Predicting RNAcompete binding from RNA bind-n-seq data

Project in Deep Learning in Computational Biology

Outline

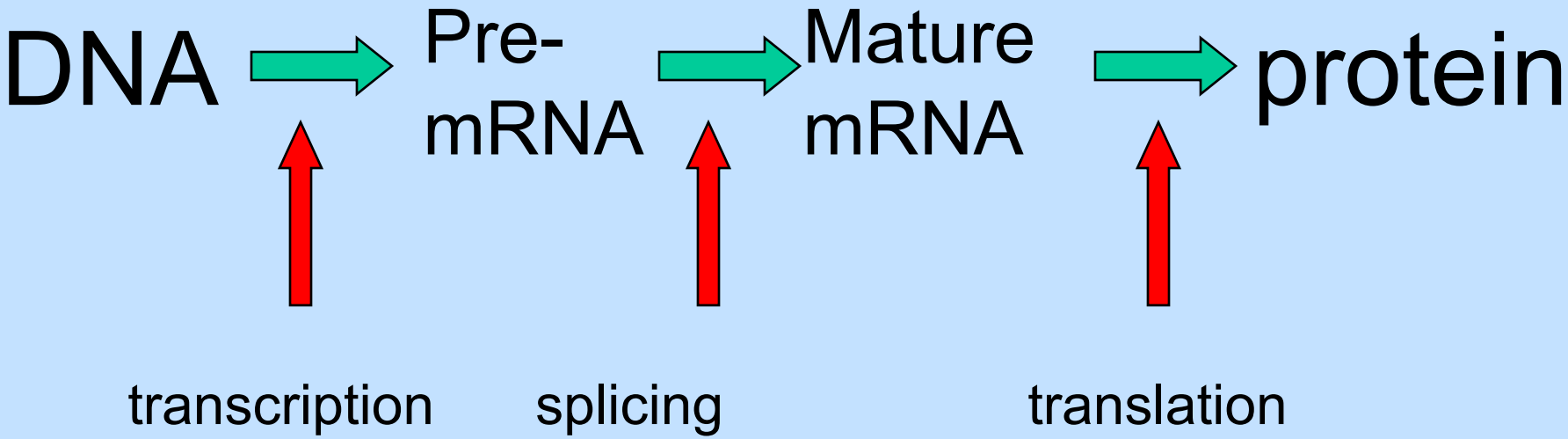
1. Some biological background again…

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

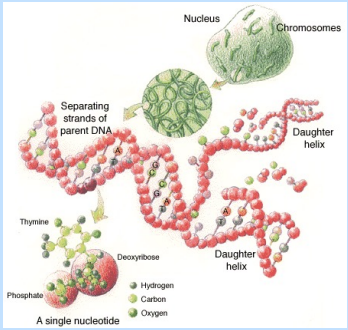
Forward/sense strand: AACTTGCG

Reverse-complement/anti-sense strand: TTGAACGC

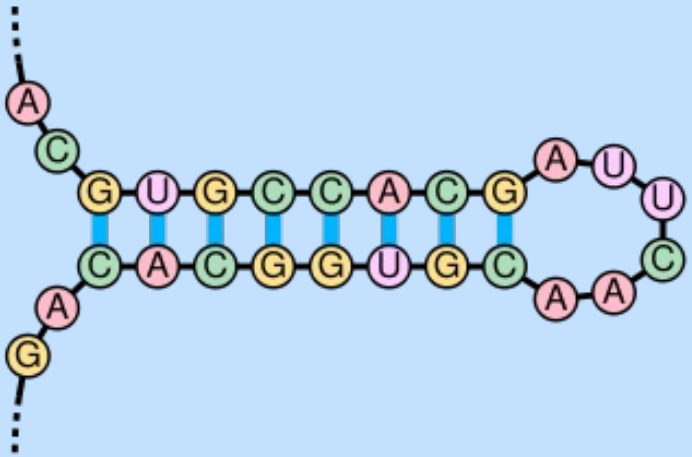
- Directional: from 5’ to 3’:

(upstream) ➡ AACTTGCGATACTCCTA ➡ (downstream)

5’ end 3’ end



RNA (Ribonucleic acid)



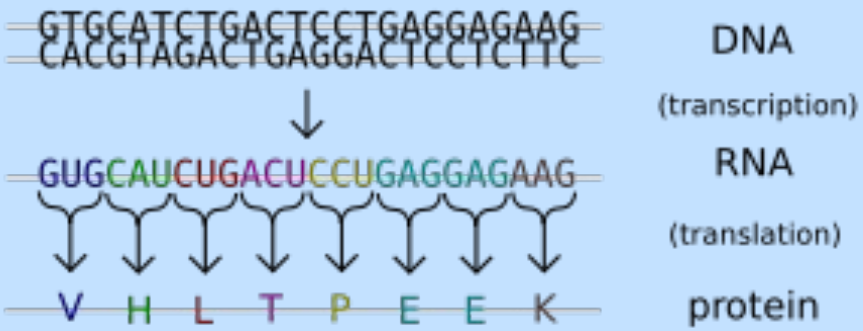
- <u>Bases</u>:
  - Adenine (A)
  - Guanine (G)
  - Cytosine (C)
  - Uracil (U); replaces T

