

Computational Biology and Data Mining



JGU

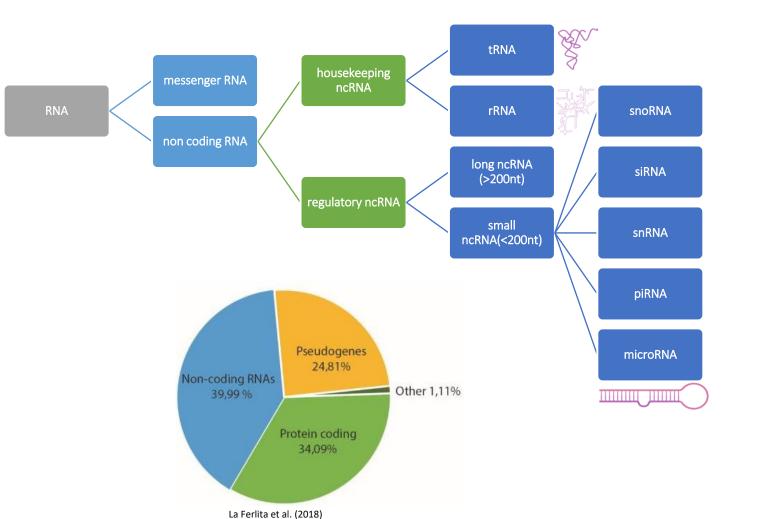
microRNA prediction

Mert Cihan

12.12.2022

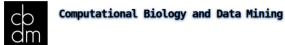
Transcriptome

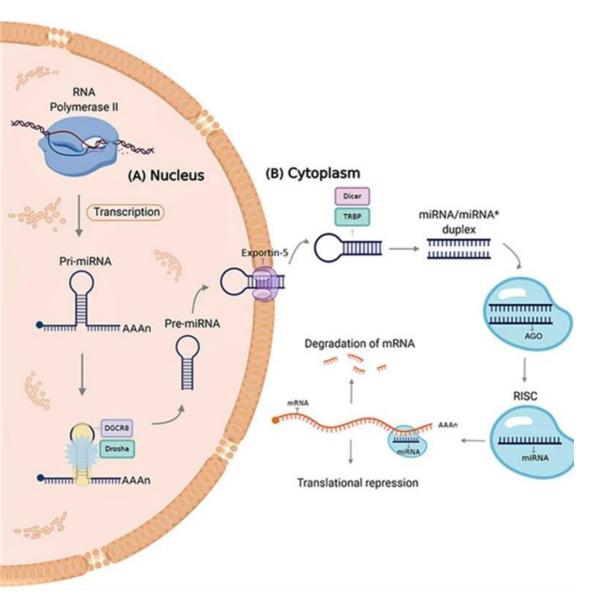




- Most transcripts in the human genome are non-coding
- Non-coding RNA essential for gene regulation at transcriptional and post-transcriptional level
- Heterochromatin formation, histone modification, DNA methylation

