Inferring microRNA regulation: A proteome perspective

Dan Ofer, Michal Linial*

Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

D.O. dan.ofer@mail.huji.ac.il; ORCID: 0000-0001-5136-8014

M.L. michall@cc.huji.ac.il; ORCID: 0000-0002-9357-4526

Abstract

Post-transcriptional regulation in multicellular organisms is mediated by microRNAs. However, the mechanisms that determine if a gene is regulated by miRNAs are poorly understood. Previous works focused mostly on miRNA seed matches and other features of the 3'-UTR of transcripts. These common approaches relied on knowledge of the miRNA families, and computational approaches still yield poor, inconsistent results, with many false positives. In this work, we present a different paradigm for predicting miRNA-regulated genes based on proteins. In a novel, automated machine learning framework, we use sequence as well as diverse functional annotations to train models on multiple organisms using experimentally validated data. We present insights from tens of millions of features extracted and ranked from different modalities. We show high predictive performance per organism and in generalization across species. We provide a list of novel predictions for *Danio rerio* (zebrafish) and *Arabidopsis thaliana* (mouse-ear cress). We compare genomic models, and observe that our protein model outperforms genomics, while a unified model improves on both. While most membranous and disease related proteins are regulated by miRNAs, we observe the G-protein coupled receptor (GPCR) family is an exception, being mostly unregulated by miRNAs, and we raise possible explanations for this. We further show that the evolutionary conservation among duplicated genes does not imply a coherence in miRNA regulation. We conclude that duplicated genes diverge in their tendency to be miRNA regulated. However, protein function is informative across species in predicting post-transcriptional miRNA regulation in living cells.

Keywords: AI model, autoML, GPCR, machine learning, miRTarBase, paralogs, post transcriptional regulation, TargetScan.

1. Introduction

MicroRNAs (miRNAs) post-transcriptionally regulate genes across all animals and plants. miRNAs are a class of short (~22 nucleotide) noncoding RNAs (ncRNAa). Mature miRNAs act via complementarity with their target mRNAs. This pairing takes place mostly in the 3'-UTR of the transcripts (Romero-Cordoba et al., 2014). In mammals, such binding leads to translational repression of the target and direct or indirect degradation of the miRNA-targeted transcript via deadenylation and decapping of its target (Valencia-Sanchez et al., 2006; O'Brien et al., 2018). miRNAs play key roles in a broad range of cellular processes and the response to changes in the environment (Leung and Sharp, 2010). The miRNA profile is tissue-specific, and an indicator of cell identity. Their ability to maintain cell and tissue homeostasis is critical, with many miRNA genes implicated in human diseases such as metabolic, inflammatory, and neurodegenerative diseases (Vishnoi and Rani, 2017 11). In cancer samples, the miRNA composition changes along with the tumorigenic process. Therefore, the miRNA profile carries useful diagnostic and prognostic potential for tumor typing and patient survival.

With the maturation of deep sequencing methodologies for small RNA identification, the number of reported mature miRNAs has drastically increased. The exhaustive catalog of miRNA (miRBase v. 22) (Kozomara et al., 2019) reports on 1917 genes that account for 2625 mature miRNAs from humans, and 1234 and 1978 genes and mature miRNAs from mice, respectively (Quillet et al., 2019). With a set of strict criteria imposed by miRBase, only a quarter of the listed miRNAs from humans are labeled with high confidence. Many of the rest have yet to be experimentally confirmed (Alles et al., 2019 2). From the standpoint of miRNA targets, it has been demonstrated that many human genes are under selective pressure to maintain miRNA pairings (Friedman et al., 2009). Despite an increase in the number of validated miRNAs, the estimated number of regulated genes remained between 60–80% of all human protein-coding genes (Sayed and Abdellatif, 2011 12; Huang et al., 2019 15).

In the last 15 years, computational miRNA-target prediction algorithms and tools have been developed (Sethupathy et al., 2006 7; Yue et al., 2009 5; Riffo-Campos et al., 2016 6). Almost all of these predicting tools are based on features derived solely from the genomic sequence. Major features include seed complementary, evolution conservation, free energy, and the position of miRNA binding sites (MBS) at the 3'-UTR. Most tools suffer from a large number of false positives, poor accuracy and sensitivity, and show a great degree of inconsistency among them (Min and Yoon, 2010).

High throughput methodologies (e.g., CLIP-seq, CLASH, CLEAR-seq) were used to conduct hundreds of experiments to infer miRNA-mRNA interactions (Li et al., 2014; Karagkouni et al., 2018). These experimental methods allowed us to assess the reliability of the different miRNA-mRNA prediction tools. In general, the match between the experimental results and the computational predicting methods is poor. Experimental observations (e.g., CLIP data) and sequence-derived information about miRNAs and mRNAs are used to determine whether a specific transcript is a genuine target of miRNAs. In this study, we address the question of whether a gene is a target of regulation by any miRNA based on their protein products, using a supervised machine learning approach. i.e., we predict if a protein is subject to direct regulation by "any" miRNA. The underlying notion is that the coding regions of most genes are under strong negative selection forces and potentially include information that determines the essentiality of a gene under miRNA regulation, irrespectively of a specific combination of miRNAs. We use miRTarBase 2020 (Huang et al., 2020) as an experimentally validated ground truth dataset. We trained the system using experimentally validated resources for human, mouse, and other model organisms and reached high performance on the task of predicting gene-miRNA interaction, using

primarily protein and minimal sequence level attributes. We included multimodal inputs from proteins and generalized across different species. We also evaluated genomic information, and compared it to the proteomic model, and a unified model that learns from both the gene and protein sequence. We present an in-depth analysis of both novel and established features, extracted automatically using an AI-assisted machine learning framework (SparkBeyond). We present a list of candidate miRNA gene predictions from less studied organisms. Our model highlights the value of information embedded in the functional proteome in revealing the complexity of regulation by miRNAs.

2. Methods

2.1.   Database: miRNA-Target interactions

We used miRTarBase 2020 (V9) as a gold standard for miRNA-Target interactions (MTIs) (Huang et al., 2020). miRTarBase compiled experimentally validated MTIs, mostly from mouse and human. The entire database collected 4.5M data points, based on CLIP-seq experiments, as evidence for human MTIs (covers ~3000 miRNAs and 17,400 genes), and 0.7M mouse MTIs (covers 2250 miRNAs and 14,300 genes). It is used as a ground truth for training. The dataset was downloaded from miRTarBase 2000 (Huang et al., 2020). The experimental results in miRTarBase 2020 (V9) have associated "weak" or "strong" evidence. Weak support refers to data collected from high-throughput experiments (e.g., CLIP-based NGS experiments, pSILAC proteomics), while strong evidence is compiled from targeted experiments such as quantitative RT-PCR (qRT-PCR), Western blots, and reporter assays. We defined a target as positive (i.e., miRNA regulated) by having "strong" experimental evidence or at least 2 unique "weak evidence" experiments. The remaining genes from miRTarBase with a single "weak" experimental evidence were labeled as "likely positives" (0). These were treated as positives for the purposes of downstream analyses, unless otherwise stated. We note that excluding these "weakly labeled" samples improved modeling performance across all organisms (not shown). All other genes were marked as "negatives", i.e., not targeted by the specific miRNAs ("-1").

2.2.   Database: Proteome

Proteins were downloaded from UniprotKB for all organisms analysed in this study. Only manually reviewed SwissProt proteins were used (Breuza et al., 2016), except for C. elegans, where all the proteins were used, after excluding fragmented sequences. Proteins annotated with no experimental evidence for their existence by UniProtKB-SwissProt were excluded, as by definition there could be no experimental evidence for their miRNA regulation. Altogether 45,846 proteins were analysed. We also analyzed the proteomes of *Danio rerio* (zebrafish) and *Arabidopsis thaliana* (mouse-ear cress).

