

Predicting RNA binding of new
RNA-binding proteins

Project in
Deep Learning in
Computational Biology

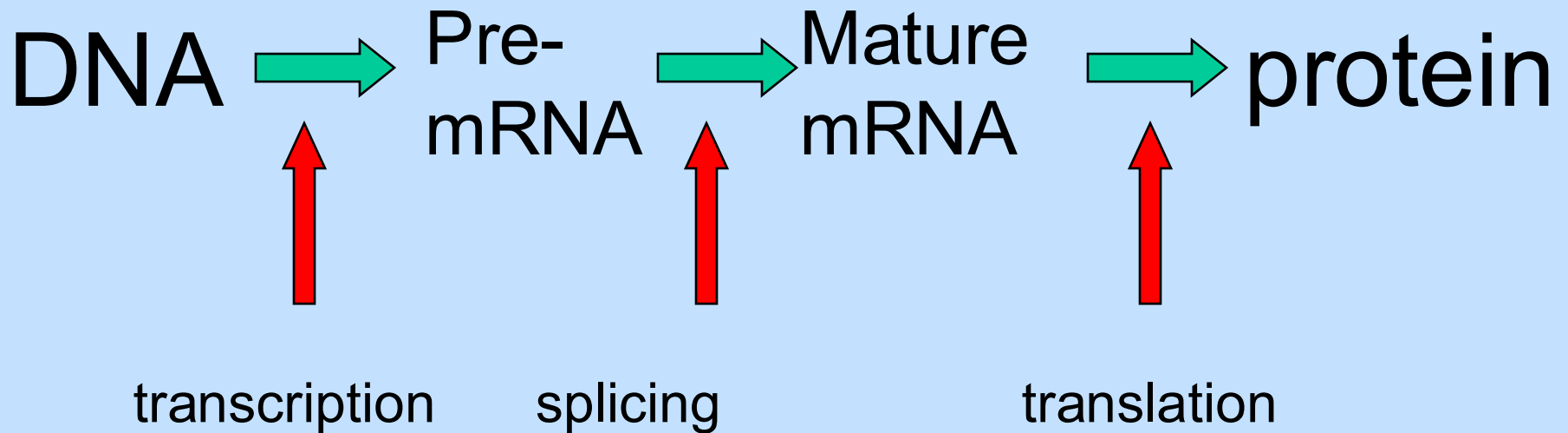
Outline

1. Some biological background again...
2. The project

1. Background

Slides with Ron Shamir and Chaim Linhart,
Computational Genomics TAU

Gene: from DNA to protein



DNA

- DNA: a “string” over the alphabet of 4 **bases** (nucleotides): { **A, C, G, T** }
- Resides in chromosomes
- Complementary **strands**: A-T ; C-G

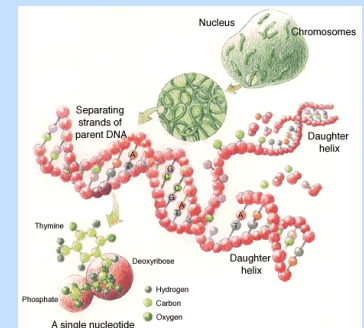
Forward/sense strand:

Reverse-complement/anti-sense strand:

AACTTGCG
TTGAACGC

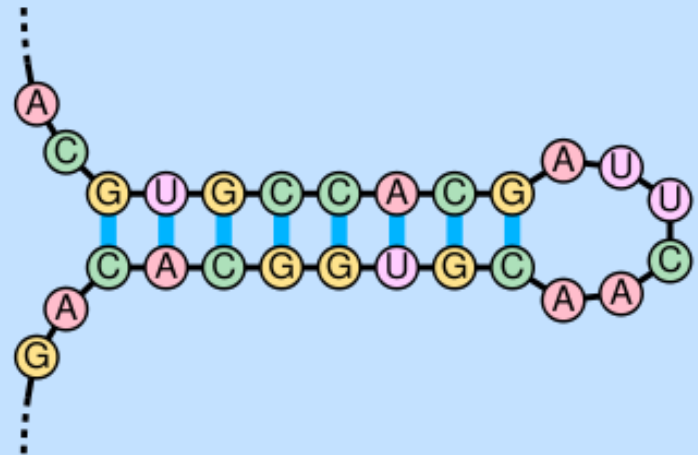
- Directional: from 5' to 3':

