Review of NLP Applications to Protein Science

Somadina Mbadiwe

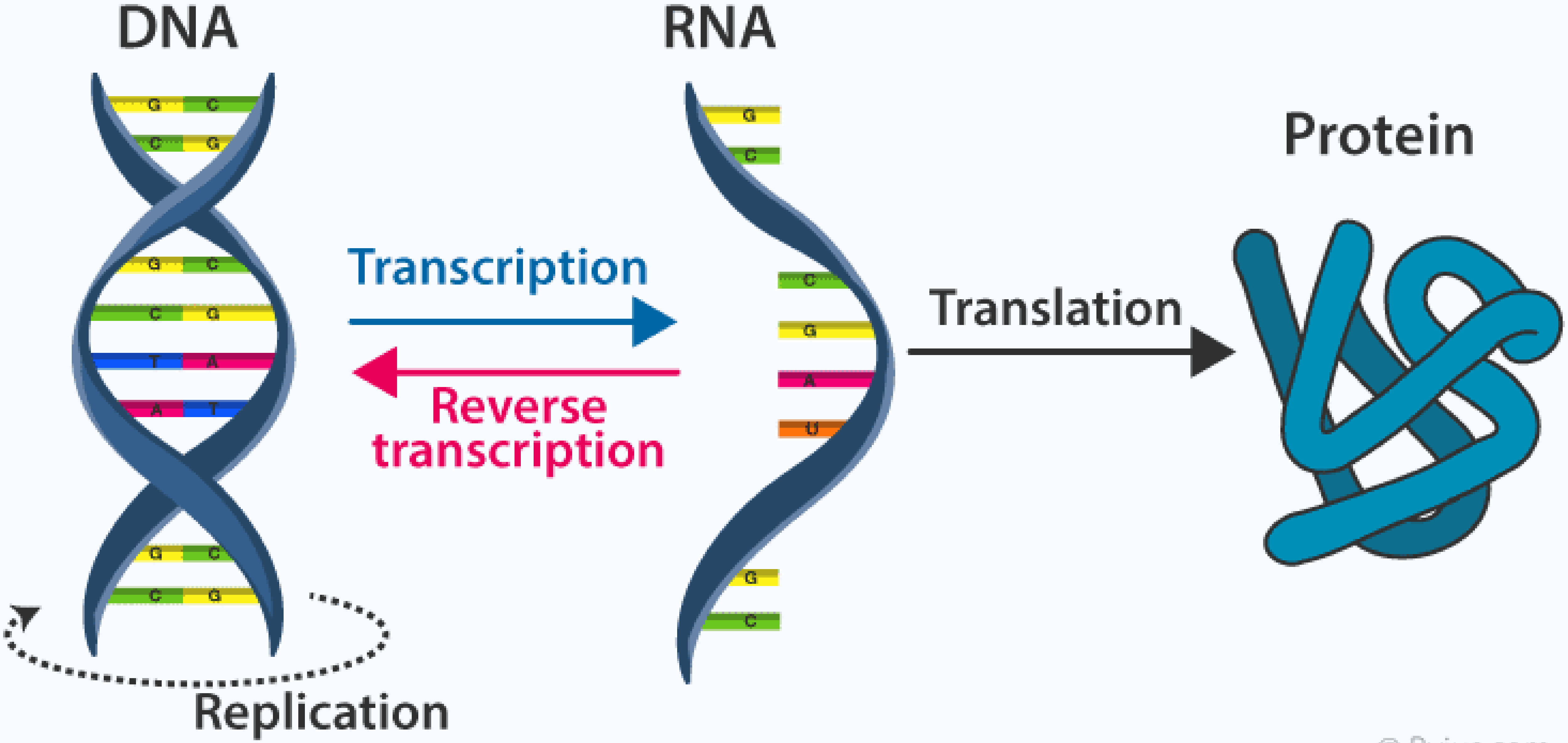
CS 593A – Fall 2020 – WVU

November 30, 2020

Why Proteins?

- Responsible for almost all biological processes critical to life.
  - Hemoglobin: carries oxygen to your cells,
  - Insulin: regulates blood glucose level
  - Rhodopsin: required for vision in dim light
- Useful in industrial settings
  - Enzymes (e.g.: Proteases, Amylases, Lipases, Cellulase) break down stains into smaller pieces to make stains easier to remove.