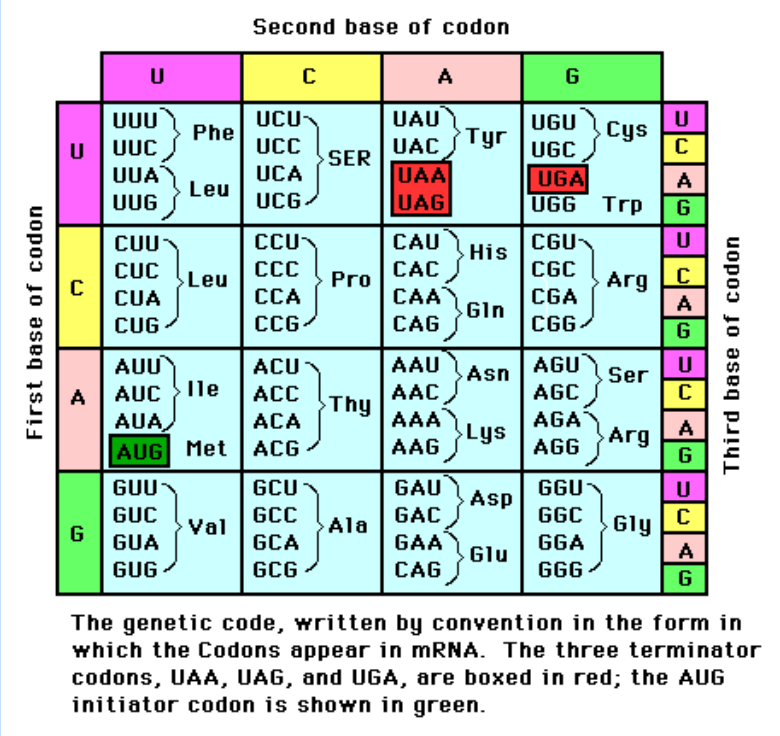
- Oriented from 5′ to 3′.

- 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

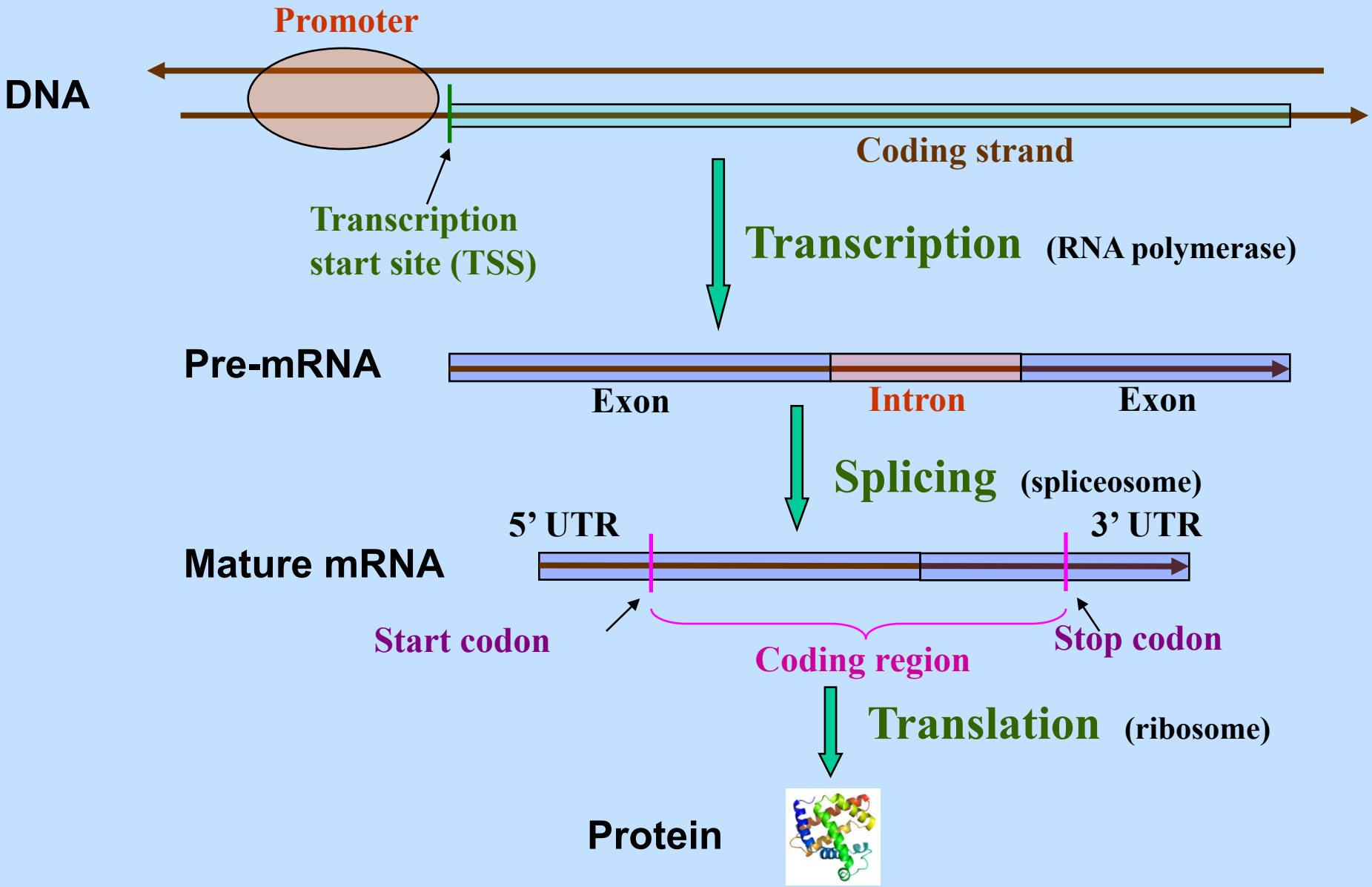




The genetic code, written by convention in the form in which the Codons appear in mRNA. The three terminator codons, UAA, UAG, and UGA, are boxed in red; the AUG initiator codon is shown in green.

