Ben Sarid 201445830, Yuval Weizman 318540721, Osher Paz 208298976

Introduction to Bioinformatics and Genomics
Assignment 1 - Report

Part 1 – Sequence alignment and RNA structure

1.1 – Sequence alignment

1.1.1

S_1 = CACAGCATTT

S_2 = TACGAGGAGT

1.1.2

Global alignment DP table:

		T	A	C	G	A	G	G	A	G	T
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
C	-1	-1	-2	-1	-2	-3	-4	-5	-6	-7	-8
A	-2	-2	0	-1	-2	-1	-2	-3	-4	-5	-6
C	-3	-3	-1	1	0	-1	-2	-3	-4	-5	-6
A	-4	-4	-2	0	0	1	0	-1	-2	-3	-4
G	-5	-5	-3	-1	1	0	2	1	0	-1	-2
C	-6	-6	-4	-2	0	0	1	1	0	-1	-2
A	-7	-7	-5	-3	-1	1	0	0	2	1	0
T	-8	-6	-6	-4	-2	0	0	-1	1	1	2
T	-9	-7	-7	-5	-3	-1	-1	-1	0	0	2
T	-10	-8	-8	-6	-4	-2	-2	-2	-1	-1	1

Optimal global alignment score: 1

Alignment #1:

Aligned S1: CAC-AGCATTT

Aligned S2: TACGAGGA-GT

Ben Sarid 201445830, Yuval Weizman 318540721, Osher Paz 208298976

Traceback path (i,j) from start to end:

[(1, 1), (2, 2), (3, 3), (3, 4), (4, 5), (5, 6), (6, 7), (7, 8), (8, 8), (9, 9), (10, 10)]

Alignment #2:

Aligned S1: CAC-AGCATTT

Aligned S2: TACGAGGAG-T

Traceback path (i,j) from start to end:

[(1, 1), (2, 2), (3, 3), (3, 4), (4, 5), (5, 6), (6, 7), (7, 8), (8, 9), (9, 9), (10, 10)]

1.1.3

End-space-free alignment DP table:

		T	A	C	G	A	G	G	A	G	T
-	0	0	0	0	0	0	0	0	0	0	0
C	0	-1	-1	1	0	-1	-1	-1	-1	-1	-1
A	0	-1	0	0	0	1	0	-1	0	-1	-2
C	0	-1	-1	1	0	0	0	-1	-1	-1	-2
A	0	-1	0	0	0	1	0	-1	0	-1	-2
G	0	-1	-1	-1	1	0	2	1	0	1	0
C	0	-1	-2	0	0	0	1	1	0	0	0
A	0	-1	0	-1	-1	1	0	0	2	1	0
T	0	1	0	-1	-2	0	0	-1	1	1	2
T	0	1	0	-1	-2	-1	-1	-1	0	0	2
T	0	1	0	-1	-2	-2	-2	-2	-1	-1	1

Optimal end-space-free alignment score: 2

Best cell (i,j): (5, 6)

Final end-space-free alignment:

Aligned S1: CAC-AG

Aligned S2: TACGAG

Ben Sarid 201445830, Yuval Weizman 318540721, Osher Paz 208298976

Traceback path (i,j) from start to end:

[(1, 1), (2, 2), (3, 3), (3, 4), (4, 5), (5, 6)]

1.2 – Folding algorithms

1.2.1 Nussinov RNA secondary structure prediction algorithm:

(1) Generated RNA Sequence (S):

ACAUGGAUUGAACUUGCGCC

Length: 20
Counts: A=5, C=5, G=5, U=5

(2)

2(i). Full DP Table

	A	C	A	U	G	G	A	U	U	G	A	A	C	U	U	G	C	G	C	C
A	0	0	0	1	2	2	2	3	4	4	4	4	5	6	6	6	7	7	8	9
C	0	0	0	1	2	2	2	3	3	3	4	4	5	5	5	5	6	6	7	8
A	0	0	0	1	1	1	1	2	2	2	3	3	4	4	4	4	5	5	6	7
U	0	0	0	0	0	0	1	1	1	1	2	3	3	3	3	4	4	5	5	6
G	0	0	0	0	0	0	0	1	1	1	2	2	3	3	3	3	4	4	5	6
G	0	0	0	0	0	0	0	1	1	1	2	2	3	3	3	3	4	4	5	6
A	0	0	0	0	0	0	0	1	1	1	2	2	2	3	3	3	4	4	5	6
U	0	0	0	0	0	0	0	0	0	0	1	2	2	2	2	3	3	4	4	5
U	0	0	0	0	0	0	0	0	0	0	1	1	1	2	2	2	3	3	4	5
G	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	2	3	3	4	5
A	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	3	4	4
A	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	2	2	3	3
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	2	2
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	2
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	2
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	2
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

2(ii). Traceback Path

Cell (0,19) -> Bifurcation split at 8
Cell (0,8) -> (1,7) : Match A-U
Cell (1,7) -> Bifurcation split at 4
Cell (1,4) -> (2,3) : Match C-G
Cell (2,3) -> (3,2) : Match A-U
Cell (5,7) -> (6,7) : i (G) is unpaired
Cell (6,7) -> (7,6) : Match A-U
Cell (9,19) -> (10,18) : Match G-C
Cell (10,18) -> Bifurcation split at 14
Cell (10,14) -> (11,13) : Match A-U
Cell (11,13) -> (12,12) : Match A-U
Cell (15,18) -> (16,17) : Match G-C
Cell (16,17) -> (17,16) : Match C-G

Ben Sarid 201445830, Yuval Weizman 318540721, Osher Paz 208298976

2(iii). Predicted Structure
Dot-Bracket Representation:

ACAUGGAUUGAACUUGCGCC

((()).())(((.))(()))

