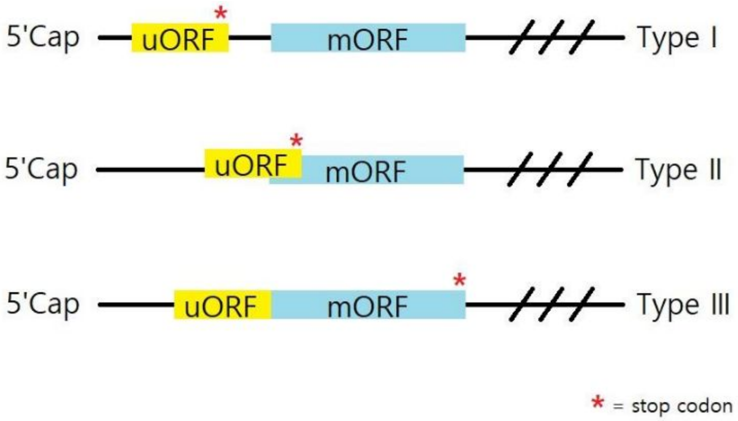
Structural characterization of uORFs using Deep Learning prediction methods

Content

1. Locate uORFs
2. Statistics on uORFs
3. Gene Ontology
4. Secondary structure prediction - Simple CNN
5. Secondary structure prediction - AlphaFold
6. Simple CNN vs AlphaFold
7. Prediction of intrinsic disorder

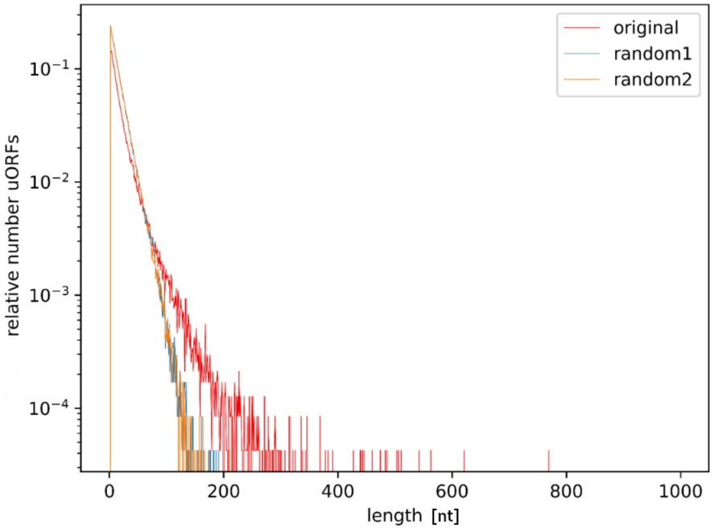
1. Locate uORFs

- Present in 5’ UTR
- Used arabidopsis thaliana dataset
- Three Types (I, II, II)
- ~50% of genes have one or more uORFs
- at 23,000 genes there was 60,000 uORFs
- 90% (Type I) 9.99% (Type II) 0.01% (Type III)



