universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Analysis of the connection between biological function of genes and the binding patterns of their respective mRNAs with RNA-binding proteins“

verfasst von / submitted by

David Eisele BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2022 / Vienna, 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 834
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Molekulare Biologie
Betreut von / Supervisor:	Univ.-Prof. Dr Bojan Zagrovic, BA

Summary

Due to the rise of high-throughput approaches, which provide new, biologically relevant data for vast numbers of genes, categorizing these genes into functional groups has become increasingly important. While there already exist well-structured gene ontologies (GO), it is still a challenge to define the functional relationships between protein coding genes. The establishment and evaluation of quantitative metrics for assessing functional relationships relies on correlations with additional information derived from gene expression or sequence data. The present thesis explores the possibility that CLIP-seq (crosslinking and immunoprecipitation with sequencing) data could be used to group cellular mRNAs according to their binding patterns with RNA-binding proteins (RBPs) and critically compares such classification with the established ways of grouping genes. For this, we relate the mRNA-RBP interaction pattern data with the expression pattern data from the Human Protein Atlas, the sequence-based protein families from the Pfam database and Gene Ontology (GO) tags which group genes according to molecular function, cellular component and biological process. Furthermore, we identify which tags fit best with the binding pattern grouping and attempt to find a known functional connection between tags and binding patterns. We demonstrate that mRNAs with more similar binding patterns are also more likely to share Pfam, GO and tissue expression tags and identify tags that coincide best with the respective binding patterns. The present work represents and early attempt to quantitatively link mRNA-RBP interaction patterns with biological function and opens up several venues for future exploration.

Zusammenfassung

Durch den Aufstieg von high-throughput Methoden, die neue, biologisch relevante Daten für eine große Anzahl von Genen zur Verfügung stellen, ist die Kategorisierung dieser Gene in funktionelle Gruppen extrem wichtig geworden. Obwohl es bereits gut strukturierte Gen Ontologien (GO) gibt, ist es noch immer eine Herausforderung, funktionelle Beziehungen zwischen Protein-codierenden Genen zu definieren. Die Feststellung und Evaluierung von quantitativen Metriken, um funktionelle Beziehungen festzustellen, bauen auf Korrelationen mit Zusatzinformationen, die aus Genexpressions- oder Sequenzdaten entstehen, auf. Die jetzige These untersucht die Möglichkeit, dass CLIP-seq (Crosslinking und Immunoprecipitation mit Sequenzierung) Daten genutzt werden könnten, um zelluläre mRNAs nach ihren Bindungsmustern von mRNA-bindenden Proteinen (RBPs) einzuteilen und diese Einteilung kritisch mit etablierten Einteilungen von Genen zu vergleichen. Dafür haben wir die mRNA-RBP Interaktionsmusterdaten mit Expressionsmusterdaten von The Human Protein Atlas, den sequenzbasierten Proteinfamilien der Pfam Datenbank und Gene

Ontology tags, die Gene nach Molekulare Funktion, Zellulären Komponent und Biologischen Prozess gruppieren, zugeordnet. Außerdem haben wir festgestellt, welche tags am besten mit dem Bindungsmuster übereinstimmen und zu versuchen, eine bekannte funktionelle Verbindungen zwischen tags und Bindungsmustern herzustellen. Wir demonstrieren, dass mRNAs mit ähnlicheren Bindungsmustern auch eine höhere Wahrscheinlichkeit haben, Pfam-, GO- und Zellexpressions-tags zu teilen und identifiziern tags, die am Besten mit den entsprechenden Bindungsmustern zusammenhängen. Die jetzige Arbeit stellt einen frühen Versuch dar, mRNA-RBP Interaktionsmuster mit biologischen Funktionen in Verbindung zu bringen und eröffnet mehrere Wege für zukünftige Erforschung.

Acknowledgments

Special thanks to Thomas Kapral, Bojan Zagrovic, Anton Polyansky and Stefanie Brandstetter for their support, advice and helpful comments.

Introduction

Classifying Gene Function

Most research on human genes focuses on approximately only 2000 of the approximately 19,000 human genes [1]. The biological function of the remaining 17,000 genes are clearly underexplored or not explored at all [1]. One way of uncovering these unknown functions is to find commonalities between the well understood genes and the underexplored genes according to some well-defined property. This may then allow one to group genes together and predict certain features, related to the biological function of the unknown genes by association. Typically, genes and/or gene products are grouped by their function [2][3][4][5], sequence similarity/evolutionary relatedness [2][6], the correlation of their expression profiles [7][8] or their protein-protein interactome [9], as further detailed bellow.

Function based methods

The first and most obvious way of grouping genes is function. It does, however, require a preexisting understanding of at least some functions in question. Knowing that two genes share a particular function, may increase the chances that they also have other, additional functions in common, enabling extrapolative predictions. However, the problem is to experimentally determine a set of overlapping functions, a challenge that may not always be possible. Moreover, having one or a few shared functions does not provide any guarantees that other functions will be shared. Another issue is to create a database that where all genes/proteins possessing the function have been correctly annotated to possess it. There are databases for very specific functions such as Postar3 that includes RNA-binding proteins that have been confirmed by CLIP [10] and also very general ones that attempt to give as much information about a protein as possible (all known functions, sequence, structure, location, etc.) like Uniprot [11]. Since there are also non-protein coding RNAs with known functions, their functions also need to be collected and annotated, as in the GeneCaRNA subsection of GeneCards [12]. Gene Ontology also has tags related to molecular function in addition to cellular component and biological process tags [3][4][5]. As GO is hierarchical there are tags that are more closely related to each other and tags that are functionally more distant. This can be a challenge for functionally grouping tags in cases where grouping tags of related function would be helpful. A paper by del Pozo et al. developed a method to detect functional distances between GO tags [13].

Function is connected to both sequence information and expression levels since proteins with similar sequence usually possess similar function [2] and genes expressed at similar levels in the same tissue are more likely to possess similar biological function [14]. There are also proteins with similar function that do not share a similar sequence. These proteins usually evolved their similar function independently of each other meaning that these proteins will be seen as similar by a function based method and dissimilar by a sequence based methods.

Sequence based methods

Another way of identifying genes that are functionally related to each other is by sequence similarity, which is a well-established approach that also works well for establishing evolutionary relatedness [2]. Importantly, genes/proteins that evolved the same function independently are frequently not similar in sequence. This fact is well understood and there are many cases of domains with the same or similar function being very different in sequence [2][6]. Conversely, if the motif sequences are similar, sequence based approaches frequently allow one to not only predict biological function, but also makes it possible to infer the mechanism by which the function occurs.

Well-established databases for identifying protein domains/motifs by their sequence are Pfam [2] and Prosite [6]. The approaches how these methods identify those sequence elements are, however, different. Pfam builds a profile hidden Markov model from a seed alignment and searches the sequences in the database to identify those that satisfy the family threshold [2]. Prosite on the other hand is built to generate profiles, which are score matrices specific for every position and are best suited to predict structural properties of entire domains, and patterns, short sequences that are best suited for identifying a specific protein function in proteins with a highly similar sequence [15].

Expression based methods

The third way of grouping genes is by their expression patterns. Microarrays have enabled us to get genome-wide expression data in a vast number of different cell types [7]. These methods are useful for identifying differences in gene expression between tissues or between healthy and diseased cells [8]. In this way, we can, for example, identify genes enriched in cell types with a certain function. Genes that are enriched/depleted in the same cell types can be viewed as similar and thus grouped together. The issue, however, is that due to a high number of genes and each gene being tested for enrichment within a certain cell type requiring a hypothesis testing (enriched or not enriched), there is a risk of false positives (genes identified as enriched when they are not) and false negatives (genes identified as not enriched when they actually are) [8]. The problem is that many of the established multiple hypothesis corrections to reduce false positives do not work well, because

microarray data does not contain independent variables or the methods are considered too stringent [8].

Interactome based methods

The final way to group genes that will be discussed here is by their interactome. The interactome is often used to refer to the protein-protein interaction (PPI) network [9]. However, more broadly, every molecular interaction within a cell can be considered a part of a wider interactome. The basic idea of the protein interactome is that proteins that interact with each other are likely to co-localize and are both involved in the same process [9]. For example, the Human Reference Protein Interactome mapping project managed to identify 53,000 protein-protein interactions [9]. Interactome based methods could be seen as a sub-type of function-based clustering since being able to interact with another protein/molecule can be viewed as a function. The GO tag for nucleotide binding for example is a Molecular Function tag [3][4][5].

RNA-RBP interaction patterns

This project aims to find whether grouping mRNAs by their RNA-binding-protein (RBP) interaction profiles can be used as an alternative grouping method to find commonalities between genes, and whether such an approach will return a grouping in agreement with the classical, well-established ways of grouping genes. It can be seen as a type of interactome-based grouping method. However, while many interactome groupings focus on protein-protein interactions, the present approach focuses on RNA-RBP interactions instead. This approach is biologically relevant because RBPs are known to regulate and process mRNAs, thus contributing to their maturation [16]. This means that the RBP binding pattern contains information about which processing RNAs undergo and how they are regulated so mRNAs with similar binding patterns have a higher likelihood of having similar regulation and processing. In addition, we study both the biological significance and intrinsic limitations of such an approach by identifying which of the gene groups best fit with the grouping obtained by their binding pattern only.

The canonical view of the interaction between an RNA and an RBP is that RBPs possess one or more, mostly structured RNA binding domains with a length of under 100 residues [17]. The most common of these is the RNA recognition motif (RRM) [17][18], while others include: K homology domain (KH), DEAD box helicase domain, zinc Finger, PUM, PUF, PUA, THUMP, YTH, dsRBDs, CSD, S1, Sm, LAM, PAZ and PIWI [17]. Another common feature is that RBPs are often enriched in intrinsically disordered regions which also often have RNA binding capability [17]. Notably, it is common for an RBP to possesses multiple different binding domains. Some of these binding domains bind cooperatively, while others bind independently of each other [19][20].

Typically, the flexibility and length of the so-called linkers (sequences between two RNA binding domains) have an effect on whether the binding is independent or cooperative [19][20]. While there are proteins that bind both RNA and DNA [17], in most cases, these respective functions are carried out by different sets of proteins [17]. RNA and DNA are very different in structure and thus some protein-nucleotide interactions are only possible in one but not the other. In particular, DNA is significantly more rigid and lacks the 2' OH on the sugar moiety which many RBPs interact with when binding RNA. This means that such a protein cannot interact with DNA, at least not by this mechanism [20]. Another difference is that even when RNA does form a basepaired structure – which it usually does not do over its whole length – it is usually not the B-form helix typical for DNA. This means that even if a protein is suited to bind dsDNA, it is not automatically guaranteed to bind dsRNA as well [21]. Finally, in the past decade or so, the researchers have also discovered a number of novel RNA binders without known RNA binding domains, expanding our understanding of RBPs [18]. Not only are these newly discovered RBPs involved in different biological processes, but the interaction mechanism appears to be RNA-driven rather than protein-driven [18]. These new findings show that RBPs are a very diverse group of proteins and it is reasonable to assume the RNAs bound by one class of RBPs may be functionally distinct from another class of RNAs interacting with a different set of RBPs, a hypothesis explored in this work.

CLIP

In recent years, CLIP-seq has been established as a powerful method for detecting RNA-RBP interactions on a cell-wide level [22]. The name CLIP-seq (or CLIP for short) stands for ***C***ross-***l***inking ***i***mmuno***p***recipitation, referring to cross-linking via UV radiation and immunoprecipitation to purify the cross-linked complexes [23], and sequencing, which has become much more efficient thanks to the development of high-throughput methods invented in recent years. Thanks to these, it is now possible to get the CLIP data of entire transcriptomes, enabling the creation of rich interaction maps [24]. Since we can get CLIP data of the entire transcriptome, we should be able to capture a large number of the all RBPs interacting with the RNAs within the transcriptome. There might be some exceptions for RBPs with very low concentration that do not meet all potential interaction partners or RBPs with poor cross-linking capabilities/interacting with RNAs with poor cross-linking capabilities, but in general, most RBPs interacting with the transcribed RNAs should be captured. If this assumption is correct, a large number of all RBPs, including those involved in RNA regulation and processing, should be captured by the CLIP data [16].

A typical CLIP experiment usually consists of the following steps: UV radiation cross-links RNA uridines and to a lesser degree guanosines in a cell to RBPs in their close proximity. Importantly, the proximity cross-linking occurs at short enough distances such that an occurrence of a cross-link

is seen as evidence of direct interaction with at least some atoms being within a bonding distance. In the next step, the cells are lysed and the RNAs digested into fragments using RNase. The cross-linked RNA-protein complexes are then purified via immunoprecipitation and, in a second purification step, via SDS-PAGE. Finally, the protein is digested with proteinase K, so only the RNA remains which is then used as a template to create cDNA, which can then be sequenced to identify the RNA it comes from [23]. There are multiple different variants of CLIP, the most important being HITS-CLIP [25], PAR-CLIP [26], iCLIP [27] and eCLIP [28].

The CLIP data used in this analysis was taken from the Postar3 database [10] and includes data from different cell cultures and tissues, which was generated with the different CLIP methods named above. Data from experiments involving overexpression and induction was filtered out in order to avoid potential artifacts due to the concentration dependence of RNA-protein interactions [22]. The colleague in the Zagrovic Group Thomas Kapral, MSc developed a method to make the results obtained by using different CLIP methods, different peak callers and different tissues comparable (see Methods: CLIP). Specifically, for each gene, the top 10, top 20 and top 50 most significantly interacting RBPs were determined, if available. In this approach, only the best binders were considered to be actual binders, making the specific binders and the promiscuous binders easier to compare. The reason for this is that a promiscuous RNA and a specific RNA that bind the same RBPs most strongly should not be considered very distinct merely because the promiscuous RNA binds some additional RBPs weakly. The approach also mitigates the problem that a higher concentration increases the probability of rarer interactions being detected [22] since only the highly ranked, strongest interactions are considered.