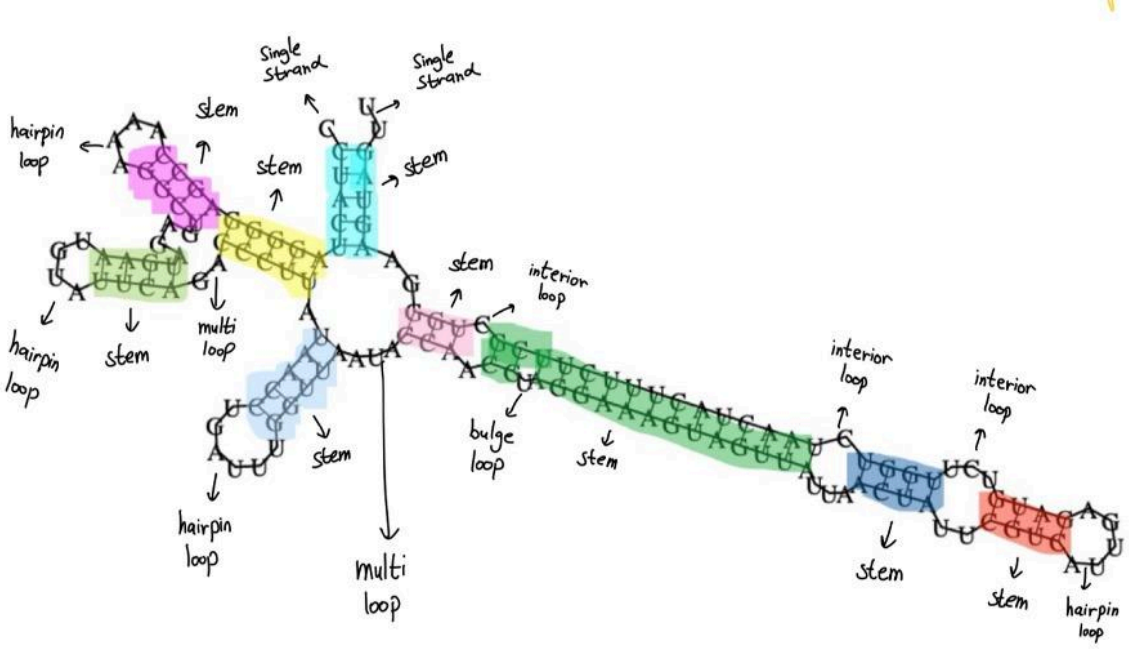
Visual Marking of Base Pairs:

Idx	Base	Pair
0	A	Pairs with index 8 (U)
1	C	Pairs with index 4 (G)
2	A	Pairs with index 3 (U)
3	U	Pairs with index 2 (A)
4	G	Pairs with index 1 (C)
5	G	-
6	A	Pairs with index 7 (U)
7	U	Pairs with index 6 (A)
8	U	Pairs with index 0 (A)
9	G	Pairs with index 19 (C)
10	A	Pairs with index 14 (U)
11	A	Pairs with index 13 (U)
12	C	-
13	U	Pairs with index 11 (A)
14	U	Pairs with index 10 (A)
15	G	Pairs with index 18 (C)
16	C	Pairs with index 17 (G)
17	G	Pairs with index 16 (C)
18	C	Pairs with index 15 (G)
19	C	Pairs with index 9 (G)

1.2.2 Vienna RNAfold application:

The minimum free energy of **-43.60 kcal/mol**

The dot-bracket representation:



1.3 – Database search & Genome Browse

1.3.1 Retrieving general information about *lin-41*

(i) Genomic location, gene function, and number of transcripts

Ben Sarid 201445830, Yuval Weizman 318540721, Osher Paz 208298976

Gene: *lin-41*
Species: *Caenorhabditis elegans* (strain Bristol N2)
Gene IDs: WormBase WBGene00003026, locus tag CELE_C12C8.3 (NCBI)

- Genomic location

According to NCBI Gene, ***lin-41* is located on chromosome I, positions 9,334,800–9,342,496** on the complement (reverse) strand of the reference sequence NC_003279.8. The gene spans 16 exons in this genomic region. (NCBI)

- Gene function (summary)

lin-41 encodes a TRIM-NHL RNA-binding protein that functions primarily as a translation repressor. It:

  - enables mRNA binding and translation repressor activity,
  - post-transcriptionally silences mRNAs by either inhibiting translation or promoting mRNA degradation,
  - participates in heterochronic regulation of development, including nematode male tail tip morphogenesis, oocyte development, and positive regulation of brood size,
  - is localized mainly in the cytoplasm and ribonucleoprotein granules. (NCBI)
- Number of transcripts

Different databases annotate slightly different numbers of isoforms.

  - NCBI RefSeq lists 2 reviewed mRNA transcripts for *lin-41*: NM_001025827.7 and NM_001025828.7. (NCBI)
  - Ensembl/WormBase (which are the primary genome resources for *C. elegans*) annotate 4 splice variants for this gene. (ensembl.org)

In this report I follow the **Ensembl/WormBase** annotation and describe these **4** transcripts in detail.

(ii) Transcript IDs, protein length, and number of exons

Transcript ID	Gene (WBGene)	Protein length (aa)	Exon count*
C12C8.3a.1	WBGene00003026	1147 aa	16*
C12C8.3a.2	WBGene00003026	1147 aa	16*

C12C8.3b.1	WBGene00003026	1143 aa	16*
C12C8.3b.2	WBGene00003026	1143 aa	15*

Websites / databases used

- Ensembl / WormBase (via Ensembl Genomes) – gene summary for WBGene00003026 (*lin-41*), transcript count. (ensembl.org)

1.3.2 Exploring *lin-41* on the genome browser

I examined the *C. elegans* **lin-41** gene (WBGene00003026, locus tag **C12C8.3**) on a genome browser.

Genome browser used

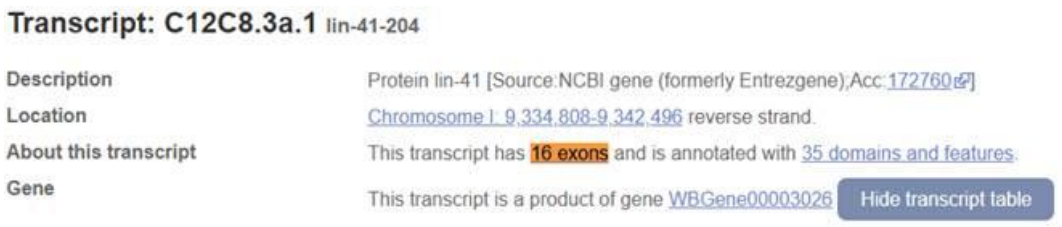
- **WormBase/Ensembl genome browser** for the *C. elegans* gene structure (exons and transcripts). (ensembl.org)
- **NCBI Gene / Genome Data Viewer** for an additional view of the genomic context and for the "Gene neighbors" list. (NCBI)

(i) Exons for each transcript

According to WormBase/Ensembl, **lin-41 (WBGene00003026)** has **4 transcripts**:

Transcript ID	Name	bp	Protein	Biotype	UniProt Match	Flags
C12C8.3a.1	lin-41-204	4796	1147aa	Protein coding	Q9U489	Ensembl Canonical APPRIS P4
C12C8.3b.1	lin-41-203	4785	1143aa	Protein coding	-	APPRIS ALT2
C12C8.3a.2	lin-41-201	4631	1147aa	Protein coding	Q9U489	APPRIS P4
C12C8.3b.2	lin-41-202	4613	1143aa	Protein coding	-	APPRIS ALT2