CENTRAL DOGMA : DNA TO RNA TO PROTEIN



DNA

RNA

Protein

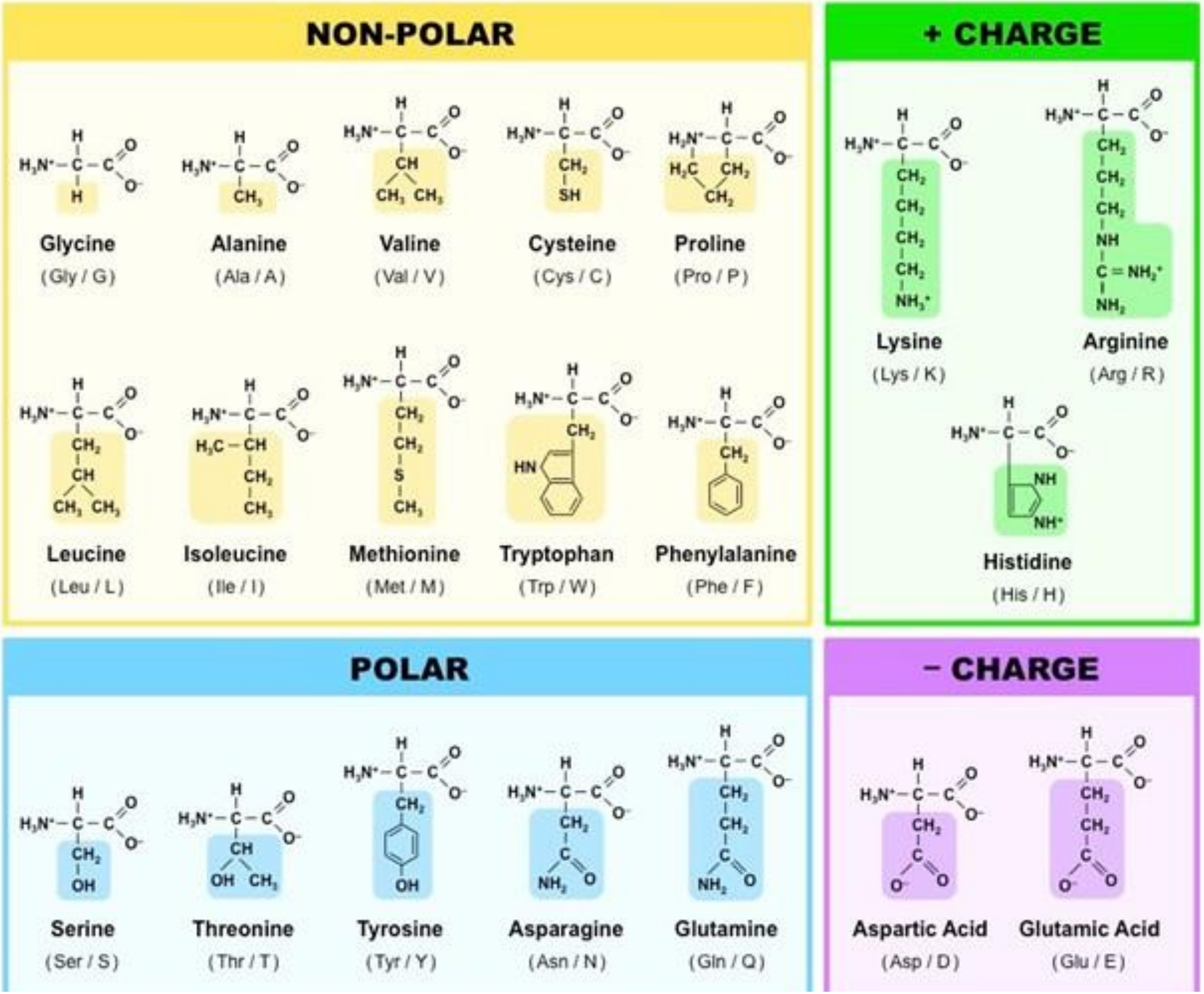
Transcription

Reverse transcription

Translation

Replication

Amino Acids are the building blocks of proteins

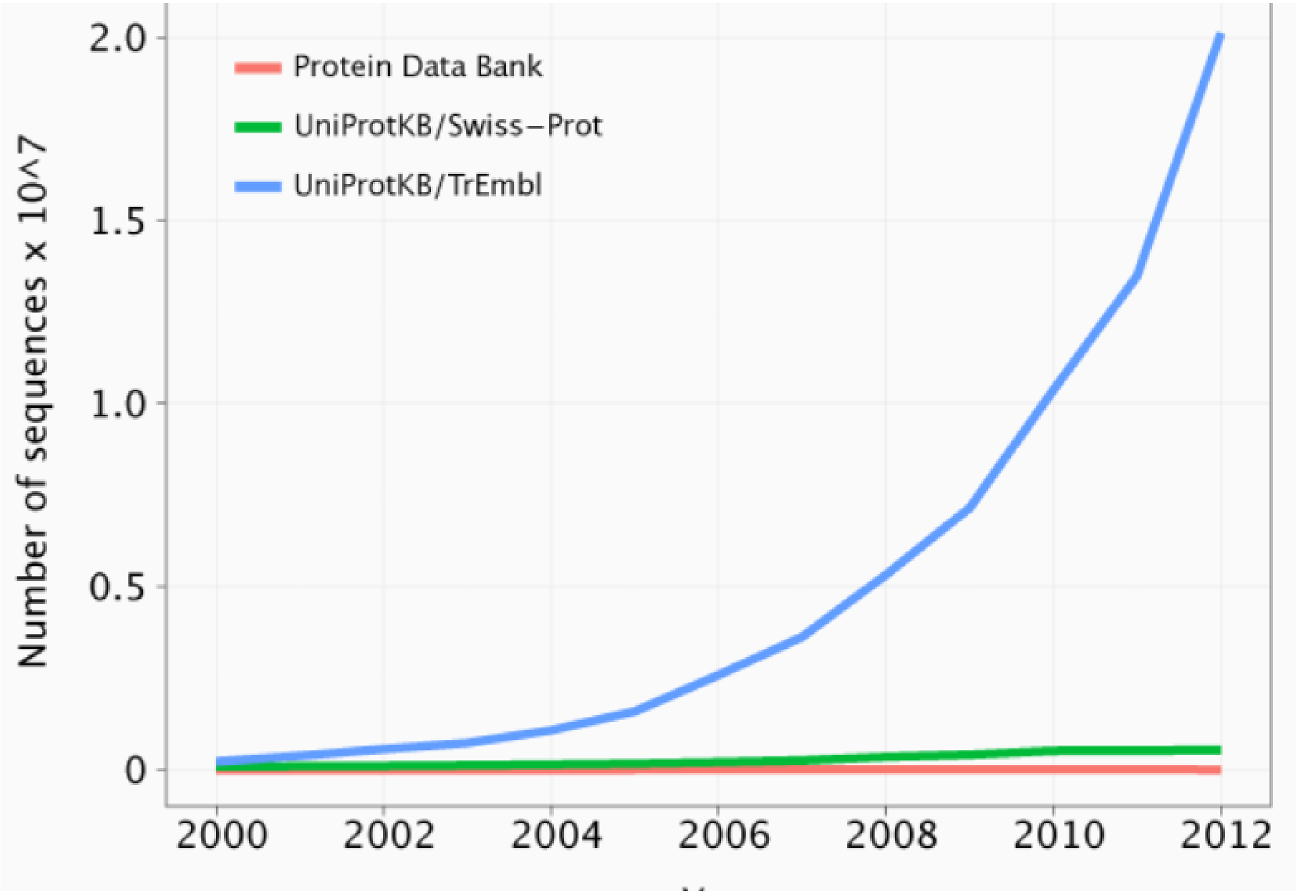


Protein As A Language

- Represented as a sequence of characters
- 25 letters for its “alphabet”
  - 20 characters for the standard (essential) amino acids,
  - 2 for the non-standard amino acids selenocysteine and pyrrolysine,
  - 2 for ambiguous amino acids, and
  - 1 for when the amino acid is unknown
- …But no concept of “words”, “sentences”, “paragraph” in the way we know it. Just letters.

