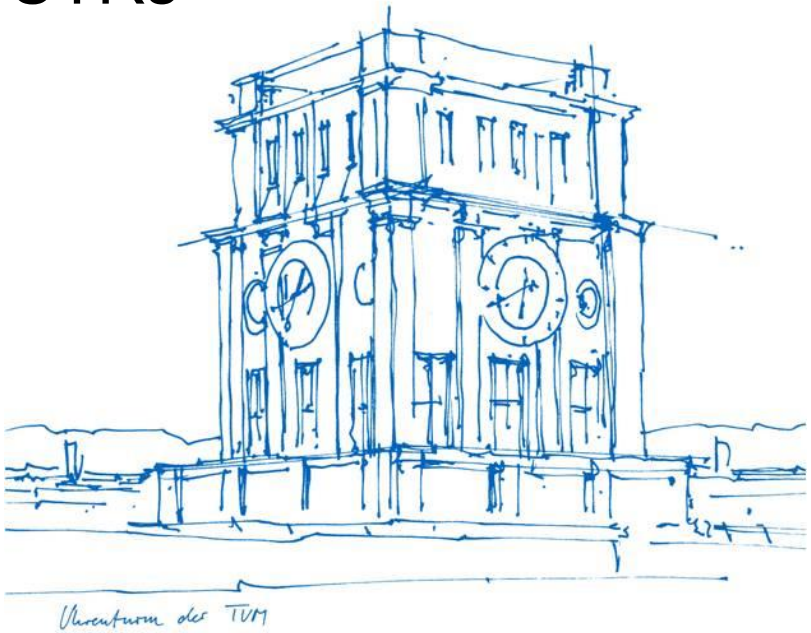
Evaluating the performance of an evolutionary masked language model of mammalian 3‘UTRs

Anas Shahzad, Lukas Groß, Nicole Martin

Supervisors: Dr. Matthias Heinig, Dr. Sergey Vilov

Technical University Munich

Lehrstuhl Computational Molecular Medicine



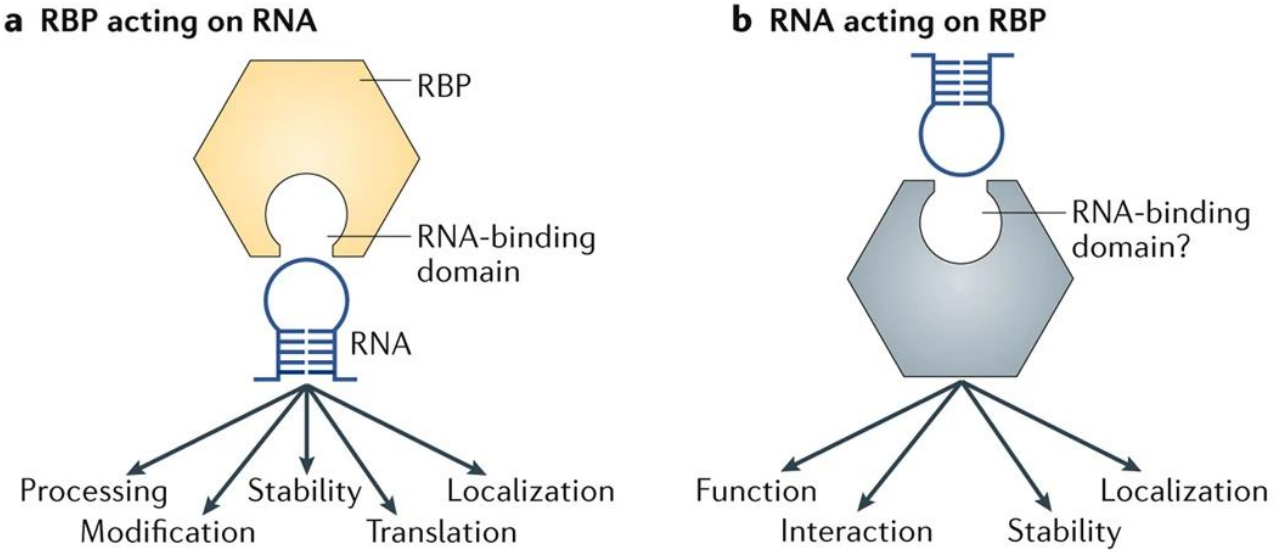
Garching, 18. July 2023

3‘UTR Region

UTR = Untranslated Region

https://upload.wikimedia.org/wikipedia/commons/b/ba/MRNA_structure.svg

RNA Binding Proteins (RBPs)