(upstream) \Rightarrow AACTTGCGATACTCCTA \Rightarrow (downstream)
 5' end 3' end



RNA (Ribonucleic acid)

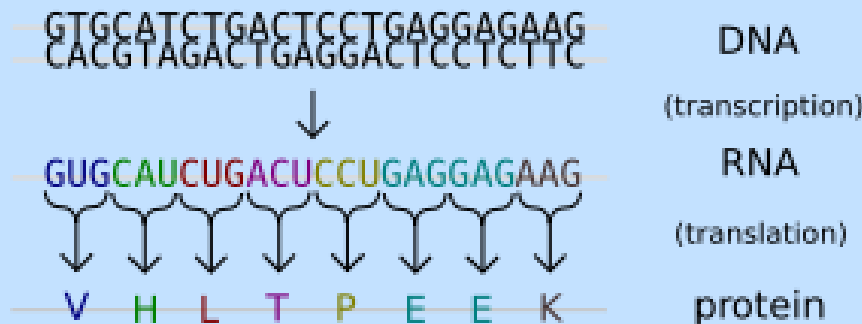
- Bases:
 - Adenine (A)
 - Guanine (G)
 - Cytosine (C)
 - Uracil (U); replaces T



- Oriented from 5' to 3'.
- Single-stranded => flexible backbone => secondary structure => catalytic role.

Translation

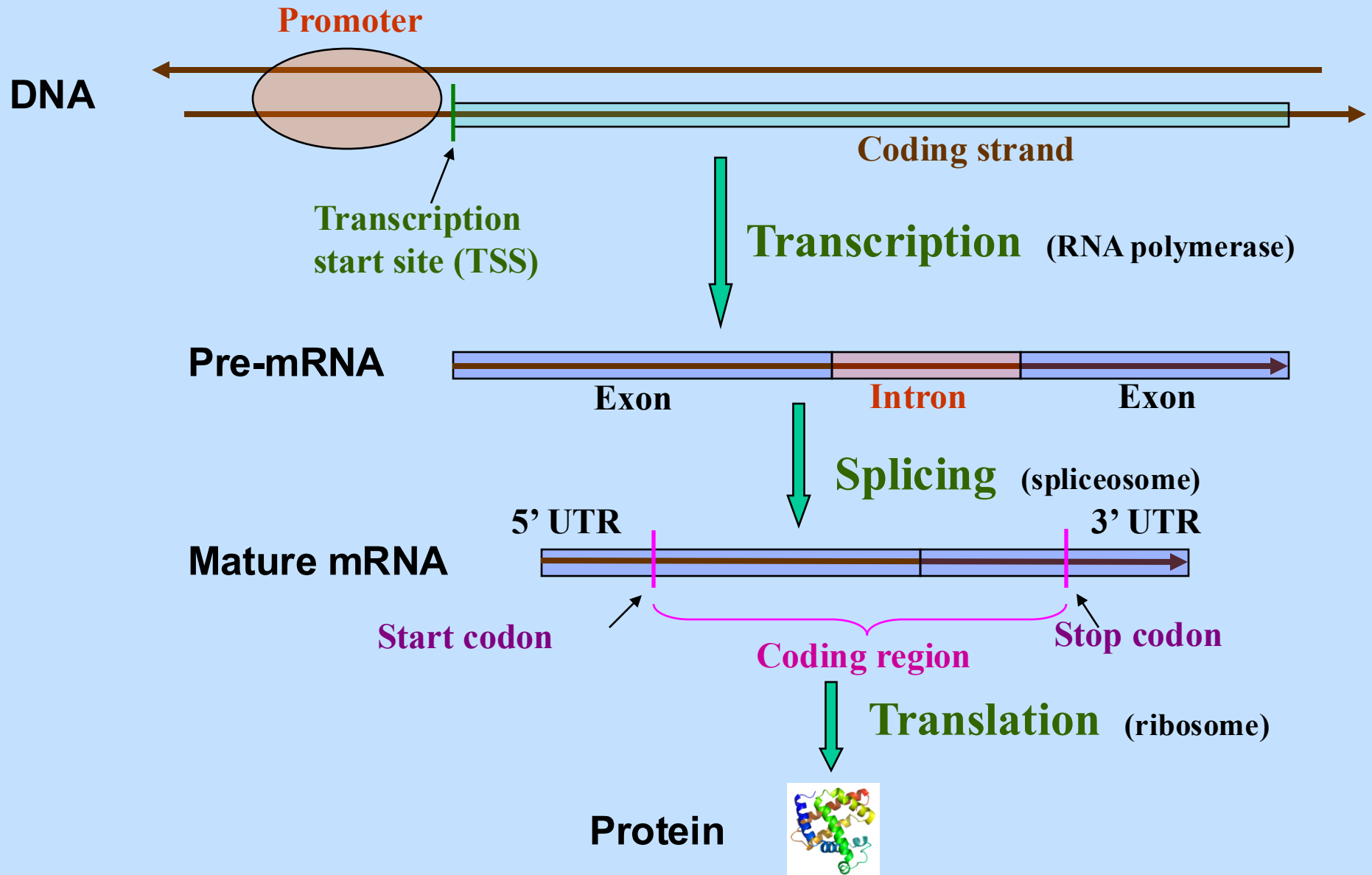
- **Codon** - a triplet of bases, codes a specific amino acid (except the stop codons); many-to-1 relation
- **Stop codons** - signal termination of the protein synthesis process