We identify genes with their matched proteins. Proteins from UniProtKB were mapped to the genes listed in miRTarBase, TargetScan and TreeFam according to their primary gene name. To connect genomics with protein identifiers, we mapped human genes by their primary gene name. For human proteins, 76% of the proteins were successfully and uniquely mapped across the different resources. A negligible amount of proteins with no primary gene name were dropped.

2.3.   Extracted features

A wide range of metadata about each protein from UniprotKB was used as proteins' features. These included the proteins' amino acid sequence (e.g., amino acid composition, counts, n-grams), molecular weight, protein length, functional keywords (e.g., secreted, membranous), gene ontology (GO) annotations for all three branches: molecular function, cell localization and biological process), pharmaceutical uses, tissue specificity, protein family, post-translational modification, involvement in

disease, compositional bias, non-terminal residues and more. In addition to the information associated with the proteins by UniProtKB, we also derived engineered features based on the primary features. notably amino acid composition and k-mers, n-grams (e.g., combinations of keywords), counts of known annotations (Ofer and Linial, 2015).

As a separated set, we extracted limited genomic data for the human proteome using the Biomart querying system of Ensembl (Yates et al., 2016). The derived features included: The length of the UTR (5' and 3'), the counts of alternative splicing, chromosomal position, nucleotide counts, k-mers and their frequency, k-mer features extracted from the 3'-UTR genomic sequence, and transcript length.

2.4. AutoML and feature extraction

Feature extraction, engineering, selection (Ofer et al., 2021), and ML model selection, parameter tuning and training was performed using the SparkBeyond autoML framework (See patent/US20170017900A1). Previous work has shown the benefit autoML models, in order to comprehensively and automatically find possible predictive signals in complex data, including in biology and healthcare (Cohen et al., 2021). The system automatically extracts and ranks a wide range of compositional features from training data. The system applies hyperparameter tuning and evaluation of machine learning models. In this study, the SparkBeyond framework was applied to genomic, proteomic and annotation data. Across the different problem formulations, the system generated on average ~22 million candidate features per organism, prior to selection. A maximum of 300 features are selected and used for the machine learning models, based on the training data. Features include textual features (n-grams, k-mers, tokenization), counts, aggregations (e.g. max, min, average, decile), interactions (e.g. length of a sequence divided by weight), missing value imputations, similarities and more.

We use RIG (Relative Information Gain) as a measure of feature importance. RIG refers to the information gain measured as a reduction in entropy produced from partitioning a set with attributes a and finding the optimal candidate that produces the highest value.

$$IG(T, a) = \mathrm{H}(T) - \mathrm{H}(T|a), IG(T, a) = \mathrm{H}(T) - \mathrm{H}(T|a),$$

where T is a random variable and $\mathrm{H}(T|a)\mathrm{H}(T|a)$ is the entropy of T given the value of attribute a. It encapsulates both the uplift of a feature (the increase in a class's likelihood, given a binary partition induced by the feature), and the support (the number of samples covered by the feature). A feature with a high RIG is expected to be relevant for any model, given that it will have good support and lift.

We report as final evaluation on a held-out test set, comprising 20% of the data. Note that the feature extraction, selection and model evaluation and tuning is performed only against a subset of the training data, to avoid the risk of overfitting and model leakage. To improve interpretability, we limited the system to prefer "simple" features, at a slight cost to performance. Features are ranked by their marginalized, non-redundant mutual information score, as well as a custom regularization scheme to favor semantically simpler features (i.e., less composite functions). Performance in human only data was based on 20% held out stratified collection of 3810 test samples. The protein sequence only model uses just the statistics from the primary sequence (e.g., length, n-grams, amino acid composition (Ofer et al., 2021)), without any of the additional annotations or metadata.

We used standard definitions for the model performance including precision and recall. In addition we report accuracy = (TP + TN)/(P + N), and the F1-Score = 2 TP/(2 TP + FP + FN) using routine notations of T (true) and F (false), P (positive) and N (negative).

2.5.   Software

Data processing and analysis used the Python Pandas (VanderPlas, 2016), and Scikit-learn software packages (Raschka and Mirjalili, 2019).

Code, figures and data available at: https://github.com/LinialLab/microRNA-Protein-Regulation.

3. Results

3.1.   Problem definition

In the context of cells and tissues, genes that respond to miRNA regulation comprise direct targets (i.e., miRNA binds to the 3'-UTR of a gene's mRNA, affecting transcript stability and inhibiting translation) and indirect effects (e.g., miRNA that downregulates a transcription factor (TF), leading to attenuation of transcription of a set of TF-responsible genes). In addition, quantitative competition on miRNA binding sites, under the paradigm of ceRNA, indirectly causes a shift in target site occupancy (Lai et al., 2016). In this study, we question whether a coding gene is subject to direct regulation by any miRNA. To answer this question, we considered features from a complete set of validated proteins within each of the studied organisms (coined "reviewed proteome" by UniProtKB-SwissProt). Our problem setting was defined as binary classification, using supervised machine learning models. Scikit-learn's linear logistic regression model was automatically selected as the best model across the different runs by the SparkBeyond autoML framework (see Methods), outperforming dozens of other attested architectures. Importantly, functional annotation provided by UniProtKB-SwissProt does not include miRNA-related knowledge, and potential target "leaks" were carefully excluded or filtered for (e.g. evidence levels for a gene's transcript).

We trained models on data and miRNA regulation annotations from different organisms. Supplementary **Table S1** summarizes the data used for the studied organisms' proteomes. We combine diverse protein functional annotations along with traditional sequence and biophysical features, as well as quantifying the relative contribution of universal genomic sequence-based features (i.e., not miRNA family specific). We identified key features that contribute to the models and suggest shared principles in miRNA regulation across species.

3.2.   Inconsistency in existing miRNA target predictions

Existing tools for predicting miRNA-gene interactions demonstrate poor consistency between tools and major resources. It is anticipated that it is mostly due to the very large number of false positives. The question of what makes transcripts in any organism a good target could not be answered based on current tools (Min and Yoon, 2010). In an effort to reduce the flood of false positives, a statistical framework across different predicting algorithms was developed, with the notion that miRNAs work together in a commutative fashion (Balaga et al., 2012).

Experimentally validated targets (as derived from miRTarBase 2020) are expected to be of higher quality and consistency, and thus are used as the "ground truth" annotations. However, such annotations suffer from inherent biases. For example, it is likely that highly expressed transcripts will be detected more often than lowly expressed ones. Similarly, miRNAs that are expressed under defined conditions might be underrepresented experimentally. Obviously, some organisms are studied more than others (e.g., human and mouse), resulting in a biased view of how many miRNA targets there are in most organisms.