Hentze, M et al. (2018)

Previous Work

Species-aware DNA language modeling

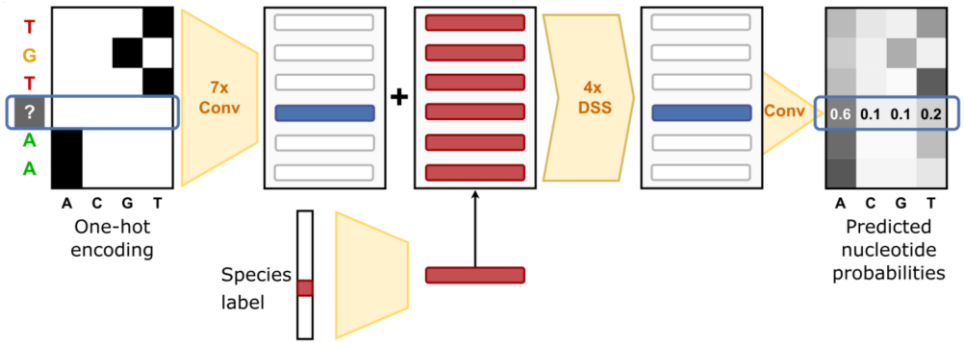
Dennis Gankin*[1], Alexander Karollus*[1], Martin Grosshauser[1], Kristian Klemon[1], Johannes Hingerl[1], Julien Gagneur[1,2,3,4,5]

Modelling of 3'UTR of fungal genomes

- Species-aware MLM
- Species-agnostic MLM
- DNABERT
- Dinucleotide
- 11-mer

→ Species-aware MLM performed the best

Species-aware Masked Language Model (MLM)



Goal of this Project

Continuation of „Species-aware DNA language modelling”

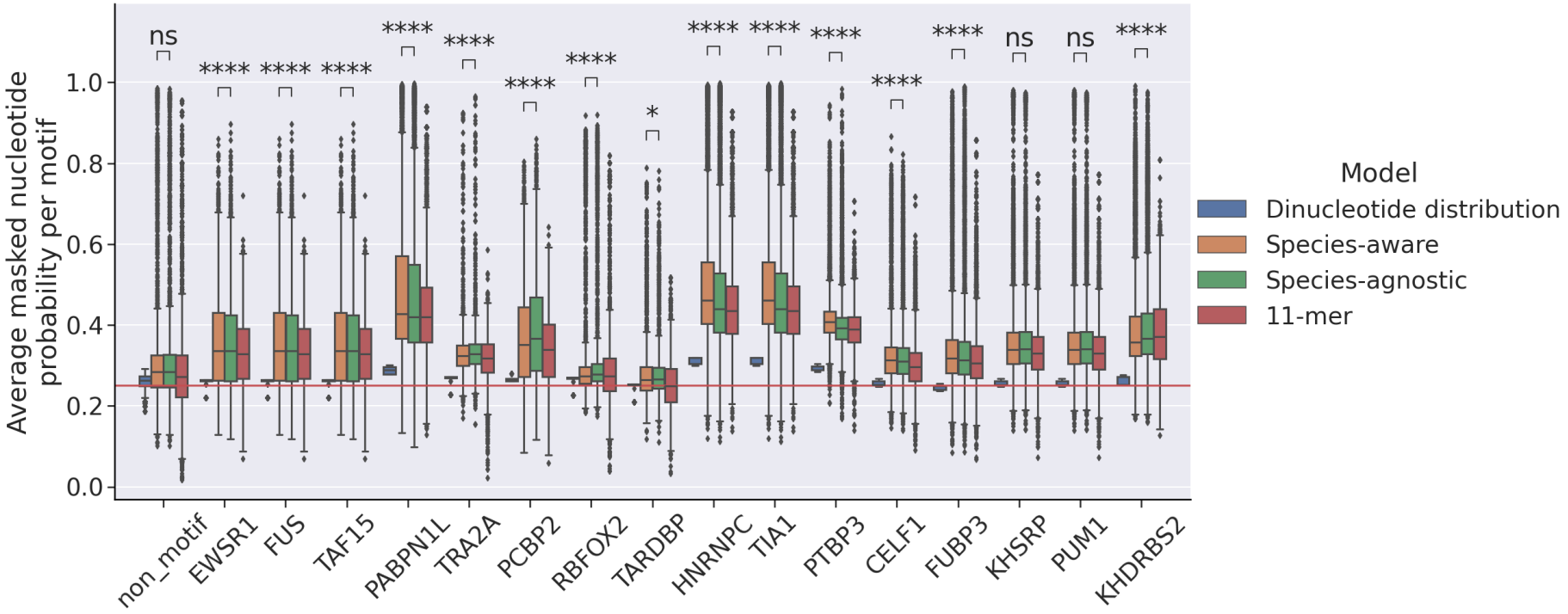
1) Applied these models to human data
   - Species-aware MLM
   - Species-agnostic MLM
   - Dinucleotide
   - 11-mer