Clustering

Older sources recommend k-means clustering over hierarchical clustering [29], but in recent years, k-means has fallen out of favor [30]. While some criticize that the number of clusters must be fixed beforehand, it is also acknowledged that k-means is superior to hierarchical clustering for small datasets [30]. For this analysis, we analyze rather large datasets, so the advantage of k-means clustering in small datasets is irrelevant to us. The fixed cluster number would not be a problem since we fix the number of clusters anyway to make comparison of different clustering procedures easier. However, k-means is a heuristic approach, meaning the results can vary between runs. This was the main reason why k-means was not used. Namely, one intention of the project was to compare different clusterings to each other and it could be problematic if it is not clear whether the difference is simply caused by natural variation of results in a heuristic approach or by the two different clusterings actually being different.

Agglomerative Hierarchical Clustering [31], on the other hand, always delivers the same result since it is not a heuristic approach. It is used to group the most similar RNAs together in a bottom-up approach. Each RNA starts as its own cluster and is then merged into a new cluster with the RNAs that most closely resembles its binding pattern. That way, the most similar RNAs are grouped together and more distinct RNAs are put in a different cluster. These clusters are then checked for enrichment of Pfam, GO and tissue expression tags. The aim is to find which tags are most highly enriched and thus explain a potential similarity in RBP-binding-pattern-based grouping and the respective other grouping (Pfam, GO, tissue). Ideally, a known link between an RBP-binding pattern and the tag it correlates to can be established, thus demonstrating a mechanistic reason why one is predictive of the other.

Materials and Methods

Software

The coding language Python and the Python modules sklearn [31], scipy [32], pandas [33], numpy [34], matplotlib [35], pickle [36] and collections were used in data analysis. Sklearn was primarily used for its Agglomerative Hierarchical Clustering, pandas for generating and modifying dataframes containing the interaction profile data, matplotlib for visualization and pickle for creating binary files of Python dictionaries.

Resources

The following ways of grouping genes were used for comparison in this analysis: Pfam tags [2], Gene Ontology IDs [3][4][5] and Protein Atlas tissue expression levels [7].

Gene Ontology tags

Gene Ontology (GO) is the largest available source of information on gene functions [3][4]. GO tags can refer to a biological process, molecular function or cellular component and every gene known to fit a given tag is associated with it. For example, a gene of an RBP will be assigned the molecular function tag referring to that function. GO terms/tags are hierarchical with a broad “parent” term having more specific “child” terms and often those “child” terms have “child” terms of their own. This means there are broad tags like the biological process tag “signaling” that is the parent of tags like “cell-cell signaling” and “extracellular matrix-cell signaling”. More broad tags will thus be more commonly found between genes than very specific tags [3][4][5]. Unfortunately, there are sometimes redundant IDs referring to the same function (RNA binding for example being tagged by GO:0003723, GO:0000498 and GO:0044822). A problem arises when genes with the same function are tagged with different GO IDs, making it appear that they do not share the same function when they actually do [37]. Another issue is that GO can, of course, by its very nature only cover known, but not unknown functions, so an RBP not known to bind RNA will lack a tag referencing it. This means GO tags give reliable information which functions a gene has but the information which functions it lacks is less reliable [3][4][5]. A list of all GO tags associated with the genes in the analysis was generated using the background file generator of Fuento, the Functional Enrichment Tool [38]. To do that the gene names had to be converted into uniprot IDs which was done by using the *gene name to uniprot ID* conversion tool on the Uniprot website [11].

Pfam tags

Protein family (Pfam) database is a major collection of protein families derived from multiple-sequence alignments and Hidden Markov Models [2]. Pfam can be used to predicts the function of protein domains based on the presence of specific aminoacid sequences. Importantly, in Pfam, one sequence is only meant to belong to one single family. This means that different families will never be identified by the same sequence but a protein can possess multiple different sequences belonging to different families. The family sequences have a known function so an unknown protein having the same Pfam sequence can be assumed to share that function as well [2]. During the duration of the present project, the Pfam database was migrated to InterPro where the Pfam database data along with other databases is still available.

Human Protein Atlas Tissue tags

The Human Protein Atlas maps the human proteome in each tissue and groups their expression level into *High*, *Medium*, *Low* and *not-detected* based on immunohistological micro-array data [7]. The reliability of each expression level is categorized as *Enhanced* (“The antibody has enhanced validation and there is no contradicting data, such as literature describing experimental evidence for a different location.“), *Supported* (“There is no enhanced validation of the antibody, but the annotated localization is reported in literature.”), *Approved* (“The localization of the protein has not been previously described and was detected by only one antibody without additional antibody validation.”) or *Uncertain* (“The antibody-staining pattern contradicts experimental data or expression is not detected at RNA level.“) [7]. In the present analysis only proteins with *High* expression levels that were either *Enhanced*, *Supported* or *Approved* were considered to be part of the RNA’s tissue tags.

CLIP data curation

The Postar3 CLIP data came from various different cell cultures and cell types so it was essential to determine which of them had gene expression patterns similar enough to be comparable [10]. Similar expression levels are needed for the CLIP results from different cells to be comparable since cross-linking in CLIP is dosage dependent, meaning that if a protein is more abundant within a cell, it is more likely to be closer to an RNA by random chance than if the protein was less abundant [22]. To determine the expression levels, tissue and cell culture expression data from the Protein Atlas was used [7]. The results showed that HeLa, HEK 293, PC-3, Hep G2, K562, SH-SY5Y and hTERT-HME1 cell cultures all clustered together in a Euclidean Ward clustering of their expression patterns, while heart muscle and adrenal glands form a more distinct cluster, with the brain cells

representing the most distinct group. Because of this, the CLIP data coming from cell cultures was included in the analysis while adrenal gland, heart muscle and brain cells were not (**Figure 1**).

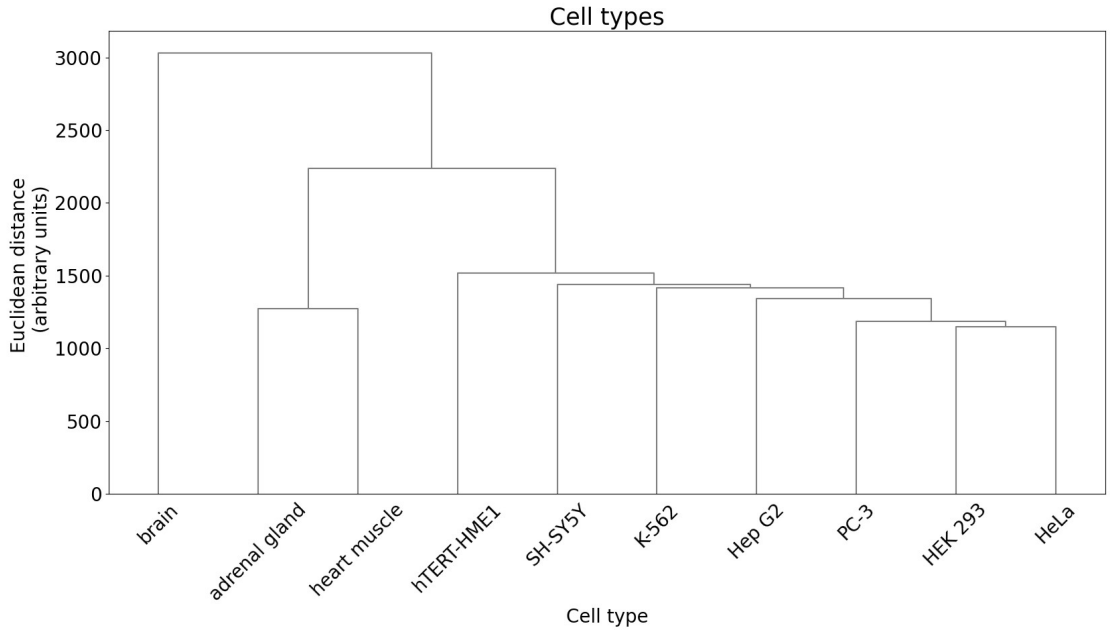


Figure 1: *Euclidean Ward clustering of cell cultures and cell types based on the gene expression pattern (expression data from the Human Proteome Atlas [7])*

The first step to make the CLIP data comparable was to rank all RNA-protein pairs within the same cell culture and peak caller according to their respective peak scores (different methods determine the score differently). Should two pairs have an identical peak score, the number of peaks is used as a tie breaker. In case there are multiple experimental results for the same RNA-protein pairs, the average rank is given to it.

In the next step, the ranks over all the different cell cultures and peak callers within a CLIP method are averaged to obtain the average ranks for a given CLIP method. The final rank is determined by averaging the different CLIP methods' averages with equal weights (such tat a CLIP method that has fewer experiments still has the same impact on the final ranking as the one that has data from more experiments). The highest ranked interaction partners of each RNA (the 10/20/50 highest ranks) are then selected and labeled as interactors. All other proteins with weaker interactions than those top proteins are counted as non-interactors. The intention behind only considering the best binders is to make more promiscuous binders more easily comparable to specific binders because if the top binders of a promiscuous RNA are similar to the top binders of a specific RNA, they can and should be considered to be similar in some of their most interesting aspects. This method also reduces the impact of RNA concentration on the detected interactions (a higher concentration

increases the probability of rarer interactions being detected) [22] by only considering the most highly ranked, strongest interactions.

To see if this approach is valid, the top 10, top 20 and top 50 results are compared to the result of including not only the top interactors, but all of them. To do that, a pairwise comparison of all RNAs is made to determine which RNAs share one or more tags (this is done separately for the Pfam, GO and tissue tags). For each RNA, the other RNAs were binned into 100 bins of 100 RNAs based on the similarity of their interaction profile (referred to as a fingerprint). This fingerprint is a vector of 0s and 1s, where 1 means there is an interaction and 0 that there is not. When 2 RNAs shared one or more tags, it was checked in which bin the similarity of their fingerprint falls and that bin then had +1 added to its score. The bins in which the RNAs did not share tags would have 0 added to their score. After this was done for all possible pairs, the average score of each bin over all RNA-pairs was calculated. If the score is highest for the initial bins and declines for the lower bins, this could be taken as providing support for the claim that similar tags mean similar RBP-binding patterns. If, however, the score is not highest for the first bin and does not decline after, it means that the groupings based on those types of tags and those specific RBP binding patterns do not group the same RNAs together. In an ideal case, the first bin should have a high score that drops off sharply for all the other bins.

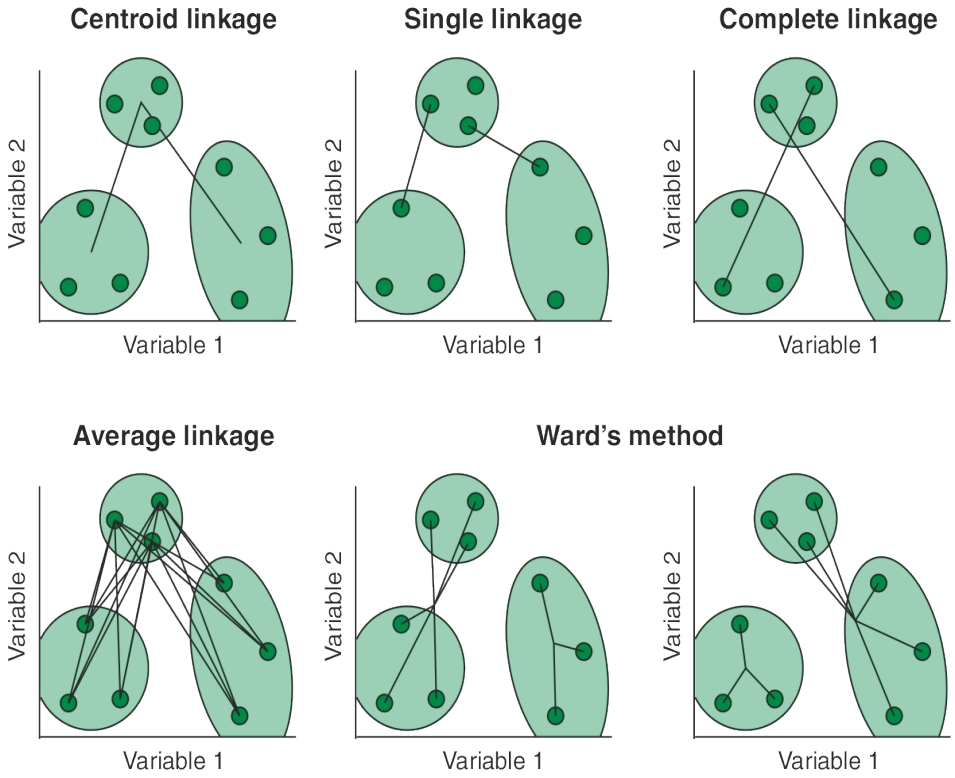
Clustering and Adjacency Lists

Agglomerative Hierarchical Clustering from the Python module sci-kit learn was used to group the genes together based on their RNA-protein interaction pattern. It is a bottom-up clustering method where each element starts out as its own cluster and then gets added to the same cluster as the most similar elements until the desired number of clusters is reached [31] (**Figure 2**). The standard distance metric used by the module is the Euclidean distance which is the shortest distance between 2 points in the n-dimensional space (n in this case being the number of proteins in the interaction matrix). Other distance metrics tested, but not used for the final analysis, were 1) the Pearson correlation coefficient between the interaction patterns and 2) the intersection divided by the union of the intersection. Both of these required us to use *precomputed linkage* after being applied which returned 1 large cluster and many small clusters many containing less than 10 RNAs.