http://ntri.tamuk.edu/cell/ribosomes.html

Gene structure (eukaryotes)



Genome sequences

- Many genomes have been sequenced, including those of viruses, microbes, plants and animals.
- <u>Human:</u>
  - 23 pairs of chromosomes
  - 3+ Gbps (bps = base pairs) , only ~3% are genes
  - ~25,000 genes
- <u>Yeast:</u>
  - 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 (RBPs) – proteins that bind to RNA subsequences, called binding sites (BSs)
- BSs of a particular RBP share a common pattern, or motif
- Input:

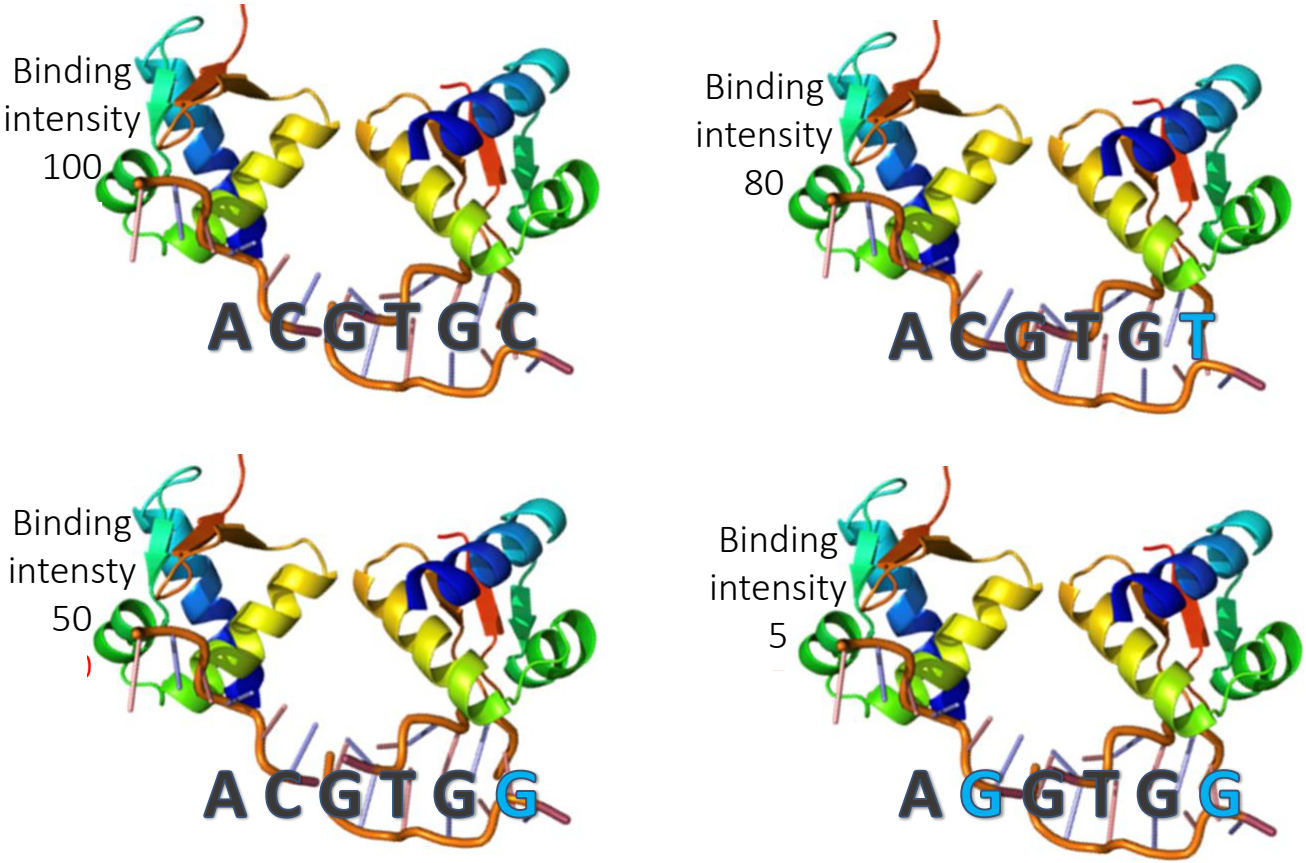
```
0       5       10      15      20      25      30      35      40      45
TCTCATCCGGTGGGAATCACTGCCGCATTTGGAGCATAAACAATGGGGGG
TACGAAGGACAAACACTTTAGAGGTAATGGAAACACAACCGGCGCATAAA
ATACAAACGAAAGCGAGAAGCTCGCAGAAGCATGGGAGTGTAAATAAGTG
GGCGCCTCATTCTCGGTTTATAAGCCAAAACCTTGTCGAGGCAACTGTCA