Figure 1A illustrates the disparity in the fraction of genes predicted to be regulated by miRNAs (as predicted by TargetScan) and those experimentally validated (by miRTarBase). We illustrate the TargetScan (release 8.0) (McGeary et al., 2019) as a reference point. It provides a score for each pair based on genomic and biochemical models of miRNA binding specificity. However, the algorithm was mostly tuned as a miRNA-mRNA predicting tool for mammals. The discrepancy in the proteomes of *D. rerio* (zebrafish) and *C. elegance* (worm) emphasizes the unproportionate number of predictions by TargetScan in view of the shortage of experimentally validated observations. **Figure 1B** shows that this inconsistency is also detected in the number of confirmed miRNA genes. The proportion of confident miRNAs reported by miRBase reaches 70% of the entire miRNA gene list. In D. melanogaster, however, it is only 57% (out of a total of 258 genes, miRBase) (Supplemental **Table S2**). In humans, the fraction of confident miRNAs is only 26% (505 out of 1917 miRNA genes) and poor confidence miRNAs are those with minimal expression level or non-canonical stem-loop structure.
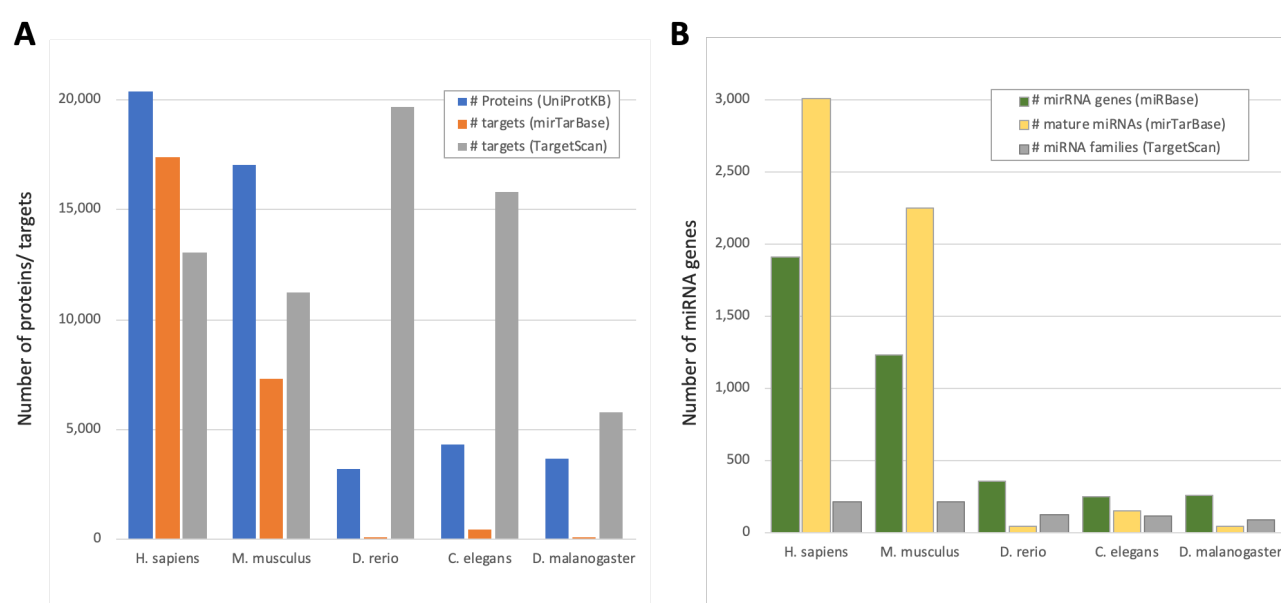

FIGURE 1. miRNAs and regulated genes by species. **(A)** Number of proteins from UniProtKB/SwissProt reported as "reviewed proteome" (filtered as in Methods), the number of reported targets from mirTarBase as high confidence (by evolutionary conservation) reported by TargetScan, for 5 model organisms. The latter two targets include all proteins, not just SwissProt. **(B)**. Number of miRNA stem-loop genes and mature miRNAs reported by miRBase, and the number of miRNA families from TargetScan for or the same model organisms as in (A).

3.3.　Prediction of miRNA regulation across organisms

For each organism, we trained a model on 80% of its curated proteins (proteome filtered collection, see Methods) and presented results on the disjointed remaining 20% test set. Labels (miRNA regulation/non-regulation) were derived from experimentally validated miRNA regulation data (miRTarBase 2020). We identified 76% of human genes and only 37% of mice as regulated (validated) genes, with lower rates for other organisms.

Table 1 shows the performance of the protein models for different model organisms. Models were only based on annotated protein data. Model results are shown for the test set (which was not used in model training). The total number of instances is for data after filtering and removal of genes that failed in the

mapping of primary gene names. The percentage of validated miRNA regulation refers to experimentally validated targets. The performance is presented according to the area under the receiver operating characteristic curve (rocAUC), precision, and recall (of the minority class) on the test set. We concluded that, despite the low coverage of annotated proteins in the fruit fly and worm proteomes, the machine learning-based model successfully characterized genes that are apparently regulated by miRNAs. It also emphasized the commonalities and differences in miRNA regulation across model organisms.

Table 1 also tests whether the strict selection of the trained set for only curated and well-annotated proteins limits the performance of the miRNA-regulatory predictor. C.Elegans for example, has only a very small number of validated miRNA targets and reviewed proteins, limiting our ability to learn from this small intersection. To this end, we partitioned the C. elegans into two sets, the entire proteome (8,577) and a smaller reviewed set (a total of 2,106). We showed a substantial improvement in the rocAUC for models trained on the larger set, despite its lower quality annotation. In mammals, it has been shown that most genes are directly regulated (based on CLIP experiments). However, the extent of miRNA regulation in invertebrate organisms is unknown.

Table 1: miRNA predictions per organisms

Species	Precision (%)[a]	Recall (%)[a]	Total proteins	Validated as miRNA regulated (%)	rocAUC
H. sapiens	63.4	27.8	18,808	76.0	76.9
M. musculus	57.1	37	16,355	37.0	67.6
C. elegans (reviewed)	0	0	2,527	2.0	62.9
C. elegans (all proteins)	50	11	8,577	1.2	79.2
D. malanogaster	33.3	10	3,140	1.5	67.5

[a]Precision, Recall values are for the predicted minority class, at default cutoff.

3.4. Generalizing between species

In addition to models per organism, we evaluated the ability of the models and features to generalize between species. We trained a model on all human proteins and evaluated it using all mouse proteins as a test set, and vice versa (mouse to human; **Table 2**). We observe excellent stability, with performance dropping only slightly compared to a dedicated model trained on the species' own data. This supports our use of the models to predict between different species, arguing that functional attributes generalize well between species.

Table 2: miRNA predictions between organisms

rocAUC score	Trained on Human	Trained on Mouse
Evaluated on Human	76.9	64.6
Evaluated on Mouse	75.0	67.6

3.5.   Protein features predict miRNA regulation

We observed a number of predictive attributes that were consistent across different iterations, and even different organisms. There exist clear global attributes in favor (or disfavor) of miRNA regulation. Features were ranked according to the calculated RIG (relative information gain) values, with higher RIG implying lower uncertainty for the target under the feature's induced partition, i.e., greater information about it (see Methods). A full ranked list of features and their statistical properties is in supplementary **Table S1.**

