is the residual sum of squares

Then we calculate the probability that a single observation from an F distribution with degrees of freedom $p2 - p1$, $n - p2$ will fall between $[0, F]$. To compute this, quantity we have used standard routines in MATLAB.

Appendix B

Sensitivity Analysis

Parameter sensitivity refers to the sensitivity of particular model variables to variations in the parameters [8]. The goal is to quantify how sensitive particular model variables are to variations in parameters. In other words, what happens to the output x_i if we change ever-so-slightly, the parameter θ_j? An output is "more sensitive" to a parameter if perturbations to the parameter have larger effects on the output and it is "less sensitive" if perturbations of the parameter results in an indiscernible change in the output [252, 253].

For a system with state variable x_i, and parameter θ_j, we can write:

$$s_{ij}(t) = \frac{\partial x_i(t)}{\partial \theta_j} \tag{B.1}$$

If a model is more sensitive to a certain parameter, that parameter is said to be more identifiable [8, 254].

Appendix C

FPexp: A class for analysis of data from ribosome footprinting experiments

FPexp is a structure implemented in MATLAB designed for easy analysis and manipulation for data from ribosomal footprinting experiments. It contains methods for calculations pertinent to gene expression analysis. Using these calculations, FPexp can return subsets of genes with user defined specifications and generate graphical representations of the data (or subsets of the data).

C.1 Data input

FPexp can handle mapped read data without restriction to organism (number or types of genes). Data about the organism, names of the genes and lengths of the genes is stored in a separate structure, geneLib. In addition to specifying the geneLib, each data set should contain the following:

- for each gene, the number of mapped mRNA reads
- for each gene, the number of mapped footprint reads
- the total number of mapped mRNA reads
- the total number of mapped FP reads
- the length of the transcriptome, T (See Section 2.2.6)

A FPexp object is created by using the built-in constructor:

```
FPexp(geneLib, transcriptome, mRNARRD, fpRRD,mRNAtMRC, fptMRC)
```

A new object should be created for each footprinting experiment (i.e. for LB_2012_1, LB_2012_2, Heat47, one should create 3 objects)

C.2 Calculations

C.2.1 Number of mRNA copies

This describes the transcriptional program of the cell. It reflects the process of transcription as well as degradation.

$$\text{\# of mRNA copies per cell} = \frac{\text{normalized mRNA reads}}{\text{gene length}} T \tag{C.1}$$

C.2.2 Ribosome density

This describes the translational program of the cell. It reflects the process of translation initiation and elongation. However, it does not reflect on the effects of termination and protein degradation.

$$\text{Ribosome Density}_{\text{mRNA}} = \frac{\text{normalized ribosome reads}}{\text{\# mRNA copies}} \tag{C.2}$$

$$\text{Ribosome Density}_{\text{gene}} = \frac{\text{normalized ribosome reads}}{\text{normalized mRNA reads}} \tag{C.3}$$

C.2.3 Transcriptional intensity

This is a proxy for copy number when T (the transcriptome number) is not available.

$$\text{Transcriptional intensity} = \frac{\text{normalized mRNA reads}}{\text{gene length}} \tag{C.4}$$

C.2.4 Translational burden

This describes potentially how much “work” a particular gene has on the translational machinery – that is, a gene that has many copies or is very long can keep many ribosomes busy.

$$\text{Transcriptional intensity} = \text{normalized mRNA reads} \cdot \text{gene length} \tag{C.5}$$

C.3 Manipulations and Subset selectors

C.3.1 Upgrade/Ignore genes that have a low number of reads

As with all experiments, there is a considerable amount of error in the data. This is especially a problem for low-valued data because when we take the error into account, these low-valued data could in reality be 0 or several fold larger. This is especially a problem when we are calculating the fold change of mRNA or FP reads between two experiments. Thus, to avoid false positives, we define a threshold (60 reads) and use this threshold to adjust our calculations in the following way:

- If the # of reads from both data sets are <60, ignore.
- If the # of read for one data set is <60 (and the other is >60), upgrade the <60 to 60.

This way, we reduce the number of false positives.

The inputs to this method are two FPexp objects and the desired threshold. It returns two new FPexp objects with the appropriate reads upgraded and the indices of the genes that should be ignored.

```
[Temp1, Temp2, ignore] = applyThreshold(FPexp1, FPexp2, threshold)    % for mRNA reads
[Temp1, Temp2, ignore] = applyThresholdFP(FPexp1, FPexp2, threshold)  % for FP reads
```

C.3.2 Finding the largest fold changes between two data sets, for mRNA and footprints

For differential gene expression analysis, we would like to identify the genes which have dramatic changes between different conditions. We provide two methods to isolate these genes:

Return the X genes which have largest fold changes

Here we simply calculate the fold changes between the two conditions, and return the X genes that result in the largest fold change.

```
[index, names, values] = largestFoldChanges(FPexp1, FPexp2, threshold, numGenesDesired)
[index, names, values] = largestFoldChangesFP(FPexp1, FPexp2, numGenesDesired)
```

A limitation of this method is that it does not account for the biological variation that may occur (i.e. this method might return more genes than those with statistically relevant fold changes.

Use the biological replicates as a null model generator

Assuming that the fold changes between the two biological replicates (in our example, LB_2012_1 and LB_2012_2) is representative of the random changes and measurement errors inherent in the experiments, we use the distribution of fold changes between the two sets to generate the upper limit of statistically relevant fold change.

```
[index, names, values] = fcAboveNull(std1, std2, exp1, exp2, threshold, percentile)
[index, names, values] = fcAboveNullFP(std1, std2, exp1, exp2, threshold, percentile)
```

C.3.3 Sorting by ribosome density

We also have a method for sorting genes into 4 or 10 pools based on their ribosome density. These functions return the indices of the genes in each quartile (or tentile).

```
[q1index q2index q3index q4index] = riboDenseQuart(FPexp)
[t1i t2i t3i t4i t5i t6i t7i t8i t9i t10i] = riboDenseTenth(FPexp)
```

Similar functions exist for sorting by any of the earlier defined calculations (copy number, gene length, translational burden, etc.)

C.3.4 Returning a random subset of genes

Used for generating bootstrapped data sets.

```
index = randIndex(FPexp, numberDesired)
```

Appendix D

EMinhib: A class for analysis of data from MaxQuant files

EMinhib is a structure implemented in MATLAB designed for analysis and manipulation for protein degradation data from MaxQuant. It contains methods for data manipulation, filtering and normalization.

D.1 Data input

An EMinhib object should be created for each experiment (or replica). Each data set should contain the following properties:

```
properties
    ogData              % source description
    dataDesc            % + data processing, etc
    protein             % protein identifier
    fullUniP
    geneName
    inhibRatio
    inhibCount
    cntlRatio
    cntlCount
    dataTime
    isIDbySite          % 1 if protein is only ID by site
    isReverse           % 1 if protein is reverse
    isContaminant       % 1 if protein is contaminant
```

A EMinhib object can be created by declaring an empty object and defining the properties:

```
myObject = EMinhib();
myObject.ogData = ...
```

A new object should be created for each experiment or replica.

D.2 Data access

Two lookup functions:

```
whatsIndex(EMinhib, proteinID)
```

Both functions take inputs of a EMinhib object and a UniPro name. The first returns the index number of the protein in the object and the second returns the measurement ratios.

D.3 Data filtering

The following functions are for removing data that is unusable.

Filtering for IDbySite / reverse / contaminants:

```
newEMinhib = removeFakes(EMinhib)
```

Filtering for data by number quantification counts (low counts means less reliable data):

```
newEMinhib = filterBelowMinCts(EMinhib, minCounts, flag)
newEMinhib = aboveMinCts(EMinhib, minCounts, flag)
```

Filtering for data below a certain threshold:

```
newEMinhib = applyThreshold(EMinhib, lowLimitRatio)
newEMinhib = applyThresholdForward(EMinhib, lowLimitRatio)
```

Both functions replace data below threshold with NaN. The second function replaces the data that is below threshold plus every future measurement.

Filtering for proteins that have data at each time point measured:

```
newEMinhib = completeDataOnly(EMinhib, minDataPts, flag)
```

D.4 Data standardization

Standardization of naming:

```
newEMinhib = firstProID(EMinhib)
alphaObj = sortAlpha(oldObj)
```

The first function reduces long UniPro groups to the first name. The second function sorts the proteins in the object alphabetically.

