HeLa Proteome: Cracking the code of RNA-dependency

Characterization of proteins based on biophysical properties

Mirjam Biollaz, Hasset Gessese, Jette Klempt-Gießing, Alicia Weeber

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

IPMB

Introduction

Biological background:
RNA forms complexes with proteins known as ribonucleoprotein complexes. These play a central role in both RNA metabolism and the regulation of gene expression. They are collectively referred to as RNA-dependent proteins (Fig.1). This group includes RNA-binding proteins (RBPs), which can bind directly to RNA, as well as RBP-binding proteins, whose interactome depends on RNA without directly binding to it[2].

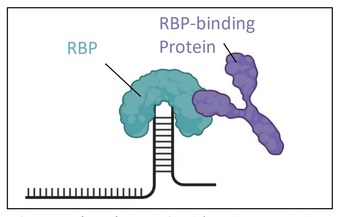


Fig. 1: RNA-dependent proteins. *Schematic representation of the binding behavior of RBPs and RBP-binding proteins.*

Dataset Generation :
Proteins from HeLa cells synchronized in interphase were analyzed. For this, a HeLa cell lysate was either treated with an RNase cocktail or left untreated as a control. Both samples were applied in triplicates to a sucrose density gradient, separated into 25 fractions based on density and size, and the protein content in each fraction was subsequently quantified using mass spectrometry[2] (Figure 2).

Dataset Structure:
- Dimensions: 7,086 × 150
- Rows: One protein per row
- Columns: Control and RNase × 25 fractions × 3 replicates

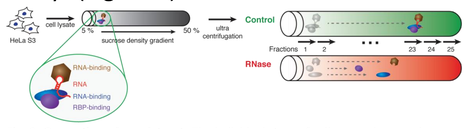
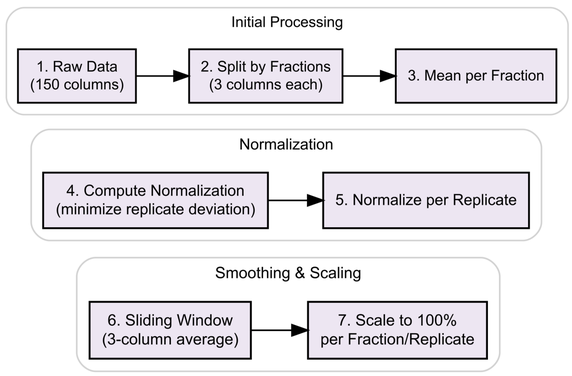


Fig. 2: Generation of our dataset. *The individual steps of cell processing are shown, as well as the behavior of the proteins during centrifugation*

1 Clean-up



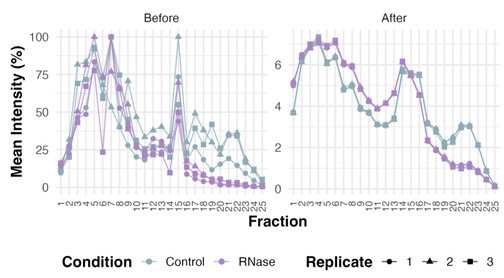


Fig. 3: Mean protein intensities per fraction before and after normalization
Normalization using mean-centering and sliding window approaches improves the alignment of protein intensity distributions across replicates

Reproducibility
The Pearson correlation coefficients between replicates are very high (mostly > 0.98, close to 1.00) for all replicate pairs indicates strong linear agreement across replicates, reflecting high reproducibility in measurements.
- The small spread in the lower tails suggests few outliers
- All three replicate pairs exhibit similar distributions, implying no replicate stands out as poor quality or inconsistent.

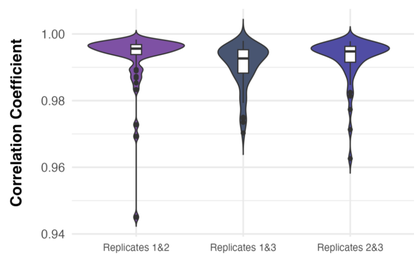


Fig. 4: Reproducibility between the replicates. *Distributions of each replicate throughout all fractions.*

2 Shift-Analysis

Goal: Assign each protein to a shift category based on criteria we defined ourselves

First Step: Finding maxima and shoulder points in the protein distribution
For each protein, the local and absolute maxima are determined from the mean values of the replicates (threshold = 2). In addition to the normal peaks, the plateau maxima and the edge maxima are also considered.
Shoulder points are obtained that represent a group of adjacent fractions (at least 4 fractions next to each other) and are not identified as maxima but still showed a large amount of protein.