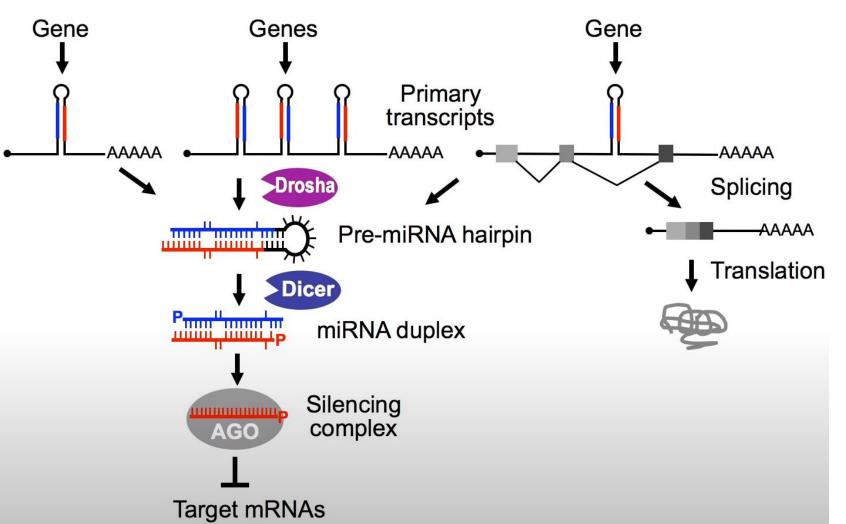
microRNA biogenesis





- microRNAs are small non-coding RNAs (22nt)
- Approximately 2500 human microRNAs
- Key-proteins: Drosha, Dicer, Argonaute
- Binding of microRNA seed in silencing complex to complementary 3' untranslated region of mRNA
- Lead to translational repression/ mRNA degrada tion
- microRNA families often enriched in targets of tr anscription factors (redundant functions)

microRNA genes



- Approximately 2300 microRNA genes
- Different primary transcripts (up to 6)
- Multiple microRNAs can be within one ORF
- Multifunctional transcripts (MIRTRONS)
- Non-concial biogenesis can avoid Drosha/Dicer dependency

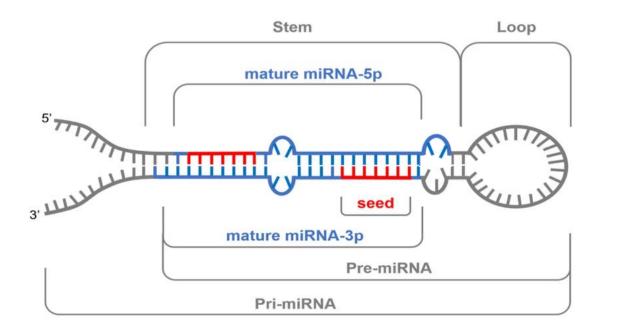
microRNA-mRNA

Offset 6mer

6mer

8mer

7mer-A1 7mer-m8



	5' · · · · · · · · · · · 3'
	5' · · · · · · · · · · · · · · · · · · ·
3'UTR	5' · · · · · · · · · · · · · · · · · · ·
	5' · · · · · · · · · · · · · · · · · · ·
	5' · · · · · · · · · · · · · · · · · · ·
mature miRNA	3' NNNNNNNNNNNNNNNNNNN 5'
	seed

 Drosha 3' cut leaves a 2 nt overlap and Dicer 5' cut leaves a 2nt overlap

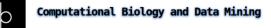
cþ am **Computational Biology and Data Mining**

- Strong propensity for one of both mature strands
- Many microRNA binding sites in 3'UTRs are broadly conserved
- Multiple microRNA binding sites for the same microRNA family
- >60% of protein coding genes are targeted

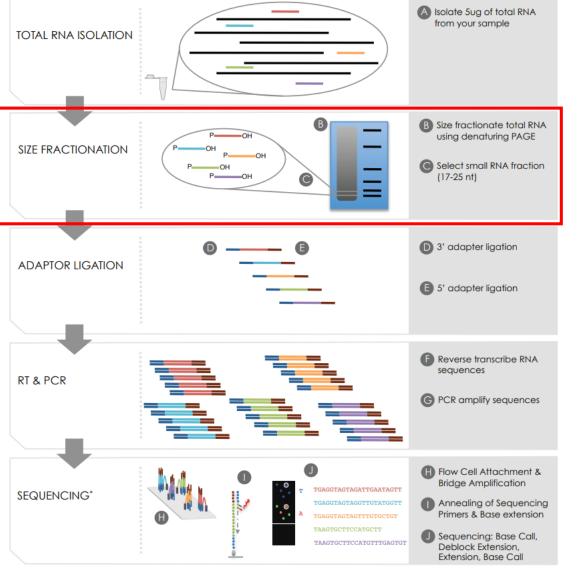
Experimental microRNA detection

METHOD	DISADVANTAGES	ADVANTAGES
Northern blotting	Low-throughput-Low sensitivity- Laborious and very time consuming	High specificity-Readily available and easy-to-perform
In situ hybridization	Low-throughput/Semi-quantitative	Monitor cellular and sub-cellular distributions/spatiotemporal expression profile
Reverse transcription(RT- qPCR)	Cannot identify novel miRNAs	High sensitivity and specificity - Can be used for absolute quantification
Microarray	Low-sensitivity and specificity/ Cannot identify novel miRNAs	Comparing the relative abundance of specific miRNAs/low cost
Next generation sequencing	Substantial computational work required	Very high sensitivity-High accuracy in distinguishing variants of miRNAs