Trained on 3’UTR sequences of 240 different mammalian species

→ Homo sapiens as a holdout set

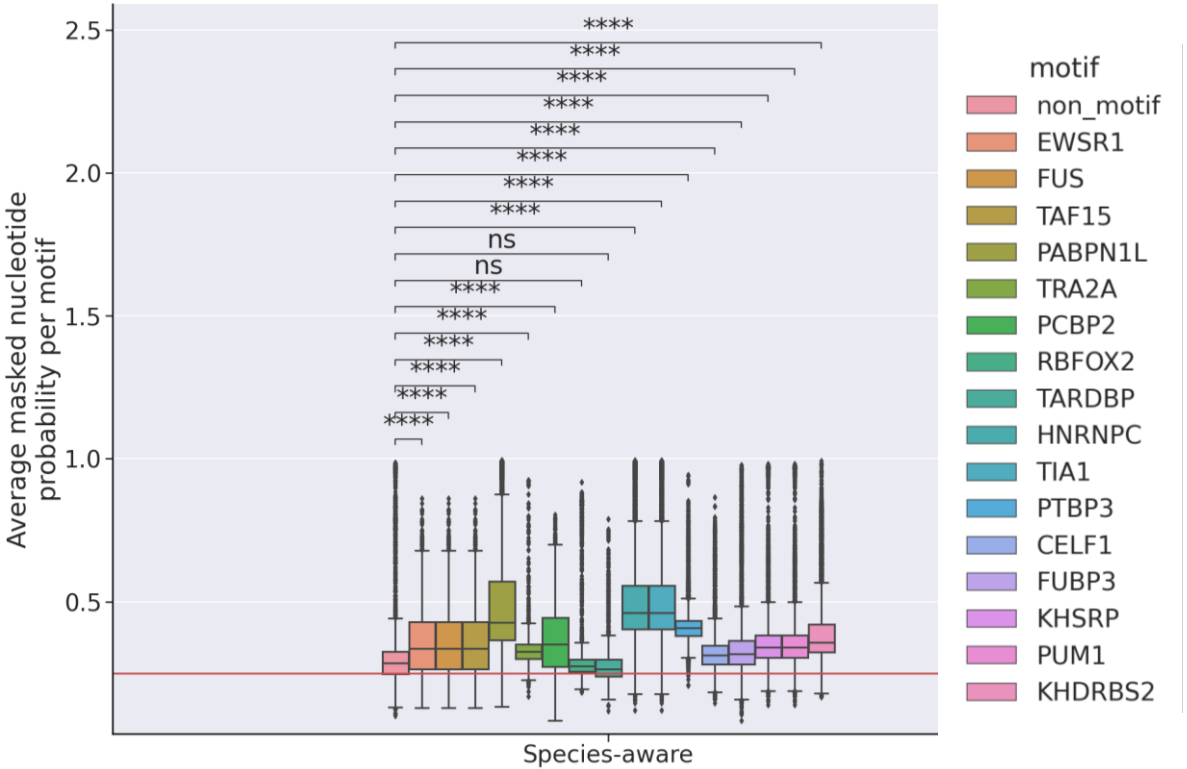
2) Explain performance of species-aware MLM
   - Aligned against the homo sapiens genome → Evolutionary information (conservation)
   - Compare and investigate performance on specific motifs