Second Step: Classification of the protein shifts
Determine the properties of the protein distributions for each protein:

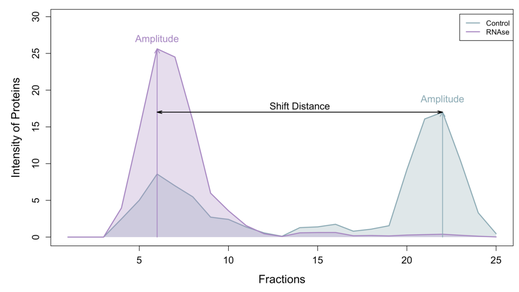


Fig. 5: Protein distribution. *Shows how the protein behaves within the fractions after centrifugation, comparing control and RNase treatment.*

Properties of the proteins using the example ACOT9_HUMAN:
I. Number of maxima for control and RNAse: 2, 1
II. Distance of maxima: 0, -16 (RNAse - control)
III. Shift-Index (sum of all distances): -16
IV. Sum of RNase amplitude gain: 17
V. area under the peaks: 25, 50 (Control) + 75 (RNase)
VI. p-value of the amplitude difference: $7{,}4 * 10^{-4}$

Paired two-sample T-test was performed:
- Same variance → Student t-test
- Two-tailed
- P-value < 0,05 → Significant

Based on these properties, a custom function is developed that defines specific criteria a protein must meet in order to show a shift. For each protein, it is assessed how many of these criteria are fulfilled. This results in an individual shift score which indicates how many different categories apply to a given protein. In consequence of this score, the proteins are then categorized (table 1).

Table 1: Assignment of shift categories to the shift score

Shift-Score	Shift-Category
≥ 5	Strong shift
≥ 4	Moderate shift
< 4	No shift

The result of this analysis is shown in figure 6. Using our function, all of our proteins are successfully assigned to a shift category. The majority of them show no shifts and can be categorized as non-RBPs while the remaining proteins are RBPs.

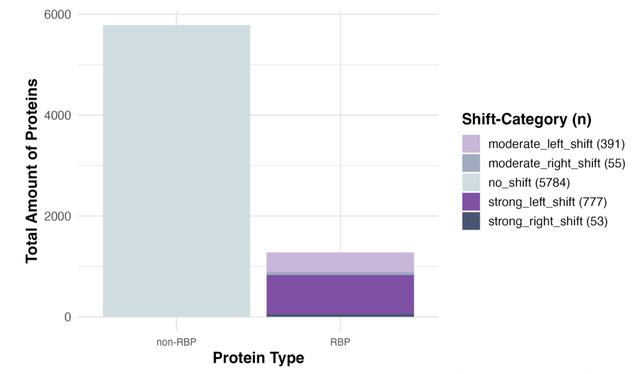
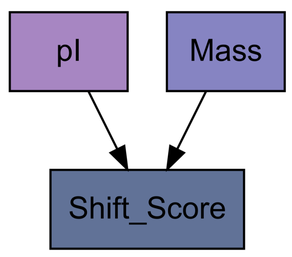


Fig. 8: Result of shift category classification. Shows the number of proteins in each shift category based on our function.

3 Modelling – further analysis of RNA dependent proteins

I. Linear regression: Isoelectric point and mass as predictors for RBP-binding proteins

Looking for different biophysical characteristics based on databases[1,3] as predictors for the shift score, the pI and mass turned out to have a significant correlation with the shift score:
- pI: $< 2 * 10^{-16}$
- mass: $1{,}85 * 10^{-6}$



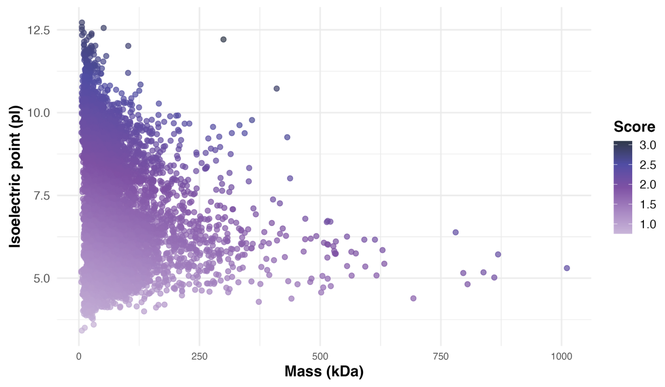


Fig. 7: Correlations between mass, pI and shift score. *Shown is the distribution of proteins by mass and isoelectric point with each protein colored according to the corresponding shift score.*

II. Data reduction and cluster analysis: shift categories remain separable after PCA based on regression variables
K-means clustering based on the results of the linear regression and coloring each shift category in the PCA scatterplot shows that mass and pI contribute to the shift score without distorting the underlying classification.

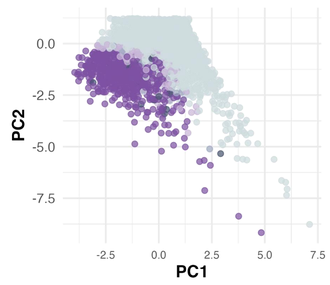


Fig. 8.1: PCA: Clustering by pI, mass and shift score. *The k-means clustering plot (k=4) is colored by the shift category of the proteins. The different shift categories are separated.*

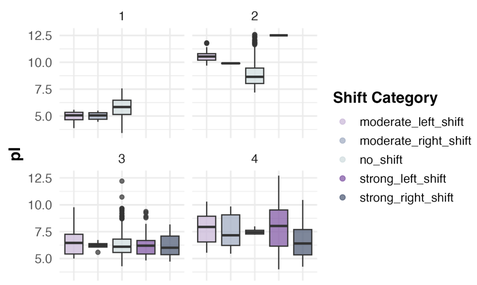


Fig. 8.2: pI values per shift category in each cluster (k=4). *It is shown that the left shift proteins are increasingly present in the fourth cluster.*

Obtaining the pI for each shift category showed a significant difference between shifting and not shifting proteins.