Exons Per Transcript:



Transcript: C12C8.3b.1 lin-41-203

Description	Protein lin-41 [Source:NCBI gene (formerly Entrezgene);Acc:172760]
Location	Chromosome I: 9,334,807-9,342,496 reverse strand.
About this transcript	This transcript has 16 exons and is annotated with 36 domains and features.
Gene	This transcript is a product of gene WBGene00003026 Hide transcript table

Transcript: C12C8.3b.2 lin-41-202

Description	Protein lin-41 [Source:NCBI gene (formerly Entrezgene);Acc:172760]
Location	Chromosome I: 9,334,806-9,341,919 reverse strand.
About this transcript	This transcript has 15 exons and is annotated with 36 domains and features.
Gene	This transcript is a product of gene WBGene00003026 Hide transcript table

Transcript: C12C8.3a.2 lin-41-201

Description	Protein lin-41 [Source:NCBI gene (formerly Entrezgene);Acc:172760]
Location	Chromosome I: 9,334,800-9,341,919 reverse strand.
About this transcript	This transcript has 15 exons and is annotated with 35 domains and features.
Gene	This transcript is a product of gene WBGene00003026 Hide transcript table

(ii) Neighboring genes

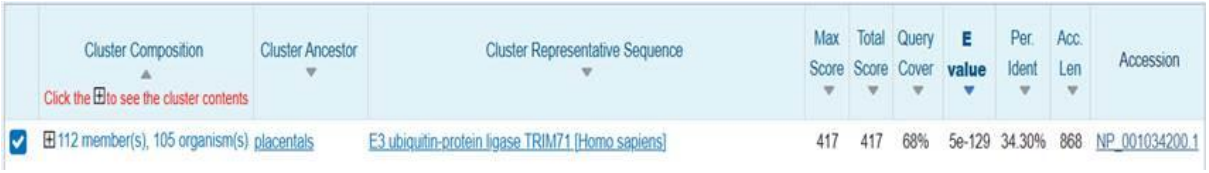
To identify neighboring genes, I used the **"Gene neighbors"** link from the NCBI lin-41 Gene page (Gene ID **172760**) and the genomic context view. (NCBI)

In the **Gene neighbors** table, the genes overlapping or immediately flanking lin-41 (C12C8.3) on chromosome I are:

Name/Gene ID	Description	Location	Aliases
C12C8.8 ID: 24104202	ncRNA [*Caenorhabditis elegans*]	Chromosome I, NC_003279.8 (9332534..9332649, complement)	CELE_
C12C8.6 ID: 13181434	ncRNA [*Caenorhabditis elegans*]	Chromosome I, NC_003279.8 (9332952..9333033, complement)	CELE_
mir-79 ID: 259856	ncRNA [*Caenorhabditis elegans*]	Chromosome I, NC_003279.8 (9332946..9333043)	CELE_C12C8.4
M04C9.2 ID: 187439	Aspartoacylase [*Caenorhabditis elegans*]	Chromosome I, NC_003279.8 (9344724..9349576, complement)	CELE_M04C9.2
M04C9.1 ID: 187438	Pyroglutamyl-peptidase 1 [*Caenorhabditis elegans*]	Chromosome I, NC_003279.8 (9350269..9352700, complement)	CELE_M04C9.1
lin-41 ID: 172760	Protein lin-41 [*Caenorhabditis elegans*]	Chromosome I, NC_003279.8 (9334800..9342496, complement)	CELE_C12C8.3

1.3.3. Identifying the human homolog of lin-41

(i)



Cluster Composition Click the ⊞ to see the cluster contents	Cluster Ancestor	Cluster Representative Sequence	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
⊞ 112 member(s), 105 organism(s)	placentals	E3 ubiquitin-protein ligase TRIM71 [Homo sapiens]	417	417	68%	5e-129	34.30%	868	NP_001034200.1

(ii)

The human homolog of *C. elegans* lin-41 is **TRIM71 (E3 ubiquitin-protein ligase TRIM71)**

Part 2 – Comparative analysis of miRNA precursors across species

2.1 Preparing and exploring the data

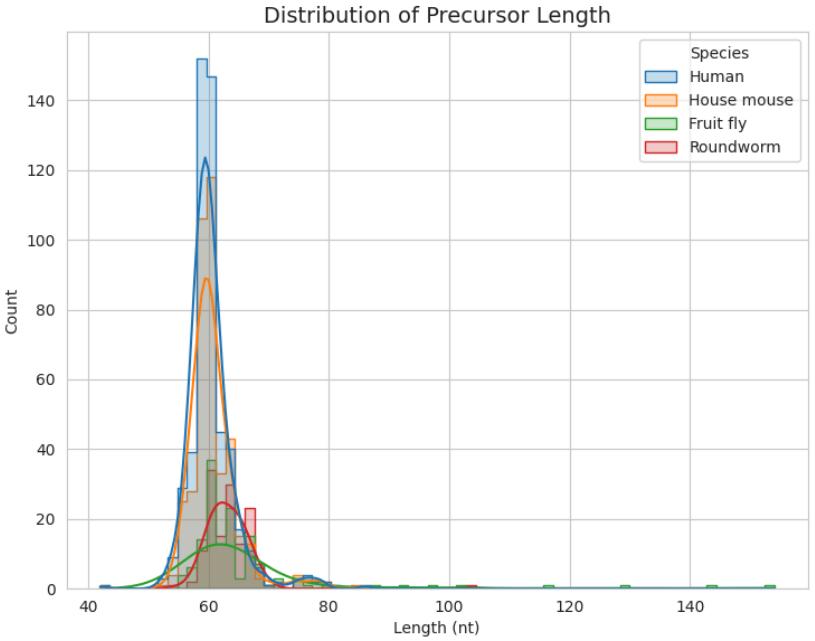
Using the miRGeneDB, we downloaded the FASTA files of precursor, mature, and star sequences for each of the four organisms: Human (Homo sapiens), House mouse (Mus musculus), Fruit fly (Drosophila melanogaster), and Roundworm (Caenorhabditis elegans). We also downloaded the GFF file and the table that appears in the browse tab for each organism.
We wrote two scripts to automate the downloads - *getfiles.py* and *gettables.py*.
With the data downloaded, we wrote a script, *mergedata.py*, to merge the files into a single excel with a sheet for each organism.