Compare Motif Reconstruction Ability Across Different Models for Selected RBPs



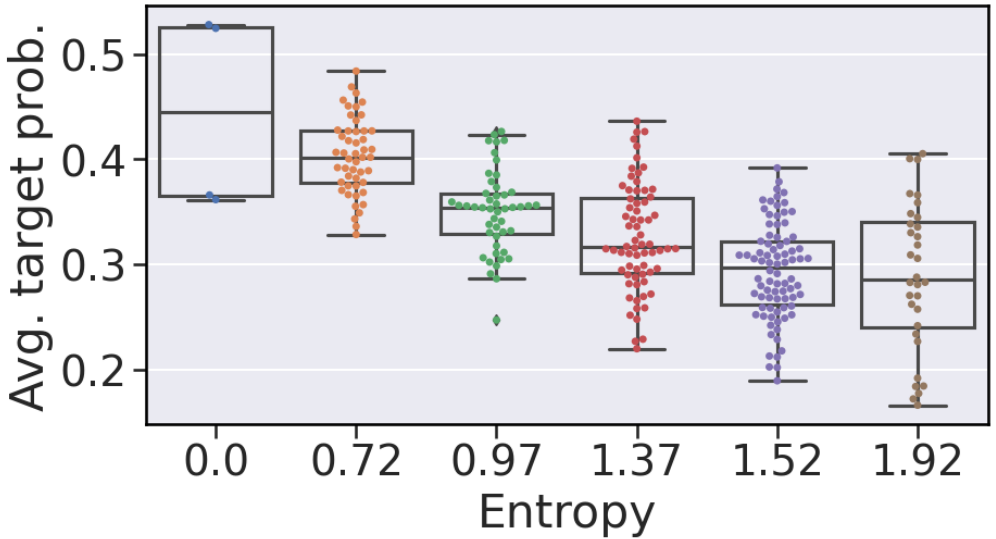
RBP binding sites were selected based on Dominguez et al. 2018

Test MLMs Reconstruction Ability against Random Motifs



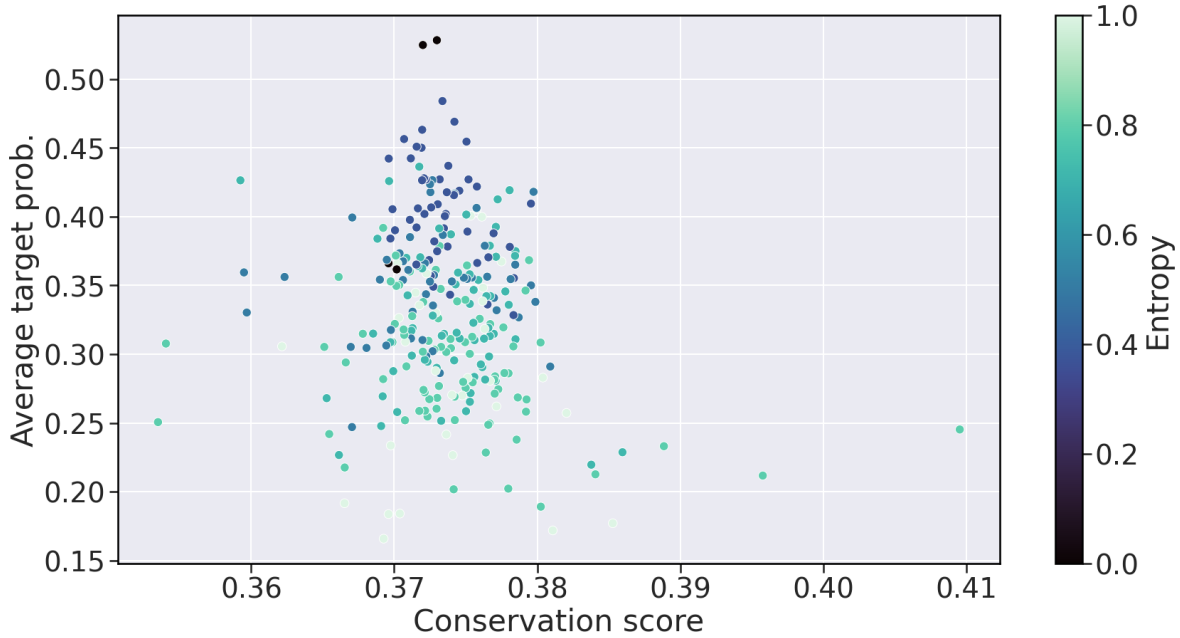
MLMs reconstruct most RBP-Motifs statistically significantly better than random 5-mers