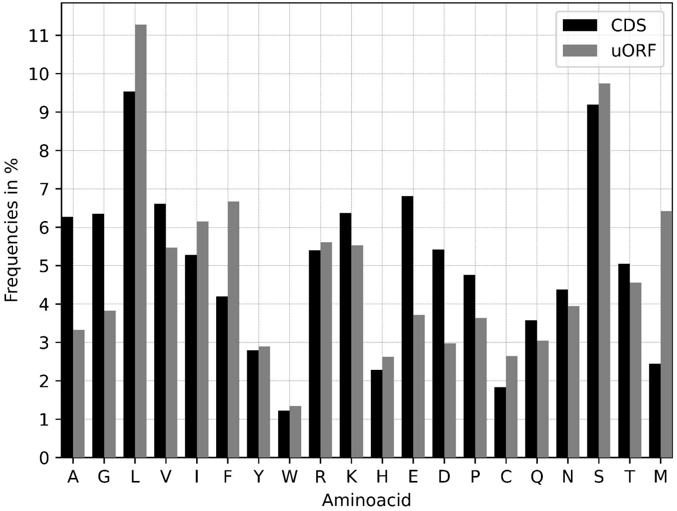
2. Statistics on uORFs 1/2

- Length of uORFs compared to uORFs found in shuffled sequences



2. Statistics on uORFs 2/2

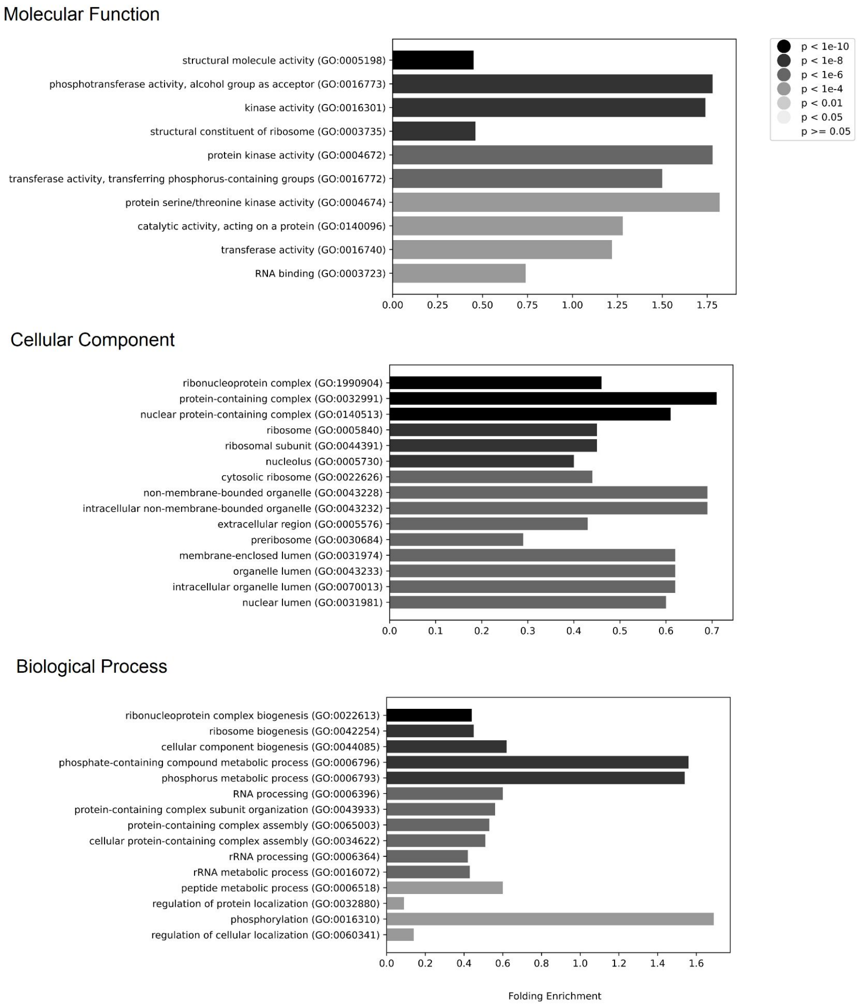
- Frequencies of amino acid residues in CDS/uORF sorted by physico-chemical properties



3. Gene Ontology

- Compared genes with

  no uORFs vs. at least one of uORFs



Molecular Function

Cellular Component

Biological Process

4. Secondary structure prediction - Simple CNN 1/3

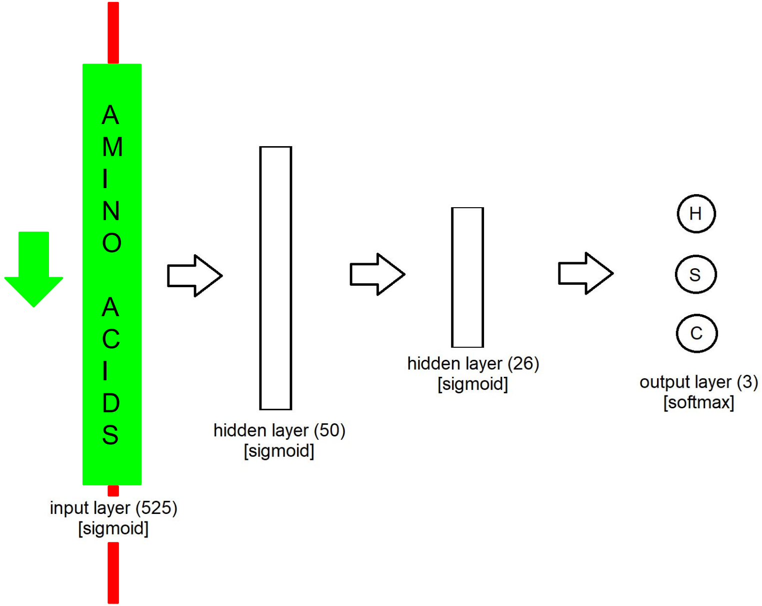
- Used data from Protein Data Bank (.pdb) files
- ~8000 proteins converted to dssp files
- Dssp files sorted to three state classification: Helix, Sheet, Coil
- Example:

Amino acid sequence	Ile	Leu	Leu	Glu	Asp	Pro	…
Structure	H	H	H	C	C	S	…

4. Secondary structure prediction - Simple CNN 2/3

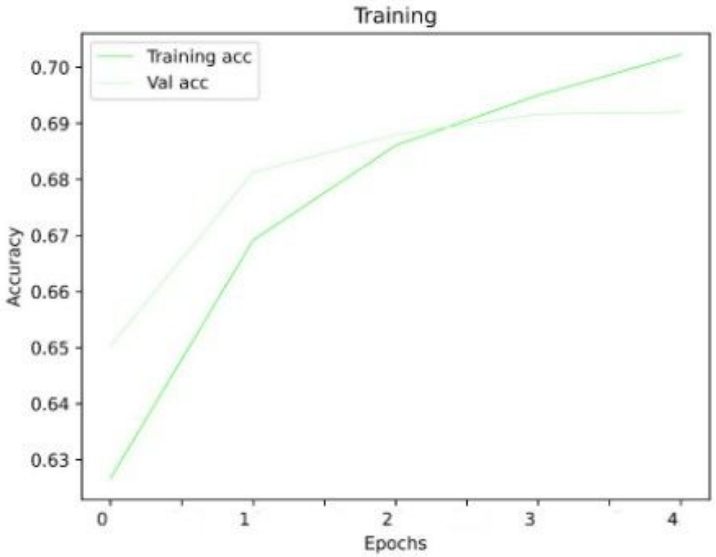
- One-hot-encoding for all different amino acids (Ile=0b00001, Leu=0b00010, …)
- Using sliding window to iterate over **all sequences** during training
- Output class shows predicted class for

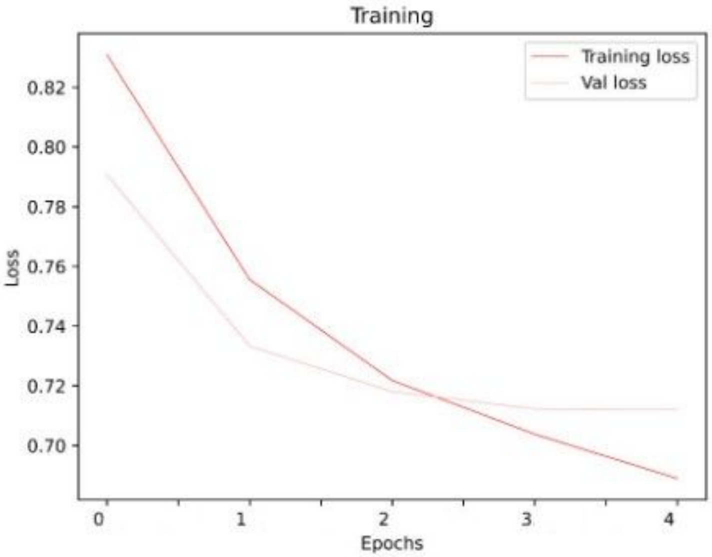
  the central amino acid in the **sliding window**



4. Secondary structure prediction - Simple CNN 3/3

- Reached an **accuracy of 0.69** in predicting between 3 possible output classes