- MirGeneDB ID, MiRBase ID, Family, Seed, Chromosome, Start, End, and Strand are taken from the browse tab. Specifically, chromosome, start, end, and strand are verified against the GFF files that also include this data.
- Precursor, mature, and star FASTA sequences are taken from their respective FASTA file. We checked for version suffixes like “-v” or “-v1” to make sure we don’t miss a match.
- In situations where the mature sequence contains two options, both 3p and 5p, and the star sequence is empty, we choose the sequence that includes the seed as the mature sequence, and assign the other as the star one.
- Mature location (5p / 3p) is extracted from the mature FASTA.

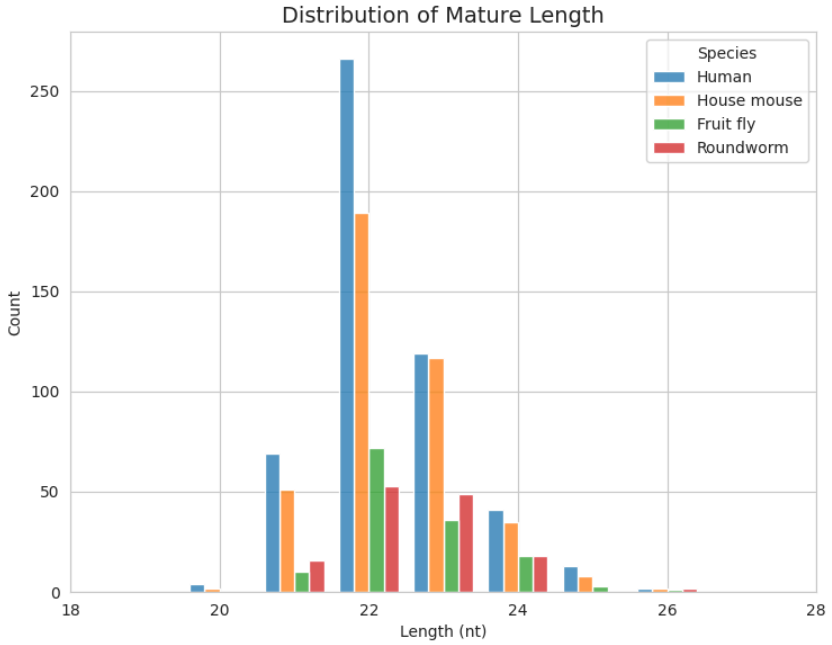
- Length for precursor, mature and star sequences is calculated using a character count of the sequence.

The jupyter notebook *2.1-initialstatistics.ipynb* is used to generate graphs and visualizations regarding the data:

Distribution of precursor length

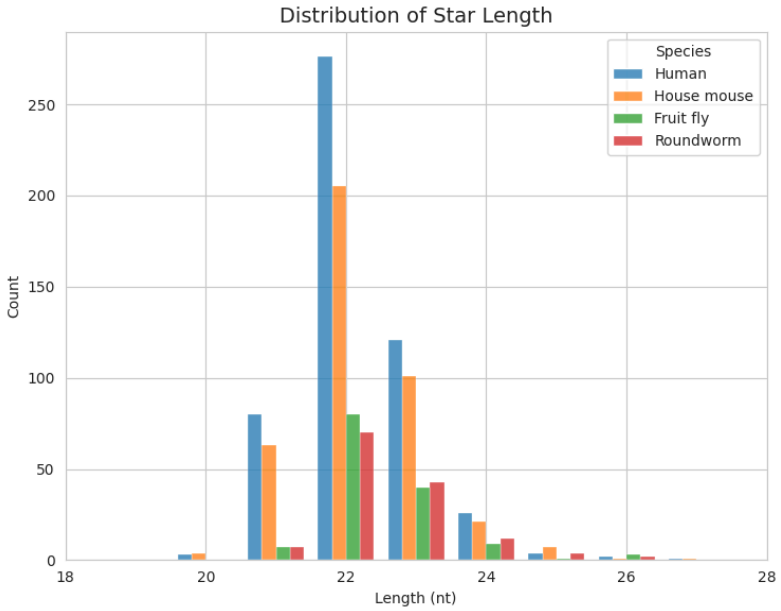


Distribution of mature length

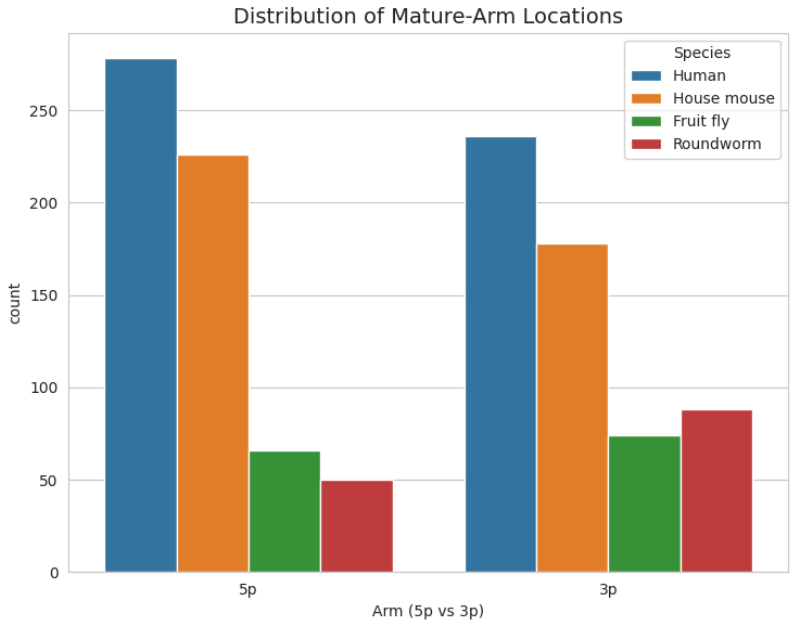


Distribution of star length

Ben Sarid 201445830, Yuval Weizman 318540721, Osher Paz 208298976

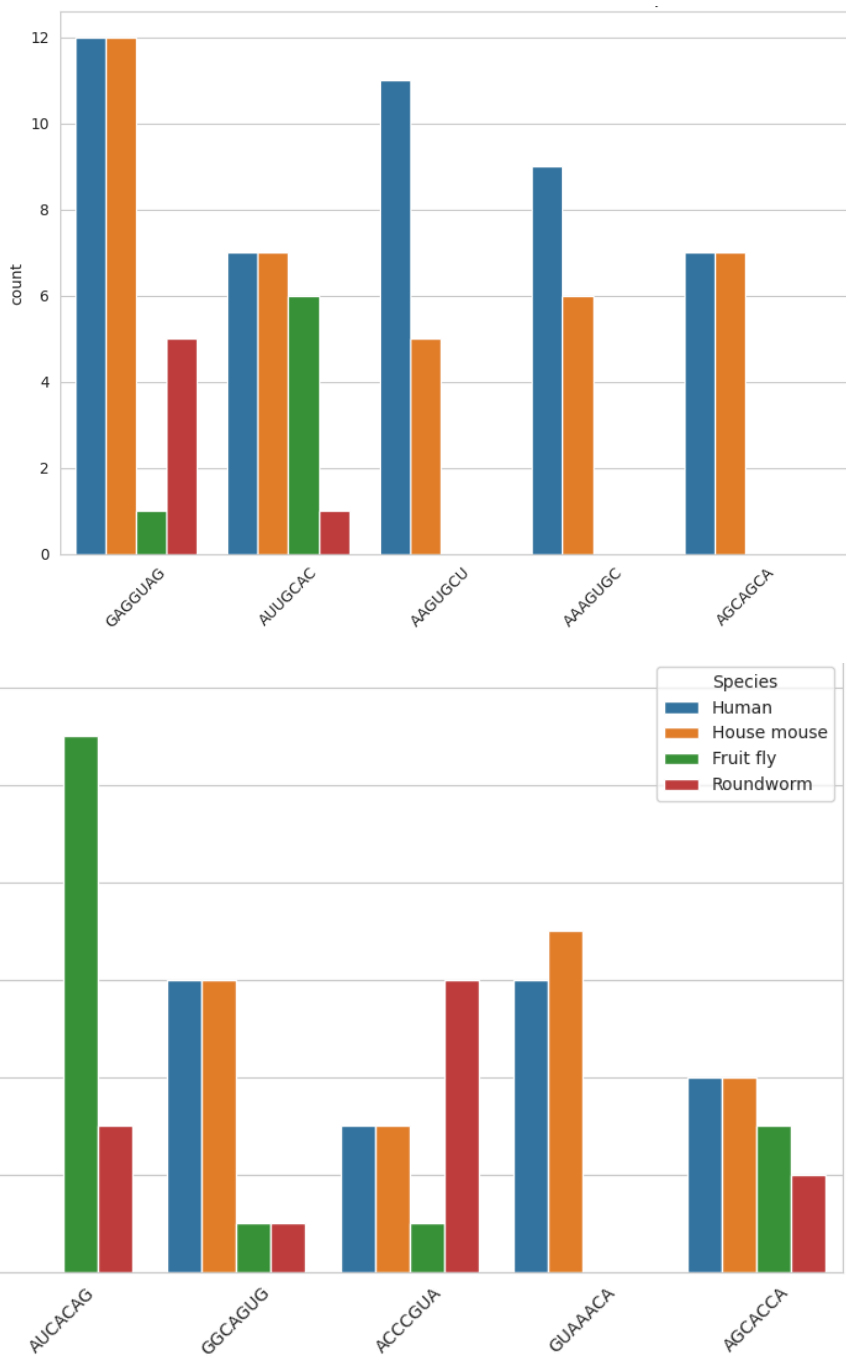


Distribution of mature-arm locations (5p vs. 3p)



Ten most common seed families

Ben Sarid 201445830, Yuval Weizman 318540721, Osher Paz 208298976



We also extracted summary statistics by species:

Species	# of miRNA genes
Fruit fly	140
House mouse	404
Human	514
Roundworm	138

	Precursor length		Mature length		Star length	
Species	**mean**	**std**	**mean**	**std**	**mean**	**std**
Fruit fly	65.92142	14.17012	22.53571	0.93231	22.47142	0.88507
House mouse	60.69059	4.145710	22.40594	0.93647	22.19802	1.44042
Human	60.36770	3.993539	22.33268	0.94704	22.17509	1.30842
Roundworm	63.28260	4.468853	22.55797	0.95915	22.57971	0.93443

Mature arm location counts:

	Mature location	
Species	**3p**	**5p**
Fruit fly	74 (52.9%)	66 (47.1%)
House mouse	178 (44.1%)	226 (55.9%)
Human	236 (45.9%)	278 (54.1%)
Roundworm	88 (63.8%)	50 (36.2%)

Comparing the data, we can see that

- The mammals, human and house mouse, have much more miRNA genes (514 and 404, respectively) compared to the roundworm (138) and fruit fly (140), on par with our expectations regarding their biological complexity.
- The means of the mature length and the star length are very similar for all four organisms.
- Length distributions for all four species are somewhat similar in shape, with the count varying with proportion to the number of miRNA genes each species has.
- Star length std for our two mammals, human and house mouse, is almost doubled than that of the roundworm and fruit fly.
- Mean of the precursor length is highest for fruit fly, lowest for human and house mouse.
- Mature location of 3p is over 50% for fruit fly and roundworm, but under 50% for human and house mouse.

- GAGGUAG is the most common seed family, with 12 appearances in human and house mouse
- In general, we see the most common seed families belong to the species human and house mouse, because of the much larger number of miRNA genes those two species have.
- Still, a very common seed family is AUCACAG, which is mostly in the fruit fly, a little in the roundworm, but not at all in human or house mouse.

2.2 Predicting RNA secondary structure

RNAfold, from the ViennaRNA package, was used to compute the secondary structure for each miRNA gene. To efficiently run it on all sequences we used a local ViennaRNA library with C-function calls, applied it using the pandas *apply* method on all rows, and extracted both the structure and the MFE at the same time.
Running the algorithm for all sequences, of all species, took 3.9s.

2.3 Extracting Structural Features

The structural features we extracted and the ways we did it were:

1. MFE (free energy) - using ViennaRNA
2. Number of base-pairs in the mature miRNA region - looking at the mature region, counting the number of (and) characters
3. Number of base-pairs in the star miRNA region - looking at the starregion, counting the number of (and) characters
4. Loop size (size of the terminal loop) - calculating the distance between the innermost base pair
5. Number of GC base-pairs across the precursor - iterating through every predicted base pair in the structure, counting how many were (G-C) pairs.
6. Number of AU base-pairs across the precursor - iterating through every predicted base pair in the structure, counting how many were (A-U) pairs.
7. Number of GU base-pairs across the precursor - iterating through every predicted base pair in the structure, counting how many were (G-U) pairs.
8. Number of unmatched nts on the mature side - extracting the mature side in the structure, counting .
9. Number of unmatched nts on the star side - extracting the star side in the structure, counting .
10. Number of unmatched nts on the precursor - looking at the entire structure, counting .
11. Maximum consecutive unmatched nts length mature side - extracting the mature side in the structure, splitting by brackets, and counting the maximum number of dots
12. Maximum consecutive unmatched nts length star side - extracting the star side in the structure, splitting by brackets, and counting the maximum number of dots

Ben Sarid 201445830, Yuval Weizman 318540721, Osher Paz 208298976

As an example, we chose the miRNA gene Hsa-Mir-374-P1 of the human. Its precursor sequence is:
UUAUAAUACAACCUGAUAAGUGUUAUAGCACUUAUCAGAUUGUAUUGUAAUU

Its strand is minus (-) and its mature location is 5p.