Experimental microRNA detection



MIRNA-SEQ LIBRARY PREPARATION



- A profile of all small RNAs and miRNAs in the transcriptome
- Small RNA targets are enriched through size selection using size-exclusion gels or commercially developed kits
- Follow RNA-seq procedure
- Bias because of degradated RNA fragments

Ex.1 Annotate microRNAs

Computational Biology and Data Mining

Exercise 1: Find the (hsa-) microRNA sequence using blast

>seq1

ACATTTACCTAGCAGAAGAAAAATCGTGTTTACGAAGGTGGTTTTCGCAGGGCGAAGCTAATTCGTGCAACTTCCCCAAATGTG GGAAGCTCGACTGCATAATTTGTGGTAGTGGGAGACTGCGTTCGCTCTTTTCCCCCG

>seq2

TCCAAACAGACACTGATGGCACCTTCTGCCATTTAGGAATTTGTTTTAAAACAGACATTTGTCTAGATATTTCCTTTGTGGCCTCC TCCCCATCAAAAGTCAATCAAACATCG

>seq3

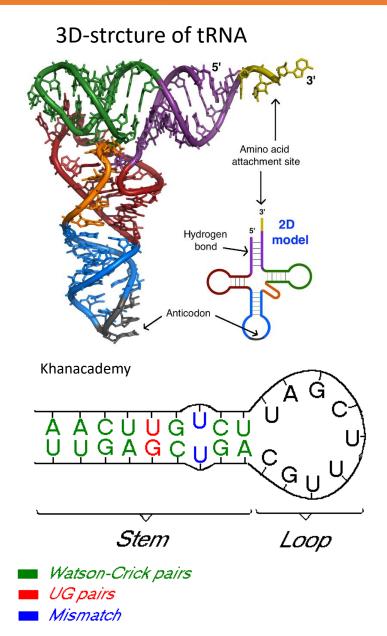
GUGGCCUCGUUCAAGUAAUCCAGGAUAGGCUGUGCAGGUCCCAAUGGGCCUAUUCUUGGUUACUUGCACGGGGACGC

Sol. 1 Annotate microRNAs



		Your sear	ch is limited to records that	include: Home sa	piens (taxid:9606)				
Enter Query S	equence		Job Title	3 sequences (seq1)			Filter Results		
Enter accession nu	Imber(s), gi(s), or FASTA sequence(s) 😯 Clear	Query subrange 😯	RID	T13Y627V013 Search exp	ires on 12-08 20:06 pn	Download All 🗸			—
			Results for	3:lcl Query 27773 seq3###(7	7bp)	~	Organism only top 20 w		exclude
			Program	BLASTN ? <u>Citation</u> ~				, binomial, taxid or group	name
		То	Database	refseq_rna <u>See details</u> ✓ + <u>Add organism</u>					
			Query ID	lcl Query_27773			Percent Identity	E value	Query Coverage
Or, upload file	Choose File No file chosen		Description	seq3###			to	to	to
Job Title	3 sequences (seq1)		Molecule type	rna					Filter Reset
	Enter a descriptive title for your BLAST search ?	Query Length	77						
Align two or mor			Other reports Descriptions	Distance tree of results	MSA viewer ?	Taxonomy			
Choose Searc	h Set		Descriptions	Graphic Summary	Alignments	laxonomy			
Database	Standard databases (nr etc.): O rRNA/ITS datab	ases 🔘 Genomic + transo	Sequences	producing significant al	ignments		Downlo	ad 🎽 Select colum	ns ∨ Show 100 ♥ 😯
	Reference RNA sequences (refseq rna)	× (?)	🗹 select all	1 sequences selected			GenE	<u>Bank Graphics Distance</u>	ce tree of results MSA Viewer
Organism Optional	9606	e:		Descriptio	n		Scientific Name	Max Total Query Score Score Cover V	E Per. Acc. Ident Len
optional	Enter organism common name, binomial, or tax id. Only 20	top taxa will be shown 😮	Homo sapie	<u>ns microRNA 26a-1 (MIR26A1), m</u>	icroRNA	<u>Homo sapi</u>	ens	143 143 100% 3	3e-33 100.00% 77 <u>NR_029499.1</u>