While multiple linkage methods were explored, the linkage method used in the final analysis was Ward's method which keeps the aggregate squared deviation minimal. It does that by calculating the centroid (the mean coordinate of all coordinates within a cluster) and the deviation from the centroid for each possible cluster merge. The clusters with the lowest centroid deviation are merged. Other linkage methods that exist, but were not used in this project are: single linkage (merges clusters based on the distance of their closest two element), absolute linkage (merges clusters based

on the distance of their most distant two element) and average linkage (merges clusters based on the average pairwise distance of their elements) [31].

Figure 2 from *Machine Learning with R, the tidyverse, and mlr* by Rhys *[39]*



Centroid linkage

The distance between the centroids of the cluster is calculated. The clusters with the closest centroid distance are merged [39].

Single linkage

The distance between the closest elements between clusters is calculated. The cluster whose closest elements are closest are merged [39].

Complete linkage

The distance between the most distant elements between clusters is calculated. The cluster whose closest elements are closest are merged [39].

Average linkage

The distance between the elements of two clusters is calculated. The clusters with the lowest average pairwise distance are merged [39].

Ward's method

The centroid and squared variance for both possible merges are calculated. The merge with the lower squared deviation is picked [39].

Adjacency Lists follow a similar goal as the clustering does: grouping similar genes together and exclude dissimilar genes from the list. The main difference is that there can be overlap between two adjacency lists, while the clusters never overlap. An adjacency list includes all RNAs an RNA has a similar binding pattern to and is calculated for each RNA separately. Most of these adjacency lists are over 90% identical, so only the longest lists distinct from each other were analyzed to represent all shorter lists they are similar to. The goal again was to check these RBP-binding pattern based groupings for enrichment in tags to demonstrate which tags are the best predictors of RNA-binding pattern and vice versa.

Results

Categorization

As previously explained, the goal of the preset project was to find out if mRNAs with more similar RBP-binding patterns are also more similar when it comes to their Pfam, GO and tissue expression tags. The top 10 (Top10), top 20 (Top20) and top 50 (Top50) interacting RBPs for each RNA were determined and a fingerprint, an N-dimensional interaction vector consisting of 0s (no interaction with that RBP) and 1s (interaction with that RBP), was generated for each RNA. The results were compared to the result of including not only the top interactors, but all of them (from now on referred to as *All*). For those interaction vectors, the number of interactions between RNAs could vary greatly with some interacting with fewer than 100 RBPs and others with thousands.

The RNA pairs were grouped into 100 bins of 100 RNAs based on the similarity of their interaction fingerprint with similar pairs located in one of the top bins and dissimilar ones in the bottom bins. If both RNAs within a pair shared a tag, +1 was added to the bin's score. The bins in which the RNAs did not share tags had 0 added to their score. The average score over all pairs within a bin was then calculated and plotted (**Figure 3b-c**). Due to the Pfam tags being more rare in general, the average intersection scores were very low compared to the Tissue Expression and GO tags.

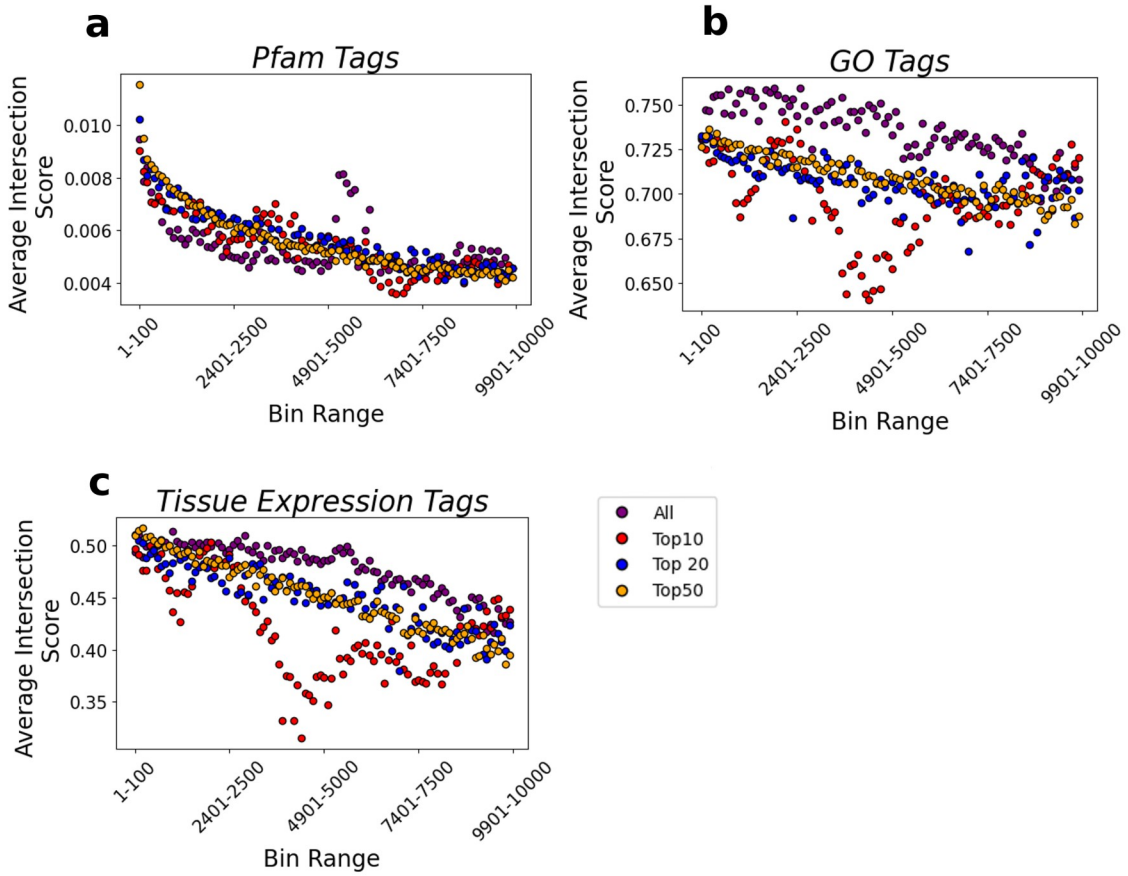


Figure 3: *Average intersection scores of each bin for Pfam (**a**), GO (**b**) and tissue expression (**c**). High scores in the top bins and low scores in the bottom bins indicate that RNA-RBP interaction pattern similarity and the presence of the respective tag (Pfam (**a**), GO (**b**) and tissue expression (**c**)) group genes similarly.*

Almost all bins of the GO (**Figure 3b**) and tissue tags (**Figure 3c**) – regardless if Top10, Top20, Top50 or All – have an average intersection score of 0.5 or higher meaning that over half of the RNAs share one or more tags while the Pfam tags have much lower intersection scores. Their highest average intersection score is around 0.011 meaning that 1.1% of RNAs in the first bin share one or more tags (which is 3 times higher than the lower bins).

Including low affinity binders

For the GO (**Figure 3b**) and tissue tags (**Figure 3c**), the ranking containing all interactors has the highest interaction scores overall. However, it is not the bins containing the highest ranked RNAs that possess the best scores, but rather mid-ranked ones. For the Pfam tags (**Figure 3a**), the highest ranked bin does give the highest score but again there are mid-sized bin outliers with an unusually high score in that some mid-ranked bins have sudden jumps in their score. This effect could come from RNAs with few protein interaction partners being considered more distant from those more

promiscuous ones, even when the promiscuous protein binds all proteins that the more specific one does too since the absence of other bindings increases their Euclidean distance.

Excluding low affinity binders

The first bin of the Top10 has the highest intersection score only for the Pfam tags (**Figure 3a**). Similarly to the All scores (where low affinity binders were excluded), there are sudden score jumps for mid ranked bins, but this is less extreme. In GO (**Figure 3b**) and tissue tags (**Figure 3c**), the mid-ranked bins have the highest score. This means that the Top10 similarity score is not suited to predict an increased likelihood of observing the presence of GO or tissue tags.

In the Top20 data, the Pfam tags (**Figure 3a**) are also the only tag type where the first bin of the Top20 has the highest intersection score. In GO (**Figure 3b**) and tissue tags (**Figure 3c**), on the other hand, the mid-ranked bins have the highest score. There are some moderate score jumps, but they are less extreme than in the Top10 data. This means the Top20 similarity score does not yield an increased likelihood of the presence of GO or tissue tags. However, the unusually high mid-ranked scores are lower than they were for the Top10 data and the sudden jumps and drops in intersection score were less extreme.

The Top50 score (**Figures 4b-c**) for all 3 tag types generally seems to be the highest at the highest ranked bins and then declines. For Pfam tags (**Figure 3c**) the score decline appears rather steep while the GO (**Figure 3b**) and tissue (**Figure 3c**) decline more gradually in an almost linear fashion. The decline from first bin to last bin is largest for the Pfam tags with the top bin’s score being approximately 3 times larger than the bottom bin’s score. The highest ranked GO bin has an intersection score of 0.73 and the lowest bin 0.69 (a difference of 5.5%), while the highest ranked tissue bin has an intersection score of 0.51 and the lowest bin 0.39 (a difference of 23.5%). The GO maximum and the tissue maximum are within 2 standard deviations of the mean.

Excluding tags

The high interaction scores of GO and tissue tags are caused by a small number of tags (31 GO tags and 45 tissue tags) that occur very often (over 1000 times) and are thus found in many RNAs, including those that are extremely different in their RBP-binding patterns. To avoid these tags influencing the results, they were removed and the binning-analysis was repeated without them. This was done for the Top50 data only as the results already showed that the other options are not as well suited: both the Top10 and Top20 results of GO and tissue tags had sudden jumps in their intersection score for mid-sized bins. It is unlikely that these jumps can be explained just by the tags occurring over 1000 times since even Pfam had sudden score spikes in the Top10 and Top20 despite not having any tags occurring over 1000 times. Including All tags and not just the top interactors is

also a poor choice for the similar reasons as for the Top10 and Top20. In addition, it also gives promiscuous RNAs a higher chance to result in an intersection. As they bind more RBPs, the promiscuous RNAs have a higher chance of exhibiting an intersection with other RNAs, which can lead to the low-similarity bins getting a sudden spike in their intersection score (**Figure 3a**)

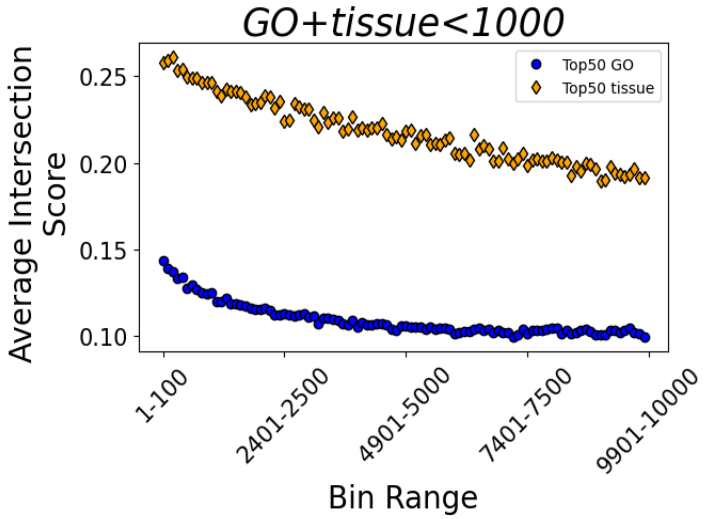


Figure 4: *GO and Tissue Top50 intersection scores with GO and tissue tags occurring over 1000 times removed*

The absolute intersection score for both tissue and GO tags drops rather drastically as compared to including the overabundant tags, meaning that many of the prior intersection scores were due to a few extremely abundant tags. The relative difference between the highest and the lowest bin score has increased from 5.5% to 28.6% for the GO tags and from 23.5% to 26.9% for tissue tags. The GO tag maximum is, thus, within 4 standard deviations of its mean and the tissue tag maximum within 3 standard deviations. This demonstrates that removing the highly abundant tags moved the maximum and the mean further apart.

Clustering

The Top10, Top20 and Top50 interactor data was split once into 10 and once into 50 clusters using sklearn's Agglomerative Hierarchical Clustering with Ward's linkage (**Figure 1***)* [31][39]. It was then calculated if any of the tags were enriched within a cluster by comparing it to how often one would expect a tag to occur by random chance within a cluster. This was carried out by randomly assigning n tags to each cluster (n being the number of tags the real cluster contains) and repeating that randomization 2000 times. Rare tags occurring 50 times or fewer were filtered out as they are too infrequent to give interesting enrichments. Another problem was that their inclusion usually shifted the distribution far away from a normal distribution, thus making the z-scores less straightforward to interpret. From the randomization, a distribution of how often a tag occurs within a cluster by chance was obtained. From this, one calculated the mean and the standard deviation for

each tag within each cluster. With that and the actual frequency of the tag, z-scores could be calculated that were used to identify the 10 most enriched tags within each cluster.

The enrichment of each tag with a high z-score within its cluster was calculated by comparing the expected occurrence of the tag within the cluster based on its size to the observed occurrence of the tag. For the 10 cluster split, the tags enriched at least 5-fold were investigated more closely, while for the 50 cluster data those that increased at least 10-fold were investigated more closely (*SuppMat 1-3*). The Top50 clusterings show more enriched tags than the Top10 and Top20, which is consistent with the finding that in the Top50, the highest ranked bins also shared more interactions.