Return a new object that only contains proteins that are defined by the user. The input list is defined by the index number of the protein. This can be generated by function whatsIndex().

```
uiProOnly(EMinhib, where, description)
```

In an array of EMinhib objects, return a new array only containing proteins that are found in all objects.

```
newArray = olArray(arrayObjs)
```

D.5 Functions for data normalization

To find the most stable proteins across all three replicas, we need a measure that tells us which proteins have consistently high signals. To do this, we assign a ”score” to each protein, defined as

$$\mathrm{score}_i = \sum_{t\in\{1,2,4,8,16,32\}} \mathrm{PercentileRank}_i(t)\,, \tag{D.1}$$

where $i = 1, 2, \cdots$ is the index of the protein and $\mathrm{PercentileRank}_i(t)$ maps the rank of each protein's signal strength (from smallest to largest, at time t) to the interval (0,1].

For example, a percentile rank of 0.75 means that the signal for protein i is greater or equal to 75% of the other proteins' signals at time t. The max score is 6 because there are six time points in the experiment and the maximum percentile rank for any protein at a certain time point is 1. Each protein has 3 scores, one from each replica.

In EMinhib, this is implemented through the following functions:

```
calcPCTL(EMinhib, flag)
topAVscorers(EMinhib, numProDesired, flag)
```

The first calculates the percentile rank for each protein within one replica. The second takes all the replica into account and calculates the score.

From the three scores, we calculate the deviation of the score from the maximum score:

$$\mathrm{dev}_i = (6 - \mathrm{score}_{i,\mathrm{rep1}})^2 + (6 - \mathrm{score}_{i,\mathrm{rep2}})^2 + (6 - \mathrm{score}_{i,\mathrm{rep3}})^2 \,. \qquad \text{(D.2)}$$

Candidates for normalization are those proteins with the lowest deviations. Notice that this measure penalizes proteins that score well in only one or two replicas and badly in the remaining replica(s).

In EMinhib, this is implemented through the following function:

```
[proteinIDs, deviations] = lowestScoreDev(arrayObjs, numDesired)
```

While our normalization scheme is independent of prior information and other experiments, it is based on several key assumptions:

1. All groups of cells (heavy/medium/light) produce and degrade proteins equally.
2. Proteins degrade at different rates (which can be differentiated in the time scale of our experiment)
3. Proteins degrading the slowest have the highest ratios (and thus will have the lowest deviations)
4. Slowest proteins don't degrade at all (at least in the time scale of our experiment)

The fourth assumption can be eliminated if we have outside knowledge about the degradation pattern of that specific protein through an additional experiment, such as radioactive labeling.

The final normalization step is implemented through two functions

```
calcMultFactors(EMinhib, proteinIDs, flag)
userIPnorm(EMinhib, multipliers, flag)
```

The first calculates the normalization factors from a list of proteins that are deemed to be the normalization candidates (proteinIDs). The output of the first function goes into the second function, which normalizes the data in the object.

Bibliography

[1] Celine Sin, Davide Chiarugi, and Angelo Valleriani. Single-molecule modeling of mRNA degradation by miRNA: Lessons from data. *Bmc Systems Biology*, 9:S2, June 2015.

[2] Alexander Bartholomaus, Ivan Fedyunin, Peter Feist, Celine Sin, Gong Zhang, Angelo Valleriani, and Zoya Ignatova. Bacteria differently regulate mRNA abundance to specifically respond to various stresses. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2063), March 2016.

[3] Celine Sin, Davide Chiarugi, and Angelo Valleriani. Quantitative assessment of ribosome drop-off in E. coli. *Nucleic Acids Research*, 44(6):2528–2537, April 2016.

[4] Erik McShane, Celine Sin, Henrik Zauber, Jonathan N. Wells, Neysan Donnelly, Xi Wang, Jingyi Hou, Wei Chen, Zuzana Storchova, Marsh Joseph A., Angelo Valleriani, and Matthias Selbach. Global quantification of cellular protein degradation kinetics. *Submitted to Cell, currently in revisions after first referee reports.*, 2016.

[5] Celine Sin, Davide Chiarugi, and Angelo Valleriani. Degradation parameters from pulse-chase experiments. *Plos One*, 11(5):e0155028, May 2016.

[6] KP Burnham and DRB Anderson. *Model selection and multimodel inference.* Springer-Verlag New York, New York, 2002.

[7] D. N. Wheatley, M. R. Giddings, and M. S. Inglis. Kinetics of degradation of short-lived and long-lived proteins in cultured mammalian-cells. *Cell Biology International Reports*, 4(12):1081–1090, 1980.

[8] J. DiStefano. Dynamic systems biology modeling and simulation. *Dynamic Systems Biology Modeling and Simulation*, 2014.

[9] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–&, 1970.

[10] Joerg E. Braun, Eric Huntzinger, Maria Fauser, and Elisa Izaurralde. GW182 proteins directly recruit cytoplasmic deadenylase complexes to miRNA targets. *Molecular Cell*, 44(1):120–133, October 2011.

[11] J. J. Furth, M. Anders, and J. Hurwitz. Role of deoxyribonucleic acid in ribonucleic acid synthesis .1. purification and properties of ribonucleic acid polymerase. *Journal of Biological Chemistry*, 237(8):2611–&, 1962.

[12] J. Hurwitz. The discovery of RNA polymerase. *Journal of Biological Chemistry*, 280(52):42477–42485, December 2005.

[13] Xiangyue Wu and Gary Brewer. The regulation of mRNA stability in mammalian cells: 2.0. *Gene*, 500(1):10–21, May 2012.

[14] M. Brengues, D. Teixeira, and R. Parker. Movement of eukaryotic mRNAs between polysomes and cytoplasmic processing bodies. *Science*, 310(5747):486–489, October 2005.

[15] Eric Huntzinger and Elisa Izaurralde. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature Reviews Genetics*, 12(2):99–110, February 2011.

[16] Marc Robert Fabian, Nahum Sonenberg, and Witold Filipowicz. Regulation of mRNA translation and stability by microRNAs. *Annual Review of Biochemistry, Vol 79*, 79:351–379, 2010.

[17] A. Ishihama. Functional modulation of Escherichia coli RNA polymerase. *Annual Review of Microbiology*, 54:499–518, 2000.

[18] D. S. Latchman. Transcription factors: An overview. *International Journal of Biochemistry & Cell Biology*, 29(12):1305–1312, December 1997.

[19] Aswin Sai Narain Seshasayee, Karthikeyan Sivaraman, and Nicholas M. Luscombe. An overview of prokaryotic transcription factors. *Handbook of Transcription Factors*, 52:7–23, 2011.

[20] Y. A. Ovchinnikov, G. S. Monastyrskaya, V. V. Gubanov, S. O. Guryev, I. S. Salomatina, T. M. Shuvaeva, V. M. Lipkin, and E. D. Sverdlov. The primary structure of Escherichia coli RNA-polymerase - nucleotide-sequence of the Rpoc gene and amino-acid-sequence of the beta'-subunit. *Nucleic Acids Research*, 10 (13):4035–4044, 1982.

[21] H. Kabata, O. Kurosawa, A. R. A. I. I., M. Washizu, S. A. Margarson, R. E. Glass, and N. Shimamoto. Visualization of single molecules of RNA-polymerase sliding along DNA. *Science*, 262(5139):1561–1563, December 1993.

[22] P. L. DeHaseth, M. L. Zupancic, and M. T. Record. RNA polymerase-promoter interactions: the comings and goings of RNA polymerase. *Journal of Bacteriology*, 180(12):3019–3025, June 1998.

[23] A. A. Travers and R. R. Burgess. Cyclic re-use of RNA polymerase sigma factor. *Nature*, 222(5193):537–&, 1969.

[24] A. N. Kapanidis, E. Margeat, T. A. Laurence, S. Doose, S. O. Ho, J. Mukhopadhyay, E. Kortkhonjia, V. Mekler, R. H. Ebright, and S. Weiss. Retention of transcription initiation factor sigma(70) in transcription elongation: Single-molecule analysis. *Molecular Cell*, 20(3):347–356, November 2005.

[25] T. M. Gruber and C. A. Gross. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annual Review of Microbiology*, 57:441–466, 2003.

[26] M. S. B. Paget and J. D. Helmann. Protein family review - the sigma(70) family of sigma factors. *Genome Biology*, 4(1):203, 2003.

[27] A. D. Grossman, D. B. Straus, W. A. Walter, and C. A. Gross. Sigma-32 synthesis can regulate the synthesis of heat-shock proteins in Escherichia coli. *Genes & Development*, 1(2):179–184, April 1987.