RNA secondary structure



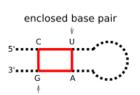
- Not only proteins and DNA have secondary structures
- Structured RNA is involved in all aspects of gene expression
- mRNAs are often structured in terminal regions
- Watson-Crick-pairing is preferred, but not as prominent as in DNA
- Mismatches created wobbles (distinct hydrogen bonds)
- Mismatches can have similar stability

RNA secondary structure

• RNA folding has a great relationship with minimum free energy (mfe)

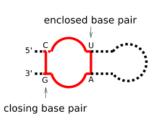
5'- A C G C Table 2. Thermodynamic	ĠŲĂ ĊĂŬ	ĊĢ ĠĊ	AA U - ₅'
and propagation in 1 M Na Propagation sequence	ΔH°, kcal/mol	ΔS°, eu	ΔG_{37}° kcal/mol
AĂ ŲU	-6.6	-18.4	-0.9
ĂŬ ŲA	-5.7	-15.5	-0.9
	-8.1	-22.6	-1.1
	-10.5	-27.8	-1.8
CU GA	-7.6	-19.2	-1.7
GĂ ÇU	-13.3	-35.5	-2.3
GU ÇA	-10.2	-26.2	-2.1
ସେଥି ମିର୍ଚ୍ଚ	-8.0	-19.4	-2.0
d C C C C C C C	-14.2	-34.9	-3.4
GG ÇC	-12.2	-29.7	-2.9
Initiation Symmetry correction (self-complementary) Symmetry correction	(0) 0	-10.8 -1.4	3.4 0.4
(non-self- complementary)	0	0	0

Freier et al. (1986)

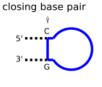


closing base pair

stacking pair



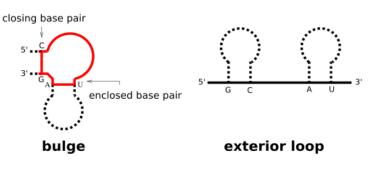
interior loop



hairpin loop

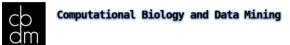
s' ... 3' ... closing base pair

multi loop



•Turner 1999 RNA parameters •Mathews 1999 DNA parameters •Andronescu 2007 RNA parameters •Mathews 2004 DNA parameters

Ex.2 microRNA folding



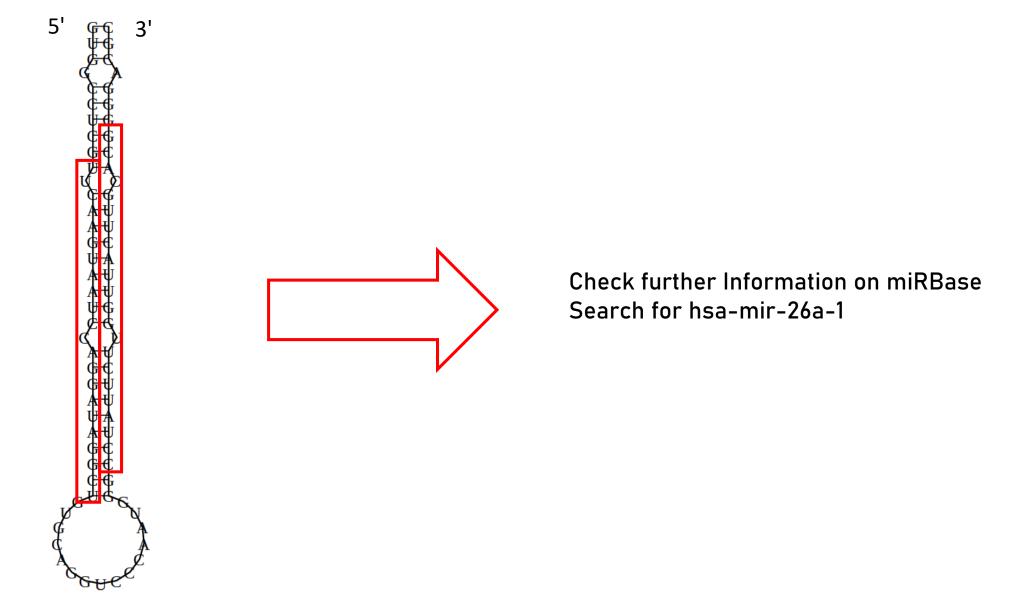
- 1. Use the *RNAfold webserver* to compute the secondary structure of prior annotated microRNA sequence.
- Find the 22 nt mature microRNA-3p and -5p sequences.
 (*Tip: Dicer 5' cuts directly at the hairpin loop*)