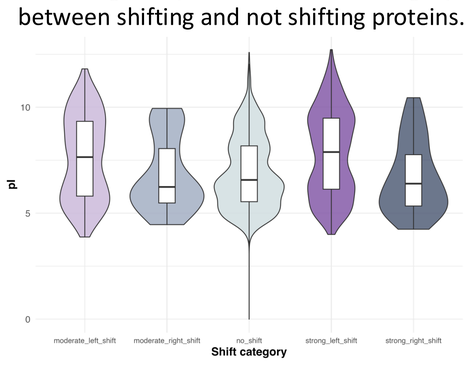
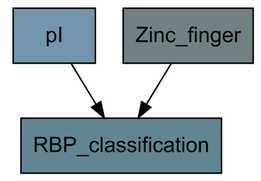


Fig. 9: Violin plot of the pI distribution in each shift category. The median is marked in each category with a line that shows that the median pI of moderate and left shift proteins is higher than those of no or right shift proteins.

Mean of pI based on RNA dependency:
- RBP: pI = 7.47
- non_RBP: pI = 6.91

III. Logistic regression: Zinc finger motives may indicate RNA dependency
The goal is to analyze more characteristics of RNA dependent proteins along with pI (as a known influential trait). Zinc finger motives shows a significant correlation with the RBP classification.



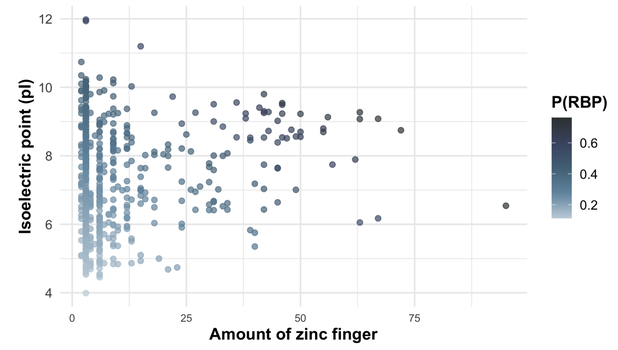
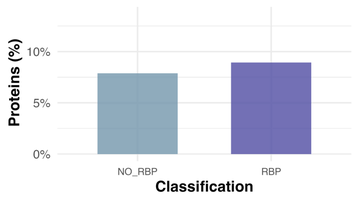


Fig.10: Predicted probability for RBP based on pI and mass. *Depicted is the correlation between pI and mass . The predicted probability for RBP classification is colored.*

Zinc fingers are relatively more common in classified RBPs



The difference of Zinc finger motives in RBP classification was statistically tested with wilcoxon rank sum test.

Fig. 11: Relative amount of proteins with zinc finger motive based on RBP classification. *In the RBP category there are 8.9% with zinc finger motives and within the non RBPs there are 7.9%.*

Discussion

- Through protein classification into shift categories, we could distinguish RBPs from non-RBPs. Compared to previous studies[1,3], 53% of RBPs were classified as non-RBPs, and 18% of non-RBPs as RBPs. The test statistics show high specificity (88%) but low sensitivity (36%), suggesting limited detection of all proteins by the function. This is a limited assumption, as studies only separate RBPs and non-RBPs. Many proteins may be RNA-dependent without directly binding RNA.

- The results of the linear regression indicate that the isoelectric point differs in RBPs compared to non-RBPs. The pI is higher in RBPs than in non-RBPs. That coincides with the expectation that RBPs are charged more positively in order to interact with negatively charged RNA. The higher pI in RBPs is also found in literature proving the accuracy of the analysis[2].
- The conclusion from the logistic regression is that zinc finger motives probably occur more often in RBPs than in non-RBPs. That result is also found in literature and fits the expectation that the motive is one way to bind to RNA[4]. To further analyze the characteristics of RNA-dependent proteins more data would have been needed as the amount of information for each characteristic often is not sufficient in the data that was used.

References

1. Ahmad S, da Costa Gonzales L J, Bowler-Barnett E H, Rice D L, Kim M, Wijerathne S, Luciani A, Kandasaamy S, Luo J, Watkins X, Turner E, Martin M J, UniProt Consortium The UniProt website API: facilitating programmatic access to protein knowledge Nucleic Acids Research, gkaf394 (2025) (https://doi.org/10.1093/nar/gkaf394)
2. Caudron-Herger et al., R-DeeP: Proteome-wide and Quantitative Identification of RNA-Dependent Proteins by Density Gradient Ultracentrifugation. 2019, Molecular Cell 75, 184-199
3. Caudron-Herger et al., RBP2GO: A comprehensive, pan-species database on RNA-binding proteins, their interactions and functions 2021, Nucleic Acids Research, gkaa1040
4. Corley M, Burns MC, Yeo GW. How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. Cell 2020; 184(6):1276–1290