As **Figures 4b-c** demonstrate, the Top50 approach (only considering the 50 strongest interacting RBPs of an RNA as interactors) consistently proved to be the best choice for defining similar RNAs that then also shared common tags. Thus, we focused on those results for further analysis as they should provide the most reliable data with the most enrichment.

Pfam tag enrichment and z-scores

In one cluster where the Top50 data was split into 50 clusters, the Pfam tag 7tm_1 was is enriched over 85-fold and had a z-score of 12.9 which is the highest enrichment of any tag. The cluster accounts for over 54% of all RNAs with this tag despite the cluster only containing 0.6% of all RNAs. The 7tm_1 tag refers to the G protein-coupled transmembrane receptors of the rhodopsin family. Importantly and somewhat surprisingly, only very few RNAs with the tag within the cluster bind the same protein with PTBP1 being the most commonly bound with 24/85. This indicates that while the RNAs within the cluster are more similar in their binding pattern compared to RNAs outside of it, the cluster still appears to contain a wide range of different interaction patterns.

In the same cluster, Trypsin (z-score=7.4, enrichment: 80-fold), V-set (Immunoglobulin V-set domain, z=6.7, e=46-fold), Lectin_C (Lectin C-type domain, z=6.2, e:74-fold), Ion_trans (Ion transport protein, z=5.0, e:57-fold) and Homeodomain (z=4.3, e:46-fold) were seen to be enriched as well.

Two other Pfam tags that were strongly enriched in two different clusters were RRM_1 (RNA recognition motif, z=17.9, e:18-fold) and Helicase_C (Helicase conserved C-terminal domain, z=13.7, e:20-fold). Both of these tags are classic RNA binding domains[18]. We found that all 24 RRM_1 tagged RNAs in the RRM_1 enriched cluster bound GRWD1 and 22 of them bound UCHL5, ZNF622 and BCLAF1. For Helicase_C we found that 20/24 of the RNAs with the tag bind UCHL5 and 19 bind FMR1, GRWD1, ZNF622 and FXR2.

Gene Ontology tags enrichment and z-scores

The GO tag with the highest z-scores across 11 clusters is GO:0005515 (protein binding) with the z-scores being between 24.1 and 49.5. The enrichment of this tag is below 5-fold in every cluster and tags with far lower z-scores possess significantly higher enrichment. However, the tags with the highest enrichment within a cluster usually have poor z-scores. Another tag with high z-scores across multiple clusters (z as high as 20) is GO:0016020 (membrane). Both of these tags are fairly abundant.

When looking at the data with the tags occurring over 1000 times excluded, GO:0005762 (mitochondrial large ribosomal subunit) and GO:0005743 (mitochondrial inner membrane) have the highest z-scores in multiple clusters (z between 11 and 17). Due to them being enriched in multiple clusters, the individual enrichment in each cluster is rather low, but the z-score still indicates its significance.

Protein Atlas Tissue Expression tags enrichment and z-scores

Both the skin 1 vascular mural cells (z=15.1, e:18-fold) and skin 2 vascular mural cells (z=12.8, e:16-fold) tags as well as skin 1 cells in spinous layer (z=14.1, e:14-fold) show the best enrichment and z-scores within the same cluster. All RNAs in the enriched cluster with the skin 1 vascular mural cells, skin 2 vascular mural cells or skin 1 cells in spinous layer tag bind the protein GRWD1. Other commonly bound proteins include: DDX24, ZNF622, BCLAF1 and UCHL5 for skin1 vascular mural cells, BCLAF1, UCHL5, ZNF622 and DDX24 for skin2 vascular mural cells and BCLAF1, RBM22, ZNF622, DDX24 and UCHL5 for skin 1 cells in spinous layer. If we only consider tags, occurring less than 1000 times, the tags with the highest z-score and enrichment remain the same. This is because all tissue tags with high z-scores and enrichment occurred fewer than 1000 times.

Adjacency List

An adjacency list for each RNA was generated, including all RNAs whose protein binding pattern is similar to the RNA (including itself). The Euclidean distance was used as a similarity measure and a cutoff of 3.5 and 4 was used when determining what counts as being adjacent. First only the longest adjacency lists which were very similar (they shared over 80% of the RNAs inside them) were analyzed. The result is not especially surprising since 2 RNAs with similar interaction patterns are also more likely to share similarity to a third that is similar to both. This is relevant because similar lists will obviously have the same tags enriched: thus, looking at adjacency lists that differ in their RNA content is more interesting since we can expect them to be enriched in different tags. To focus only on the lists that are most distinct from each other, we filtered for the longest lists with an overlap of <60% to the other longest lists and looked if any enrichment could be detected. This increases the likelihood of the different lists containing different RNAs with different tag-enrichment. To determine if there is an enrichment, we treated the adjacency list as one single cluster and every RNA outside of it as another cluster and filtered out all rare tags occurring fewer than 50 times. We next randomized the tags in both clusters 5000 times by randomly distributing all tags over both clusters while keeping the number of tags constant to get a distribution of how often certain tags would be expected by random chance. Finally, we calculated z-scores for each tag and picked out the 10 tags with the highest z-score for further analysis. The goal was to find if different tags are enriched in lists sharing fewer tags.

More Stringent Cutoff: Euclidean Distance of 3.5

The largest list at the 3.5 cutoff is 4825 RNAs long. The enrichment graphs can be seen in **Figure 5.1a-c**. The two Pfam tags with the highest z-score were 7tm_1 (z-score=12.8) and 7tm_4 (z=8.1). Out of the 124 RNAs in the cluster with the 7tm_1 tag, 52 bound the protein PTBP1 and 45 bound FIP1L1 and 10 out of 35 RNAs with the 7tm_4 tag bound PTBP1.

The GO tags with the highest z-scores were GO:0005654 (z=19.1, nucleoplasm), GO:0003723 (z=15.2, RNA binding) and GO:0005829 (z=15.1, cytosol). 59 out of 101 RNAs within the cluster bound the RBP ELAVL1. 221 out of 388 with the nucleoplasm tag bound ELAVL1 and 194 out of 388 bound FIP1L1. The cytosol tag has the same proteins bound most often (ELAVL1: 386/735, FIP1L1: 353/735).

For the tissue tags, the tags with the highest z-scores were testis elongated or late spermatids (z=18.5), where 75/149 of the RNAs with the tag were bound by ELAVL1 and retina photoreceptor cells (z=14.7), where 16/32 of the RNAs with the tag were bound by ELAVL1.

For the 3.5 cutoff, the largest list that shared less than 60% with the longest list was around 1400 RNAs long. There were no other lists longer than 1000 RNAs and had <60% identity with the longest and the second longest lists. The clearest tag enrichment for that list can be seen in the Pfam (**Figure 5.2 a**) and GO tags (**Figure 5.2 b**), while for the tissue tags (**Figure 5.2 c**) the enrichment is harder to detect. The tags with the strongest z-scores in that adjacency list over all are KRAB (z-score= 7.26, 80/80 RNAs with the tag bind NUDT21 and ELAVL1 and 79/80 bind PTBP1), zf-C2H2 (z= 6.44, NUDT21 bound by 118/120, FIP1L1 and ELAVL1 117/120), Homeodomain (z=6.22, 46/47 bind RBM15 and 45/47 bind RBM15B), Cadherin (z=3.44, FIP1L1, PTBP1 and NUDT21 bound by 21/22, ELAVL1: 20/22) and Cadherin_2 (z=3.80, NUDT21 and ELAVL1: 15/15, FIP1L1 and PTBP1: 14/15) (**Figure6.2a**). The only enriched Pfam tag that this list shares with the longest list is the Homeodomain tag.

The GO tags have even better z-scores with GO:0000981 (DNA-binding transcription factor activity, RNA polymerase II-specific), GO:0000978 (RNA polymerase II cis-regulatory region sequence-specific DNA binding), and GO:0006357 (regulation of transcription by RNA polymerase II**)** having z-scores above 10 and all being related to RNA polymerase II.

189/198 RNAs in the cluster with the GO:0000981 tag bound ELAVL1 and 188/198 bound NUDT21. RNAs with the GO:0000978 tag also bound the ELAVL1 (172/182) and NUDT21 (174/182) RBPs often and the same is true for the GO:0006357 tag (ELAVL1: 215/226 and NUDT21: 214/226)

The tissue tags with the highest z-scores were fallopian tube ciliated cells (tip of cilia) (z= 4.87, NUDT21: 52/86, ELAVL1: 52/86) and fallopian tube ciliated cells (cilia axoneme) (z=4.79, PTBP1: 33/52, ELAVL1: 31/52).

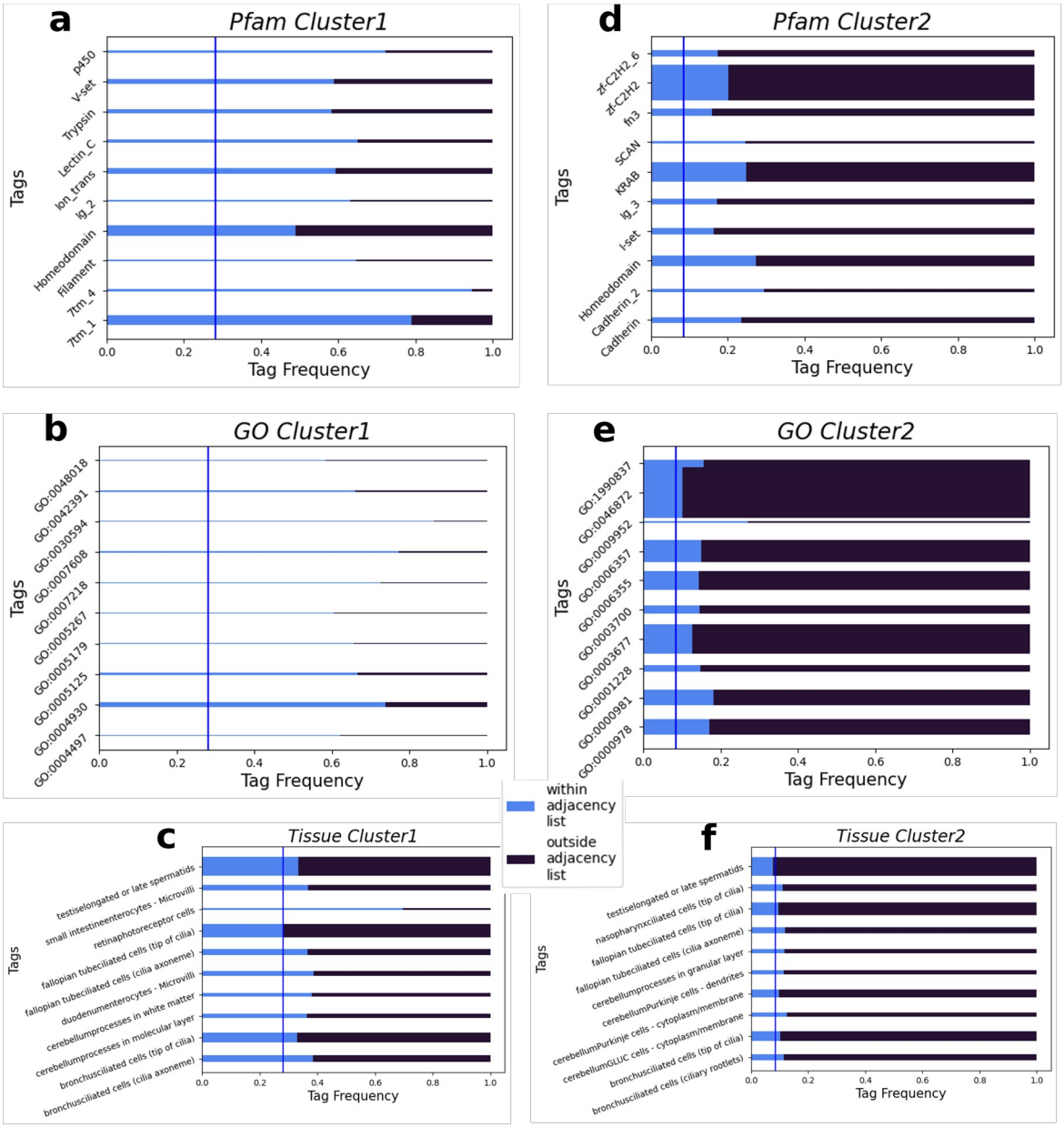


Figure 5: *Cutoff 3.5*

Enrichment of tags with the highest z-scores in each adjacency list at a cutoff of 3.5 in Euclidean distance is depicted. The light-blue bars indicate the frequency of the tag within the adjacency list, the dark-blue bar indicates the frequency of the tag outside of the adjacency list. The blue vertical line indicates the expected frequency of the tag if the tag was neither enriched nor depleted in the adjacency list. The thickness of the bars is proportional to the abundance of the tag i.e. thicker bars represent more frequent tags.

a) *Pfam tags with the highest z-score in the longest adjacency list;* ***b)*** *GO tags;* ***c)*** *tissue tags;* ***d)*** *Pfam tags with the highest z-score in the second longest adjacency list that has <60% similarity to the longest list;* ***e)*** *GO tags;* ***f)*** *tissue tags*

Less Stringent Cutoff: Euclidean Distance of 4

The largest list at the cutoff of 4 included approximately one third of all RNAs and the longest list being 60% different from the overall longest was around 1400 RNAs long. Again, there were no other lists longer than 1000 RNAs that had less than 60% identity with the longest and second longest list. This time, however, all 3 types of tags had poor z-scores with most of them being negative. This means that while the graphs show some enrichment, it is not only statistically insignificant, but less from what would be expected by random chance.