		Second base of codon					
		U	C	A	G		
First base of codon	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } SER UCA } UCG }	UAU } Tyr UAC } UAA } UAG }	UGU } Cys UGC } UGA } UGG } Trp	U C A G	Third base of codon
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G	
	A	AUU } Ile AUC } AUA } AUG } Met	ACU } ACC } Thy ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G	
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G	

The genetic code, written by convention in the form in which the Codons appear in mRNA. The three terminator codons, UAA, UAG, and UGA, are boxed in red; the AUG initiator codon is shown in green.

Gene structure (eukaryotes)



Genome sequences

- Many genomes have been sequenced, including those of viruses, microbes, plants and animals.
- Human:
 - 23 pairs of chromosomes
 - 3+ Gbps (bps = base pairs) , only ~3% are genes
 - ~25,000 genes
- Yeast:
 - 16 chromosomes
 - 20 Mbps
 - 6,500 genes

Regulation of Expression

- Each **cell** contains an identical copy of the whole genome - but utilizes only a subset of the genes to perform diverse, unique tasks
- Most genes are highly regulated - their **expression** is limited to specific tissues, developmental stages, physiological condition
- Main regulatory mechanisms:
 - transcriptional regulation
 - **post-transcriptional regulation**

Post-transcriptional regulation

- Post-transcription is regulated primarily by RNA-binding proteins (**RBP**s) – proteins that bind to RNA subsequences, called binding sites (**BS**s)
- BSs of a particular RBP share a common pattern, or **motif**

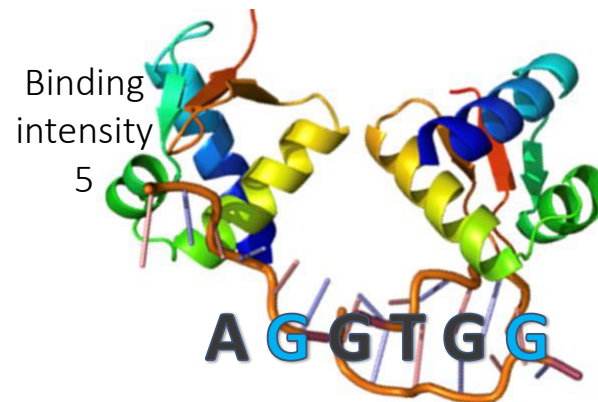
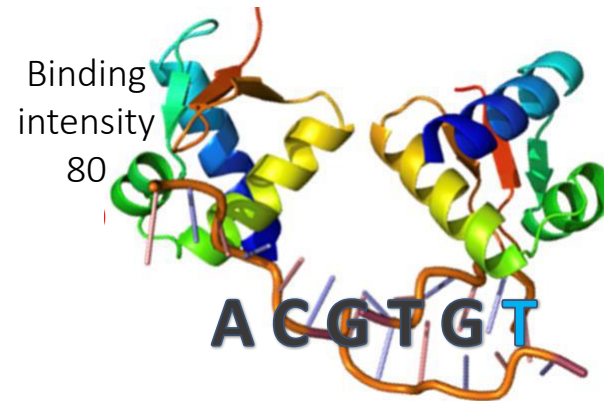
• Input:

```
0       5       10      15      20      25      30      35      40      45
TCTCATCCGGTGGGAATCACTGCCGCATTTGGAGCATAAACAATGGGGGG
TACGAAGGACAAACACTTTAGAGGTAATGGAAACACAACCGGCGCATAAA
ATACAAACGAAAGCGAGAAGCTCGCAGAAGCATGGGAGTGTAAATAAGTG
GGCGCCTCATTCTCGGTTTATAAGCCAAAACCTTGTCGAGGCAACTGTCA
TCAAATGATGCTAGCCGTCGGAATCTGGCGAGTGCATAAAAAGAGTCAAC
```