The mature sequence is:
UUAUAAUACAACCUGAUAAGUG

The star sequence is:
CUUAUCAGAUUGUAUUGUAAUU

The folded structure is:
((((((((((.((((((((((....)))))))))).))))))))))..

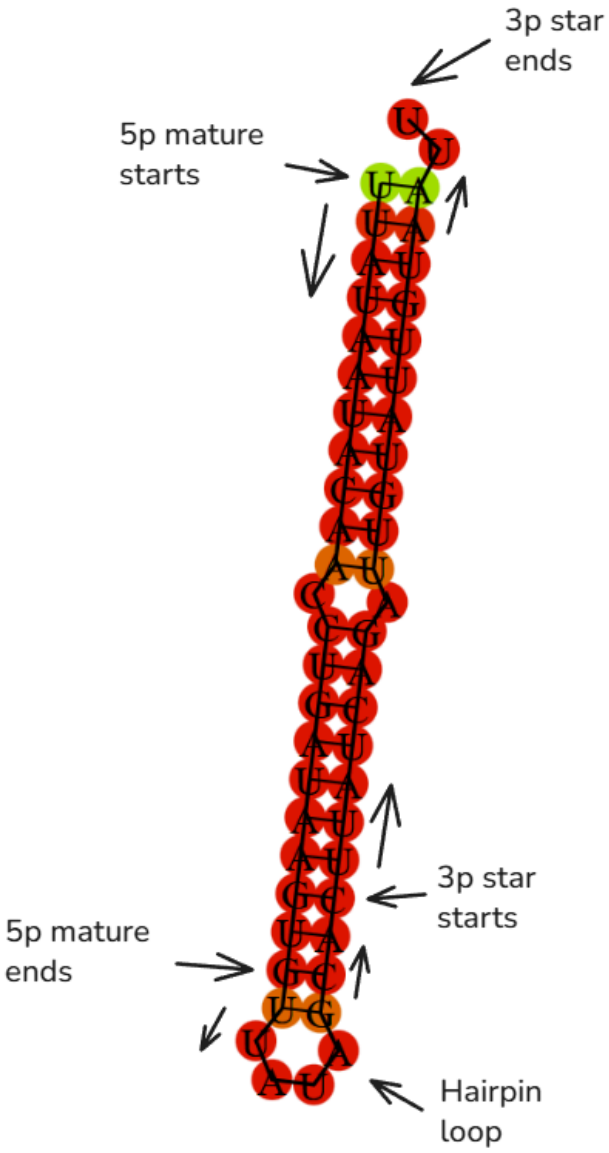
((((((((((.((((((((((....))))))))).))))))))))..

Green is 5p, the mature region. Red is 3p, the star region.

MFE	Mature_BP_Count	Star_BP_Count	Terminal_Loop_Size	GC_Pairs_Precursor	AU_Pairs_Precursor
-25.399	21	19	4	5	15

GU_Pairs_Precursor	Mature_Unmatched_Count	Star_Unmatched_Count	Precursor_Unmatched_Count	Mature_Max_Consecutive_Unmatched	Star_Max_Consecutive_Unmatched
2	1	3	8	1	2

Ben Sarid 201445830, Yuval Weizman 318540721, Osher Paz 208298976



We can see MFE is -25.399, a stable number due to our many base pairs (22).
We can see mature count base pairs is 21, and star count base pairs is 19. This is visible at the bottom of the drawing, where indeed the star region starts 2 nt after where the mature region ends in parallel.
The loop at the bottom is 4 nts.
GC, AU, GU count is countable in the main part and its two stems. 5, 15, 2 respectively.
Mature unmatched nt is 1. That's the interior loop (left part).
Star unmatched nt is 3. That's the interior loop (right part) + the single-stranded tail at the top.
Precursor unmatched nt is 8. That's the mature unmatched nt + star unmatched nt + loop.
Mature max consecutive unmatched is the interior loop, 1.
Star max consecutive unmatched is the tail at the top, 2.

Ben Sarid 201445830, Yuval Weizman 318540721, Osher Paz 208298976

2.4 Machine learning: Classifying miRNA precursors across species - Methods

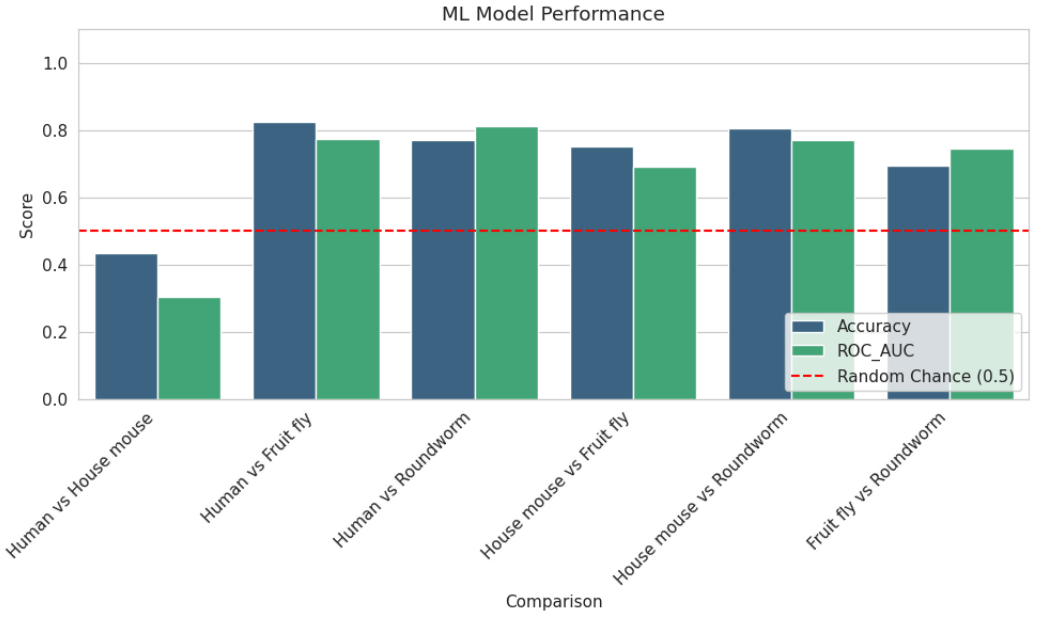
To assess whether structural features alone can differentiate miRNA precursors between species, we designed a pairwise classification experiment. We implemented a Random Forest Classifier (100 trees, random seed fixed for reproducibility) for all six pairwise species comparisons.