A high RIG value encapsulates a high degree of confidence with strong statistical significance for the discrimination power (chi2 p-value <1.0E-04 on the test set). **Figure 2** shows the partition of the top 15 features with high uplift in discriminating for miRNA regulation (**Figure 2A**) and not being miRNA regulated (**Figure 2B**). For example, long protein length, identified with an optimal threshold of under 349 amino acids long, with sequences shorter than this being 1.38-fold more likely to be being associated with a gene that is not regulated by miRNAs (marked as prediction = 0). In the case of protein length, there are no missing values (i.e., all proteins have a sequence from which length is derived). However, the support for other features is often very limited. For example, there are only 232 proteins in the training set with the "sensory" keyword, which accounts for 2.4% of all proteins. For this selected set, the tendency to not be miRNA regulated is substantial (2.57-fold). Among the top features is membership in GPCR family 1 (associated with the olfactory receptors) which are 2.41 times less likely to be regulated (**Figures 2,3**). Features directly associated with protein signaling, localization and stability, such as post translational modification and alternative splicing, are significant for predicting miRNA regulation. For example, proteins involved in the ubiquitination process ("ubl") are in favor of miRNA regulation (by 1.26-fold), suggesting that gene regulation may involve multiple regulatory mechanisms, including tagging for degradation. Other annotations (from Gene Ontology), such as subcellular location (e.g., cytosol, nucleoplasm) and interaction with RNAs (e.g. RNA binding or ribonucleoproteins) also contributed to the models' success.
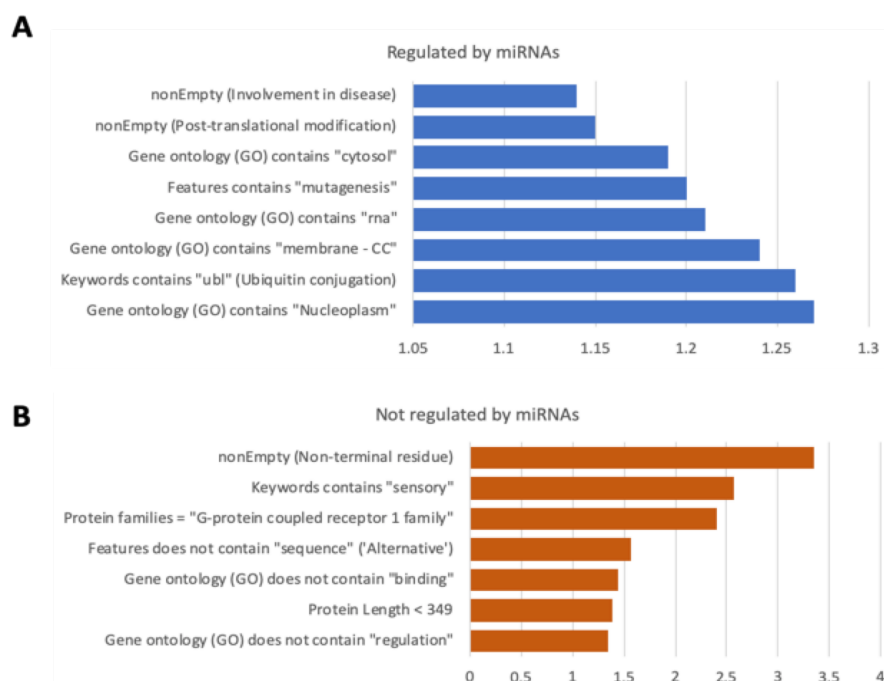
FIGURE 2. Selection of impactful features from the human protein model. **(A)** miRNA regulated, (B) not regulated by miRNA. The x-axis is the lift of features, relative to class prior baseline (~76% regulated, 24% not regulated). All above features have an uncorrected p-value at least under 0.0001, using a Chi2 test on the test data. Full list of features and statistics in supplementary **Table S3**, along with ranking by information gain (RIG). Note the different scale of the effect in A and B. CC: Cellular component branch of GO annotations.

Features related to amino acid sequence composition were also informative, e.g., having more than a single methionine, or an especially high percentage of lysine (K). However, these characteristics are most likely a result of the underlying codon frequency (Supplemental **Table S1**), or relate to specific structural/functional families. We expect that some of these features may be confounded by the aforementioned higher-level functional properties.

Table 3: Tissue specificity features

Feature - tissue specificity	Dominant prediction	RIG[a]	Support (%)	Lift 1[b]	Lift 0[b]
Contains "muscle"	1	0.012	13.22	**1.19**	0.55
Contains "heart"	1	0.011	13.96	**1.18**	0.57
Contains "brain"	1	0.008	17.48	**1.14**	0.67
Contains "testis"	0	0.007	0.92	0.33	**2.57**

[a]Relative information gain. [b]Lift indicates the effect fold for the feature on the prediction of being (Lift 1) and not being (Lift 0) miRNA regulated.

Another notable set of features relates to tissue specificity. **Table 3** shows the contributions of the different tissues and organs. This set of features aligns with the accepted notion that miRNA expression profiles are tissue-specific and effective at distinguishing between tissues (Rasnic et al., 2017). However, it is unknown whether some tissues are more amenable to regulation than others. Most tissue specificity features (**Table 3**) were in favor of miRNA regulation (prediction =1), contributing to the discriminative power by ~1.2-fold lift and a RIG of ~0.01. Interestingly, the testis was an exception, with a 2.6-fold effect against miRNA regulation. The statistical results (**Table 3**) were also validated by a Chi2-based hypothesis test, measuring the probability of observing a Chi2 deviation between the expected and observed labels of this extreme or greater (p-value), on the test set. All p-values were <0.001.

3.6. Functional groups coherence in miRNA regulation

We tested functional groups according to the UniProtKB family relations. The most significant family groups, characterized by their average regulated fraction are shown (**Figure 3**). Examples with statistically significant statistics are listed. There are 475 protein kinases, of which 85% have been shown to be under miRNA regulation. The key signaling proteins of small GTPases (e.g., Ras, Rho) are also shown to be regulated (160 proteins, 92% are miRNA regulated). In other protein families, such as Histone H2A & H2B, none of the 17 proteins are regulated by miRNA. Supplementary **Table S4** lists protein families along with their prediction statistics.
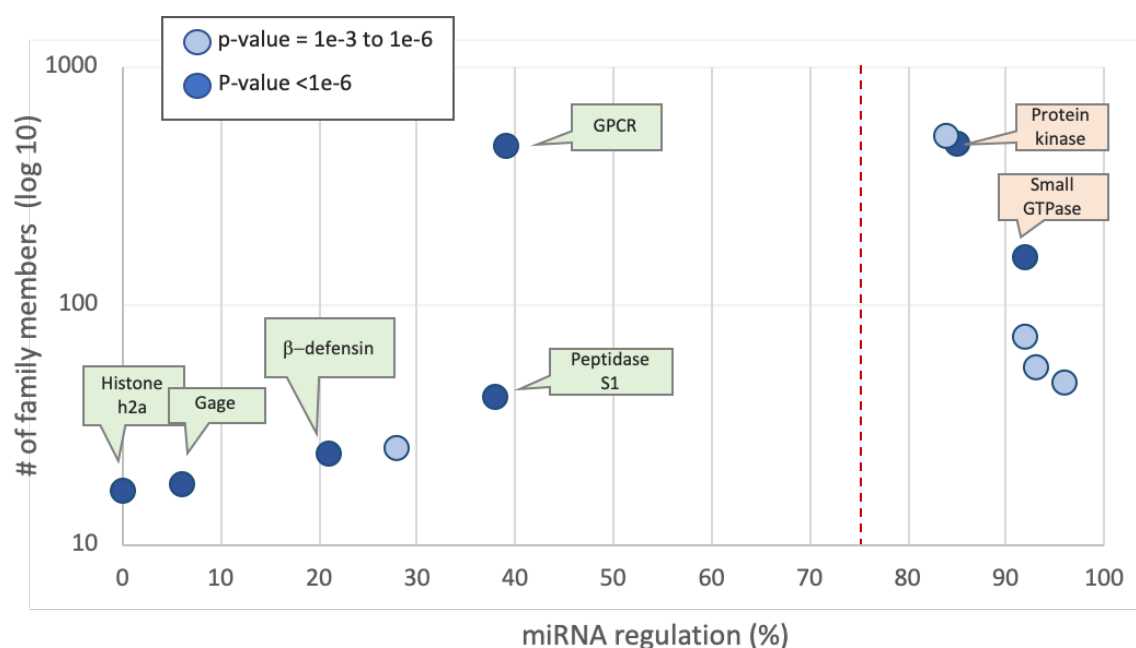
FIGURE 3. Protein families by mean percentage of miRNA regulation prediction of their members. Families defined by UniProtKB. Examples from human families are partitioned by the prior rate for miRNA regulation of 75% (dashed line). The p-value for each family member are marked by the two-sided binomial test. The most significant families (p-value <1E-6 are indicated) are listed, colored green and orange by their prediction as 0 and 1, respectively. Supplementary **Table S4** lists protein families along with their prediction statistics.