• Output:

```
0       5       10      15      20      25      30      35      40      45
TCTCATCCGGTGGGAATCACTGCCGCATTTGGAGCATAACAATGGGGGG
TACGAAGGACAAACACTTTAGAGGTAATGGAAACACAACGGCGCATAAA
ATACAAACGAAAGCGAGAAGCTCGCAGAAGCATGGAGTGTAATAAGTG
GGCGCCTCATTCTCGGTTTATAAGCCAAAACCTTGTCGAGGCAACTGTCA
TCAAATGATGCTAGCCGTCGGAATCTGGCGAGTGCATAAAAAGAGTCAAC
```

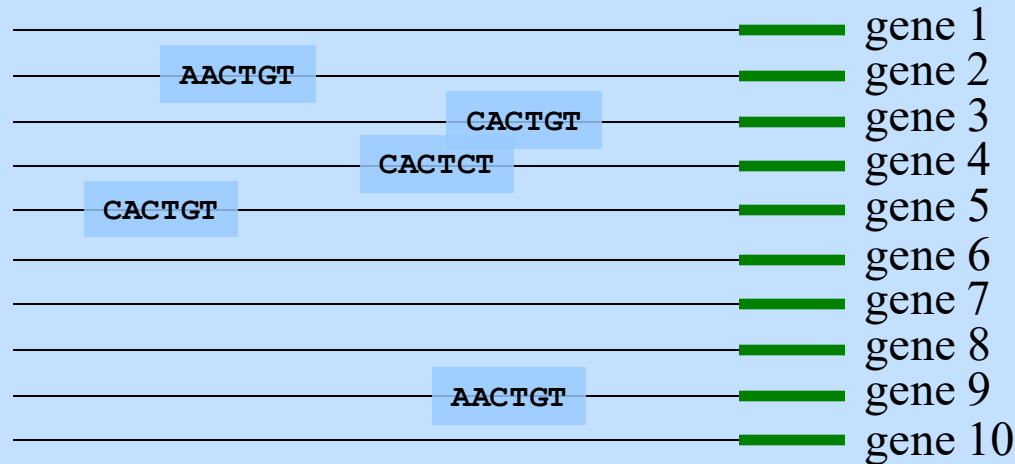
Protein RNA-binding



RBP BS motif models - strings

- Consensus (“degenerate”) string:

$\begin{matrix} A & & C \\ C & A C T & G T \end{matrix}$



- List of k-mers (weighted or unweighted)

How can we use NN for these models?

RBP BS models - PWM

- Position weight matrix (PWM): each position has weights for the 4 possible letters (A, C, G, T)

- For example:

How can we use NN
for this model?