Design Choices:

1. **Classifier:** Random Forest was selected for its ability to model non-linear biological relationships, its robustness to outliers, and its native capability to quantify feature importance.
2. **Performance Evaluation:** For each pairwise comparison, the dataset was partitioned into a stratified 80/20 train-test split. Stratification was used to maintain the original class proportions (species ratio) in both training and testing sets, preventing bias due to differing sample sizes (Human is approximately 514, while fruit fly is 140, for example). Model performance was evaluated using Accuracy and ROC-AUC, to ensure performance regardless of class imbalance.
3. **Feature Analysis:** We used the Random Forest's built-in Gini Importance (Mean Decrease in Impurity) to calculate a relevance score for each structural feature, identifying which structural properties were most critical for distinguishing between the species.

2.5 Summary and Interpretation

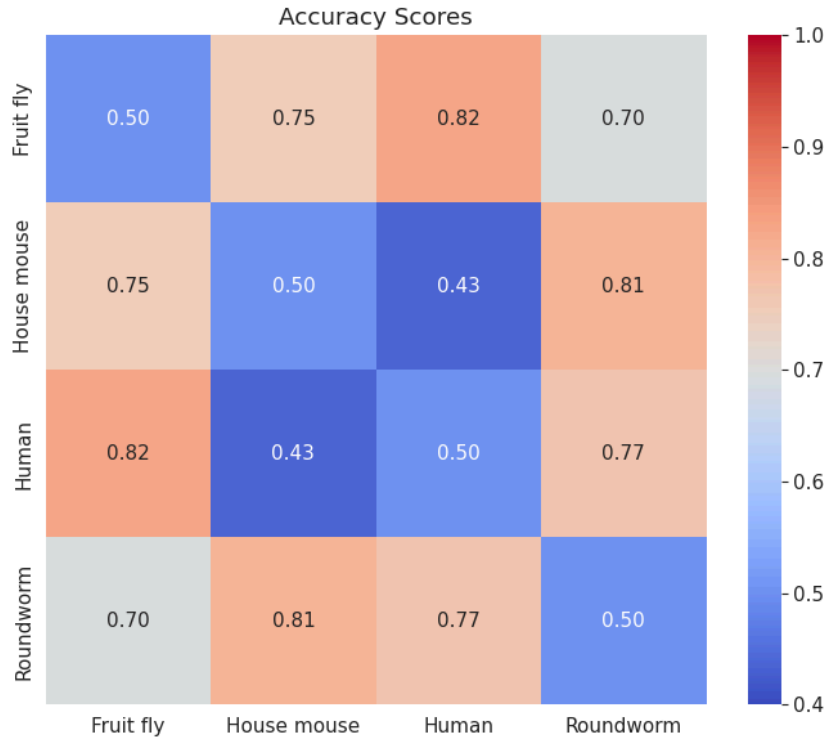
We present our results:



Human vs House mouse:
Accuracy=0.4348
ROC-AUC=0.3031
Human vs Fruit fly:

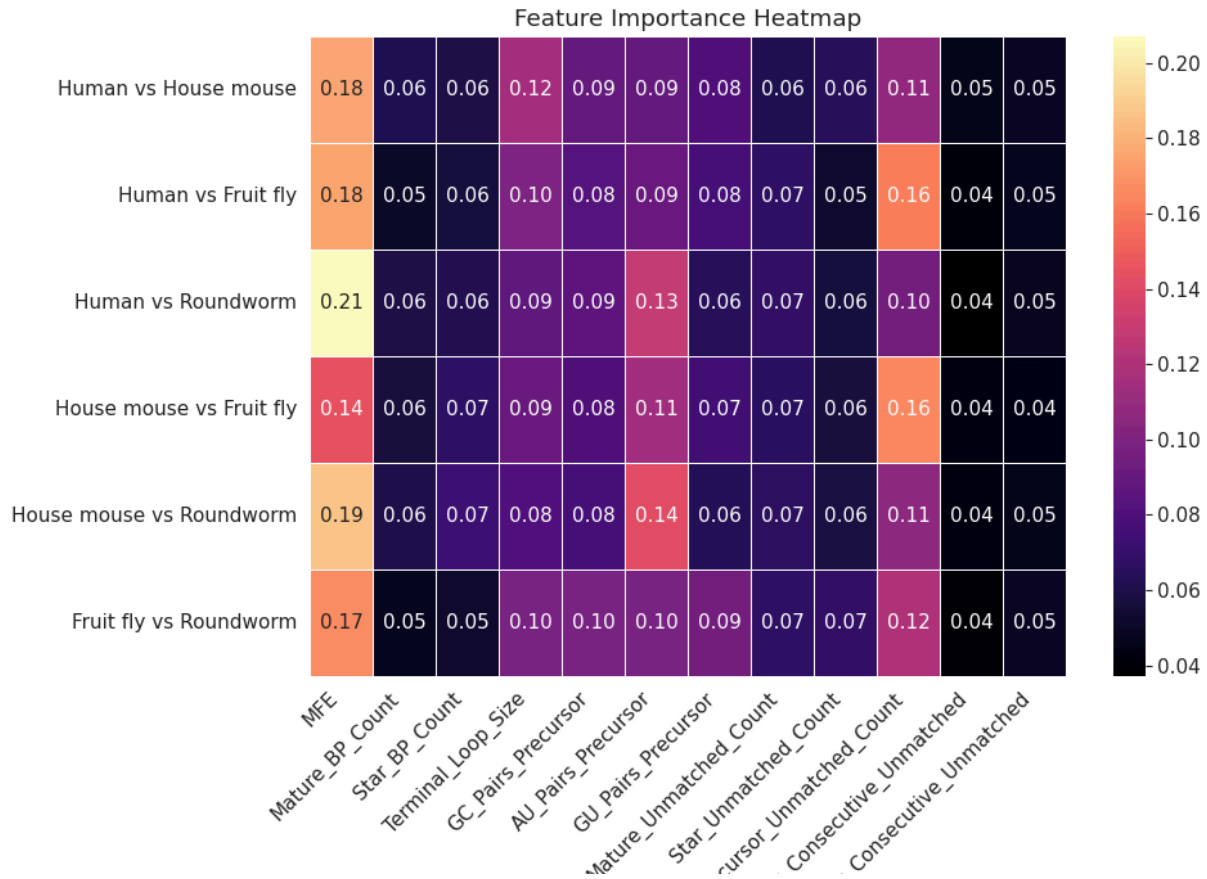
Accuracy=0.8244
ROC-AUC=0.7731
Human vs Roundworm:
Accuracy=0.7710
ROC-AUC=0.8121
House mouse vs Fruit fly:
Accuracy=0.7523
ROC-AUC=0.6903
House mouse vs Roundworm:
Accuracy=0.8073
ROC-AUC=0.7720
Fruit fly vs Roundworm:
Accuracy=0.6964
ROC-AUC=0.7455