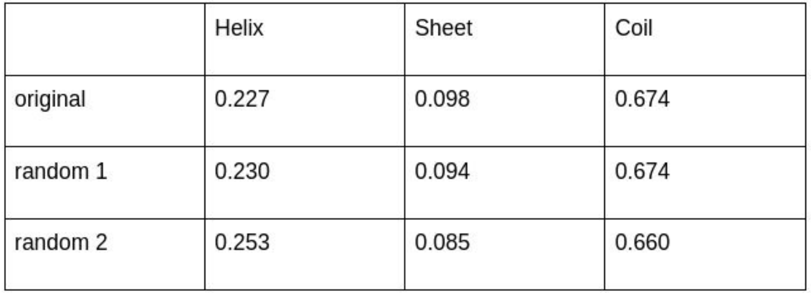


5. Secondary structure prediction - AlphaFold

- Google Colab (modified to run multiple sequences at one time)
- Returns .pdb file -> Convert to dssp -> Convert to 3 classes (H, S, C)
- Count occurrence of structures

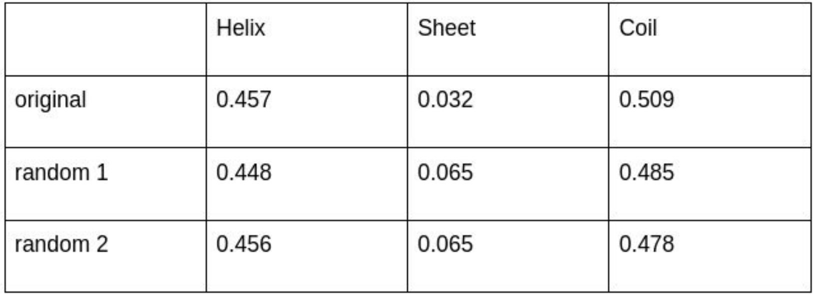
6. Simple CNN vs. AlphaFold

Simple CNN (1000 uORFs)



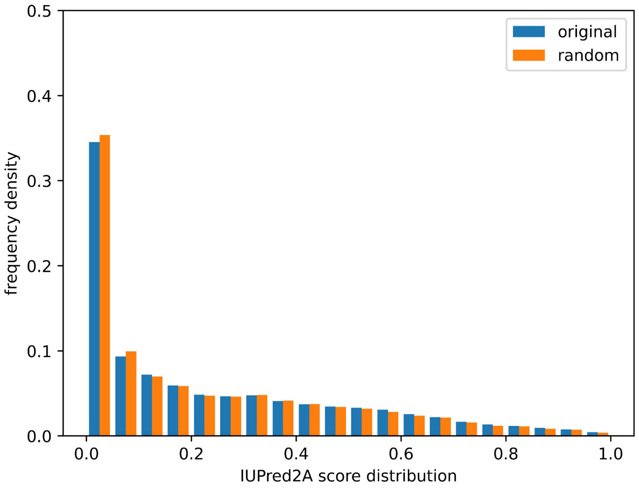
	Helix	Sheet	Coil
original	0.227	0.098	0.674
random 1	0.230	0.094	0.674
random 2	0.253	0.085	0.660

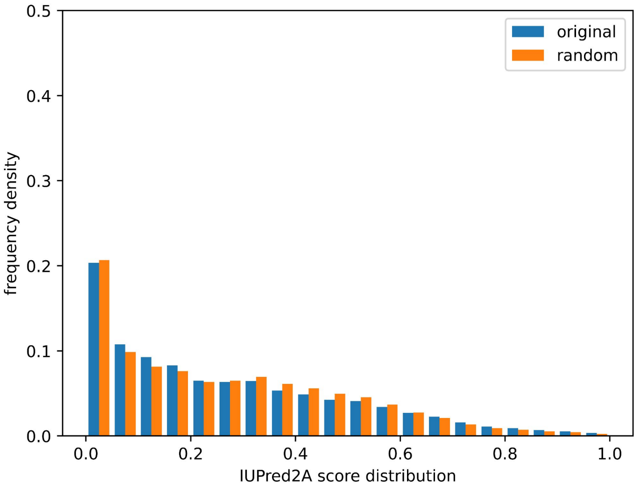
AlphaFold (200 uORFs)



	Helix	Sheet	Coil
original	0.457	0.032	0.509
random 1	0.448	0.065	0.485
random 2	0.456	0.065	0.478

7. Prediction of intrinsic disorder


uORF


CDS

Conclusion

- Observed unique properties of uORFs
- Results show uORFs **maybe** try to avoid defined structures
- Machine learning approaches trained by proteins!
- Further research needed (different species, etc.)