	1	2	3	4	5	6
A	0.1	0.8	0	0.7	0.2	0
C	0	0.1	0.5	0.1	0.4	0.6
G	0	0	0.5	0.1	0.4	0.1
T	0.9	0.1	0	0.1	0	0.3

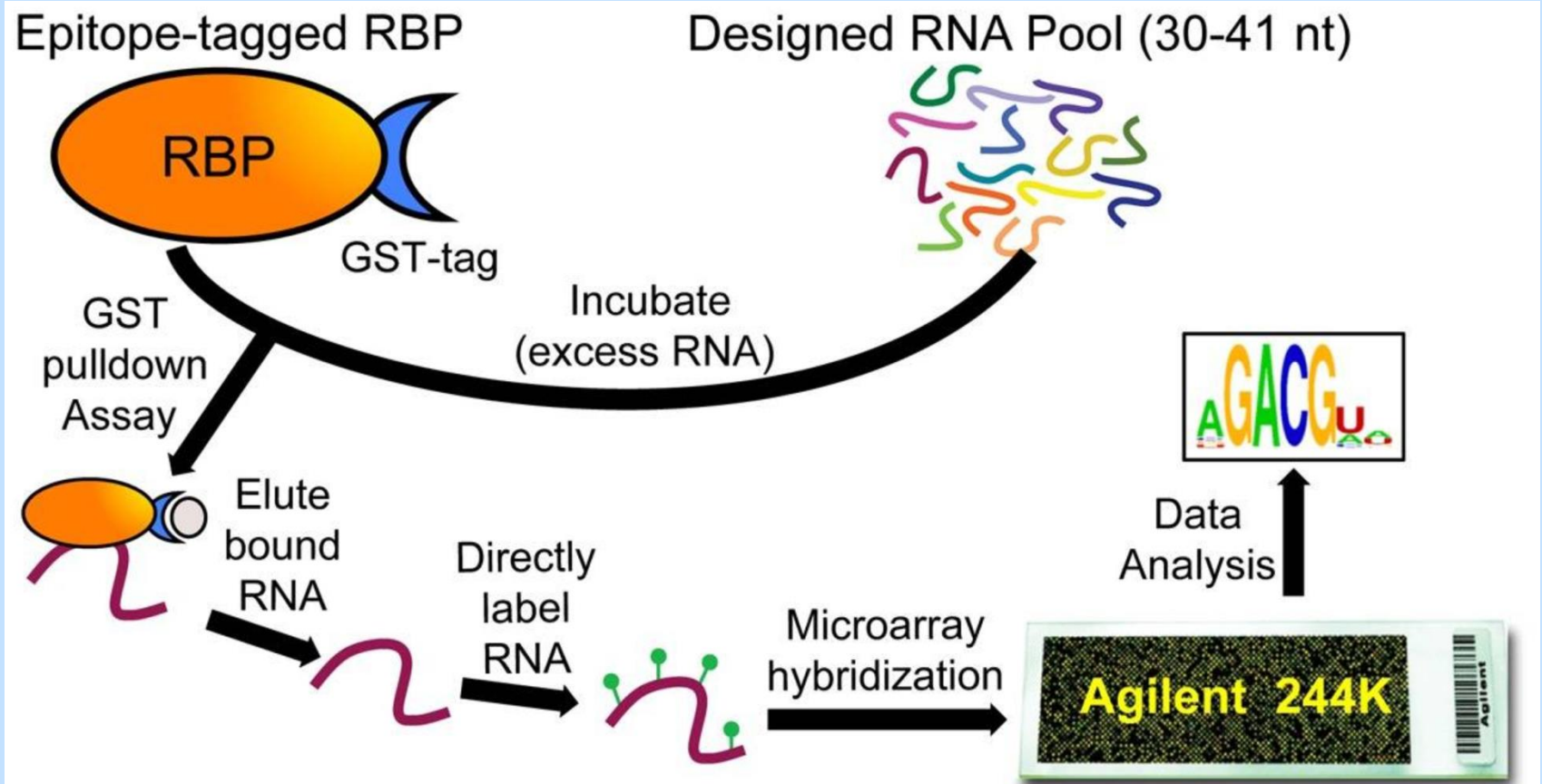
- Logo format:



RNAcompete

Ray et al, Nat. Biotech 2009

- Generate an RNA pool covering all possible 9-mers, each at least 16 times
- Detect RBP binding to specific 9-mers



RNAcompete

b

HuR

```

AGACCUUCAUGUUCUGUUUCUUUUACUGCUUUGGUUGU
AGAAGUUUGUUUUGGUUUUUGCUUUUCCUCGUGUGCA
AGAUUUUUUUACUUUCAUGUCUAAUCACCUUUGGGGAA
AGGUUUCGUUUUGUGUACUUCUCCCUAGUUUCUUUGGC
AGGAGUUAGUUUGGUUUUACUUUGUGUCCGCCUACGGU
AGAAGUAGGGAAGUUUUUUUCGUUGAUUUAGUGGGCCU
AGAUUUUUUUCGUUUAUAAUAAUGCGUUUAGUGCCAC
AGAUUCAGAUUUAGCACCACUGUUGUAGUAUGUUUUUGU
AGAAUUGAUUUUUUUUUGUUGUCUCACCCAUCUGCGUU
AGACUGCUAACUUUUUUUUUGUCGUCCUACGUUGCA
    
```