[28] G. Dreyfuss, V. N. Kim, and N. Kataoka. Messenger-RNA-binding proteins and the messages they carry. *Nature Reviews Molecular Cell Biology*, 3(3):195–205, March 2002.

[29] Soumaya Laalami, Lena Zig, and Harald Putzer. Initiation of mRNA decay in bacteria. *Cellular and Molecular Life Sciences*, 71(10):1799–1828, May 2014.

[30] Georg Stoecklin and Nancy Kedersha. Relationship of GW/P-bodies with stress granules. *Ten Years of Progress in Gw/p Body Research*, 768:197–211, 2013.

[31] Jonathan Houseley and David Tollervey. The many pathways of RNA degradation. *Cell*, 136(4):763–776, February 2009.

[32] Jeff Ross. mRNA turnover. In *eLS*. John Wiley & Sons, Ltd, 2001.

[33] M. P. Deutscher. Degradation of RNA in bacteria: comparison of mRNA and stable RNA. *Nucleic Acids Research*, 34(2):659–666, 2006.

[34] A. Wilczynska and M. Bushell. The complexity of miRNA-mediated repression. *Cell Death and Differentiation*, 22(1):22–33, January 2015.

[35] Tadashi Nishihara, Latifa Zekri, Joerg E. Braun, and Elisa Izaurralde. miRISC recruits decapping factors to mirna targets to enhance their degradation. *Nucleic Acids Research*, 41(18):8692–8705, October 2013.

[36] Joerg E. Braun, Eric Huntzinger, and Elisa Izaurralde. The role of GW182 proteins in miRNA-mediated gene silencing. *Ten Years of Progress in GW/P Body Research*, 768:147–163, 2013.

[37] Olaf Isken and Lynne E. Maquat. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes & Development*, 21(15): 1833–1856, August 2007.

[38] P. Anderson and N. Kedersha. Visibly stressed: the role of eIF2, TIA-1, and stress granules in protein translation. *Cell Stress & Chaperones*, 7(2):213–221, April 2002.

[39] L. Nover, K. D. Scharf, and D. Neumann. Formation of cytoplasmic heat-shock granules in tomato cell-cultures and leaves. *Molecular and Cellular Biology*, 3(9): 1648–1655, 1983.

[40] N. Gilks, N. Kedersha, M. Ayodele, L. Shen, G. Stoecklin, L. M. Dember, and P. Anderson. Stress granule assembly is mediated by prion-like aggregation of TIA-1. *Molecular Biology of the Cell*, 15(12):5383–5398, December 2004.

[41] N. Kedersha and P. Anderson. Stress granules: sites of mRNA triage that regulate mRNA stability and translatability. *Biochemical Society Transactions*, 30:Molec & Cellular Pharmol Grp, November 2002.

[42] Ana Eulalio, Isabelle Behm-Ansmant, and Elisa Izaurralde. P bodies: at the crossroads of post-transcriptional pathways. *Nature Reviews Molecular Cell Biology*, 8 (1):9–22, January 2007.

[43] F. H. Crick, S. Brenner, Watstobi.rj, and L. Barnett. General nature of genetic code for proteins. *Nature*, 192(480):1227–&, 1961.

[44] R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, Marquise.m, S. H. Merrill, J. R. Penswick, and A. Zamir. Structure of a ribonucleic acid. *Science*, 147(3664): 1462–&, 1965.

[45] Yuichi Taniguchi, Paul J. Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X. Sunney Xie. Quantifying E-coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538, July 2010.

[46] S. G. E. Andersson and C. G. Kurland. Codon preferences in free-living microorganisms. *Microbiological Reviews*, 54(2):198–210, June 1990.

[47] J. R. Warner. The economics of ribosome biosynthesis in yeast. *Trends in Biochemical Sciences*, 24(11):437–440, November 1999.

[48] Chris A. Brackley, M. Carmen Romano, and Marco Thiel. The dynamics of supply and demand in mRNA translation. *Plos Computational Biology*, 7(10):e1002203, October 2011.

[49] Matthew Scott, Stefan Klumpp, Eduard M. Mateescu, and Terence Hwa. Emergence of robust growth laws from optimal regulation of ribosome synthesis. *Molecular Systems Biology*, 10(8), August 2014.

[50] Yury S. Polikanov, Gregor M. Blaha, and Thomas A. Steitz. How hibernation factors RMF, HPF, and YfiA turn off protein synthesis. *Science*, 336(6083):915–918, May 2012.

[51] Cristian Del Campo, Alexander Bartholomaeus, Ivan Fedyunin, and Zoya Ignatova. Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function. *Plos Genetics*, 11(10):e1005613, October 2015.

[52] Jack A. Dunkle and Jamie H. D. Cate. Ribosome structure and dynamics during translocation and termination. *Annual Review of Biophysics, Vol 39*, 39:227–244, 2010.

[53] Hans Bremer and Patrick P Dennis. Modulation of chemical composition and other parameters of the cell at different exponential growth rates. *EcoSal Plus*, 3 (1), 2008.

[54] U. Vogel and K. F. Jensen. The rna chain elongation rate in Escherichia coli depends on the growth-rate. *Journal of Bacteriology*, 176(10):2807–2813, May 1994.

[55] C. G. Kurland. Translational accuracy and the fitness of bacteria. *Annual Review of Genetics*, 26:29–50, 1992.

[56] Grzegorz Kudla, Andrew W. Murray, David Tollervey, and Joshua B. Plotkin. Coding-sequence determinants of gene expression in Escherichia coli. *Science*, 324 (5924):255–258, April 2009.

[57] J. Ross Buchan and Ian Stansfield. Halting a cellular production line: responses to ribosomal pausing during translation. *Biology of the Cell*, 99(9):475–487, September 2007.

[58] Justin Gardin, Rukhsana Yeasmin, Alisa Yurovsky, Ying Cai, Steve Skiena, and Bruce Futcher. Measurement of average decoding rates of the 61 sense codons in vivo. *eLife*, 3, October 2014.

[59] J. Elf, D. Nilsson, T. Tenson, and M. Ehrenberg. Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science*, 300(5626):1718–1722, June 2003.

[60] J. Elf and M. Ehrenberg. What makes ribosome-mediated transcriptional attenuation sensitive to amino acid limitation? *Plos Computational Biology*, 1(1):e2, June 2005.

[61] Gong Zhang, Magdalena Hubalewska, and Zoya Ignatova. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature Structural & Molecular Biology*, 16(3):274–280, March 2009.

[62] Jeremy E. Wilusz. Controlling translation via modulation of tRNA levels. *Wiley Interdisciplinary Reviews-rna*, 6(4):453–470, July 2015.

[63] J. R. Menninger. Metabolic role of peptidyl-transfer-RNA hydrolase .3. peptidyl transfer-rna dissociates during protein-synthesis from ribosomes of Escherichia coli. *Journal of Biological Chemistry*, 251(11):3392–3398, 1976.

[64] Angelo Valleriani, Gong Zhang, Apoorva Nagar, Zoya Ignatova, and Reinhard Lipowsky. Length-dependent translation of messenger RNA by ribosomes. *Physical Review E*, 83(4):042903, April 2011.

[65] F. Jorgensen and C. G. Kurland. Processivity errors of gene-expression in Escherichia coli. *Journal of Molecular Biology*, 215(4):511–521, October 1990.

[66] J. L. Manley. Synthesis and degradation of termination and premature-termination fragments of beta-galactosidase invitro and invivo. *Journal of Molecular Biology*, 125(4):407–432, 1978.

[67] K. Tsung, S. Inouye, and M. Inouye. Factors affecting the efficiency of protein-synthesis in Escherichia coli - production of a polypeptide of more than 6000 amino-acid residues. *Journal of Biological Chemistry*, 264(8):4428–4433, March 1989.

[68] Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S. Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624–635, April 2014.

[69] Gene-Wei Li, Eugene Oh, and Jonathan S. Weissman. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, 484 (7395):538–U172, April 2012.

[70] Eugene Oh, Annemarie H. Becker, Arzu Sandikci, Damon Huber, Rachna Chaba, Felix Gloge, Robert J. Nichols, Athanasios Typas, Carol A. Gross, Guenter Kramer, Jonathan S. Weissman, and Bernd Bukau. Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell*, 147 (6):1295–1308, December 2011.

[71] Ada Yonath. Ribosomes: Ribozymes that survived evolution pressures but is paralyzed by tiny antibiotics. In *Macromolecular Crystallography*, pages 195–208. Springer, 2012.