Why NLP is great for protein studies

- Sequence – Structure gap



Observe how the green line (the protein sequences associated with a known or predicted function) is very close to the red line (the number of known protein structures). However, there is a growing gap between the red and the blue line (the number of protein sequences).

Protein-Based NLP Models

- ProtTrans (Elnaggar, 2020 (published this Summer))
  - Trained on data containing up to 393 billion amino acids (words) from 2.1 billion protein sequences (22- and 112-times the entire English Wikipedia)
  - Demos on HuggingFace:
    - Uniref100: https://huggingface.co/Rostlab/prot_bert
    - BFD: https://huggingface.co/Rostlab/prot_bert_bfd
    - BFD (T5, XL): https://huggingface.co/Rostlab/prot_t5_xl_bfd

Protein Datasets

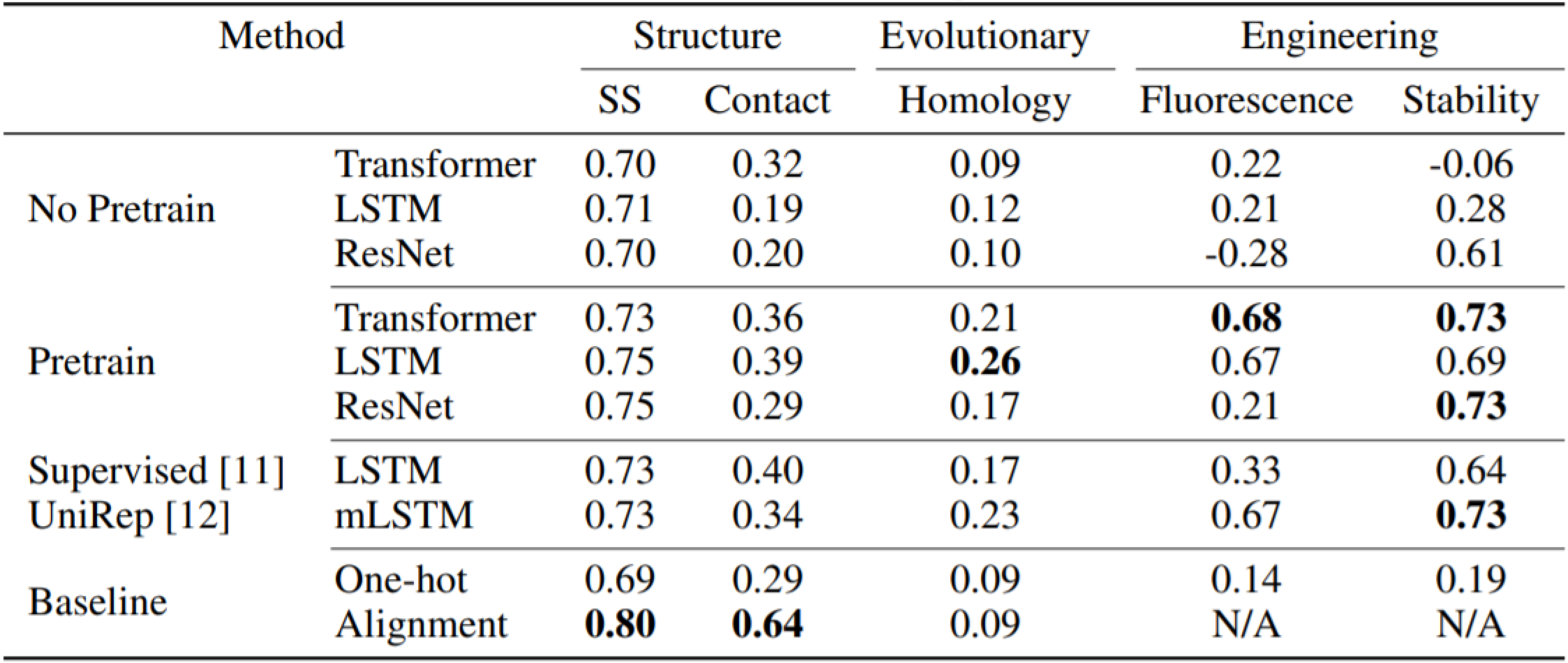
- ProteinNet (AlQuraishi, 2019)
    - Standardized data set for machine learning of protein structure.
    - Provides protein sequences, structures (secondary and tertiary), multiple sequence alignments (MSAs), position-specific scoring matrices (PSSMs), and standardized training / validation / test splits.
    - Publicly available
- Protein Data Bank (PDB)
- UniProt
- Pfam
- Uniref100