		Stem	Loop
RNAfold WebServer	1 Enter Input 2 Vesuts		
	[Home New job Help]	mature miRNA-5p	
The RNAfold web server will predict secondary structures of single stranded RNA or DNA sequences. Current limits are 7,500 nt for partition function calculations and 10,000 nt for minimum free energy only predictions.			
Simply paste or upload your sequence below and click Proceed. To get more information on the meaning of the options click the 🥑 symbols. You can test the server using this sample sequence.			
Paste or type your sequence here:	[clear]	5'	
		TTT I I I I I I I I I I I I I I I I I I	ATTA
Show constraint folding			A
Or update at left in FASTA format. Choose File. No file chosen			(A
Fold algorithms and basic options			1
minimum free energy (MFE) and partition function (g)			6 7
o minimum free energy (MFE) only go		ULL O	VV
no GU pairs at the end of helices 🥑 avoid tooleted base pairs 🥪		3' MINING Seed	CULP .
		3.	
Show advanced options		1010	
interactive RNA secondary structure plot 🥑		mature miRNA-3p	
RNA secondary structure plots with reliability annotation (Partition function folding only) 🥹			
2 Mountain pilot 🥹		Pre-miRNA	
Notification via e-mail upon completion of the job (optional): your e-mail		Пеншина	
	Proceed »		
		Pri-miRNA	
Institute for Theoretical Chemistry University of Venma mail@dx.univle.ac.at			

Sol.2 microRNA folding

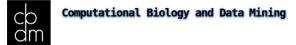


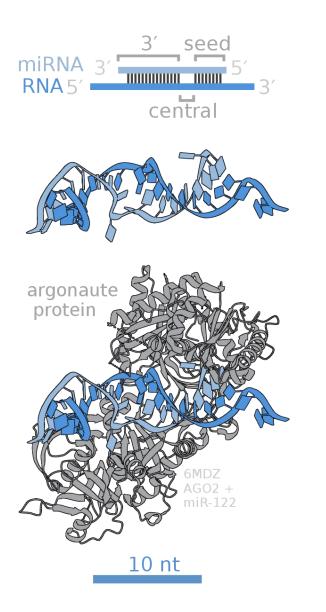
Computational Biology and Data Mining



MFE secondary structure

Finding microRNA targets

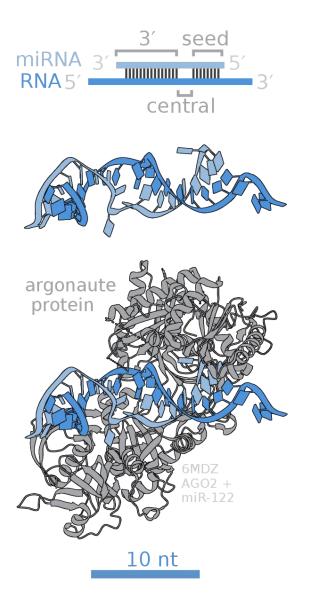




- Mature microRNA is loaded into argonaute protein
- microRNA seed (nt 2-8 from 5' end in mature microRNA) binds to complementary 3'UTR sequence
- 3'UTRs of mRNAs can be really long; on average 800 nt, but up to 10,000 nt
- Searching only complementary sequences leads to many false positives
- Integrate further information for scoring of miRNA-mRNA interactions