[72] Bjoern Schwanhaeusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, May 2011.

[73] Juergen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, December 2008.

[74] Carlus Deneke, Reinhard Lipowsky, and Angelo Valleriani. Complex degradation processes lead to non-exponential decay patterns and age-dependent decay rates of messenger RNA. *Plos One*, 8(2):e55442, February 2013.

[75] Ulf Andersson Orom, Finn Cilius Nielsen, and Anders H. Lund. MicroRNA-10a binds the 5 ' UTR of ribosomal protein mRNAs and enhances their translation. *Molecular Cell*, 30(4):460–471, May 2008.

[76] Shobha Vasudevan, Yingchun Tong, and Joan A. Steitz. Switching from repression to activation: MicroRNAs can up-regulate translation. *Science*, 318(5858):1931–1934, December 2007.

[77] Elisa Izaurralde. Elucidating the temporal order of silencing. *Embo Reports*, 13 (8):662–663, August 2012.

[78] Emanuela Repetto, Paola Briata, Nathalie Kuziner, Brian D. Harfe, Michael T. McManus, Roberto Gherzi, Michael G. Rosenfeld, and Michele Trabucchi. Let-7b/c enhance the stability of a tissue-specific mRNA during mammalian organogenesis as part of a feedback loop involving KSRP. *Plos Genetics*, 8(7):e1002823, July 2012.

[79] Fatima Cairrao, Anason S. Halees, Khalid S. A. Khabar, Dominique Morello, and Nathalie Vanzo. AU-rich elements regulate drosophila gene expression. *Molecular and Cellular Biology*, 29(10):2636–2643, May 2009.

[80] Feng Ma, Xingguang Liu, Dong Li, Pin Wang, Nan Li, Liwei Lu, and Xuetao Cao. MicroRNA-466l upregulates IL-10 expression in TLR-triggered macrophages by antagonizing RNA-binding protein tristetraprolin-mediated IL-10 mRNA degradation. *Journal of Immunology*, 184(11):6053–6059, June 2010.

[81] Q. Jing, S. Huang, S. Guth, T. Zarubin, A. Motoyama, J. M. Chen, F. Di Padova, S. C. Lin, H. Gram, and J. H. Han. Involvement of microRNA in AU-rich element-mediated mRNA instability. *Cell*, 120(5):623–634, March 2005.

[82] Yvonne Tay, John Rinn, and Pier Paolo Pandolfi. The multilayered complexity of ceRNA crosstalk and competition. *Nature*, 505(7483):344–352, January 2014.

[83] Reena V. Kartha and Subbaya Subramanian. Competing endogenous RNAs (ceRNAs): new entrants to the intricacies of gene regulation. *Frontiers in Genetics*, 5, January 2014.

[84] Joel G. Belasco. All things must pass: contrasts and commonalities in eukaryotic and bacterial mRNA decay. *Nature Reviews Molecular Cell Biology*, 11(7):467–478, July 2010.

[85] Jian Lu and Andrew G. Clark. Impact of microRNA regulation on variation in human gene expression. *Genome Research*, 22(7):1243–1254, July 2012.

[86] Carlus Deneke, Reinhard Lipowsky, and Angelo Valleriani. Effect of ribosome shielding on mRNA stability. *Physical biology*, 10(4):046008, 2013.

[87] Qianqian Wu, Kate Smith-Miles, Tianshou Zhou, and Tianhai Tian. Stochastic modelling of biochemical systems of multi-step reactions using a simplified two-variable model. *Bmc Systems Biology*, 7:S14, October 2013.

[88] A. Valleriani, S. Liepelt, and R. Lipowsky. Dwell time distributions for kinesin's mechanical steps. *Epl*, 82(2):28011, April 2008.

[89] Peter Keller and Angelo Valleriani. Single-molecule stochastic times in a reversible bimolecular reaction. *Journal of Chemical Physics*, 137(8):084106, August 2012.

[90] Angelo Valleriani, Xin Li, and Anatoly B. Kolomeisky. Unveiling the hidden structure of complex stochastic biochemical networks. *Journal of Chemical Physics*, 140 (6):064101, February 2014.

[91] Stephanie Helfer, Johanna Schott, Georg Stoecklin, and Klaus Foerstemann. AU-rich element-mediated mRNA decay can occur independently of the miRNA machinery in mouse embryonic fibroblasts and drosophila S2-cells. *Plos One*, 7(1): e28907, January 2012.

[92] Latifa Zekri, Duygu Kuzuoglu-Oeztuerk, and Elisa Izaurralde. GW182 proteins cause PABP dissociation from silenced miRNA targets in the absence of deadenylation. *Embo Journal*, 32(7):1052–1065, April 2013.

[93] Eulalia de Nadal, Gustav Ammerer, and Francesc Posas. Controlling gene expression in response to stress. *Nature Reviews Genetics*, 12(12):833–845, December 2011.

[94] Malin Akerfelt, Richard I. Morimoto, and Lea Sistonen. Heat shock factors: integrators of cell stress, development and lifespan. *Nature Reviews Molecular Cell Biology*, 11(8):545–555, August 2010.

[95] Thusitha S. Gunasekera, Laszlo N. Csonka, and Oleg Paliy. Genome-wide transcriptional responses of Escherichia coli k-12 to continuous osmotic and heat stresses. *Journal of Bacteriology*, 190(10):3712–3720, May 2008.

[96] Luis Lopez-Maury, Samuel Marguerat, and Juerg Baehler. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews Genetics*, 9(8):583–593, August 2008.

[97] Klaus Richter, Martin Haslbeck, and Johannes Buchner. The heat shock response: Life on the verge of death. *Molecular Cell*, 40(2):253–266, October 2010.

[98] T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292 (5518):929–934, May 2001.

[99] Suzanne Komili and Pamela A. Silver. Coupling and coordination in gene expression processes: a systems biology view. *Nature Reviews Genetics*, 9(1):38–48, January 2008.

[100] Christine Vogel and Edward M. Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4):227–232, April 2012.

[101] Nicholas T. Ingolia, Sina Ghaemmaghami, John R. S. Newman, and Jonathan S. Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223, April 2009.

[102] A. J. Cozzone and G. S. Stent. Movement of ribosomes over messenger-RNA in polysomes of Rel+ and Rel- Escherichia coli strains. *Journal of Molecular Biology*, 76(1):163–179, 1973.

[103] Huili Guo, Nicholas T. Ingolia, Jonathan S. Weissman, and David P. Bartel. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–U66, August 2010.

[104] Shaomei He, Omri Wurtzel, Kanwar Singh, Jeff L. Froula, Suzan Yilmaz, Susannah G. Tringe, Zhong Wang, Feng Chen, Erika A. Lindquist, Rotem Sorek, and Philip Hugenholtz. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nature Methods*, 7(10):807–U58, October 2010.

[105] Nicholas T. Ingolia. Genome-wide translational profiling by ribosome footprinting. *Methods in Enzymology, Vol 470: Guide to Yeast Genetics*, 2010.

[106] Lichun Jiang, Felix Schlesinger, Carrie A. Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R. Gingeras, and Brian Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, 21(9):1543–1551, September 2011.

[107] Gong Zhang, Ivan Fedyunin, Sebastian Kirchner, Chuanle Xiao, Angelo Valleriani, and Zoya Ignatova. FANSe: an accurate algorithm for quantitative mapping of large scale sequencing reads. *Nucleic Acids Research*, 40(11):e83, June 2012.

[108] Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5(7):621–628, July 2008.

[109] Marie-Agnes Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Celine Keime, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloe, Caroline Le Gall, Brigitte Schaeffer, Stephane Le Crom, Mickael Guedj, and Florence Jaffrezic. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6): 671–683, November 2013.

[110] James H. Bullard, Elizabeth Purdom, Kasper D. Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *Bmc Bioinformatics*, 11:94, February 2010.

[111] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-methodological*, 57(1):289–300, 1995.

[112] Ronny Lorenz, Stephan H. Bernhart, Christian Hoener Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6:26, November 2011.

[113] Kajetan Bentele, Paul Saffert, Robert Rauscher, Zoya Ignatova, and Nils Bluethgen. Efficient translation initiation dictates codon usage at gene start. *Molecular Systems Biology*, 9:675, June 2013.

[114] Lars B. Scharff, Liam Childs, Dirk Walther, and Ralph Bock. Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. *Plos Genetics*, 7(6):e1002155, June 2011.