Compare Entropy against Performance of species-aware MLM



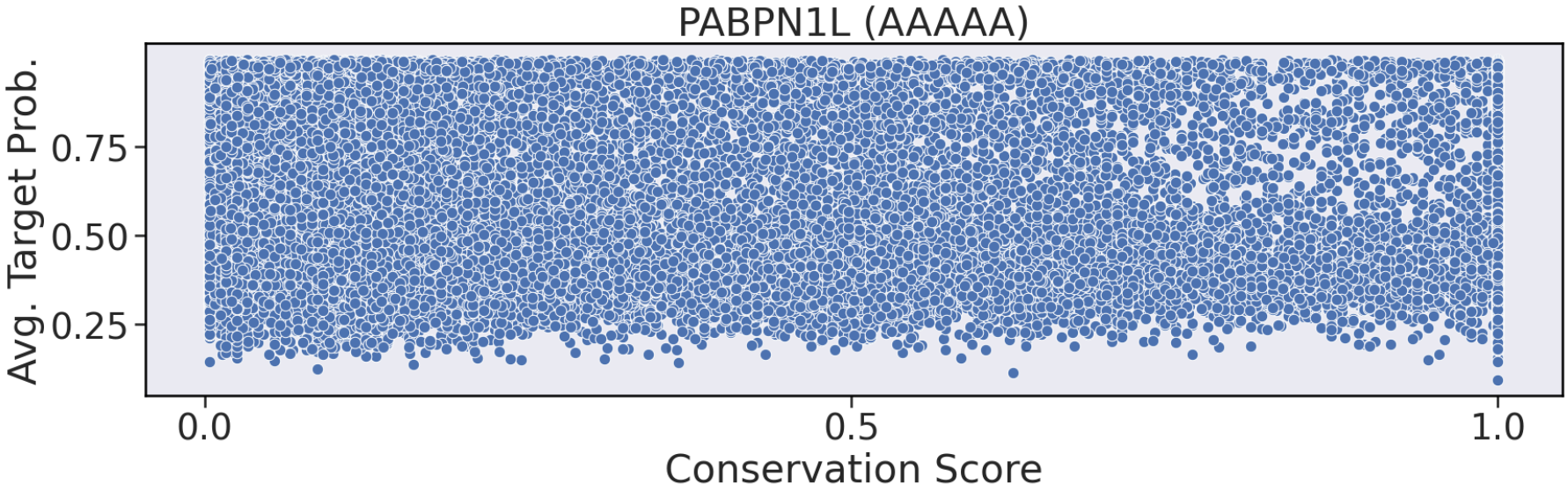
Average target probability decreases with increasing entropy