3.7.   miRNA predictions in novel organisms

We examined putative predictions for less annotated model organisms such as *D. rario* (zebrafish) and *A. thaliana* (mouse-ear cress). 355 miRNA genes in zebrafish (216 high confidence by miRBase) and 326 miRNAs in A. thaliana (177 high confidence by miRBase) were identified from genome and RNA-seq experiments (Zhan and Lukens, 2010). However, these organisms have a low number of experimentally validated miRNA targets, with just 187 and 70 for *D. rario* and *A. thaliana*, respectively. This is despite a comparable fraction of their genes being regulated according to computational predictions (e.g., **Figure 1A**). We focused our analysis on the zebrafish and its 25,919 known proteins as retrieved from TargetScanFish. We included in the prediction scheme all non-fragmented proteins, irrespective of their annotation status. Our model, trained exclusively on annotated human genes, predicted miRNA regulation in 22,759 (87%) zebrafish genes, and 30,502 (83%) mouse-ear cress genes.

Compared to the validated shortlist of known miRNA targets, our "positive" predictions match the quantities observed in the well-studied humans and mice (~82%). Our results also have a better concordance with TargetScan labels (Pearson correlation = 0.11) than with the small, non-represented experimentally validated samples (Pearson correlation = 0.027), further hinting that to a large degree, genes that are likely miRNA candidates have not yet been validated. Filtering the putative TargetScanFish to include only candidates from conserved miRNA families yields similar, improved results with a correlation of 0.12, supporting the above hypothesis. Supplementary **Table S5** lists all predictions for *D. rario* and *A. thaliana*.

3.8.   Shared miRNA regulation by protein function

Our predictions suggest that miRNA regulation is prominent also in organisms lacking extensive experimental studies. The information provided to the predictor was restricted to proteomic and global functional information (see Methods). We list genes from *D. rario* that were strongly predicted to be under miRNA regulation scheme but lack experimental evidence. Supplementary **Figure S1** shows few homologous genes of D. rario genes that share function with human and mouse and were also among the top predictions of miRNA regulation, with no known experimental evidence. For example, park7 gene shares 91.5% identity at the protein level between humans and mice, and is a clear homolog of park7 in zebrafish (with 83.1 and 81% sequence identity in human and mouse, respectively). However, applying TargetScan applied in a stringent mode for including only reliablly conserved miRNAs shows that while there is no miRNA binding site in mouse transcript (ENSMUST00000030805.8), and only a single miRNA binding site in human ENST00000493678.1 transcript, but TargetScan predicts 13 different miRNAs (total of 18 binding sites) in the regulation of Park7 homolog (Maillard deglycase) from zebrafish. Additional examples include the important cancer driver genes pkd2 and apc (Supplementary **Figure S1**).

Table 4 lists gene candidates with a high probability of being miRNA regulated in zebrafish but were not identified as such by TargetScanFish. Note that these highly predicted genes act in the nuclei during development, undergoing post translational modifications. However, the prediction score is not a direct outcome of a long 3'-UTR. As noted, this specific model did not use any genomic/ 3'-UTR length information.

Table 4: Zebrafish miRNA predictions

Gene name (primary)	Protein name (description)	miRNA regulation score	3' UTR length (bp)	UniprotKB Keywords
hnrnpub	Nuclear ribonucleoprotein U-like protein	91.8	1789	Methylation. Phosphoprotein. Ribonucleoprotein. Viral nucleoprotein. Virion
utp25	U3 small nucleolar RNA-associated protein 25 homolog (UTP25 small subunit processor component)	90.8	367	Developmental protein. Nucleus. Phosphoprotein.
tbxta	T-box transcription factor T-A (Brachyury protein homolog-A) (Zf-T-A)	89.8	863	Activator. DNA-binding. Developmental protein. Nucleus Transcription. Transcription regulation. Wnt signaling pathway
ttn.1	Titin, tandem duplicate 1	89.6	1,001	ATP-binding. Coiled coil. Disulfide bond. Immunoglobulin domain. Kinase. Nucleotide-binding Repeat. Transferase

mib1	RING-type E3 ubiquitin transferase (EC 2.3.2.27)	89.5	1,556	ANK repeat. Cell membrane. Coiled coil. Cytoplasm. Cytoskeleton. Developmental protein. Membrane. Metal-binding. Notch signaling pathway Repeat. Transferase. Ubl conjugation pathway. Zinc-finger

3.9.   Integrating genomic information into a unified model

Most computational prediction tools and algorithms for miRNA-mRNA interactions are based on sequence pairing in the 3'-UTRs. We added an additional set of genomic features extracted from the gene's summary statistics (length, CDS length, nucleotide composition), the 5'-UTR and 3'-UTR length, and the number of splicing variants. The performance of a model using only the genomic features was inferior to that of the protein-informed model (**Figure 4**). Note that our model does not consider the number of miRNA binding sites, rather it includes the 3'-UTR length (which is obviously associated with an increased probability of binding miRNAs). The model outperforms a naive boolean heuristic based on the existence of a 3'-UTR. The most important information at the genomic level is that a gene has a long sequence (e.g., total transcript length > 540 nt) and/or a long 3'-UTR. Another highly informative feature of a gene is its chromosome position (i.e., start and end). Specifically, genes located towards the "start" of the chromosome were less likely to be miRNA targets. Notably, none of the extracted genomic features relate to any information on specific miRNA sequences, the presence of a match with a seed, or energetic data on their pairing (e.g., n-grams complementary to a list of seeds were not included). Combining the modalities into a single, unified model outperformed any individual modality (**Figure 4**).
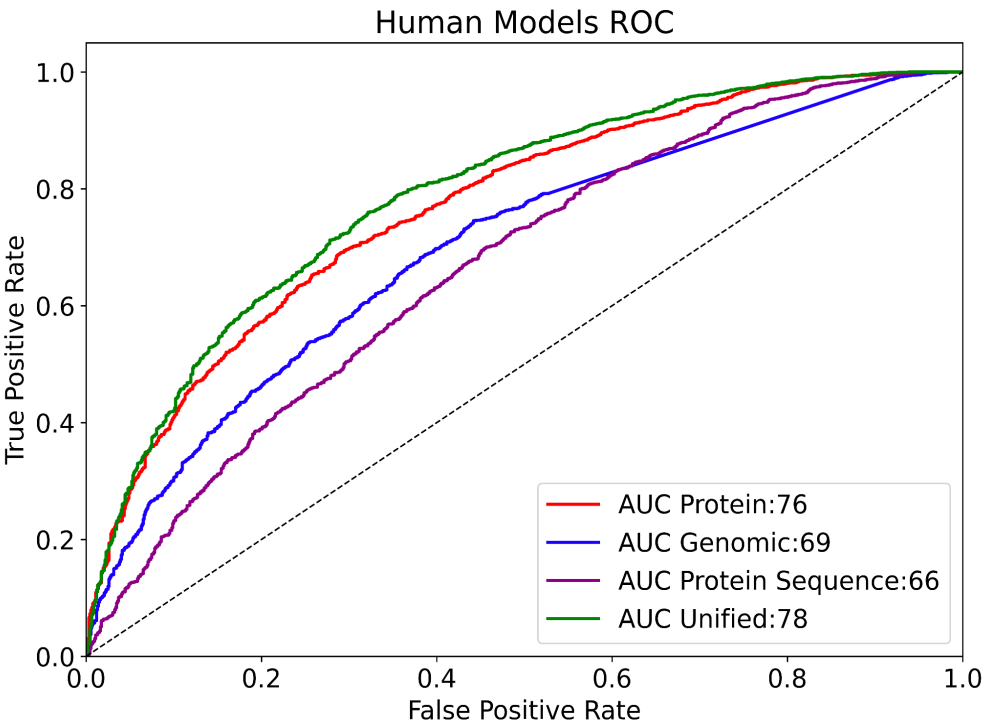
FIGURE 4. Comparing the rocAUC (Area Under the Receiver Operating Characteristic Curve) of different models on human data. Results shown for the test set (3810 samples, 75.8% miRNA regulated). "Protein" model includes functional annotations, keywords, primary sequence and protein families. "Protein sequence" model used only protein sequence and derived features. Unified model includes all protein and genomic features. For detailed performance see Supplementary **Table S5**.