[115] H. Bremer, P. Dennis, and M. Ehrenberg. Free RNA polymerase and modeling global transcription in Escherichia coli. *Biochimie*, 85(6):597–609, June 2003.

[116] Daniel Zenklusen, Daniel R. Larson, and Robert H. Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural & Molecular Biology*, 15(12):1263–1271, December 2008.

[117] Samuel Marguerat, Alexander Schmidt, Sandra Codlin, Wei Chen, Ruedi Aebersold, and Juerg Baehler. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell*, 151(3):671–683, October 2012.

[118] H. Tao, C. Bausch, C. Richmond, F. R. Blattner, and T. Conway. Functional genomics: Expression analysis of Escherichia coli growing on minimal and rich media. *Journal of Bacteriology*, 181(20):6425–6440, October 1999.

[119] Daniel Hebenstreit, Miaoqing Fang, Muxin Gu, Varodom Charoensawan, Alexander van Oudenaarden, and Sarah A. Teichmann. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Molecular Systems Biology*, 7:497, June 2011.

[120] Carlus Deneke, Sophia Rudorf, and Angelo Valleriani. Transient phenomena in gene expression after induction of transcription. *Plos One*, 7(4):e35044, April 2012.

[121] Eric Guisbert, Takashi Yura, Virgil A. Rhodius, and Carol A. Gross. Convergence of molecular, modeling, and systems approaches for an understanding of the Escherichia coli heat shock response. *Microbiology and Molecular Biology Reviews*, 72(3):545–+, September 2008.

[122] I. R. Booth, J. Cairney, L. Sutherland, and C. F. Higgins. Enteric bacteria and osmotic-stress - an integrated homeostatic system. *Journal of Applied Bacteriology*, 65:S35–S49, 1988.

[123] J. M. Wood, E. Bremer, L. N. Csonka, R. Kraemer, B. Poolman, T. van der Heide, and L. T. Smith. Osmosensing and osmoregulatory compatible solute accumulation by bacteria. *Comparative Biochemistry and Physiology A-molecular*

and Integrative Physiology, 130(3):European Soc Comparat Physiol & Biochem, October 2001.

[124] A. Weber and K. Jung. Profiling early osmostress-dependent gene expression in Escherichia coli using DNA microarrays. *Journal of Bacteriology*, 184(19):5502–5507, October 2002.

[125] Botao Liu, Yan Han, and Shu-Bing Qian. Cotranslational response to proteotoxic stress by elongation pausing of ribosomes. *Molecular Cell*, 49(3):453–463, February 2013.

[126] Reut Shalgi, Jessica A. Hurt, Irina Krykbaeva, Mikko Taipale, Susan Lindquist, and Christopher B. Burge. Widespread regulation of translation by elongation pausing in heat shock. *Molecular Cell*, 49(3):439–452, February 2013.

[127] R. A. Mooney, S. A. Darst, and R. Landick. Sigma and RNA polymerase: An on-again, off-again relationship? *Molecular Cell*, 20(3):335–345, November 2005.

[128] Axel Mogk, Damon Huber, and Bernd Bukau. Integrating protein homeostasis strategies in prokaryotes. *Cold Spring Harbor Perspectives in Biology*, 3(4): a004366, April 2011.

[129] Jens Kortmann and Franz Narberhaus. Bacterial RNA thermometers: molecular zippers and switches. *Nature Reviews Microbiology*, 10(4):255–265, April 2012.

[130] F. Arsene, T. Tomoyasu, and B. Bukau. The heat shock response of Escherichia coli. *International Journal of Food Microbiology*, 55(1-3):Conseil Reg Bretagne; Conseil Gen Finistere; European Community; VilleEOLEOLQuimper; Univ Bretagne Occidentale; Technopole Quimper Cornouaille;EOLEOLCNRS; ADRIA; Soc Francaise Microbiol, April 2000.

[131] J. W. Erickson and C. A. Gross. Identification of the sigma-e subunit of Escherichia coli RNA-polymerase - a 2nd alternate sigma-factor involved in high-temperature gene-expression. *Genes & Development*, 3(9):1462–1471, September 1989.

[132] Daniel B. Goodman, George M. Church, and Sriram Kosuri. Causes and effects of N-Terminal codon bias in bacterial genes. *Science*, 342(6157):475–479, October 2013.

[133] Richard J Jackson, Ann Kaminski, and Tuija AA Pöyry. Coupled termination-reinitiation events in mRNA translation. *Cold Spring Harbor Monograph Archive*, 48:197–223, 2007.

[134] J. Ma, A. Campbell, and S. Karlin. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *Journal of Bacteriology*, 184(20):5733–5745, October 2002.

[135] Arnim Weber, Stephanie A. Koegl, and Kirsten Jung. Time-dependent proteome alterations under osmotic stress during aerobic and anaerobic growth in Escherichia coli. *Journal of Bacteriology*, 188(20):7165–7175, October 2006.

[136] H. Bremer and P. Dennis. Feedback control of ribosome function in Escherichia coli. *Biochimie*, 90(3):493–499, March 2008.

[137] J. R. Warner, J. Vilardell, and J. H. Sohn. Economics of ribosome biosynthesis. *Cold Spring Harbor Symposia on Quantitative Biology*, 66:567–574, 2001.

[138] Premal Shah, Yang Ding, Malwina Niemczyk, Grzegorz Kudla, and Joshua B. Plotkin. Rate-limiting steps in yeast protein translation. *Cell*, 153(7):1589–1601, June 2013.

[139] Matthew Scott, Carl W. Gunderson, Eduard M. Mateescu, Zhongge Zhang, and Terence Hwa. Interdependence of cell growth and gene expression: Origins and consequences. *Science*, 330(6007):1099–1102, November 2010.

[140] M. Violet Lee, Scott E. Topper, Shane L. Hubler, James Hose, Craig D. Wenger, Joshua J. Coon, and Audrey P. Gasch. A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Molecular Systems Biology*, 7: 514, July 2011.

[141] V. Ramakrishnan. Ribosome structure and the mechanism of translation. *Cell*, 108(4):557–572, February 2002.

[142] K. C. Keiler, P. R. H. Waller, and R. T. Sauer. Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger rna. *Science*, 271 (5251):990–993, February 1996.

[143] Kenneth C. Keiler. Mechanisms of ribosome rescue in bacteria. *Nature Reviews Microbiology*, 13(5):285–297, May 2015.

[144] Hani S. Zaher and Rachel Green. A primary role for Release Factor 3 in quality control during translation elongation in Escherichia coli. *Cell*, 147(2):396–408, October 2011.

[145] Yuhei Chadani, Katsuhiko Ono, Shin-ichiro Ozawa, Yuichiro Takahashi, Kazuyuki Takai, Hideaki Nanamiya, Yuzuru Tozawa, Kazuhiro Kutsukake, and Tatsuhiko Abo. Ribosome rescue by Escherichia coli ArfA (YhdL) in the absence of trans-translation system. *Molecular Microbiology*, 78(4):796–808, November 2010.

[146] Yuhei Chadani, Katsuhiko Ono, Kazuhiro Kutsukake, and Tatsuhiko Abo. Escherichia coli YaeJ protein mediates a novel ribosome-rescue pathway distinct from SsrA- and ArfA-mediated pathways. *Molecular Microbiology*, 80(3):772–785, May 2011.

[147] Hani S. Zaher and Rachel Green. Quality control by the ribosome following peptide bond formation. *Nature*, 457(7226):161–U51, January 2009.

[148] M. A. Gilchrist and A. Wagner. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *Journal of Theoretical Biology*, 239 (4):417–434, April 2006.

[149] CG Kurland and Riitta Mikkola. The impact of nutritional state on the microevolution of ribosomes. In *Starvation in bacteria*, pages 225–237. Springer, 1993.

[150] S. D. Hooper and O. G. Berg. Gradients in nucleotide and codon usage along Escherichia coli genes. *Nucleic Acids Research*, 28(18):3517–3523, September 2000.

[151] Olga L Gurvich, Pavel V Baranov, Jiadong Zhou, Andrew W Hammer, Raymond F Gesteland, and John F Atkins. Sequences that direct significant levels of frameshifting are frequent in coding regions of escherichia coli. *The EMBO Journal*, 22(21):5941–5950, 2003.

[152] Pavel V Baranov, John F Atkins, and Martina M Yordanova. Augmented genetic decoding: global, local and temporal alterations of decoding processes and codon meaning. *Nature Reviews Genetics*, 2015.