TCAAATGATGCTAGCCGTCGGAATCTGGCGAGTGCATAAAAAGAGTCAAC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

- Output:

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

http://docs.seqan.de/seqan/1.2/PAGE_Motif+_Search.html

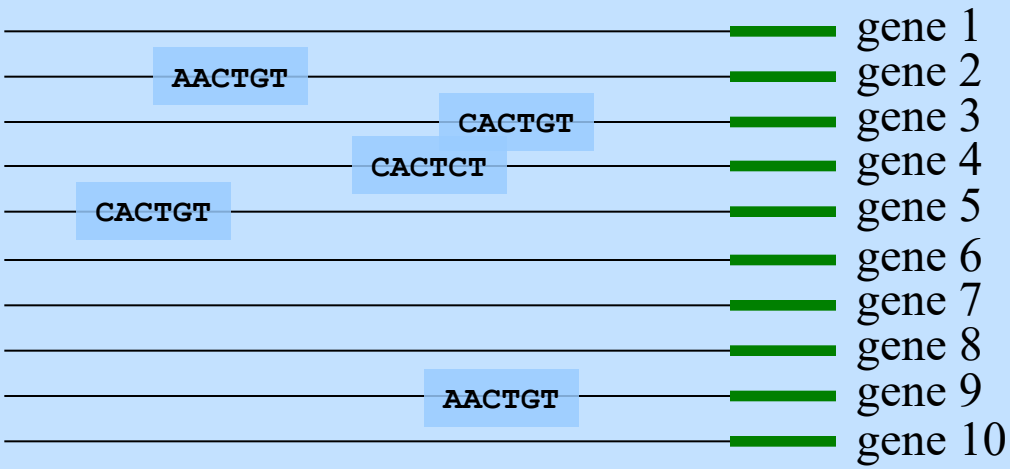
Protein RNA-binding



RBP BS motif models - strings

- Consensus ("degenerate") string:

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



- 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:

	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

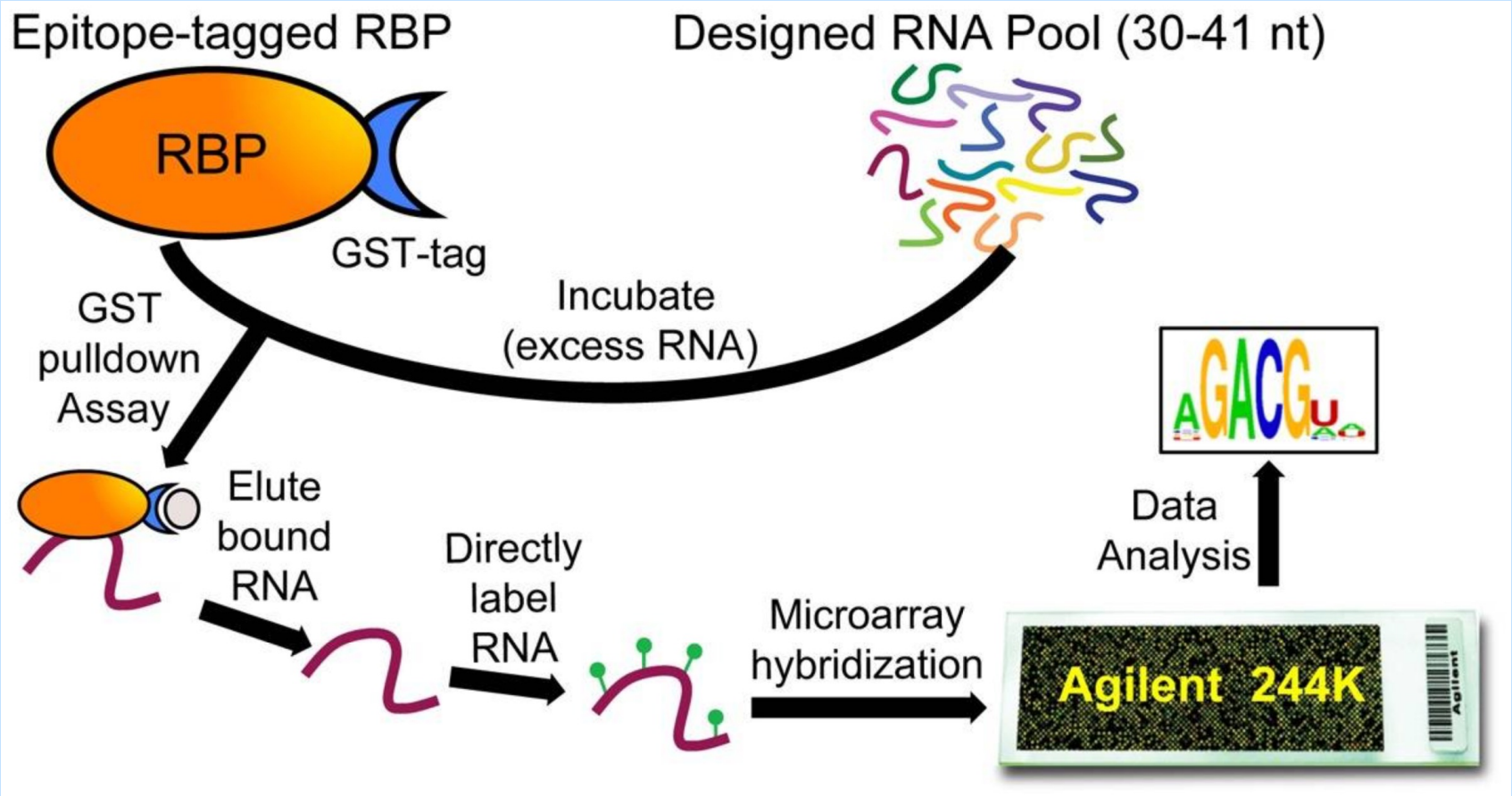
How can we use NN for this model?

- 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

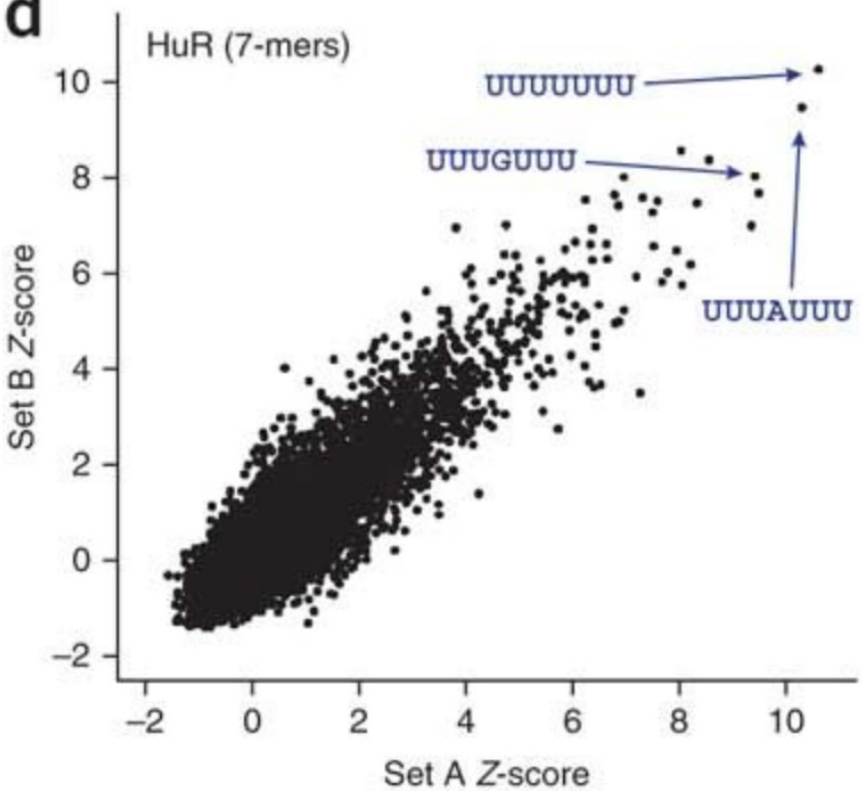
b HuR

AGACCUUCAUGUUCUGUUUCUUUUACUGCUUUGGUUGU
AGAAGUUUGUUUUGGUUUUUGCUUUCUCCUCGUGUGCA
AGAUUUUUUACUUUCAUGUCUAACUCACCUUUGGGGAA
AGGUUUCGUUUUGUGUACUUCUCCCUAGUUUCUUUGGC
AGGAGUUAGUUUGGUUUUACUUUGUGUCCGCCUACGGU