3.10. Evolutionary related genes display regulation coherence

We further tested whether genes that share evolutionary ancestry (i.e., orthologs) share miRNA regulation coherence. From a sequence perspective, the 3'-UTR is not under strong purifying selection across species, and in general, conservation is minimal among orthologs. To test the cases in which miRNA regulation has been specialized, we used TreeFam as a source for family relationship groups (Schreiber et al., 2014). There are 8,819 unique TreeFam families and 5,377 protein families in this data. The question we asked is whether genes that belong to the same family share miRNA regulation. We matched the TreeFam groups and UniProtKB-SwissProt defined protein families according to the primary gene names from the reviewed human proteome. About 90% (16,995) of the genes had a matching TreeFam gene family, and 72% had a matching protein family. We note that the remaining 10% of genes that lacked a matching TreeFam family were twice as likely to be unregulated (49% unregulated vs. only 24% overall). The distribution of protein family sizes in humans is shown in **Figure 5A**. Note that the majority of families had only one member and thus lacked the notion of paralogs. A small number of proteins belong to multiple families. **Figure 5B** schematically describes the notion of paralogs.

Within TreeFam families that consisted of at least 2 genes (11,356 genes), we tested the null hypothesis, for which there is no coherence in miRNA regulation in proteins belonging to the same paralogous groups. We found that 61% of TreeFam families, and 56% of protein families were coherent and split between the families that are all regulated (or not) by miRNAs. We further validated our hypothesis, that genes belonging to the same family tend to share the same regulation. The statistical test checked the greater success of assigning miRNA regulation according to the mode of each TreeFam family, for all proteins belonging to groups with at least 2 members. A one-sided ("greater than") binomial test yielded a p-value of 1e-61.

3.11. GPCRs are rarely regulated by miRNAs

Most GPCRs are not regulated by miRNAs. We investigated whether this pattern holds true for all of the major GPCR families (470 proteins, 6 classes). In addition, we tested whether close paralogs of the GPCR within each class are coherent in their regulation mode of miRNA. **Figure 5C** shows the dendrogram of the Secretin subfamily with functional partition to receptor types (colored coded). We confirmed that there is no direct relation between the functional relationships and the regulation mode (e.g., CRFR1/2, VIPR1/2). The same phenomenon applied to Class T (Taste 2), where 7 out of 25 (28%) are known to be regulated (**Figure 5D**). We conclude that the selective pressure to maintain the same regulation is weakened in duplicated genes, allowing for innovation and accelerated evolution that ultimately leads to a divergence in regulation.
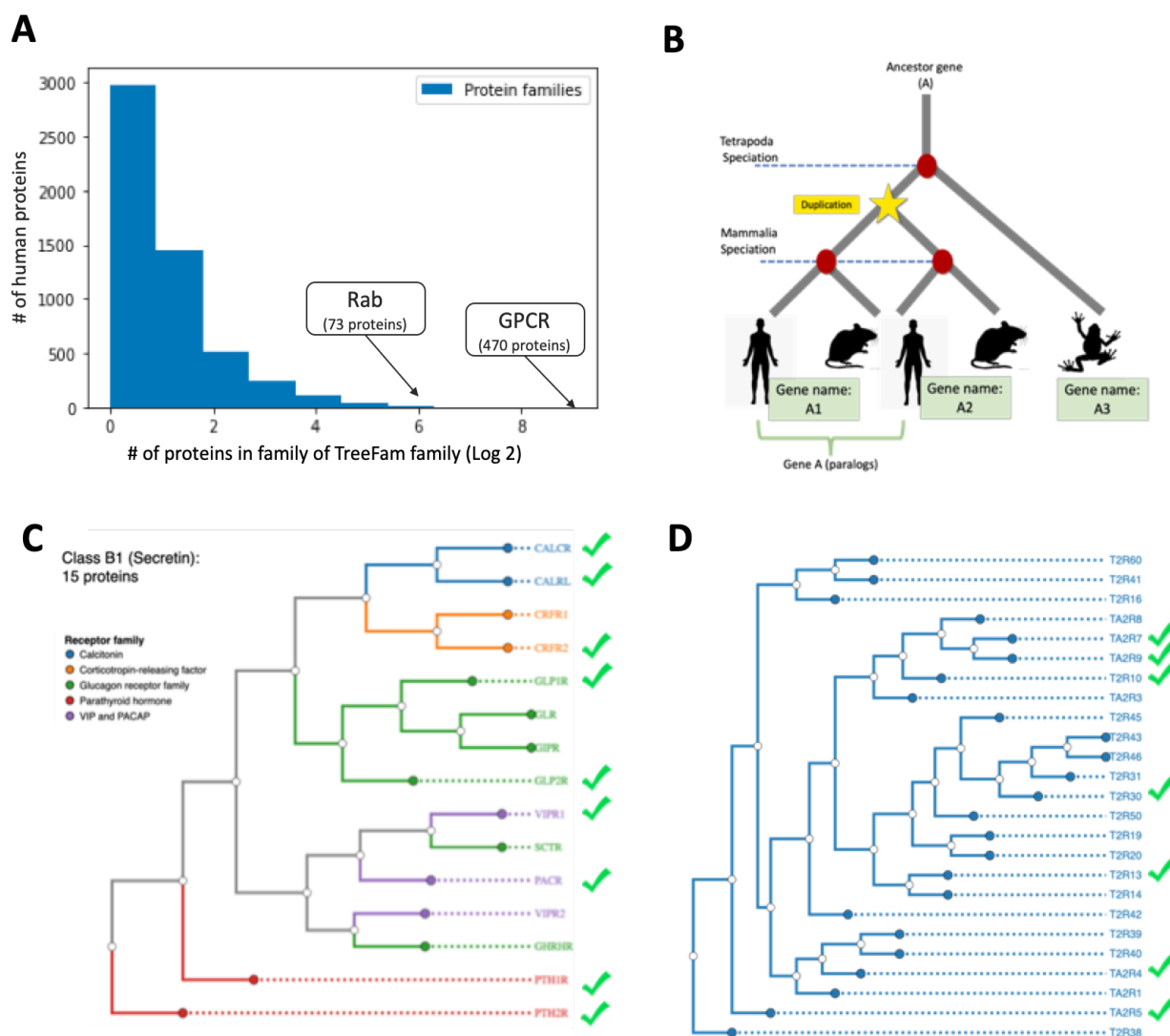
13

FIGURE 5. miRNA regulation of human paralogs. **(A)** The count of protein family members (log2 scale) with the number of proteins belonging to a family (x-axis). The largest family with 470 members of G-protein coupled receptor (GPCR) is marked. **(B)** Schematic view of homologs and paralog of ancestor gene A. The duplication event prior to specialization defines paralogs. **(C)** Phylograms of the GPCR Class B1 (Secretin) with 15 gene paralogs with 9 regulated by miRNAs. The receptor groups are color coded (e.g., Glucagon receptor). **(D)** Members of GPCR of Class T (Taste 2) with 25 proteins, among them only 7 proteins were predicted as miRNA regulated (marked in green symbol) the rest of the proteins were predicted as not under miRNA regulation. The source of proteins in C and D and their annotation is according to GPCRdb (Pandy-Szekeres et al., 2022). The trees are generated from whole sequence phylogenetic trees within each GPCR class using unweighted pair group method with arithmetic mean (UPGMA, 10 replicates).