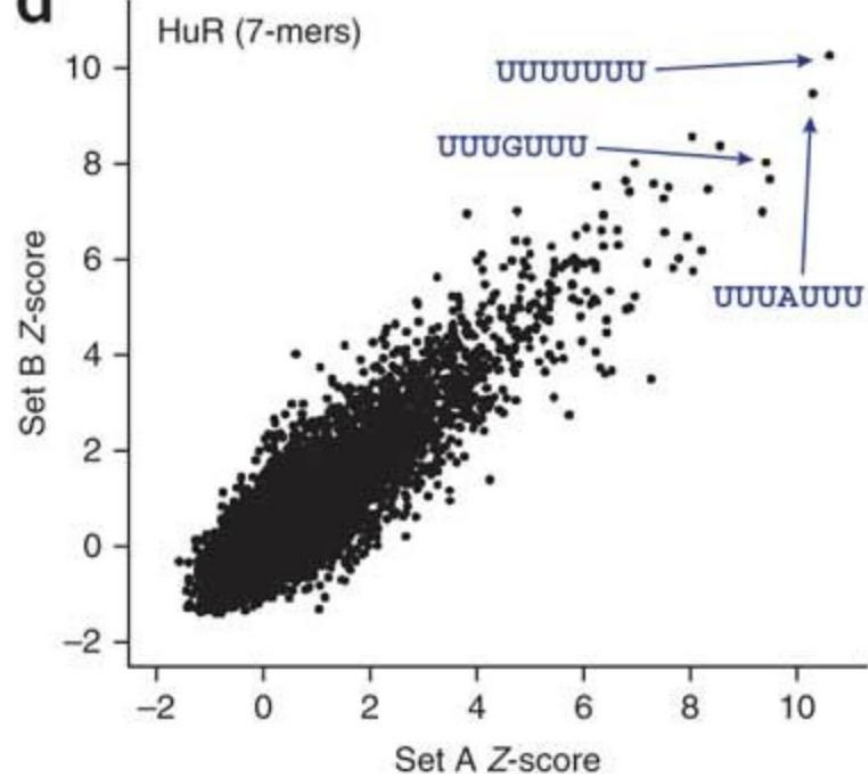
c

Vts1

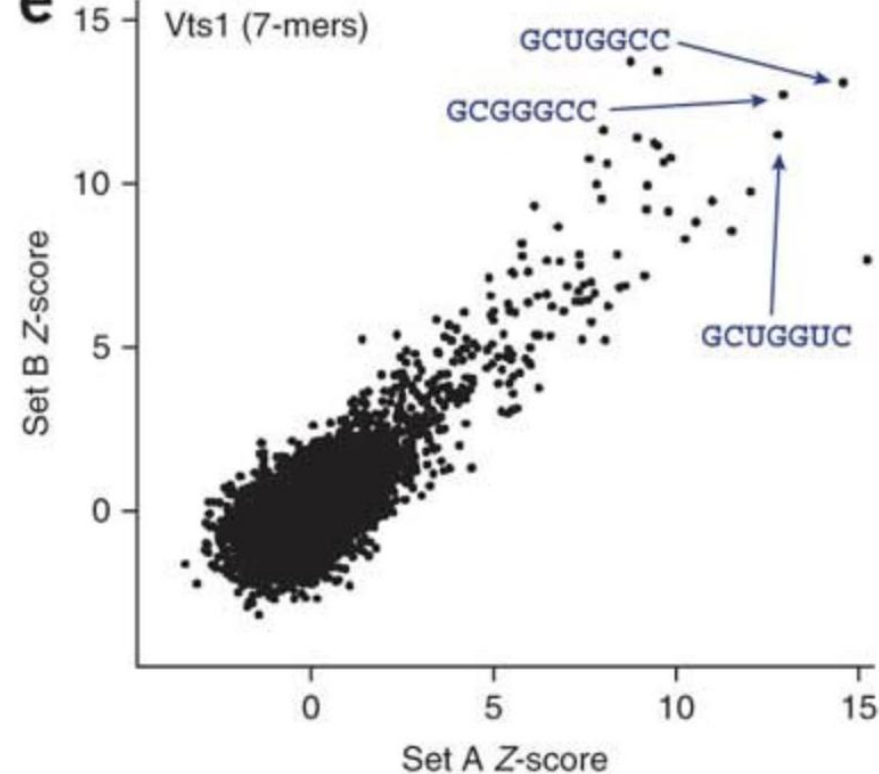
```

AGGGCCAAUGCGGUGCAGGCGCUGCGUUGGC
AGGUUCAAAACGCAUGCGGGUCAAGCUACGAACGCCACU
AGAGAUGGAUUGGGCUGGCACCGGUCCAUC
AGGGCGUUUAACGGCGGGUCCCGUUAGGCGC
AGAGAGUGCAUAGCUGGCUUUCUAUGCGCUC
AGGGCCAAUGCGGUGCAGGCGCUGCGUUGGC
AGAGUGUCAGGUACAUAAGUAGCUCAAUGAGUUCGUU
AGAGUCCAGAACGGCAGGCACGUUUUGGAC
AGAGACAGGAGCUUAUUCAUUGAUCAGCGAUCGCG
AGAGGUUGAGACGGCAGGUCCCGUCUCGGCC
    
```