AGAAGUAGGGAAGUUUUUUUCGUUGAUUUAGUGGGCCU
AGAUUAUUUCUCGUUAUAAUAAUGCGUUUAGUGCCCAC
AGAUUCAGAUUUAGCACCACUGUUGUAGUAUGUUUUGU
AGAAUUGAUUUAUUUUUGUUGUCUCACCCAUCUGCGUU
AGACUGCUAACUUAUUUUUUUGUCGUCCUAACGUUGCA

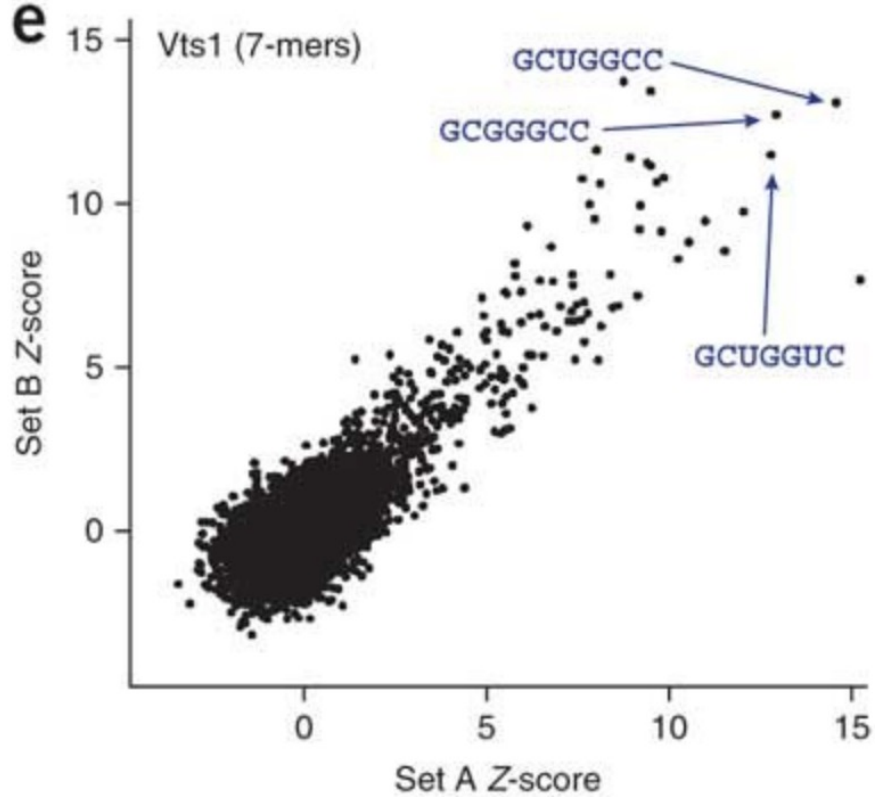
c Vts1

AGGGCCAAUGCGGUGCAGGCGCUGCGUUGGC
AGGUUCAAACGCAUGCGGGUCAAGCUACGAACGCCACU
AGAGAUGGAUUGGGCUGGCACCGGUCCAUC
AGGGCGUUUAACGGCGGGUCCCGUUAGGCGC
AGAGAGUGCAUAGCUGGCUUCUAUGCGCUC
AGGGCCAAUGCGGUGCAGGCGCUGCGUUGGC
AGAGUGUCAGGUACAUAAAGUAGCUCAAUGAGUUCGUU
AGAGUCCAGAACGGCAGGCACGUUUUGGAC
AGAGACAGGAGCUUAUUCAUUGAUCAGCGAUCGCG
AGAGGUUGAGACGGCAGGUCCCGUCUCGGCC

d



e



RNAcompete - implementation

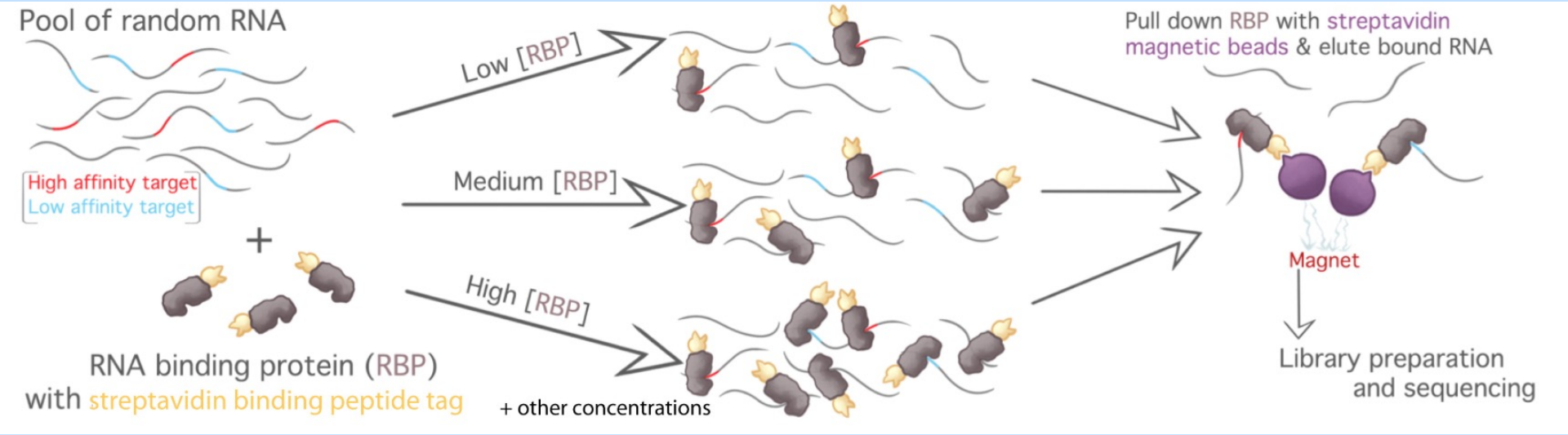
- 30-41nt variable region
- A sequence library covering of all 9-mers, each at least 16 times
- ~240K probes

RNA bind-n-seq

Lambert *et al*, Molecular Cell. 2014

- Start with a pool of random oligos
- With several protein concentrations:
  - Synthesize random oligos
  - Sequence them
  - Let the protein bind to the oligos
  - Filter out bound oligos
  - Sequence them

RNA bind-n-seq



The computational challenge

- Input: RBNS data (4-6 sequence files) of one RBP and a list of RNAcompete probes (1 sequence file)
- Goal: Predict RNA binding intensity for each RNAcompete probe

- Intuition: learning a binding model in one technology to predict binding in another

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 RNA Bind-n-Seq data of different RBPs
- Learn a binding model for each RBP and use it to score RNAcompete probes
- Main challenges:
  - Prediction performance
  - Training runtime

RNA bind-n-seq Input

- 4-6 sequence files with hundred of thousands of lines, each containing an oligo sequence and its number of occurrences