4. Discussion

In this study, we address the question of miRNA regulation as a binary problem of prediction without considering the binding capacity, sequence specificity of each of the miRNA individually or the nature of the regulation. The information we use is mainly derived from the protein sequence and its associated annotations. For example, we showed that proteins located in the membrane or the nucleus have a higher tendency to be regulated by miRNAs (Supplemental **Table S1**).

14

Experimental evidence on miRNA regulation is scarce and fragmented. Even in rats the number of stem-loop miRNA genes is only 40% of the number reported in mice, with only 323 miRNAs marked as high confidence. In this study, we provided a machine learning model that accurately predicts validated miRNA regulation in novel (to the model) organisms with minimal experimental results, without requiring known miRNA genes seeds (**Table 4).** We propose to use sequenced genomes to determine the proteome and elementary genomics properties for poorly studied organisms. The capacity of a trained model to transfer successfully across organisms is a key feature in building universal models, capable of covering all domains of life, as demonstrated in proteins language models (Ofer and Linial, 2015; Ofer et al., 2021). In addition, we anticipate that training on proteins from multiple proteomes will improve not only prediction of miRNA regulation but also related tasks (Brandes et al., 2022).

Machine learning approaches were applied in the field of miRNAs for the prediction of miRNAs from genomic information, miRNA targets of both (Singh et al., 2017; Parveen et al., 2019). All these methods are based on the properties of the molecular complementarity of miRNA and miRNA binding sites. In contrast, the strength of the automated machine learning AI model is the extensive exploration of extremely high number of features (each of our models explored millions of features) from diverse sources, both validating previous discoveries and yielding potential novel insights. We tested the predictive model on unseen data (rather than on repeated sampling of the training set). The use of an entropy-based criterion (RIG, see Methods) highlighted informative features with high discriminatory power, stability and coverage. Discovery of novel sets of features that are not necessarily explainable by current knowledge, we expect to extract understanding on the biology. For example, we observed that the length of 5'-UTR, along with the length of the gene's CDS were quite informative. Several examples confirmed that gene activation by miRNAs include binding to the 5′-UTR as shown for ribosomal proteins translation during amino acid starvation (Da Sacco and Masotti, 2012).

Another feature that contributed to the performance of the model concerns the number of alternative variants at the tail of a gene (Müller et al., 2014). In our model, the combined features of alternative polyadenylation (APA) and tissue specificity (Table 3) reflect the importance of post transcriptional regulation of the 3'-UTR as over 50% of conserved miRNAs target sites reside downstream of the proximal polyadenylation site in mammalian genes (Ren et al., 2020). It was shown that many 3'-UTR APA variants are associated with genes expressed in specific tissues and conditions (Yang et al., 2022).

We showed that a unified model, combining proteomic and genomic modalities outperformed other models (**Figure 4**). In future work, we hope to experiment with models using data from multiple species simultaneously. An additional benefit of proteome-based prediction in predicting miRNA in organisms lacking experimental data is with disease related orthologs. Among the top predictions in zebrafish are genes associated with human diseases including Parkinson's disease (park7), cancer (apc, a known tumor supresion gene) and kidney failure (pkd2). The shared regulation among protein families allows investigating human diseases through miRNA regulation orthologs in simpler model organisms (Chang and Mendell, 2007).

In humans, GPCRs are the largest membranous family and represent ancient duplications and further diversification. The composition of GPCR in the plasma membrane of cells are tightly regulated in health and disease. Over 400 human GPCRs (excluding hundreds that are involved in olfaction) are divided into 6 functional classes that are responsible for sensing smell, taste, pain, mechanical stress, vision, but also aspects of adhesion and differentiation (Pándy-Szekeres et al., 2018). We showed that GPCRs are underregulated by miRNA. Protein families such as histones are not regulated by miRNAs and their 3'-UTR is extremely short. Members of the GPCR proteins have 3'-UTR with an average

15

length of 400-800 nucleotides, but still are mostly non-regulated by miRNAs. GPCRs act in almost every aspect of signal transduction and there are many levels of regulation that tune their activity including quantity, localization in the plasma membrane, recycling and endocytosis. We propose that translation regulation and attenuation of GPCR transcript stability by miRNAs do not contribute to the regulation of the GPCR superfamily.

We further tested the sensitivity of our proteome-based models in alternative problem formulations. About 5% of the human coding genes are marked as protein receptors, including many of the GPCR family members. We therefore trained a model only on human receptor proteins (a total of 998 proteins). Only 59% of these genes are known to be miRNA regulated (compared to 76% for the entire human proteome) The model reached a rocAUC of 84%, supporting the ability of the model to generalize, rather than merely predicting GPCRs as being unregulated. A top feature in this subpopulation was being involved in "olfaction", covering 23% of the receptors.

Previous studies that sought shared properties among the miRNA regulated targets proposed that such targets are enriched in protein-protein interactions. However, large protein complexes are mostly excluded from miRNA regulation {Das, 2013 # 33). Inspecting the top features from the results of our unified model show that many of the miRNA regulated proteins participate in signal transduction, post translated modification, nucleic acid binding proteins (e.g., transcription factors) and cellular trafficking (e.g., small GTPase) (**Figure 3**). We also found that nucleosomes, ribosomes, and other stable complexes are not likely to be regulated by miRNAs.

While protein functions are under purifying selection, miRNA binding sites at the 3'-UTR are fast evolving. It has been estimated that duplicated genes in humans are twice as likely to be miRNA targets. Moreover, paralogs on average have longer 3'-UTR relative to singletons {Li, 2008 #34} and the breath of regulation is greater among paralogs. Moreover, among duplicated genes that are within the same 3D topological associated domain (TAD), the coordinated expression is lower than the average non-related genes within TADs (Ibn-Salem et al., 2017). We show that protein families tend to display coherent behavior with respect to miRNA regulation, but this coherence is not visible at the level of pairs of paralogs (**Figure 5**).

A number of works have proposed miRNA regulation as a novel and very recent evolutionary innovation: while this could explain the differences in amount of miRNA regulated genes species, it fails at explaining the disparity between experimental and computational predictions. A more parsimonious explanation is that experimental validation is lacking, and that additional, more stable computational methods, that can also generalize across taxa, are needed to prioritize targets.

The role of non-coding RNA regulation to maintain cellular homeostasis applies in all organisms. We expect that a similar AI-based approach will be useful for creating a generalized model for post transcriptional regulation in living cells. Such an integrative model will use the features from miRNA regulation and extended to statistical features engineered and extracted from other long non-coding RNAs (lncRNAs), circular RNAs (circRNAs) and pseudogenes.

Abbreviations

AI, artificial intelligence; CLASH, cross-linking ligation and sequencing of hybrids; CLIP, cross-linking and immunoprecipitation; GPCR, G-protein coupled receptor; miRNA, microRNA; ML, machine learning; ncRNA, noncoding RNA, rocAUC, area under the receiver operating characteristic curve; TAD, topological associated domain; UPGMA, unweighted pair group method with arithmetic mean; UTR, untranslated region.

Supplementary Material

Supplementary **Figure S1**. TargetScanFish view of miRNA regulated prediction for disease related genes. Supplementary **Table S1,** Proteomes of model organisms ML predictions. Supplementary **Table S2,** Summary statistics of miRNA regulation by organism**;** Supplementary **Table S3,** Top 15 features from protein annotations; Supplementary **Table S4,** Protein family miRNA prediction; Supplementary **Table S5,** Samples of zebrafish miRNA regulation prediction.

5. References

Alles, J., Fehlmann, T., Fischer, U., Backes, C., Galata, V., Minet, M., et al. (2019). An estimate of the total number of true human miRNAs. Nucleic Acids Res 47(7), 3353-3364. doi: 10.1093/nar/gkz097.