The data was not normally distributed making z-scores a more difficult to interpret. To account for that, an additional check for enriched tags using the Chebyshev inequality was performed. The Chebyshev inequality states that only 25% of values can lie outside 2 standard deviation from the mean. This means that a value greater than the mean plus 2 standard deviations is guaranteed to be at least top 25% (possibly higher up)[40][41]. The results confirmed that not a single tag is 2 standard deviations above the mean despite there being values being 2 standard deviations below the mean. This was a reflection of the fact that the distribution is not a normal distribution, but left-skewed.

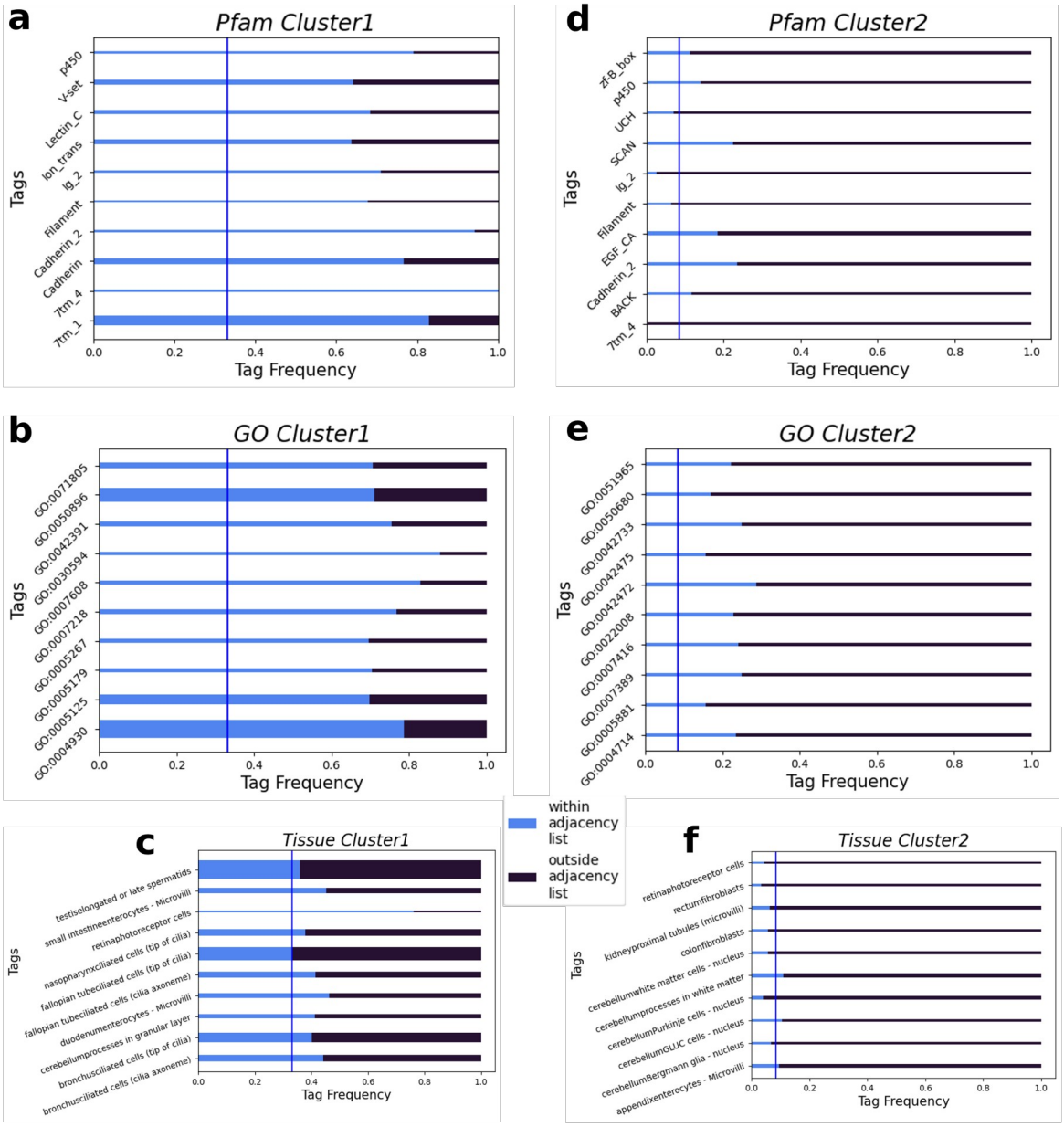


Figure 6: *Cutoff 4*

Enrichment of tags with the highest z-scores in each adjacency list at a cutoff of a Euclidean distance of 4 is depicted. The light blue bars indicate the frequency of the tag within the adjacency list, the dark blue bar indicates the frequency of the tag outside of the adjacency list. The blue vertical line indicates the expected frequency of the tag if the tag was neither enriched nor depleted in the adjacency list. The thickness of the bars is proportional to the abundance of the tag so thicker bars represent more frequent tags.

a) *Pfam tags with the highest z-score in the longest adjacency list;* ***b)*** *GO tags;* ***c)*** *tissue tags;* ***d)*** *Pfam tags with the highest z-score in the second longest adjacency list that has less than 60% similarity to the longest list;* ***e)*** *GO tags;* ***f)*** *tissue tags*

Discussion

As the main result of the present project, we have to demonstrated that mRNAs that share a more similar RBP-binding pattern are also more likely to share Pfam (**Figure 3a**), GO and tissue expression tags (**Figure 4**), if one corrects for promiscuous binders by only looking at their top50 interaction partners and, in the case of GO and tissue expression, excludes tags that occur over 1000 times. The latter requirement was relevant, since those tags are so abundant that even non-similar RNAs are likely to share them (**Figure 3b-c**). This means that interaction patterns are suited to create groupings in agreement with Pfam, GO and tissue tags. Pfam tags fit especially well with the RBP-binding pattern grouping. This could come from them being very stringently assigned and clearly defined [2]. In contrast some GO categories can also include tags that refer to very broad concepts [3][4][5] and tissue tags can include mRNAs highly expressed in many tissues [7]. It may be important in a future project to optimize the cutoff-values for GO and tissue tags further to find if the data matches even better with more specific tags. Another interesting aspect could be to study the similarity of Pfam, GO and tissue expression tag groupings and compare them to each other. Should they prove to deliver similar results to each other as well, this would be clear evidence that these methods all detect aspects of the objective similarity between proteins. Should they deliver dissimilar results despite all of them sharing similarities when it comes to the RBP- binding pattern, this may provide hints about the extent of how general and universal the role of RBP binding is, as opposed to it being more specifically related to particular functional categories.

Identification of RNAs within a cluster of a certain binding pattern that were enriched in specific tags was also successful for some of the clusters. Several highly enriched tags with extremely high z-scores were successfully identified, with the Pfam tags arguably yielding the best results. Unfortunately, a mechanistic explanation as to why the RBP-binding pattern led to an increase of those tags could not be found and only a few RBPs met the prediction of the complementarity hypothesis and bound the mRNA they are encoded by. The GO tags with high z-scores appear somewhat dubious since the tags with the highest z-scores have high z-scores within multiple clusters and are only depleted in one or two of the 50 clusters, hinting that their enrichment in a given cluster is primarily a result of their complete absence in another. The tissue expression tags with the best enrichment and z-score are all enriched within the same cluster as the Pfam RRM_1 domain. Therefore, analyzing if there could be a link between those enrichment might be an interesting aspect to study in the future.

Furthermore, the RBPs do not appear to share sequence-based Pfam tags with the RNAs they bind. This in turn could have served as evidence for the complementarity hypothesis since it would have

meant that RBPs with a certain domain bind RNAs coding for that same domain. However, since Pfam tags only refer to a sequence with known function and not the full sequence of a protein, it would perhaps be useful to compare the binding patterns to other sequence-based similarity metrics like sequence alignments. The complementarity hypothesis predicts that mRNAs binding an RBP, that binds its own mRNA, have a higher than average likelihood of being similar in sequence to the mRNA encoding the said RBP [42]. Such an analysis is already currently underway in the Zagrovic group.

Categorization

The most apparent difference between GO (**Figure 3b**) and tissue (**Figure 3c**) tags versus Pfam tags (**Figure 3a**) is that the absolute intersection scores are much higher for GO and tissue tags. The relative difference between the top and the bottom bin however is highest in the Pfam (around 3-fold) and lowest in the GO tags (5.5%). In other words, the tag type with the lowest total intersection score in the top bin also has the highest difference between the top and the bottom bin.

GO and tissue

The issue with the GO and tissue tags is that – regardless of the method (all, Top10, Top20 and Top50) – almost 70% of the least similar RNAs share 1 or more GO tags and almost 40% share 1 or more tissue tags. This means that both of them contain a sizable number of tags that are so abundant that even the least similar of RNAs have a high likelihood of sharing them. This presents a problem for analyzing the data usefully since it is extremely likely that some tags shared by RNAs with similar binding patterns are simply shared by many RNAs regardless of how similar or dissimilar they are. The difference in the top bin and the bottom bin intersection score is small especially for the GO tags. This means that the GO (**Figure 3b**), but also tissue (**Figure 3***c*) tag curves on their own are somewhat problematic when it comes to demonstrating the enrichment of tags within the top bins. The Pfam tag curves, on the other hand, demonstrate very conclusively that mRNAs in the top bin of the Top50 approach have a 3x higher intersection score than those in the bottom bin, so there is a much higher probability of the top bins sharing tags as compared to the bottom bins.

This issue of tags shared by all RNAs regardless of similarity in binding pattern was resolved by excluding tags occurring in over 1000 RNAs since those tags are not specific to begin with and are anyways poorly suited to establish similarity between genes. Since only 31 GO tags and 45 tissue tags occurred over 1000 times, this change affected few tags. The GO tags excluded were all very broad “parent” tags and their more specific “children” tags are still be included in the analysis [3][4][5]. This means that all that their exclusion does is to allow us to detect the specific tags more

clearly without losing the meaning of the excluded tags completely since their more specific children are still retained in the analysis.

While the concept of parent and children tags does not apply to the tissue tags, the exclusion of the highly abundant tags can still be justified. Namely, if a gene is highly expressed in many tissues, it is not helpful to include it when attempting to identify whether an RBP-binding pattern and a tissue-specific expression group the same genes together because it can, due to its common expression, not be specific to a small enough subset.

The exclusion of these few tags makes the absolute similarity scores drop greatly, meaning that these few tags really are responsible for most of the tag intersections detected in **Figures 4b, c**. Here, both tissue and GO tags show a higher relative difference between their top bins and their bottom bins, suggesting that a more similar RBP-binding profile increases the likelihood of two mRNAs to also be similar in the GO and tissue tags they possess – if highly abundant, unspecific tags are excluded. Even if the relative difference between the top and the bottom bins is still not as large as observed in the Pfam tags, the relative difference between the top and the bottom bins in more specific tags as compared to the unmodified dataset is still greater than it was with the highly abundant, unspecific tags included (**Figure 4**).

Pfam

The Pfam data in **Figure 3a** shows that mRNAs with similar protein binding patterns are more likely to share protein family tags than mRNAs that do not, when one corrects for more promiscuous mRNAs by excluding the less strongly bound protein interaction partners. When we include all interactions of promiscuous mRNAs, however, it appears that some mRNAs that seem to be dissimilar in their RBP-binding patterns are more similar in the associated tags than those with more similar interaction patterns. Due to this effect vanishing when only including the strongest interaction partners – especially when looking at the Top50 interacting proteins for each mRNA (**Figure 3a**) – one could argue that this demonstrates that it is caused by mRNAs that bind many different proteins. The mRNAs binding many different proteins might have higher promiscuity or they could only appear to possess higher promiscuity due to a higher concentration, increasing the likelihood of low affinity interactions occurring at a detectable level [22]. Since the effect of both is the same – an mRNA having more measured interactions thus appearing promiscuous – it does not really matter which of the two effects is at play in this context. Since only including the Top50 interactions more accurately predicts Pfam tag similarity between mRNAs as compared to including all interactions, the effect appears to be caused by mRNAs sharing the same core RBPs they bind strongly and some additional weak interactions with other RBPs. This difference should not be

understood as the mRNAs being fundamentally different in their binding patterns because in the end they do share their strongest interactions, which are arguably the most important ones. Since the uncorrected approach does not distinguish between weak and strong interactions while the Top50 approach exclusively counts the 50 strongest interactions as real interactions, one could argue that it demonstrates that the Top50 approach is superior for this type of analysis.

The Pfam tag results, however are also problematic since many individual tags are rather rare (some only occurring once in the data) and are thus not, or at least not often, shared by the proteins coded by the investigated mRNAs since their assignment is very stringent. One might be tempted to assume that if an mRNA has an almost identical binding patterns to mRNAs with a certain Pfam tag, it might also have this Pfam tag, but it was not annotated wrongfully. However, since Pfam tags are defined based on the sequence, it is impossible that a tag's presence was overlooked without the very definition of the sequence characterizing the tag being wrong or the protein coded by the mRNA not having been scanned properly for it [2]. This means that the claim that an mRNA should be assigned a Pfam tag because it has a similar RBP binding pattern to several mRNAs with the tag, requires a great amount of additional evidence before being entertained. It is, however, still worth investigating why mRNAs that do share a certain tag are more likely to have a specific binding pattern and why others with similar binding patterns are not.

As a summary, it was successfully shown that mRNAs with similar Pfam, GO and tissue expression tags are more likely to share similar RBP binding patterns which is the first key point we attempted to investigate. This, however, is only possible for the GO and tissue expression data when highly abundant tags are excluded and for all these tags it works best when only the 50 strongest interacting RBPs of each mRNA are counted as interacting.