d



e



RNAcompete - implementation

- 30-41nt variable region
- A sequence library covering of all 9-mers, each at least 16 times
- ~240K probes

The computational challenge

- Input:
 - Amino-acid sequence of a new protein (1 sequence file)
 - a list of RNAcompete probes (1 sequence file)
- Goal: Predict RNA binding intensity for each RNAcompete probe
- Intuition: learning a binding model based on a protein's amino acid sequence

The project

General goals

- Research
 - Learn about known solutions
 - Trial and error with training data
- Develop software from A-Z:
 - Design of deep neural networks
 - Implementation
 - Execution & analysis of test data
- A taste of bioinformatics
- Have fun
- Get credit...

The computational task

- Given a set of RNAcompete data of different RBPs
- Learn a model that for a given RBP can score RNAcompete probes
- Main challenges:
 - Prediction performance
 - Training runtime

RBP Input

- 1 amino-acid sequence file

<sequence>\n

```
MSSKIFVGGNNVYIGYDNQKALGLILRRFETRKGFGFVTFDDQSRGKLIPYEVVGEVVKLGYRDQRG
```

RNAcompete Input

File with ~120K lines, each containing a probe sequence of length 30-41.

<sequence>\n

```
GTAATATTACATTCCGTTTAACTGCGCGCCTACGG
TTTCTCTATGCAGATTAGCTCAATTCCCAATACCTA
GTAATATTCATACGGCCTCGACTCATTACACGGCTT
CTATATGTGCATATATTACCCTACAGGTCGGATAGC
TTCGGATTGGGTGCATAATGAAGCAGGACTATTAAA
CCGAGGTATTTGCATTTTCAGTCGTCTTCTAATGAAG
CCCAGCCGGATTTCATACACGGACCGAGCAATGG
GAAGGATGGTATGCGTTTATTTGCATAGACAACATC
GTTACTATAATTAGCAGCGCGTTTCGGGCCGTTCCAC
...
...
...
ACGTTTGCATTAATTTCCCCACATTATTATGCGGA
CCCAGTTAATTATGCGCCACGAGTCCGTCATAGTTA
ACCTCATTAAATGTATTAAGGTTGTTTAAAGTTGGG
```

- The input sequence file is sorted lexicographically
- The output is a file with a binding intensity for input sequence (in the same order)

Solution example

- Train a model that receives a pair
(amino acid RBP sequence, RNA probe sequence)

And outputs a binding intensity

- Protein representation can be using protein LLMs
- RNA representation by one-hot

Input schedule

You will be given:

200 development sets (RBP sequence + **training** RNAcompete probes with binding intensities).

44 evaluation sets (RBP sequence + **test** RNAcompete probes). You have to assign a binding intensity to each RNAcompete probe.

Output

List of binding intensities - each corresponding to a an RNA probe sequence.

The goal

- To assign binding intensities to RNAcompete probes in the RNAcompete file
- Return a file with all binding intensities to all RNAcompete probes
- For this list we can compare to real values by Pearson correlation:
$$\text{cov}(x,y)/(\text{std}(x)*\text{std}(y))$$

Implementation

- Input: the 1st argument is the RBP filename, and the 2nd is the RNAcompete filename
- Output: a file with RNA binding intensities (in the same order of the RNA sequences)
- Training runtime will be measured
- Reasonable documentation

Submission

- Electronic design document
- Electronic code submission
- 44 binding intensities files, e.g. RBP201.txt
- Executable / script for running time test

Design document

- 3-5 pages (pdf), Hebrew/English
- Briefly describe main goal, input and output of program
- Describe neural net architecture, algorithms, and scores

References

- Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins
- Ray *et al.* Nature Biotechnology, 2009
- ProbeRating: a recommender system to infer binding profiles for nucleic acid-binding proteins
- Yang *et al.* Bioinformatics, 2020