[153] Gong Zhang, Ivan Fedyunin, Oskar Miekley, Angelo Valleriani, Alessandro Moura, and Zoya Ignatova. Global and local depletion of ternary complex limits translational elongation. *Nucleic acids research*, 38(14):4778–4787, 2010.

[154] A. Valleriani, Z. Ignatova, A. Nagar, and R. Lipowsky. Turnover of messenger RNA: Polysome statistics beyond the steady state. *Epl*, 89(5):58003, March 2010.

[155] Luca Ciandrini, Ian Stansfield, and M. Carmen Romano. Ribosome traffic on mRNAs maps to gene ontology: Genome-wide quantification of translation initiation rates and polysome size regulation. *Plos Computational Biology*, 9(1):e1002866, January 2013.

[156] Shlomi Reuveni, Isaac Meilijson, Martin Kupiec, Eytan Ruppin, and Tamir Tuller. Genome-scale analysis of translation elongation with a ribosome flow model. *Plos Computational Biology*, 7(9):e1002127, September 2011.

[157] Peng Lu, Christine Vogel, Rong Wang, Xin Yao, and Edward M. Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotechnology*, 25(1):117–124, January 2007.

[158] Nicholas T. Ingolia, Gloria A. Brar, Noam Stern-Ginossar, Michael S. Harris, Gaeele J. S. Talhouarne, Sarah E. Jackson, Mark R. Wills, and Jonathan S. Weissman. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Reports*, 8(5):1365–1379, September 2014.

[159] Arvind R. Subramaniam, Brian M. Zid, and Erin K. O'Shea. An integrated approach reveals regulatory controls on bacterial translation elongation. *Cell*, 159 (5):1200–1211, November 2014.

[160] Rembrandt J. F. Haft, David H. Keating, Tyler Schwaegler, Michael S. Schwalbach, Jeffrey Vinokur, Mary Tremaine, Jason M. Peters, Matthew V. Kotlajich, Edward L. Pohlmann, Irene M. Ong, Jeffrey A. Grass, Patricia J. Kiley, and Robert Landick. Correcting direct effects of ethanol on translation and transcription machinery confers ethanol tolerance in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 111(25):E2576–E2585, June 2014.

[161] Monica S. Guo, Taylor B. Updegrove, Emily B. Gogol, Svetlana A. Shabalina, Carol A. Gross, and Gisela Storz. MicL, a new sigma(e)-dependent sRNA, combats envelope stress by repressing synthesis of Lpp, the major outer membrane lipoprotein. *Genes & Development*, 28(14):1620–1634, July 2014.

[162] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1): 207–210, January 2002.

[163] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, April 2010.

[164] R Core Team et al. *R: A language and environment for statistical computing.* Vienna, Austria, 2013.

[165] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.

[166] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–U54, April 2012.

[167] Paul Julian Kersey, James E. Allen, Mikkel Christensen, Paul Davis, Lee J. Falin, Christoph Grabmueller, Daniel Seth Toney Hughes, Jay Humphrey, Arnaud Kerhornou, Julia Khobova, Nicholas Langridge, Mark D. McDowall, Uma Maheswari, Gareth Maslen, Michael Nuhn, Chuang Kee Ong, Michael Paulini, Helder Pedro, Iliana Toneva, Mary Ann Tuli, Brandon Walts, Gareth Williams, Derek Wilson, Ken Youens-Clark, Marcela K. Monaco, Joshua Stein, Xuehong Wei, Doreen Ware, Daniel M. Bolser, Kevin Lee Howe, Eugene Kulesha, Daniel Lawson, and Daniel Michael Staines. Ensembl genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Research*, 42(D1):D546–D552, January 2014.

[168] OJ Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–&, 1961.

[169] Sarah E. Barchinger and Sarah E. Ades. Regulated proteolysis: control of the Escherichia coli sigma(E)-dependent cell envelope stress response. *Sub-cellular biochemistry*, 66:129–60, 2013.

[170] Tamir Tuller, Asaf Carmi, Kalin Vestsigian, Sivan Navon, Yuval Dorfan, John Zaborske, Tao Pan, Orna Dahan, Itay Furman, and Yitzhak Pilpel. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2):344–354, 2010.

[171] R. Schoenheimer. The dynamic state of body constituents. *The dynamic state of body constituents.*, pages 78–Pp., 1942.

[172] C. Deduve and R. Wattiaux. Functions of lysosomes. *Annual Review of Physiology*, 28:435–&, 1966.

[173] Avram Hershko and Aaron Ciechanover. The ubiquitin system for protein degradation. *Annual review of biochemistry*, 61(1):761–807, 1992.

[174] Aaron Ciechanover. Intracellular protein degradation: from a vague idea, through the lysosome and the ubiquitin–proteasome system, and onto human diseases and drug targeting (nobel lecture). *Angewandte Chemie International Edition*, 44(37): 5944–5967, 2005.

[175] A. L. Goldberg and D. I. C. E. J. F. Intracellular protein degradation in mammalian and bacterial-cells. *Annual Review of Biochemistry*, 43:835–869, 1974.

[176] R. T. Schimke and D. Doyle. Control of enzyme levels in animal tissues. *Annual Review of Biochemistry*, 39:929–&, 1970.

[177] Ryan E. Tyler, Margaret M. P. Pearce, Thomas A. Shaler, James A. Olzmann, Ethan J. Greenblatt, and Ron R. Kopito. Unassembled CD147 is an endogenous

endoplasmic reticulum-associated degradation substrate. *Molecular Biology of the Cell*, 23(24):4668–4678, December 2012.

[178] Cl Ward and R. R. Kopito. Intracellular turnover of cystic-fibrosis transmembrane conductance regulator - inefficient processing and rapid degradation of wild-type and mutant proteins. *Journal of Biological Chemistry*, 269(41):25710–25718, October 1994.

[179] Feng Wang, Larissa A. Durfee, and Jon M. Huibregtse. A cotranslational ubiquitination pathway for quality control of misfolded proteins. *Molecular Cell*, 50(3): 368–378, May 2013.

[180] Woong Kim, Eric J. Bennett, Edward L. Huttlin, Ailan Guo, Jing Li, Anthony Possemato, Mathew E. Sowa, Ramin Rad, John Rush, Michael J. Comb, J. Wade Harper, and Steven P. Gygi. Systematic and quantitative assessment of the ubiquitin-modified proteome. *Molecular Cell*, 44(2):325–340, October 2011.

[181] U. Schubert, L. C. Anton, J. Gibbs, C. C. Norbury, J. W. Yewdell, and J. R. Bennink. Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature*, 404(6779):770–774, April 2000.

[182] R. M. Vabulas and F. U. Hartl. Protein synthesis upon acute nutrient restriction relies on proteasome function. *Science*, 310(5756):1960–1963, December 2005.

[183] J. S. Andersen, Y. W. Lam, A. K. L. Leung, S. E. Ong, C. E. Lyon, A. I. Lamond, and M. Mann. Nucleolar proteome dynamics. *Nature*, 433(7021):77–83, January 2005.

[184] Mary K. Doherty, Dean E. Hammond, Michael J. Clagule, Simon J. Gaskell, and Robert J. Beynon. Turnover of the human proteome: Determination of protein intracellular stability by dynamic SILAC. *Journal of Proteome Research*, 8(1): 104–112, January 2009.

[185] Izumi V. Hinkson and Joshua E. Elias. The dynamic state of protein turnover: It's about time. *Trends in Cell Biology*, 21(5):293–303, May 2011.

[186] Marko Jovanovic, Michael S. Rooney, Philipp Mertins, Dariusz Przybylski, Nicolas Chevrier, Rahul Satija, Edwin H. Rodriguez, Alexander P. Fields, Schraga Schwartz, Raktima Raychowdhury, Maxwell R. Mumbach, Thomas Eisenhaure, Michal Rabani, Dave Gennert, Diana Lu, Toni Delorey, Jonathan S. Weissman, Steven A. Carr, Nir Hacohen, and Aviv Regev. Dynamic profiling of the protein life cycle in response to pathogens. *Science*, 347(6226):1259038, March 2015.

[187] Anders R. Kristensen, Joerg Gsponer, and Leonard J. Foster. Protein synthesis rate is the predominant regulator of protein expression during differentiation. *Molecular Systems Biology*, 9:689, September 2013.

[188] Mark Larance, Yasmeen Ahmad, Kathryn J. Kirkwood, Tony Ly, and Angus I. Lamond. Global subcellular characterization of protein degradation using quantitative proteomics. *Molecular & Cellular Proteomics*, 12(3):638–650, March 2013.