Clustering

Clustering was used to group mRNAs of similar RBP binding patterns together and to investigate whether there are any tags containing Pfam, GO or tissue expression tags with a probability above random chance. This should hopefully reveal which tags are responsible for the higher intersection scores of mRNAs in the top similarity bin. Another aim was to find an explanation as to why these tags are linked to a certain RBP binding pattern in the literature.

Pfam

Homeodomain proteins are known to act as transcription factors [43] and the Hox transcription factor Ultrabithorax (Ubx) has also been demonstrated to bind RNA as well [44]. It was analyzed if the RBPs interacting with the RNAs with the Homeodomain tag also themselves were

Homeodomain proteins, but they were not. It was then checked if any of the proteins bound within the Top50 possessed the homedomain tag, but they did not. The reason this was checked was the so called complementarity hypothesis that predicts that proteins should be prone bind the mRNA they are encoded by (and RNAs similar in sequence to it) [42]. My colleague Thomas Kapral and coworkers, for example, recently found that 230 out of 341 RNA binding proteins (RBPs) bind their own mRNAs. Moreover, the RBPs also showed a preference to bind the coding sequence of their autogenous mRNA [22]. The mRNAs coding for Homeodomain proteins in this analysis are bound by non-homeodomain proteins, but there is no evidence that they are bound by Homeodomain proteins not meeting the prediction of the complementarity hypothesis. mRNAs with the Homeodomain tag still appear to have a more similar RBP interaction patterns than expected by random chance, but according to their Pfam domains, those RBPs they interact with lack a Homeobox domain [2]. This could be interpreted as Hox coding mRNAs being bound by the same or similar non-Hox proteins, but the data shows that the most commonly bound RBP was bound by less than one third of all Homeodomain tagged RNAs (ELAVL1: 16/50). In other words, while their binding patterns are more similar inside the group than outside the group, there is still a great binding diversity within this grouping.

The RRM_1 enrichment and Helicase C enrichment were interesting since it means that certain RBPs are binding to mRNAs that code for a nucleic-acid binding protein. This could serve as evidence for the complementarity hypothesis if the RBPs that bind the mRNAs also happened to be RBPs encoded by said mRNA. Therefore, it was checked if any of the RBPs encoded by the mRNAs in the cluster possessing the Helicase C-terminal domain tag also happened to be bound to their autogenous mRNA and if that autogenous mRNA was within the cluster. There were only 3 RBPs (DDX24, DHX30, DDX21) with the tag that also had their autogenous RNA in the cluster and 2 of them (DDX24, DHX30) interact with their autogenous mRNA [2]. This sample size is too small to make any definitive statements concerning the complementarity hypothesis. For RRM_1 the same issue hold, but manifested even more extremely as only the RBPs EWSR1's RNA was found to be present within the cluster and the pair was not considered to be interacting [2]. This means that there is not evidence that RNAs coding for a RRM_1 or Helicase C domain are likely to bind an RBP with that domain.

All 24 RRM_1 tagged RNAs in the RRM_1 enriched cluster bound GRWD1 and 22 of them bound UCHL5, ZNF622 and BCLAF1. This means that possessing an RRM_1 tag seems to specifically favor RNAs that bind those proteins. For Helicase_C, a similar effect can be observed where 20/24 of the RNAs with the tag bind UCHL5 and 19 bind FMR1, GRWD1, ZNF622 and FXR2. This means that RNAs with one of these tags in the enriched cluster are highly likely to bind specific

proteins. We have to keep in mind, however, that this is a very small number of the total interaction partners each RBP has that are either outside the cluster or do not have a partner with the respective Pfam tag. So, while it is interesting that most of them bind these specific RBPs, binding one or two of these RBPs is in no way predictive of having the domain. It is, thus, important to consider the similarity of the entire RBP binding profile and not just the most commonly bound RBPs, which is what the Agglomerative Hierarchical Clustering allowed us to do. [31].

Immunoglobulin V-set domain is a tag occurring in many different protein families including the light and heavy chain of immunoglobulin, T-cell receptors, JAMs, tyrosin-protein kinases and myelin membrane adhesion proteins [2]. The proteins interacting with the mRNAs associated with the tag are not known as having the tag [2], so it is most likely not a case of the proteins binding an mRNA because it codes for the V-set domain of the mRNA the binding protein is encoded by.

The Trypsin, Lectin C-type domain and Ion transport protein tags do not have an obvious connection between their function and the protein binding patterns of the mRNAs that contain the tag since the tags are not directly related to protein binding or mRNA protein interaction. The most common RBP that bound the mRNAs within cluster where one of the tags was enriched for all 3 is ELAVL1 (with 9/27 for RNAs with the Trypsin tag, 9/27 for Lectin and 17/33 for Ion transport). There is no RBP that all of them share and ELAVL1 is one of the proteins interacting with the most mRNAs in the data. Therefore, it appears the protein binding pattern within these clusters enriched in one of the 3 tags is not very similar to each other but just more similar to each other than to the mRNAs outside the cluster. While less impressive at first glance than the RRM and Helicase C findings that show a clear preference for binding a certain RBP, the data shows that even the more subtle similarities, which lead to the clustering of mRNAs, also caused the enrichment of Trypsin, Lectin or Ion transport tags within their respective clusters. It appears that clustering is able to pick up similarities in binding patterns that are not obvious at a first glance, but still make mRNAs more similar to each other in a way that is manifested in their function or the function of the protein they encode.

Since the G protein-coupled receptor (7tm_1) is the tag with the highest enrichment and z-score, it is potentially the most interesting. However, this tag has similar issues to Lectin, Trypsin and Ion transport tags with no RBP being bound by the majority of mRNAs with the tag. Instead, the similarity of their binding pattern again appears to be more subtle. Perhaps the RNAs are only grouped as similar because they are even more different from mRNAs in other clusters. However, the astonishingly high z-score of the 7tm_1 tag makes it highly unlikely that them falling into the same cluster is a mere coincidence. We, however, have no mechanistic explanation for why grouping mRNAs by their interaction pattern with RBPs would lead to mRNAs associated with the

7tm_1 tag being grouped together. All that can be demonstrated is that in this specific case mRNA-RBP interactions group mRNAs that code for a G protein-coupled receptor together – it does, however, also group them with a large number other mRNAs. Moreover, since Pfam identifiers are defined by sequence [2], it is impossible for it to simply be an unknown function of the other mRNAs.

In conclusion, the attempt was made of finding a connection between the Pfam tags and the RBP-binding pattern in the literature. For some Pfam tags, this hypothesis could not be tested well since the mRNAs sharing a tag had very distinct binding patterns.

Gene Ontology

Due to the low difference of the intersection score in the lowest and the highest bin, GO tags are not expected to yield many reliable results. Additionally, the tags with the highest z-scores and the tags with the highest total enrichment did not match. A reason for that could be that there are tags occurring thousands of times and some occurring only around 50-times. The larger tags can already produce a high z-score when enriched slightly, while the rare tags with lower z-scores can already produce higher enrichment scores. Since they take more factors into account (mean and standard deviation), z-scores are generally more statistically robust than simple enrichments. For this reason, a high enrichment without a high z-score is uninformative. A high z-score without a high enrichment might be less informative than one with a high enrichment, but it is not completely uninformative as it still means the tag is far more frequent in the cluster than expected by chance. The protein binding tag that is enriched in multiple clusters is one of the most abundant tags, meaning it is a very common and generic function that has multiple tags hierarchically underneath it. Due to those tags being more specific, one of them having a high z-score would have been more interesting.

When tags occurring over 1000 times are excluded, the tags with the best z-scores again have high z-scores in multiple clusters and both are cellular component tags related to the mitochondrion (mitochondrial large ribosomal subunit, mitochondrial inner membrane). Due to the mitochondrion being an endosymbiont [45], it makes sense that mRNAs related to it would have different RBP-binding patterns than mRNAs of a purely eukaryotic origin. What is, however, surprising is that they would be enriched in multiple clusters instead of only one and that not more mitochondria-related GO tags present high enrichment. This indicates that these specific mitochondria-related tags might be severely underrepresented in some clusters and over-represented in others. On closer inspection, it seems that both tags either have a high z-score or are present at an expected level in most clusters but completely absent from one cluster (mitochondrial inner membrane) or two

clusters (mitochondrial large ribosomal subunit). This indicates that their high z-scores are merely a consequence of the complete depletion of the tag within a single cluster, which is quite surprising since many of their z-scores were rather high. This means we were unable to find a functional connection between the GO tags and the RBP-binding patterns.

Protein Atlas Tissues

All of the enriched tags are somehow related to the skin and are all enriched in the same cluster. All mRNAs with one of the 3 tags bind GRWD1 with most of them also binding UCHL5 and ZNF622.

GRWD1, UCHL5 and ZNF622 are also proteins associated with mRNAs with the RRM_1 tag enrichment. Unsurprisingly the skin tissue tags are enriched within the same cluster as the RRM_1 tag. This means that the cluster is not only enriched in mRNAs with RRM_1 coding sequences, but the mRNAs in it are also more likely to be highly expressed in certain skin cell types.

Excluding tags occurring over 1000 times does not change which tags have the highest enrichment and z-scores.

We were unable to find a connection between the tissue tags and the RBP-binding pattern in the literature. One could speculate that there is a connection between the RRM_1 Pfam domain and the enriched tissue tags since they are enriched in the same tag.

Adjacency Lists

The longest adjacency list in the cutoff 4 data included over 30% of all RNAs and the second largest was significantly shorter and included no enriched tags with a high z-score. This is most likely because the cutoff is too generous and adds mRNAs that are not very similar to the adjacency lists. Since the data from the larger Euclidean distance cutoff of 4 was inconclusive, the focus of the discussion will be on the 3.5 cutoff.

It appears that, in the 3.5 cutoff cluster, tags associated with mRNAs that bind NUDT21 and ELAVL1 have the highest z-scores. The issue is that binding those RBPs is enriched in both adjacency lists that are meant to be distinct from each other as the shorter list shares less than 60% of the mRNAs with the longer list. This means that either the lists are not distinct enough and the overlap is still too large or that even RNAs with very distinct binding patterns have a high likelihood to bind NUDT21 and ELAVL1. Since both NUDT21 and ELAVL1 have been found to be RBPs bound by thousands of the mRNAs investigated the latter is confirmed, but the former could still play a role.

Conclusion and outlook

The project was successful in demonstrating that grouping RNAs by their RBP-interaction profiles on average puts together genes that share Pfam tags. The same is true for tissue expression and GO tags if highly abundant, unspecific tags are removed. This means that RBP-interaction profiles have a demonstrable use for grouping genes in a biologically meaningful fashion.

We also successfully identified clusters that were enriched in a certain tag to such a degree that it is unlikely that the enrichment occurred by random chance. A mechanistic explanation for the enriched tags, however, could not be found.

This project was a first exploration of the possibility of grouping genes based on their mRNA's RBP-binding patterns and, thus, further research is needed that takes into account problems and potential pitfalls identified in this project.

The Problem of Promiscuous RBPs

There is a fundamental issue with the mRNA-RBP-interaction data that was ignored during the analysis, namely, that the data contains RBPs that bind many mRNAs and are thus less specific than others that only bind a few. As the GO and tissue expression data has demonstrated, excluding the broad, unspecific tags improved the score difference between the top and bottom bins (**Figure 4**) in the RBP-interaction data. So, if it is true that unspecific tags mask the dissimilarity of very distinct genes, perhaps unspecific RBPs cause the same effect in the interaction data. This means that RBPs that are bound by a certain percentage of mRNAs should be excluded from the analysis before the Top50 score is calculated in an attempt to make the results more robust. Perhaps it would then even make sense to take a look at the intersection score of including all bound RBPs, the top10 and the top20 again instead of only the top50 and compare their similarity to the Pfam, GO and tissue tags. As specific RBPs are less abundant, excluding non-specific RBPs would lead to similarity in the specific tags being given more weight in determining mRNA similarity.

The extent of this issue was underestimated during the project and only came to light while writing this thesis: there are RBPs in the top50 data that bind over 10,000 RNAs which is over 50% of the 17,000 RNAs studied. Overlooking this fact might have greatly influenced the data generated in this project as compared to what it could have been if commonly bound RBPs were excluded.

When excluding unspecific RBPs, it is important to pick a reasonable cutoff. An unreasonably stringent threshold will exclude a too large number of RBPs while a too lenient threshold will allow unspecific tags to stay in. This means that different thresholds should be tested and discussed to

establish what is reasonable or a threshold could be defined as reasonable based on the percentage of RBPs excluded by it.

TADs

Another future direction could be to analyze in which Topologically Associated Domains (TADs) the genes coding for the mRNAs are located in. TADs are chromatin domains that can be several hundred kb long and are spatially separated from other TADs [46]. They can be found in eukaryotes, while bacteria have domains similar to TADs referred to as Chromatin Interaction Domains (CIDs). That latter name was established because the chromatin within a CID prefers to interact with chromatin located in the same CID [47]. In TADs, it has been discovered that chromatin is more likely to interact within the same TAD than it is to interact with chromatin from a different TAD as well [46]. Should mRNAs coded within the same TAD really possess similar RBP-binding patterns, it would further highlight the scientific usefulness of grouping genes by their mRNA's RBP-binding pattern. There exists evidence that TADs might be involved in gene regulation [47]. In particular, it has been proposed that TADs play a role in regulating transcription [46] so mRNAs from the same TAD being bound by the same RBPs could hint at them not only having their transcription regulated together, but also having the same RBP involved in their regulation providing a potential explanation of how they are regulated together.

Is clustering even a suitable approach?