We can see that other than our Human vs House mouse model, which performed very poorly (around 43%), the other 5 classifiers achieved high accuracy, ranging from ~70% to 82%. ROC-AUC for those five is also high, with a similar range. We also present these results in heatmap form:



Combining our different models' feature importance, using the Random Forest's Gini Importance, we generated a heatmap of which features were most important for which model:

Ben Sarid 201445830, Yuval Weizman 318540721, Osher Paz 208298976

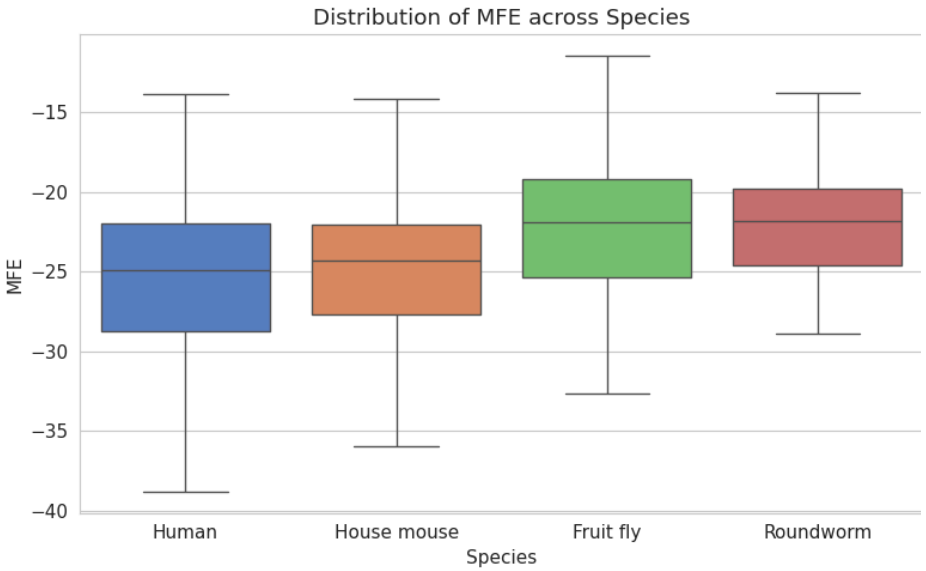


We can see that the most important features for classification, which are also the features that differed the most, were MFE, Precursor AU pairs, and Precursor unmatched nt count, with MFE specifically being the most important feature to differentiation in most models.

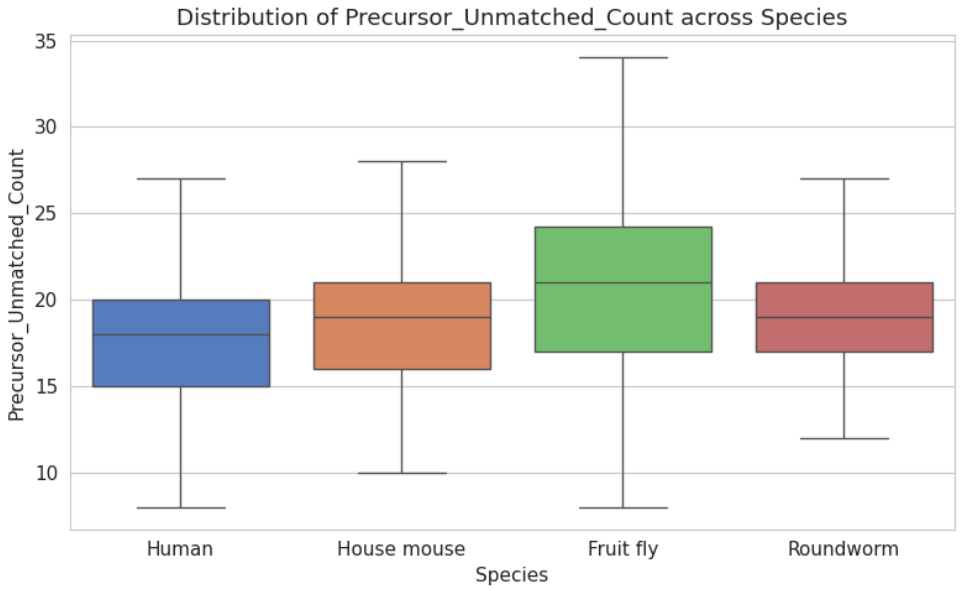
It also appears that the more distant two species were evolution-wise, the better the model was in differentiating them. We can see that between the mammals (human and house mouse) and the fruit fly, and also between the mammals and the roundworm, our models performed very well. On the other hand, between a human and a house mouse, where evolutionary distance is shorter (both are mammals, evolutionary split was generally “recent” compared to fruit fly), the model struggled greatly in differentiating them.

Ignoring our human vs house mouse model, which performed poorly, we think that the feature importance of both MFE and the Precursor unmatched nt reflect the evolutionary distance between the organisms.

Ben Sarid 201445830, Yuval Weizman 318540721, Osher Paz 208298976



We can see in this box plot how the two mammals' MFE range and 50% concentration are decidedly lower than the fruit fly and roundworm, indicating a higher complexity in the miRNA genes, resulting in a higher absolute MFE. Considering that both mammals are warm blooded and the fruit fly and roundworm cold blooded, this seems to support the idea that more stable miRNA genes (with lower MFEs) are necessary to withstand higher body temperatures.



Unmatched precursor nt count can be interpreted similarly, with more unmatched precursor nts leading to a less stable structure that can still survive in cold blooded species, but not in warm blooded ones, where higher stability is required.

In conclusion, our models did a very good job (70% accuracy and ROC-AUC and higher) differentiating between mammals, fruit flies, and roundworms, which are all evolutionary distant and exhibit very different structural properties. It failed to differentiate between the two mammals in the experiment, leading us to think that structural difference between

mammals is much smaller, and more advanced approaches are needed to differentiate between them.