[189] Matthias Selbach, Bjoern Schwanhaeusser, Nadine Thierfelder, Zhuo Fang, Raya Khanin, and Nikolaus Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, September 2008.

[190] D. C. Dieterich, A. J. Link, J. Graumann, D. A. Tirrell, and E. M. Schuman. Selective identification of newly synthesized proteins in mammalian cells using bioorthogonal noncanonical amino acid tagging (BONCAT). *Proceedings of the National Academy of Sciences of the United States of America*, 103(25):9482–9487, June 2006.

[191] K. L. Kiick, E. Saxon, D. A. Tirrell, and C. R. Bertozzi. Incorporation of azides into recombinant proteins for chemoselective modification by the Staudinger ligation. *Proceedings of the National Academy of Sciences of the United States of America*, 99(1):19–24, January 2002.

[192] Katrin Eichelbaum and Jeroen Krijgsveld. Rapid temporal dynamics of transcription, protein synthesis, and secretion during macrophage activation. *Molecular & Cellular Proteomics*, 13(3):792–810, March 2014.

[193] Katrin Eichelbaum, Markus Winter, Mauricio Berriel Diaz, Stephan Herzig, and Jeroen Krijgsveld. Selective enrichment of newly synthesized proteins for quantitative secretome analysis. *Nature Biotechnology*, 30(10):984–+, October 2012.

[194] Laurie D. Cohen, Rina Zuchman, Oksana Sorokina, Anke Mueller, Daniela C. Dieterich, J. Douglas Armstrong, Tamar Ziv, and Noam E. Ziv. Metabolic turnover of synaptic proteins: Kinetics, interdependencies and implications for synaptic maintenance. *Plos One*, 8(5):e63191, May 2013.

[195] Andrew J. M. Howden, Vincent Geoghegan, Kristin Katsch, Georgios Efstathiou, Bhaskar Bhushan, Omar Boutureira, Benjamin Thomas, David C. Trudgian, Benedikt M. Kessler, Daniela C. Dieterich, Benjamin G. Davis, and Oreste Acuto. QuaNCAT: quantitating proteome dynamics in primary cells. *Nature Methods*, 10 (4):343–+, April 2013.

[196] Susanne Tom Dieck, Lisa Kochen, Cyril Hanus, Maximilian Heumueller, Ina Bartnik, Belquis Nassim-Assir, Katrin Merk, Thorsten Mosler, Sakshi Garg, Stefanie

Bunse, David A. Tirrell, and Erin M. Schuman. Direct visualization of newly synthesized target proteins in situ. *Nature Methods*, 12(5):411–+, May 2015.

[197] H. Akaike. New look at statistical-model identification. *Ieee Transactions on Automatic Control*, AC19(6):716–723, 1974.

[198] Noboru Mizushima and Masaaki Komatsu. Autophagy: Renovation of cells and tissues. *Cell*, 147(4):728–741, November 2011.

[199] Alessandro Ori, Murat Iskar, Katarzyna Buczak, Panagiotis Kastritis, Luca Parca, Amparo Andres-Pons, Stephan Singer, Peer Bork, and Martin Beck. Spatiotemporal variation of mammalian protein complex stoichiometries. *Genome Biology*, 17:47, March 2016.

[200] A. L. Goldberg. Protein degradation and protection against misfolded or damaged proteins. *Nature*, 426(6968):895–899, December 2003.

[201] Y. Kuriyama and T. Omura. Different turnover behavior of phenobarbital-induced and normal NADPH-cytochrome-c reductases in rat liver microsomes. *Journal of Biochemistry*, 69(4):659–&, 1971.

[202] I. Blikstad, W. J. Nelson, M. O. O. N. R. T., and E. Lazarides. Synthesis and assembly of spectrin during avian erythropoiesis - stoichiometric assembly but unequal synthesis of alpha-spectrin and beta-spectrin. *Cell*, 32(4):1081–1091, 1983.

[203] P. R. Johnson, R. Swanson, L. Rakhilina, and M. Hochstrasser. Degradation signal masking by heterodimerization of MAT alpha 12 and MATa1 blocks their mutual destruction by the ubiquitin-proteasome pathway. *Cell*, 94(2):217–227, July 1998.

[204] Yun Wah Lam, Angus I. Lamond, Matthias Mann, and Jens S. Andersen. Analysis of nucleolar protein dynamics reveals the nuclear degradation of ribosomal proteins. *Current Biology*, 17(9):749–760, May 2007.

[205] Y. Minami, A. M. Weissman, L. E. Samelson, and R. D. Klausner. Building a multichain receptor - synthesis, degradation, and assembly of the T-cell antigen receptor. *Proceedings of the National Academy of Sciences of the United States of America*, 84(9):2688–2692, May 1987.

[206] Anna Shemorry, Cheol-Sang Hwang, and Alexander Varshavsky. Control of protein quality and stoichiometries by N-Terminal acetylation and the N-End rule pathway. *Molecular Cell*, 50(4):540–551, May 2013.

[207] Brandon H. Toyama, Jeffrey N. Savas, Sung Kyu Park, Michael S. Harris, Nicholas T. Ingolia, III Yates, John R., and Martin W. Hetzer. Identification

of long-lived proteins reveals exceptional stability of essential cellular structures. *Cell*, 154(5):971–982, August 2013.

[208] Christine Leibiger, Nadezda Kosyakova, Hasmik Mkrtchyan, Michael Glei, Vladimir Trifonov, and Thomas Liehr. First molecular cytogenetic high resolution characterization of the NIH 3T3 cell line by murine multicolor banding. *Journal of Histochemistry & Cytochemistry*, 61(4):306–312, 2013.

[209] Silvia Stingele, Gabriele Stoehr, Karolina Peplowska, Juergen Cox, Matthias Mann, and Zuzana Storchova. Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Molecular Systems Biology*, 8:608, September 2012.

[210] Terry Hassold, Heather Hall, and Patricia Hunt. The origin of human aneuploidy: where we have been, where we are going. *Human Molecular Genetics*, 16:R203–R208, October 2007.

[211] Stefano Santaguida and Angelika Amon. Short- and long-term effects of chromosome mis-segregation and aneuploidy. *Nature Reviews Molecular Cell Biology*, 16 (8):473–485, August 2015.

[212] Noah Dephoure, Sunyoung Hwang, Ciara O'Sullivan, Stacie E. Dodgson, Steven P. Gygi, Angelika Amon, and Eduardo M. Torres. Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. *eLife*, 3, July 2014.

[213] Tamar Geiger, Juergen Cox, and Matthias Mann. Proteomic changes resulting from gene copy number variations in cancer cells. *Plos Genetics*, 6(9):e1001090, September 2010.

[214] Stefanie Duttler, Sebastian Pechmann, and Judith Frydman. Principles of cotranslational ubiquitination and quality control at the ribosome. *Molecular Cell*, 50(3): 379–393, May 2013.

[215] Yujin E. Kim, Mark S. Hipp, Andreas Bracher, Manajit Hayer-Hartl, and F. Ulrich Hartl. Molecular chaperone functions in protein folding and proteostasis. *Annual Review of Biochemistry, Vol 82*, 82:323–355, 2013.

[216] Tanya Vavouri, Jennifer I. Semple, Rosa Garcia-Verdugo, and Ben Lehner. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell*, 138(1):198–208, July 2009.

[217] U. de Lichtenberg, L. J. Jensen, S. Brunak, and P. Bork. Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724–727, February 2005.

[218] Joseph A. Marsh, Helena Hernandez, Zoe Hall, Sebastian E. Ahnert, Tina Perica, Carol V. Robinson, and Sarah A. Teichmann. Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell*, 153(2):461–470, April 2013.

[219] Or Matalon, Amnon Horovitz, and Emmanuel D. Levy. Different subunits belonging to the same protein complex often exhibit discordant expression levels and evolutionary properties. *Current Opinion in Structural Biology*, 26:113–120, June 2014.

[220] Einav Simon and Daniel Kornitzer. Pulse-chase analysis to measure protein degradation. *Laboratory Methods in Enzymology: Protein Pt A*, 536:65–75, 2014.

[221] Romain Christiano, Nagarjuna Nagaraj, Florian Froehlich, and Tobias C. Walther. Global proteome turnover analyses of the yeasts S-cerevisiae and S-pombe. *Cell Reports*, 9(5):1959–1965, December 2014.