Song & Singh (2009) [48] argue that simply looking for enrichment within clusters is not the best way to predict protein function in yeast. This finding might not be directly applicable to the present project focusing on attempting to predict gene function based on human RNA-RBP interactions. Moreover, the presently based score was not just based on enrichment but also z-scores when whereby the observed tag distribution was compared against randomized backgrounds. The clustering methods tested by Song & Singh (2009) were also not hierarchical clustering approaches but instead clustering algorithms specialized for protein-protein interaction (PPI) networks: *NetworkBlast* [49], *Cfinder* [50], *Markov Clustering (MCL)* [51], *DPClus* [52], *Mcode* [53] and *SpectralMod* [54] (please, see SuppMat 4 in Supplementary Materials for further details about these methods). Since they are all designed for protein-protein interactions, they assume square interaction matrices, which the mRNA-RBP interaction matrix is not since there are around 17,000 mRNAs but fewer than 300 RBPs in the presently available data. This means that these methods may require extensive modifications to be used on mRNA-RBP data or may not be usable at all.

As all of the above methods use very different metrics for creating clusters than the agglomerative hierarchical clustering uses, we should not assume that the issue Song & Singh (2009) [48] found with those methods will also necessarily occur with agglomerative hierarchical clustering with Ward's linkage. It would, however, be wise to not blindly trust the enriched functions within clusters and instead test different clustering methods like Song & Singh (2009) [48] suggest. In addition, one should also test other clustering methods that have been developed in this rapidly growing field in the same way (see SuppMat 4 for more details). The tests of Song and Singh (2009) [48] essentially compared the clustering they obtained from each method to a clustering corresponding to an ideal network where all proteins were perfectly separated by their GO terms and clusterings that were obtained from randomized networks. These networks are then scored to find out if and how much better the clustering of the network is compared to the randomized network clustering results.

The z-score used to evaluate the hierarchical clustering in this thesis, while being based on the same idea of comparing the obtained clustering against randomized backgrounds, works somewhat differently because not the interactions themselves are randomized, but the tags. The difference is that in the Song & Singh's (2009) [48] approach, the tags that are annotated to the same protein stay together, while in the present approach, the tags have no such restriction and even tags always occurring together do not do so in our randomizations. Our approach could be seen as especially problematic for the hierarchical GO tags because the nonsensical case that a child tag occurs without the parent tag being present could theoretically occur. This means that while it was important to compare our clustering to random clusterings instead of solely relying on enrichments, in the future our scoring method should be carefully weighed for its benefits and drawbacks compared to Song & Singh's (2009) [48] scoring methods.

Overall, the present thesis represents an early exploration of the possibility to group genes i.e. mRNAs based on their interaction patterns with RBPs. While our work has only opened up several exciting research avenues that should be explored further, the preliminary results suggest that indeed the biological function of genes and gene products may be related to as well as, in part, be read from the RBP-interaction patterns of their associated mRNAs.

Bibliography

[1]: Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. *Large-scale investigation of the reasons why potentially important genes are ignored*. PLoS Biol.. 16(9):e2006643. 2018
[2]: Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. *Pfam: The protein families database in 2021*. Nucleic Acids Research. 49(D1):D412–D419. 2020
[3]: Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. *Gene ontology: tool for the unification of biology*. Nat Genet.. 25(1):25-29. 2000
[4]: The Gene Ontology Consortium. *The Gene Ontology resource: enriching a GOld mine*. Nucleic Acids Res. 49(D1):D325-D334. 2021
[5]: Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. *PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools*. Nucleic Acids Res.. 47(D1):D419-D426. 2019
[6]: Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N. *PROSITE, a protein domain database for functional characterization and annotation*. Nucleic Acids Res. 38(Database issue):D161-6. 2010
[7]: Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Pontén F.. *Tissue-based map of the human proteome*. Science. 347:6220. 2015
[8]: Tarca AL, Romero R, Draghici S. *Analysis of microarray experiments of gene expression profiling.*. BMC Bioinformatics. 195(2):373-88. 2006
[9]: Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, Bian W, Brignall R, Cafarelli T, Campos-Laborie FJ, Charloteaux B, Choi D, Coté AG, Daley M, Deimling S, Desbuleux A, Dricot A, Gebbia M, Hardy MF, Kishore N, Knapp JJ, Kovács IA, Lemmens I, Mee MW, Mellor JC, Pollis C, Pons C, Richardson AD, Schlabach S, Teeking B, Yadav A, Babor M, Balcha D, Basha O, Bowman-Colin C, Chin SF, Choi SG, Colabella C, Coppin G, D'Amata C, De Ridder D, De Rouck S, Duran-Frigola M, Ennajdaoui H, Goebels F, Goehring L, Gopal A, Haddad G, Hatchi E, Helmy M, Jacob Y, Kassa Y, Landini S, Li R, van Lieshout N, MacWilliams A, Markey D, Paulson JN, Rangarajan S, Rasla J, Rayhan A, Rolland T, San-Miguel A, Shen Y, Sheykhkarimli D, Sheynkman GM, Simonovsky E, Taşan M, Tejeda A, Tropepe V, Twizere JC, Wang Y, Weatheritt RJ, Weile J, Xia Y, Yang X, Yeger-Lotem E, Zhong Q, Aloy P, Bader GD, De Las Rivas J, Gaudet S, Hao T, Rak J, Tavernier J, Hill DE, Vidal M, Roth FP, Calderwood MA. *A reference map of the human binary protein interactome*. Nature. 580(7803):402-408. 2020
[10]: Zhao W, Zhang S, Zhu Y, Xi X, Bao P, Ma Z, Kapral TH, Chen S, Zagrovic B, Yang YT, Lu ZJ.. *POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins*. Nucleic Acids Res. 50(D1):D287-D294. 2022
[11]: The UniProt Consortium. *UniProt: the universal protein knowledgebase in 2021*. Nucleic Acids Research. 49(D1):D480–D489. 2021
[12]: Barshir R, Fishilevich S, Iny-Stein T, Zelig O, Mazor Y, Guan-Golan Y, Safran M, Lancet D. *GeneCaRNA: A Comprehensive Gene-centric Database of Human Non-coding RNAs in the GeneCards Suite*. J Mol Biol. 433(11):166913. 2021
[13]: del Pozo A, Pazos F & Valencia A. *Defining functional distances over Gene Ontology*. BMC Bioinformatics. 9:50. 2008
14: Chandra G & Tripathi S, *Approach for Clustering of Gene Expression Datawith Detection of Functionally Inactive Genes and Noise,*Advances in Intelligent Computing ,-,2019

[15]: Preciado AG, Peimbert M, Merino E. *Encyclopedia of Microbiology (Third Edition)*. Academic Press. Types of Data and Bioinformatic Tools:220. 2009
[16]: Van Nostrand EL, Pratt GA, Yee BA, Wheeler EC, Blue SM, Mueller J, Park SS, Garcia KE, Gelboin-Burkhart C, Nguyen TB, Rabano I, Stanton R, Sundararaman B, Wang R, Fu XD, Graveley BR & Yeo GW. *Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins*. Genome Biol. 21:90. 2020
[17]: Corley M, Burns MC, Yeo GW. *How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms*. Molecular Cell. 78(1):9-29. 2020
[18]: Beckmann BM, Castello A, Medenbach J. *The expanding universe of ribonucleoproteins: of novel RNA-binding proteins and unconventional interactions*. European Journal of Physiology. 468(6):1029–1040. 2016
19: Cléry A & Allain FHT, *From structure to function of RNA binding domains*,Landes Bioscience ,-,2012
[20]: Hoffman. *AANT: the Amino Acid-Nucleotide Interaction Database*. Nucleic Acids Res.. 32:D174-D181. 2004
[21]: Bercy M & Bockelmann U. *Hairpins under tension: RNA versus DNA*. Nucleic Acids Res.. 43(20):9928-9936. 2015
[22]: Kapral TH, Farnhammer F, Zhao W, Lu ZJ, Zagrovic B. *Widespread autogenous mRNA–protein interactions detected by CLIP-seq*. Nucleic Acids Research. 50(17):9984-9999. 2022
[23]: Hafner M, Katsantoni M, Köster T, Marks T, Mukherjee J, Staiger D, Ule J, Zavolan M. *CLIP and complementary methods.*. Nature Reviews Methods Primers. 1:20. 2021
[24]: Lee FCY & Ule J. *Advances in CLIP Technologies for Studies of Protein-RNA Interactions*. Molecular Cell. 69(3):354-369. 2018
[25]: Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, Darnell JC, Darnell RB. *HITS-CLIP yields genome-wide insights into brain alternative RNA processing*. Nature. 456(7221):464-469. 2008
[26]: Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T.. *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP*. Cell. 141(1):129-141. 2010
[27]: König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. *iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution*. Nat Struct Mol Biol. 17(7):909-915.
[28]: Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, Stanton R, Rigo F, Guttman M, Yeo GW. *Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)*. Nat Methods. 13(6):508-514. 2016
[29]: D'haeseleer P. *How does gene expression clustering work?*. Nature Biotechnology. 23(12):1499-1501. 2005
[30]: Nies HW, Zakaria Z, Mohamad MS, Chan HW,Zaki N, Sinnott RO, Napis S, Chamoso P, Omatu S & Corchado JM. *A Review of Computational Methods for ClusteringGenes with Similar Biological Functions*. Processes. 7:550. 2019
[31]: Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel P, Blondel M, Müller A, Nothman J, Louppe G, Prettenhofer P, Weiss R. Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. *Scikit-learn: Machine Learning in Python*. JMLR 12. 12(85):2825-2830. 2011
[32]: Virtanen P, Gommers R, Oliphant, TE et al.. *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*. Nature Methods. 17(3):261-272. 2020
[33]: McKinney W. *Data structures for statistical computing in python*. In Proceedings of the 9th Python in Science Conference. 445:51–56. 2010
[34]: Harris CR. *Array programming with NumPy*. Nature. 585:357–362. 2020

[35]: Hunter JD. *Matplotlib: A 2D Graphics Environment*. Computing in Science & Engineering. 9(3): 90-95. 2007
[36]: Van Rossum G. *The Python Library Reference, release 3.8.2.*. Python Software Foundation.. :. 2020
[37]: Gillis J & Pavlidis P. *Assessing identity, redundancy and confounds in Gene Ontology annotations over time.*. Bioinformatics. 29(4):476–482. 2013
[38]: Weichselbaum D, Zagrovic B, Polyansky A. *Fuento: functional enrichment for bioinformatics*. Bioinformatics. 33(16):2604–2606. 2017
39: Rhys HI, *Machine Learning with R, the tidyverse, and mlr* ,Manning ,-,2020
[40]: Amidan BG, Ferryman TA, Cooley SK. *Data outlier detection using the Chebyshev theorem*. Conference Paper in IEEE Aerospace Conference Proceedings. -:3814-3819. 2005
[41]: Balcerzak J & Pedzich PA. *A Proof of the Chebyshev theorem*. Geodesy and Cartography. -:-. 2000
[42]: Zagrovic B, Bartonek L, Polyansky AA. *RNA-protein interactions in an unstructured context*. FEBS Letters. 592(17):2901-2916. 2018
[43]: Svingen T & Tonissen KF. *Hox transcription factors and their elusive mammalian gene targets*. Heredity. 97:88-96. 2006
[44]: Carnesecchi J, Boumpas P, van Nierop Y, Sanchez P, Domsch K, Pinto HD, Borges Pinto P, Lohmann I. *The Hox transcription factor Ultrabithorax binds RNA and regulates co-transcriptional splicing through an interplay with RNA polymerase II*. Nucleic Acids Res.. 50(2):763-783. 2021
[45]: Margulis L & Chapman MJ. *Endosymbioses: cyclical andpermanent in evolution*. Trends in Mirobiology. 6(9):342-346. 1998
[46]: Dekker J. & Heard E.. *Structural and functional diversity of Topologically Associating Domains*. FEBS Letters. 589(20 Pt A):2877-2884. 2015
[47]: Beagan JA & Phillips-Cremins JE. *On the existence and functionality of topologically associating domains*. Nat Genet.. 52(1):8-16. 2020
[48]: Song J & Singh M. *How and when should interactome-derived clusters be used to predict functional modules and protein function?*. Bioinformatics. 25(23):3143–3150. 2009
[49]: Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T. *Conserved patterns of protein interaction in multiple species* . Proc Natl Acad Sci U S A. 102(6):1974-1979. 2005
[50]: Adamcsek B, Palla G, Farkas IJ, Derényi I, Vicsek T. *CFinder: locating cliques and overlapping modules in biological networks*. Bioinformatics. 22(8):1021-1023. 2006
[51]: Enright AJ, Van Dongen S, Ouzounis CA. *An efficient algorithm for large-scale detection of protein families*. Nucleic Acids Res. 30(7):1575-1584. 2002
[52]: Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S. *Development and implementation of an algorithm for detection of protein complexes in large interaction networks* . BMC Bioinformatics. 7:207.
[53]: Bader GD & Hogue CW. *An automated method for finding molecular complexes in large protein interaction networks*. BMC Bioinformatics. 4:2. 2003
[54]: Newman MEJ. *Modularity and community structure in networks* . Proc Natl Acad Sci U S A. 103(23):8577-8782. 2006
[55]: Rani RR, Ramyachitra D & Brindhadevi, A. *Detection of dynamic protein complexes through Markov Clustering based on Elephant Herd Optimization Approach*. Scientific Reports. 9:11106. 2019
[56]: Nepusz T, Yu H & Paccanaro A. *Detecting overlapping protein complexes in protein-protein interaction networks*. Nat Methods. 18;9(5):471-772. 2012
[57]: Li M, Chen J, Wang J, Hu B, Chen G. *Modifying the DPClus algorithm for identifying protein complexes based on new topological structures*. BMC Bioinformatics. 9:398. 2008
[58]: Meng X, Xiang J, Zheng R, Wu FX, Li M. *DPCMNE: Detecting Protein Complexes From Protein-Protein Interaction Networks Via Multi-Level Network Embedding* . IEEE/ACM Trans Comput Biol Bioinform.. 19(3):1592-1602. 2022

[59]: Shen X, Zhou J, Yi L, Hu X, He T, Yang J. *Identifying protein complexes based on brainstorming strategy*. Methods. 110:44-53. 2016
[60]: Xiao Q, Luo P, Li M, Wang J, Wu FX. *A Novel Core-Attachment-Based Method to Identify Dynamic Protein Complexes Based on Gene Expression Profiles and PPI Networks*. Proteomics. 19(5):e1800129. 2019

Supplementary Material:

The supplementary data (*SuppMat1-3*) includes a list of tags that are 10-fold enriched within a cluster (for data split into 50 clusters) or 5-fold enriched (for data split into 10 clusters). This data only include tags that have one of the 10 highest z-scores within their cluster.