Balaga, O., Friedman, Y., and Linial, M. (2012). Toward a combinatorial nature of microRNA regulation in human cells. Nucleic acids research 40(19), 9404-9416.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). ProteinBERT: A universal deep-learning model of protein sequence and function. Bioinformatics. doi: 10.1093/bioinformatics/btac020.

Breuza, L., Poux, S., Estreicher, A., Famiglietti, M.L., Magrane, M., Tognolli, M., et al. (2016). The UniProtKB guide to the human proteome. Database 2016.

Chang, T.-C., and Mendell, J.T. (2007). microRNAs in vertebrate physiology and human disease. Annu. Rev. Genomics Hum. Genet. 8, 215-239.

Cohen, S., Dagan, N., Cohen-Inger, N., Ofer, D., and Rokach, L. (2021). ICU Survival Prediction Incorporating Test-Time Augmentation to Improve the Accuracy of Ensemble-Based Models. IEEE Access 9, 91584-91592.

Da Sacco, L., and Masotti, A. (2012). Recent insights and novel bioinformatics tools to understand the role of microRNAs binding to 5'untranslated region. International journal of molecular sciences 14(1), 480-495.

Friedman, R.C., Farh, K.K.-H., Burge, C.B., and Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome research 19(1), 92-105.

Huang, H.Y., Lin, Y.C., Li, J., Huang, K.Y., Shrestha, S., Hong, H.C., et al. (2020). miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. Nucleic Acids Res 48(D1), D148-D154. doi: 10.1093/nar/gkz896.

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2019). HMDD v3.0: a database for experimentally supported human microRNA-disease associations. Nucleic Acids Res 47(D1), D1013-D1017. doi: 10.1093/nar/gky1010.

Ibn-Salem, J., Muro, E.M., and Andrade-Navarro, M.A. (2017). Co-regulation of paralog genes in the three-dimensional chromatin architecture. Nucleic Acids Res 45(1), 81-91. doi: 10.1093/nar/gkw813.

Karagkouni, D., Paraskevopoulou, M.D., Chatzopoulos, S., Vlachos, I.S., Tastsoglou, S., Kanellos, I., et al. (2018). DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. Nucleic Acids Res 46(D1), D239-D245. doi: 10.1093/nar/gkx1141.

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. Nucleic acids research 47(D1), D155-D162.

Lai, X., Wolkenhauer, O., and Vera, J. (2016). Understanding microRNA-mediated gene regulatory networks through mathematical modelling. Nucleic Acids Res 44(13), 6019-6035. doi: 10.1093/nar/gkw550.

Leung, A.K., and Sharp, P.A. (2010). MicroRNA functions in stress responses. Mol Cell 40(2), 205-215. doi: 10.1016/j.molcel.2010.09.027.

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2014). starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. Nucleic acids research 42(D1), D92-D97.

McGeary, S.E., Lin, K.S., Shi, C.Y., Pham, T.M., Bisaria, N., Kelley, G.M., et al. (2019). The biochemical basis of microRNA targeting efficacy. Science 366(6472). doi: 10.1126/science.aav1741.

Min, H., and Yoon, S. (2010). Got target? Computational methods for microRNA target prediction and their extension. Exp Mol Med 42(4), 233-244. doi: 10.3858/emm.2010.42.4.032.

Müller, S., Rycak, L., Afonso-Grunz, F., Winter, P., Zawada, A.M., Damrath, E., et al. (2014). APADB: a database for alternative polyadenylation and microRNA regulation events. Database 2014.

O'Brien, J., Hayder, H., Zayed, Y., and Peng, C. (2018). Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Front Endocrinol (Lausanne) 9, 402. doi: 10.3389/fendo.2018.00402.

Ofer, D., Brandes, N., and Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. Comput Struct Biotechnol J 19, 1750-1758. doi: 10.1016/j.csbj.2021.03.022.

Ofer, D., and Linial, M. (2015). ProFET: Feature engineering captures high-level protein functions. Bioinformatics 31(21), 3429-3436.

Pandy-Szekeres, G., Esguerra, M., Hauser, A.S., Caroli, J., Munk, C., Pilger, S., et al. (2022). The G protein database, GproteinDb. Nucleic Acids Res 50(D1), D518-D525. doi: 10.1093/nar/gkab852.

Pándy-Szekeres, G., Munk, C., Tsonkov, T.M., Mordalski, S., Harpsøe, K., Hauser, A.S., et al. (2018). GPCRdb in 2018: adding GPCR structure models and ligands. Nucleic acids research 46(D1), D440-D446.

Parveen, A., Mustafa, S.H., Yadav, P., and Kumar, A. (2019). Applications of Machine Learning in miRNA Discovery and Target Prediction. Curr Genomics 20(8), 537-544. doi: 10.2174/1389202921666200106111813.

Quillet, A., Saad, C., Ferry, G., Anouar, Y., Vergne, N., Lecroq, T., et al. (2019). Improving Bioinformatics Prediction of microRNA Targets by Ranks Aggregation. Front Genet 10, 1330. doi: 10.3389/fgene.2019.01330.

Raschka, S., and Mirjalili, V. (2019). Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing Ltd.

Rasnic, R., Linial, N., and Linial, M. (2017). Enhancing identification of cancer types via lowly-expressed microRNAs. Nucleic Acids Res 45(9), 5048-5060. doi: 10.1093/nar/gkx210.

Ren, F., Zhang, N., Zhang, L., Miller, E., and Pu, J.J. (2020). Alternative Polyadenylation: A new frontier in post transcriptional regulation. Biomarker Research 8(1), 1-10.

Riffo-Campos, Á.L., Riquelme, I., and Brebi-Mieville, P. (2016). Tools for sequence-based miRNA target prediction: what to choose? International journal of molecular sciences 17(12), 1987.

Romero-Cordoba, S.L., Salido-Guadarrama, I., Rodriguez-Dorantes, M., and Hidalgo-Miranda, A. (2014). miRNA biogenesis: biological impact in the development of cancer. Cancer biology & therapy 15(11), 1444-1455.

Sayed, D., and Abdellatif, M. (2011). MicroRNAs in development and disease. Physiological reviews 91(3), 827-887.

Schreiber, F., Patricio, M., Muffato, M., Pignatelli, M., and Bateman, A. (2014). TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic acids research 42(D1), D922-D925.

Sethupathy, P., Megraw, M., and Hatzigeorgiou, A.G. (2006). A guide through present computational approaches for the identification of mammalian microRNA targets. Nat Methods 3(11), 881-886. doi: 10.1038/nmeth954.

Singh, S., Benton, R.G., Singh, A., and Singh, A. (2017). Machine learning techniques in exploring microRNA gene discovery, targets, and functions. Bioinformatics in MicroRNA Research, 211-224.

Valencia-Sanchez, M.A., Liu, J., Hannon, G.J., and Parker, R. (2006). Control of translation and mRNA degradation by miRNAs and siRNAs. Genes & development 20(5), 515-524.

VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc.".

Vishnoi, A., and Rani, S. (2017). MiRNA biogenesis and regulation of diseases: an overview. MicroRNA Profiling, 1-10.

Yang, X., Tong, Y., Liu, G., Yuan, J., and Yang, Y. (2022). scAPAatlas: an atlas of alternative polyadenylation across cell types in human and mouse. Nucleic Acids Res 50(D1), D356-D364. doi: 10.1093/nar/gkab917.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., et al. (2016). Ensembl 2016. Nucleic acids research 44(D1), D710-D716.

Yue, D., Liu, H., and Huang, Y. (2009). Survey of computational algorithms for microRNA target prediction. Current genomics 10(7), 478-492.

Zhan, S., and Lukens, L. (2010). Identification of novel miRNAs and miRNA dependent developmental shifts of gene expression in Arabidopsis thaliana. PLoS One 5(4), e10157. doi: 10.1371/journal.pone.0010157.