[222] Peter Landgraf, Elmer R. Antileo, Erin M. Schuman, and Daniela C. Dieterich. BONCAT: metabolic labeling, click chemistry, and affinity purification of newly synthesized proteomes. *Methods in molecular biology (Clifton, N.J.)*, 1266:199–215, 2015.

[223] Mark Larance and Angus I. Lomond. Multidimensional proteomics for cell biology. *Nature Reviews Molecular Cell Biology*, 16(5):269–280, May 2015.

[224] Lars Doelken, Zsolt Ruzsics, Bernd Raedle, Caroline C. Friedel, Ralf Zimmer, Joerg Mages, Reinhard Hoffmann, Paul Dickinson, Thorsten Forster, Peter Ghazal, and Ulrich H. Koszinowski. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *Rna-a Publication of the Rna Society*, 14(9):1959–1972, September 2008.

[225] Hidenori Tani and Nobuyoshi Akimitsu. Genome-wide technology for determining RNA stability in mammalian cells historical perspective and recent advantages based on modified nucleotide labeling. *Rna Biology*, 9(10):1233–1238, October 2012.

[226] Qingbo Li. Advances in protein turnover analysis at the global level and biological insights. *Mass Spectrometry Reviews*, 29(5):717–736, September 2010.

[227] Lan K. Nguyen, Maciej Dobrzynski, Dirk Fey, and Boris N. Kholodenko. Polyubiquitin chain assembly and organization determine the dynamics of protein activation and degradation. *Frontiers in Physiology*, 5:UNSP 4, January 2014.

[228] C. E. Clayton. Networks of gene expression regulation in Trypanosoma brucei. *Molecular and Biochemical Parasitology*, 195(2):96–106, July 2014.

[229] Abeer Fadda, Mark Ryten, Dorothea Droll, Federico Rojas, Valentin Faerber, Jurgen R. Haanstra, Clemetine Merce, Barbara M. Bakker, Keith Matthews, and Christine Clayton. Transcriptome-wide analysis of trypanosome mRNA decay reveals complex degradation kinetics and suggests a role for co-transcriptional degradation in determining mRNA levels. *Molecular Microbiology*, 94(2):307–326, October 2014.

[230] Fiona Achcar, Abeer Fadda, Jurgen R. Haanstra, Eduard J. Kerkhoven, Dong-Hyun Kim, Alejandro E. Leroux, Theodore Papamarkou, Federico Rojas, Barbara M. Bakker, Michael P. Barrett, Christine Clayton, Mark Girolami, R. Luise Krauth-Siegel, Keith R. Matthews, and Rainer Breitling. The silicon trypanosome: A test case of iterative model extension in systems biology. *Advances in Microbial Systems Biology*, 64:115–143, 2014.

[231] Javad Noorbakhsh, Alex H. Lang, and Pankaj Mehta. Intrinsic noise of microRNA-regulated genes and the ceRNA-hypothesis. *Plos One*, 8(8):e72676, August 2013.

[232] N. E. Buchler, U. Gerland, and T. Hwa. Nonlinear protein degradation and the function of genetic circuits. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9559–9564, July 2005.

[233] Howard M Taylor and Samuel Karlin. *An introduction to stochastic modeling.* Academic press, 2014.

[234] Jay R Greenberg. High stability of mRNA in growing cultured cells. *Nature*, 240 (5376):102–&, 1972.

[235] Benjamin Neymotin, Rodoniki Athanasiadou, and David Gresham. Determination of in vivo RNA kinetics using RATE-seq. *Rna*, 20(10):1645–1652, October 2014.

[236] Johanna Schott, Sonja Reitter, Janine Philipp, Katharina Haneke, Heiner Schaefer, and Georg Stoecklin. Translational regulation of specific mRNAs controls feedback inhibition and survival during macrophage activation. *Plos Genetics*, 10 (6):e1004368, June 2014.

[237] Anne Schwabe and Frank J. Bruggeman. Contributions of cell growth and biochemical reactions to nongenetic variability of cells. *Biophysical Journal*, 107(2): 301–313, July 2014.

[238] S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as

a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics*, 1(5):376–386, May 2002.

[239] Jingyi Hou, Xi Wang, Erik McShane, Henrik Zauber, Wei Sun, Matthias Selbach, and Wei Chen. Extensive allele-specific translational regulation inhybrid mice. *Molecular Systems Biology*, 11(8):825, August 2015.

[240] J. Rappsilber, Y. Ishihama, and M. Mann. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Analytical Chemistry*, 75(3):663–670, February 2003.

[241] Murat Eravci, Christian Sommer, and Matthias Selbach. IPG strip-based peptide fractionation for shotgun proteomics. *Shotgun Proteomics: Methods and Protocols*, 1156:67–77, 2014.

[242] Jacek R. Wisniewski, Alexandre Zougman, and Matthias Mann. Combination of FASP and stagetip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *Journal of Proteome Research*, 8(12):5674–5678, December 2009.

[243] Maria E. Sheean, Erik McShane, Cyril Cheret, Jan Walcher, Thomas Mueller, Annika Wulf-Goldenberg, Soraya Hoelper, Alistair N. Garratt, Markus Krueger, Klaus Rajewsky, Dies Meijer, Walter Birchmeier, Gary R. Lewin, Matthias Selbach, and Carmen Birchmeier. Activation of MAPK overrides the termination of myelin growth and replaces Nrg1/ErbB3 signals during Schwann cell development and myelination. *Genes & Development*, 28(3):290–303, February 2014.

[244] M. Puchades, A. Westman, K. Blennow, and P. Davidsson. Removal of sodium dodecyl sulfate from protein samples prior to matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Communications in Mass Spectrometry*, 13(5): 344–349, 1999.

[245] Matthias D. Sury, Erik McShane, Luis Rodrigo Hernandez-Miranda, Carmen Birchmeier, and Matthias Selbach. Quantitative proteomics reveals dynamic interaction of c-Jun N-terminal Kinase (JNK) with RNA transport granule proteins splicing factor proline-and glutamine-rich (Sfpq) and Non-POU domain-containing octamer-binding protein (Nono) during neuronal differentiation. *Molecular & Cellular Proteomics*, 14(1):50–65, January 2015.

[246] Akira Motoyama, Tao Xu, Cristian I. Ruse, James A. Wohlschlegel, and I. I. I. Yates, John R. Anion and cation mixed-bed ion exchange for enhanced multidimensional separations of peptides and phosphopeptides. *Analytical Chemistry*, 79 (10):3623–3634, May 2007.

[247] Pietro Sormanni, Carlo Camilloni, Piero Fariselli, and Michele Vendruscolo. The s2D method: Simultaneous sequence-based prediction of the statistical populations of ordered and disordered regions in proteins. *Journal of Molecular Biology*, 427 (4):982–996, February 2015.

[248] Martyn D. Winn, Charles C. Ballard, Kevin D. Cowtan, Eleanor J. Dodson, Paul Emsley, Phil R. Evans, Ronan M. Keegan, Eugene B. Krissinel, Andrew G. W. Leslie, Airlie McCoy, Stuart J. McNicholas, Garib N. Murshudov, Navraj S. Pannu, Elizabeth A. Potterton, Harold R. Powell, Randy J. Read, Alexei Vagin, and Keith S. Wilson. Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D-biological Crystallography*, 67:235–242, April 2011.

[249] Jonathan N. Wells, L. Therese Bergendahl, and Joseph A. Marsh. Operon gene order is optimized for ordered protein complex assembly. *Cell Reports*, 14(4): 679–685, February 2016.

[250] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H. Werner Mewes. CORUM: the comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Research*, 38:D497–D501, January 2010.

[251] Yasunobu Okamura, Yuichi Aoki, Takeshi Obayashi, Shu Tadaka, Satoshi Ito, Takafumi Narise, and Kengo Kinoshita. COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic Acids Research*, 43(D1): D82–D86, January 2015.

[252] K. Thomaseth and C. Cobelli. Generalized sensitivity functions in physiological system identification. *Annals of Biomedical Engineering*, 27(5):607–616, September 1999.

[253] HT Banks, Sava Dediu, and Stacey L Ernstberger. Sensitivity functions and their uses in inverse problems. *Journal of Inverse and Ill-posed Problems jiip*, 15(7): 683–708, 2007.

[254] Yunfei Chu and Juergen Hahn. Parameter set selection via clustering of parameters into pairwise indistinguishable groups of parameters. *Industrial & Engineering Chemistry Research*, 48(13):6000–6009, July 2009.