```
<sequence> \t <count> \n
```

```
CGACTAAGGAACTACGACGA    1
AGAACTTGAGACGTTTATAT    1
ACACCCACGCATACAAAACG    1
GCCCTCACGGTGTGAGAGCG    1
TACGTGGAATTATCGAACGC    1
CGTCAACCGCCTACCGATCT    1
CTTCCGCGATCATTAATGTG    1
AACAACAACCTGAAGGACCC    1
CCTAAGAACATGTCTGACAA    1
...
...
...
ATTAGATCCCTAGATGGAGA    1
CTTTCCTGCGAACCCCTTGA    1
GGACAGTTCTACGGGGGATG    1
```

Input (random sequences)

```
CCAGATATAAAAACAGACAG    1
GCGGCAAAGCTATAATTGAT    1
CAGATACGCTGAACTAAATA    1
CTACCCCATAAGCCTATGAG    1
TGAGCGGATTGAGTAATTAA    1
CATGATTTGAAGGGTAGCAG    1
CACACGCTCCAACCCTACCC    1
ACTCTCAAGTCACGTTACTA    1
AACGGGGGGAAAATGCGAGT    1
...
...
...
TGTACTCGAAACCTGCTCCC    1
CAAGCAGCCACCGATTACGC    1
AGATAATGAAAAGGCAAACG    1
```

Concentration 5 nM

```
AGTTAAAATGTAATATTATC    1
ACAGGCTCATTCCACATGGA    1
TGCAGTTATGCATATATTCA    6
ATTAGGAAAGGAGATGAACT    1
GTATAATTTTATGCAAATCA    1
ACCGAATTCCACGGGAGCAT    1
ATCAGTTATGCATATATTCA    2
AACTCATTCAAAGTAAACCT    1
TGTACCAACAAATTATTGTT    1
...
...
...
GCTCATTCTACGTGTAGGCG    1
TGATAATGGCGATCCGCTGC    1
TATCTGTAGACAATTCGATT    1
```

Concentration 80 nM

```
TGCAGTTATGCATATATTCA    8
TGATAATTTTATGCAAAACA    1
ATCAGTTATGCATATATTCA    3
CTCACACGTTCTTTCCGCTT    1
ACGCCTATACCACCCCCTGC    1
CTCCCCCGTCAGCCTTGTCC    2
TGAAGCCTACCTACTAAAAA    1
TCCTAACCCACATGCCATTA    1
ACTCTCCTTCCCTTCTGCAC    1
...
...
...
TGCAAATGAGTTGCAGCATA    2
AACAACTGATTTGCATAATC    2
GCGACATATGAGGTGACACC    1
```

Concentration 1300 nM

RNAcompete Input

File with ~240K lines, each containing a probe sequence of length 30-41.

<sequence> \n

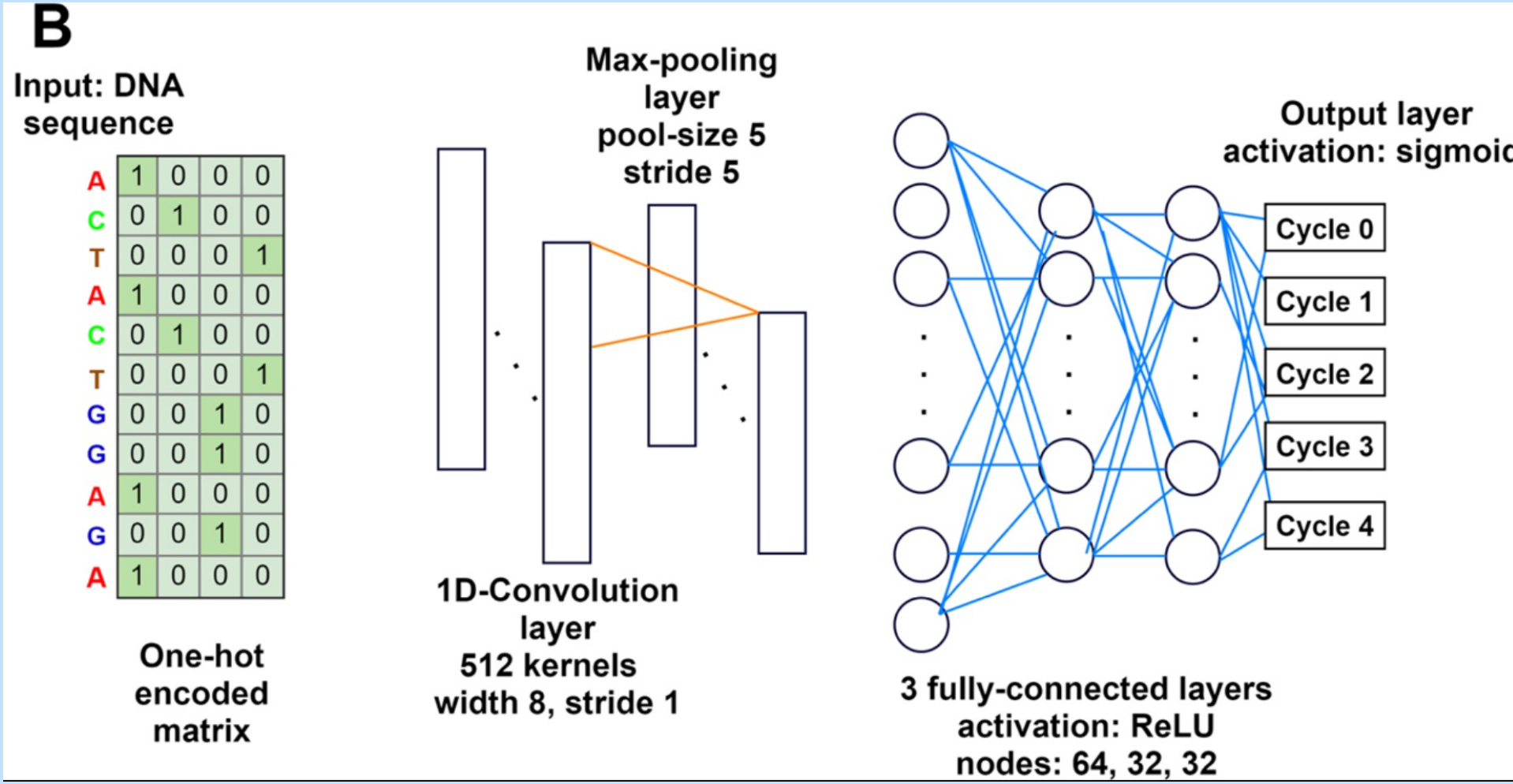
```
GTAATATTACATTCCGTTTAACCTGCGCGCCTACGG
TTTCTCTATGCAGATTAGCTCAATTCCCAATACCTA
GTAATATTCATACGGCCTCGACTCATTACACGGCTT
CTATATGTGCATATATTACCCTACAGGTCGGATAGC
TTCGGATTGGGTGCATAATGAAGCAGGACTATTAAA
CCGAGGTATTTGCATTTCAGTCGTCTTCTAATGAAG
CCCAGCCGGATTTGCATACCACGGACCGAGCAATGG
GAAGGATGGTATGCGTTTATTTGCATAGACAACATC
GTTACTATAATTAGCAGCGCGTTCGGGCCGTTCCAC
...
...
...
ACGTTTGCATTAATTTCCCCCACATTATTATGCGGA
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 (1/2)