SuppMat 1: Top 10, enriched Tags

SuppMat 1a: Pfam tags from RNAs split into 50 clusters that are at least 10x enriched

Cluster 0	Trypsin					
Cluster 2	7tm_1	Cadherin	Ion_trans	Lectin_C	Trypsin	V-set
Cluster 7	Lectin_C					
Cluster 31	DEAD					
Cluster 33	Cadherin	Cadherin_2				
Cluster 41	Cadherin	Cadherin_2				

SuppMat 1b: Pfam tags from RNAs split into 10 cluster that are at least 5x enriched

Cluster 2	7tm_1	Collagen	Ion_trans	LRR_8	Lectin_C	Trypsin	V-set
Cluster 4	PH	PHD	RhoGAP	RhoGEF			
Cluster 6	Cadherin_2	Homeodomain					

SuppMat 1c: GO, 50 clusters, 10x enriched

Cluster 0	GO:0001540	GO:0004867	GO:0005576	GO:0005615	GO:0005788	GO:0010951
	GO:0012507	GO:0031093	GO:0035580	GO:0043202	GO:0062023	GO:0070062
	GO:0072562					
Cluster 2	GO:0004867	GO:0005576	GO:0005615	GO:0010951	GO:0035580	GO:0062023
Cluster 9	GO:0035580					

SuppMat 1d: GO, 10 cluster, 5x enriched

Cluster 0	GO:0002181	GO:0003735	GO:0005747	GO:0005840	GO:0022625	GO:0022626
	GO:0032543	GO:0032981	GO:0042776	GO:0070469		

SuppMat 1e: tissues, 50 clusters, 10x enriched

Cluster 0	Colon endocrine cells	Colon enterocytes	Colon goblet cells
	Duodenum enterocytes	Placenta syncytiotrophoblasts - microvilli	Small intestine endocrine cells
	Small intestine enterocytes	Small intestine paneth cells	

SuppMat 1f: tissues, 10 cluster, 5x enriched

No enriched tags

SuppMat 2: Top 20, enriched Tags

SuppMat 2a: Pfam, 50 clusters, 10x enriched

Cluster 2	7tm_1	Cadherin	Cadherin_2	Homeodomain	Ion_trans	Lectin_C	Trypsin
	V-set						
Cluster 3	7tm_1						
Cluster 21	Cadherin	Cadherin_2	Collagen	RhoGAP			
Cluster 40	Cadherin	Cadherin_2					

SuppMat 2b: Pfam, 10 cluster, 5x enriched

Cluster 2	7tm_1	Ion_trans	LRR_8	Lectin_C	Trypsin	V-set
Cluster 5	Cadherin_2					

SuppMat 2c: GO, 50 clusters, 10x enriched

Cluster 0	GO:0004867	GO:0010951	GO:0031093	GO:0072562		
Cluster 2	GO:0004867	GO:0005201	GO:0005576	GO:0005615	GO:0010466	GO:0010951
	GO:0030414	GO:0062023	GO:0072562			
Cluster 21	GO:0005201	GO:0005604				

SuppMat 2d: GO, 10 cluster, 5x enriched

Cluster 0	GO:0002181	GO:0003735	GO:0022625	GO:0022626

SuppMat 2e: tissues, 50 clusters, 10x enriched

Cluster 4	Skin 1 lymphocytes	Skin 2 lymphocytes	Skin 2 melanocytes
Cluster 18	Lung alveolar cells type I	Lung endothelial cells	Skin 2 melanocytes
Cluster 49	Breast glandular cells	Hippocampus neuronal cells	Skin 1 lymphocytes
	Skin 2 lymphocytes	Skin 2 melanocytes	Tonsil germinal center cells

SuppMat 2f: tissues, 10 cluster, 5x enriched

No enrichment

SuppMat 3: Top 50, enriched Tags

SuppMat 3a: Pfam, 50 clusters, 10x enriched

Cluster 2	CH	RhoGEF	SH2	SH3_1	PK_Tyr_Ser-Thr	fn3
Cluster 3	7tm_1	C2	CH	Cadherin	Cadherin_2	EGF_CA
	I-set	Ig_3	Ion_trans	Pkinase	RhoGEF	SH2
	SH3_1	V-set	Trypsin	fn3	PK_Tyr_Ser-Thr	KRAB
	MFS_1	Homeodomain	zf-C2H2	Lectin_C		
Cluster 7	Cadherin	Cadherin_2				
Cluster 10	DEAD	Helicase_C	PHD			
Cluster 15	C2	Homeodomain	EGF_CA	Ig_3	PK_Tyr_Ser-Thr	SH3_1
	fn3	I-set				
Cluster 19	Cadherin_2					
Cluster 21	KRAB					
Cluster 22	RRM_1					

SuppMat 3b: Pfam, 10 cluster, 5x enriched

Cluster 0	Cadherin	Cadherin_2		
Cluster1	Cadherin	Cadherin_2		
Cluster 3	7tm_1	Lectin_C	Trypsin	V-set
Cluster7	DEAD	Helicase_C	PHD	RRM_1

SuppMat 3c: GO, 50 clusters, 10x enriched

Cluster 0	GO:0005762	GO:0032407				
Cluster 2	GO:0005518	GO:0032266	GO:0032495	GO:0036089	GO:0040008	
Cluster 3	GO:0005515	GO:0005518	GO:0005634	GO:0005737	GO:0005741	GO:0005829
	GO:0006694	GO:0007599	GO:0031902	GO:0032266	GO:0032407	GO:0032495
	GO:0040008	GO:0046872	GO:0050728	GO:0060005	GO:0090502	
Cluster 4	GO:0005762	GO:0032543	GO:0036089			
Cluster 5	GO:0005743	GO:0005762	GO:0032543			
Cluster 6	GO:0032407					
Cluster 7	GO:0036089					
Cluster 11	GO:0042754					
Cluster 15	GO:0032510	GO:0042754				
Cluster 16	GO:0032510					
Cluster 17	GO:0032407					
Cluster 18	GO:0007599					
Cluster 21	GO:0005743	GO:0005762	GO:0008033	GO:0032510	GO:0032543	GO:0042754
	GO:0051536					
Cluster 22	GO:0042754					
Cluster 24	GO:0032510					
Cluster 27	GO:0032495					
Cluster 34	GO:0004333					
Cluster 35	GO:1905313					
Cluster 37	GO:0042754					
Cluster 48	GO:0032510					

SuppMat 3d: GO, 10 cluster, 5x enriched

Cluster 1	GO:0007156	GO:0097151
Cluster 3	GO:0060658	
Cluster 7	GO:0045646	

SuppMat 3e: tissues, 50 clusters, 10x enriched

Cluster 3	Cerebellum GLUC cells - nucleus	Colon enterocytes	Colon goblet cells	Duodenum enterocytes
	Duodenum paneth cells	Rectum endocrine cells	Rectum enterocytes	Rectum goblet cells
	Testis pachytene spermatocytes			
Cluster 10	Skin 1 vascular mural cells	Skin 2 vascular mural cells		
Cluster 22	Skin 1 cells in spinous layer	Skin 1 langerhans cells	Skin 1 vascular mural cells	Skin 2 vascular mural cells
Cluster 48	Appendix germinal center cells	colonenterocytes	Colon goblet cell	Duodenum enterocytes
	Duodenum glands of Brunner	Duodenum paneth cells	Rectum enterocytes	Rectum goblet cells

SuppMat 3f: tissues, 10 cluster, 5x enriched

Cluster 7	Adrenal gland glandular cells	Bone marrow hematopoietic cells	Colon glandular cells
	Epididymis glandular cells	Esophagus squamous epithelial cells	Fallopian tube glandular cells
	Gallbladder glandular cells	Lung alveolar cells type I	Lung alveolar cells type II
	Lymph nodegerminal center cells	Pancreas exocrine glandular cells	Placenta decidual cells
	Placenta trophoblastic cells	Rectum glandular cells	Seminal vesicle glandular cells
	Stomach 1 glandular cells	Stomach 2 glandular cells	Testis Leydig cells
	Thyroid gland glandular cells	Tonsil germinal center cells	Urinary bladder urothelial cells

SuppMat 4:Summary of some clustering algorithms specialized for PPI networks

Algorithms tested by Song & Singh (2009) [48]

NetworkBlast [49] compares interaction networks in different species and identifies conserved paths and conserved clusters. Paths are meant to resemble signal transduction pathways and clusters in this context refer not merely to proteins with similar binding partners, but proteins predicted to be in a protein complex with each other. The significance of a network is calculated by comparing its likelihood score to the likelihood score of a randomized network. Functional enrichment within a cluster is calculated by using GO annotations as a reference [49]. Coding a variation of this method for RNA-RBP interactions only makes sense if we assume that all RNAs with similar RBP binding networks can be considered a complex which they – in the context of mRNAs – are not. It would also require CLIP data from multiple species to compare the interaction networks between species.

Cfinder [50] requires a binary interaction list and creates an interaction network and searches for dense subgraphs (groups). In this algorithm, one node in the network can belong to multiple groups. The groups generated are intended to be used for discovery of novel protein functions [50]. It uses a clique perlocation method. A cluster consists of multiple adjacent k-cliques and the links within those cliques. A k-clique is a subgraph with k-nodes (with k typcially being a number between 4 and 6) and k-cliques are considered adjacent to each other when they share k-1 nodes [50].

The Markov clustering (MCL) [51] approach concerns clustering of proteins based on their sequence similarity scores rather than their interaction patterns. There are, however, ways to use MCL on protein interaction networks [55]. MCL algorithms work by computing the probabilities of random walks, whereby the interaction network is split into clusters using the expansion and the inflation operator. Expansion is meant to calculate ways of getting from one node to another through random walks. Inflation is meant to boost the probability of intra-cluster walks while decreasing inter-cluster walks thus improving cluster separation. The method does not require *a priori* understanding of cluster structure [51].

DPClus [52] generates an interaction network and searches for densely connected network regions. These regions are called clusters and are supposed to be biologically functional units, thus allowing one to predict complexes and protein function. The method is also capable of detecting subclusters within a cluster by detecting sparse regions within a cluster [52].

Mcode [53] is a method for finding regions of an interaction network that are densely populated based only on connectivity data. These regions often contain proteins known to be a part of the same protein complex. Importantly, Mcode can define relevant regions that even when data from different experiments is used. During the vertex weighing, the method takes into account that vertices usually follow a power law, meaning there are a few highly connected and many sparsely connected vertices, and gives highly connected vertices even stronger weight. The most densely connected vertex functions as a seed and the predicted complex is expanded on from it. Connected vertices over a given weight threshold are added to the cluster and the vertices connected to them are also checked for their vertex weight [53].

SpectralMod [54] attempts to optimize the modularity by subdividing the interaction network. Modularity can be calculated by subtracting the expected number of edges in a randomized network from the number of the edges in the network observed. The network is divided as long as the modularity of the network improves. Once there is no longer a way to increase the modularity by further divinding the system, the final clusters have been obtained [54].

Other Algorithms

ClusterONE [56] is a clustering algorithm designed for the detection of overlapping protein complexes in protei-protein interaction networks. It was published in 2012 and uses a cohesiveness score and a greedy growth process. Typically, the protein with the largest number of connections is used as the seed and a cohesive group is grown from it using a greedy growth process. Once the growth is finished, the protein with the next largest number of connections that was not included in the first grouping is used as the next seed. This is done until all proteins are grouped. In the next step, overlapping cohesive groups are merged when their overlap is greater than 0.8. In the final step, complex candidates that contain too few proteins or have a low density are removed [56].

IPCA [57] is a modification of the DPClus algorithm. It uses a different extending model that takes the topology of complexes into account by introducing the interaction probability as a metric [57].

DPCMNE [58] first compresses the PPI network via a hierarchical compression strategy and network representation learning is performed on these smaller networks obtained. That way, protein embeddings are obtained and then concatenated to obtain an embedding of the full input network. From this concatenated embedding, protein complexes are detected using a core attachment based clustering method [58].

IPC-BSS [59] utilizes the brainstorming strategy and combines GO terms and network topology for its clustering. For the clustering, center nodes are selected and each of them consolidated with other nodes close by to obtain the initial clusters. The distance metric is a combination of GO similarity

and network topology metrics. These clusters are optimized using an improved K-means algorithm. In a final step, clusters are merged and filtered out so only the best, most robust clusters remain [59].

CO-DPC [60] is designed to find dynamic protein complexes. The algorithm uses gene expression profiles to find all active proteins and then builds dynamic PPI networks according to the co-expression principle. It is also capable of detecting dense subgraphs to use them as cores that it then expands to predict protein complexes. All resulting complexes are ranked according to their density and overlap and those less dense with large overlap were filtered out [60].