Compare Conservation Score against Performance of species-aware MLM



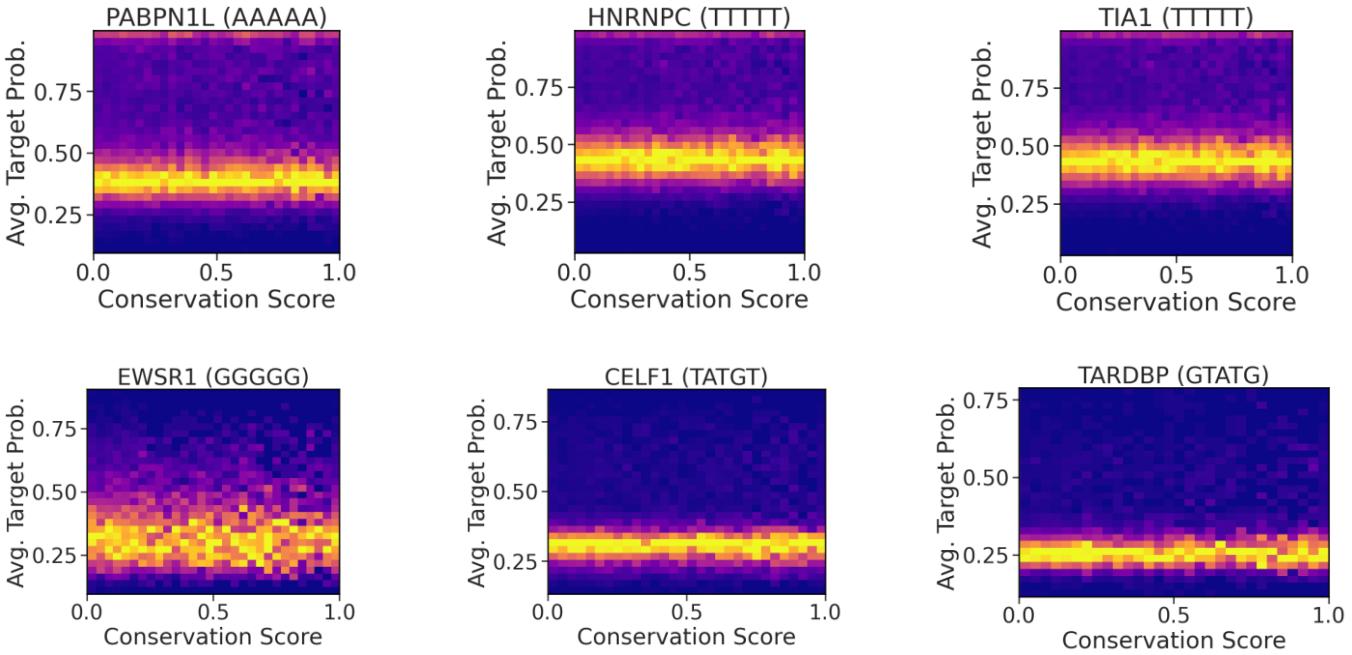
No correlation between conservation score and average target probability

Compare Conservation Score against Performance of PABPN1L for the species-aware MLM



Performance independent of conservation score

Compare Conservation Score against Performance of species-aware MLM



Performance independent of conservation score

Conclusion

Comparison of models:

- MLMs perform similar to 11-mer
- MLMs and 11-mer outperform dinucleotide model

Deep dive into species-aware MLM:

- Entropy correlates negatively to target probability
- No correlation between conservation score and target probability

→ Evolutionary independence of model's predictions

Outlook:

- Further investigation: How does the model come to its predictions?

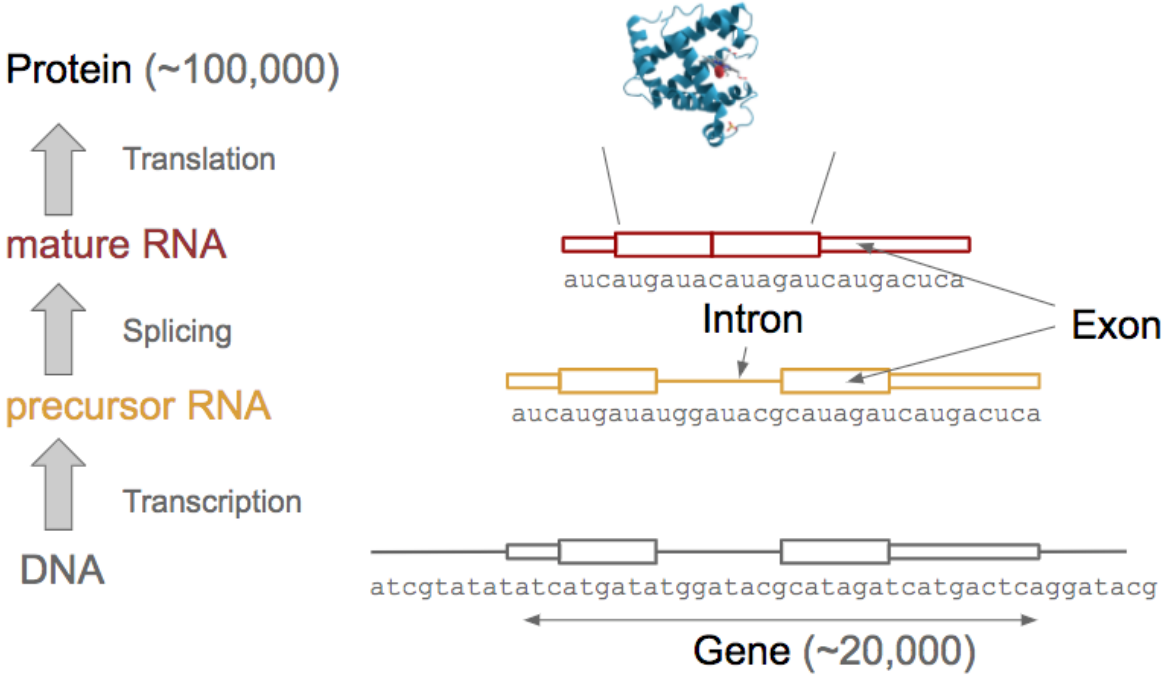
Special thanks to:
- Prof. Dr. Julien Gagneur
- Dennis Gankin, Alexander Karollus
- Dr. Matthias Heinig and Dr. Sergey Vilov
- Anna Chernysheva