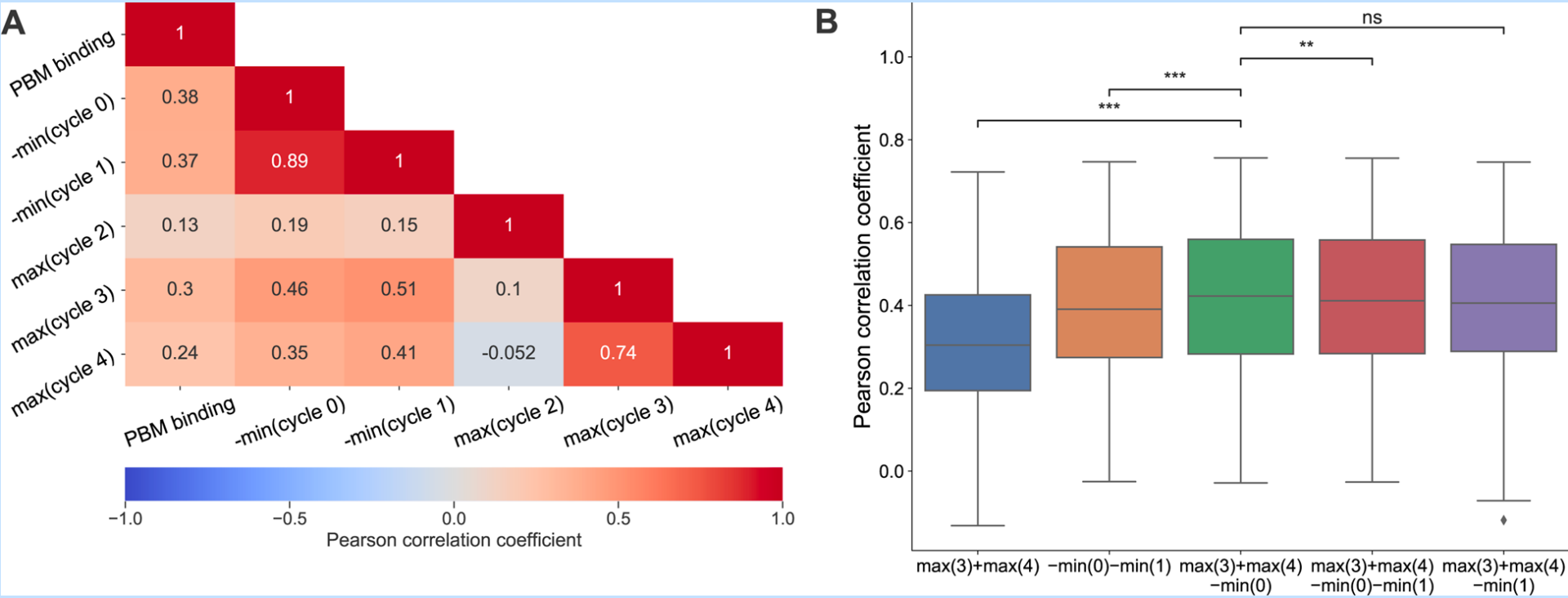
- Train an experiment simulator for each experiment



Asif and Orenstein, DeepSELEX, Bioinformatics 2020

Solution example (2/2)

- Score RNAcompete probes by aggregating sliding windows scores



Asif and Orenstein, DeepSELEX, Bioinformatics 2020

Input schedule

You will be given:

16 training sets (RBNS data + RNAcompete probes with binding intensities).

15 test sets (RBNS data + 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:

  cov(x,y)/(std(x)*std(y))

Implementation

- Input: the 1st argument is the RNAcompete filename, and 4-6 filenames of RBNS files
- 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
- 15 binding intensities files, e.g. RBP19.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

- **RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins**
- Lambert *et al.* Molecular Cell, 2014