Ex. 3 microRNA seed





1. How many human genes are potentially targeted by previously identified mature microRNAs? (www.targetscan.org/vert_80/)

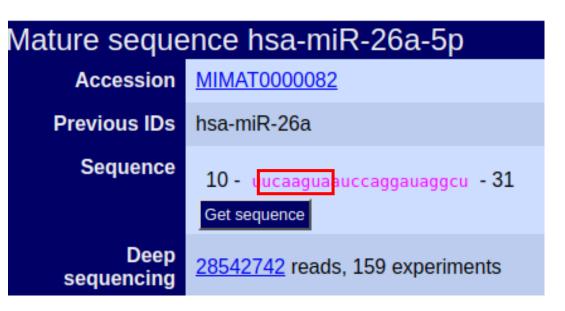
2. Which gene has the most conserved 8-mer binding sites?

3. Find the microRNA seed of previously identified mature microRNA. Does it match the positions in the mature microRNA?

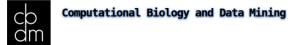
Sol. 3 microRNA seed

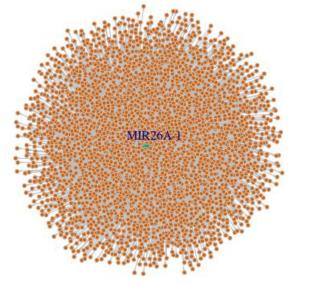
	Humon I m	iD 26 En																				
	1045 transcripts with conserved sites, containing a total of 1209 conserved sites and 579 poorly conserved sites.																					
	neuse note that these predicted targets include some tarse positives. [Neutrinore]																					
	Genes with only poorly conserved sites are not shown. [View top predicted targets, irrespective of site conservation]																					
I	Table sorted by cumulative weighted context++ score [Sort table by predicted occupancy] [Sort table by aggregate P _{CT}]																					
	The table shows at most one transcript per gene, selected for being the most prevalent, based on 3P-seq tags. [Download table]																					
		Represen-		Number of		Conserved si		sites Poor		Poorly conserved sites				Predicted occupancy			Cumulative			Previous		
	Target gene		Gene name	3P-seq tags supporting	Link to sites in UTRs	total	8mer	7mer- m8	7mer- A1	tota	l 8mer	7mer- m8	7mer- A1	6mer sites	Representative miRNA	mod	high miRNA	trans-	weighted context++ score	Total context++ score	Aggregate P _{CT}	TargetScan publica- tion(s)
I		ENST0000		UTR + 5														mirina				
1	nolymerase (RNA) III (DNA directed) polymentide G																					
	PTEN	0371953.3	phosphatase and tensin homolog	118	Sites in UTR	4	3	0	1	0	0	0	0	0	hsa-miR-26b-5p	0.1153	0.6550	2.1149	-0.55	-0.60	> 0.99	2011, '15
1																						
							/															

ACUGUUAGGGAAUUUUACUUGAA	
	8mer
UCGGAUAGGACCUAAUGAACUU	

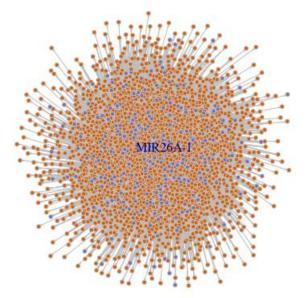


microRNA-mRNA Interactome





all interactions MIR26A (7897)



- microRNA-mRNA interactions often false-positive
- Experimental validation often indirect from high-throughput
- Integration of further information can help ensuring accuracy:

•site type

supplementary pairing

local AU

minimum distance

•3' UTR length

•TA (target site abundance)

•SPS (seed-pairing stability)

- •Conservation of 3'UTR
- Conservation of microRNA family
- •Thermodynamic stability of precursor microRNA