Questions?

Resources:

- Colomé-Tatché, M., J. Gagneur, M. Heinig, and A. Marsico (2023). Lecture script for Machine Learning for Regulatory Genomics (IN2393)
- Hong, D. and S. Jeong (2023). “3’UTR diversity: expanding repertoire of RNA alterations in human mRNAs”. In: Molecules and Cells 46.1, pp. 48–56
- Hentze, M. W., A. Castello, T. Schwarzl, and T. Preiss (2018). “A brave new world of RNA-binding proteins”. In: Nature reviews Molecular cell biology 19.5, pp. 327–341.
- Gankin, D., A. Karollus, M. Grosshauser, K. Klemon, J. Hingerl, and J. Gagneur (2023). “Species-aware DNA language modeling”. In: bioRxiv, pp. 2023–01
- Stelzer, G., N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, et al. (2016). “The GeneCards suite: from gene data mining to disease genome sequence analyses”. In: Current protocols in bioinformatics 54.1, pp. 1–30
- Dominguez, D., P. Freese, M. S. Alexis, A. Su, M. Hochman, T. Palden, C. Bazile, N. J. Lambert, E. L. Van Nostrand, G. A. Pratt, et al. (2018). “Sequence, structure, and context preferences of human RNA binding proteins”. In: Molecular cell 70.5, pp. 854–867

Biological Background



Protein	Top 5-mer motif
EWSR1	GGGGG
FUS	GGGGG
TAF15	GGGGG
HNRNPL	ACACA
PABPN1L	AAAAA
TRA2A	GAAGA
HNRNPL	ACACA
PABPN1L	AAAAA
TRA2A	GAAGA
PCBP2	CCCCC
RBFOX2	GCATG
TARDBP	GTATG
HNRNPC	TTTTT
TIA1	TTTTT
PTBP3	TTTCT
CELF1	TATGT
FUBP3	TATAT
KHSRP	TGTAT
PUM1	TGTAT
KHDRBS2	ATAAA

Motif	Alphabet sequence	Description Stelzer et al. 2016
RBM22	TCCGG	The encoded protein may play a role in cell division and may be involved in pre-mRNA splicing. *GeneCards: RBM22 Gene* 2023
RBM4	GCGTA	Enables RNA binding activity and cyclin binding activity. *GeneCards: RBM4 Gene* 2023
EIF4G2	GGTCG	Appears to play a role in the switch from cap-dependent to IRES-mediated translation during mitosis, apoptosis and viral infection. *GeneCards: EIF4G2 Gene* 2023
RBM4B	ACGCG	Enables RNA binding activity. Predicted to be involved in entrainment of circadian clock by photoperiod; mRNA splicing, via spliceosome; and regulation of gene expression. *GeneCards: RBM4B Gene* 2023
RBM45	ACGCG	This gene encodes a member of the RNA recognition motif (RRM)-type RNA-binding family of proteins. This protein has been localized to inclusion bodies in the brain and spinal cord of amyotrophic lateral sclerosis and Alzheimer's patients. *GeneCards: RBM45 Gene* 2023