Performance Evaluation: TAPE

- TAPE: **T**asks **A**ssessing **P**rotein **E**mbeddings (Roshan et al, NeurIPS 2019)
- Includes five (5) biologically relevant supervised tasks that evaluate the performance of learned protein embeddings.
  - Task 1: Secondary Structure (SS) Prediction (Structure Prediction Task)
  - Task 2: Contact Prediction (Structure Prediction Task)
  - Task 3: Remote Homology Detection (Evolutionary Understanding Task)
  - Task 4: Fluorescence Landscape Prediction (Protein Engineering Task)
  - Task 5: Stability Landscape Prediction (Protein Engineering Task)

Performance Evaluation: TAPE

Table 2: Results on downstream supervised tasks



Method		Structure		Evolutionary	Engineering	
		SS	Contact	Homology	Fluorescence	Stability
No Pretrain	Transformer	0.70	0.32	0.09	0.22	-0.06
	LSTM	0.71	0.19	0.12	0.21	0.28
	ResNet	0.70	0.20	0.10	-0.28	0.61
Pretrain	Transformer	0.73	0.36	0.21	**0.68**	**0.73**
	LSTM	0.75	0.39	**0.26**	0.67	0.69
	ResNet	0.75	0.29	0.17	0.21	**0.73**
Supervised [11]	LSTM	0.73	0.40	0.17	0.33	0.64
UniRep [12]	mLSTM	0.73	0.34	0.23	0.67	**0.73**
Baseline	One-hot	0.69	0.29	0.09	0.14	0.19
	Alignment	**0.80**	**0.64**	0.09	N/A	N/A

My Experiments: Plan

- Task
  - Predicting secondary structure (Q8) from protein sequence.
    - I’ll use 9 classes instead of 8; the ninth being for ‘Unknown / Unlabeled’.
- Dataset:
  - NetSurfP 2.0 (Klausen et al, 2019) for finetuning / training
  - CB512, CASP12, TS115 datasets (HHBlits) for testing.
- Model
  - Prot-Bert and Prot-Bert-BFD models from ProtTrans (Elnaggar et al, 2020)

Comparing Result on **CB513** Dataset.
Task: Secondary Structure Prediction (Q8)

Model	Parameters	Accuracy (%)
UniRef (Alley et al, 2019)	18M	58.4
SeqVec (Heinzinger et al, 2019)	93M	62.1
TAPE (Rao et al, 2019)	38M	58.0
NetSurfp2.0 (Klausen et al, 2019)		72.0
ProtBert (from ProtTrans)	420M	66.0
ProtBert-BFD (from ProtTrans)	420M	70.0
Prot-Bert (ours; finetuned on NetSurfp2 dataset; MAX_LEN = 512)	420M	**79.5**
Prot-Bert-BFD (ours; finetuned on NetSurfp2 dataset; MAX_LEN = 512)	420M	**79.5**

Comparing Result on **CASP12** Dataset.
Task: Secondary Structure Prediction (Q8)

Model	Parameters	Accuracy (%)
UniRef (Alley et al, 2019)	18M	
SeqVec (Heinzinger et al, 2019)	93M	
TAPE (Rao et al, 2019)	38M	58.0
NetSurfp2.0 (Klausen et al, 2019)		71.1
ProtBert (from ProtTrans)	420M	63.0
ProtBert-BFD (from ProtTrans)	420M	65.0
Prot-Bert (ours; finetuned on NetSurfp2 dataset; MAX_LEN = 512)	420M	**81.0**
Prot-Bert-BFD (ours; finetuned on NetSurfp2 dataset; MAX_LEN = 512)	420M	**81.0**

Comparing Result on **TS115** Dataset
Task: Secondary Structure Prediction (Q8)

Model	Parameters	Accuracy (%)
UniRef (Alley et al, 2019)	18M	
SeqVec (Heinzinger et al, 2019)	93M	
TAPE (Rao et al, 2019)	38M	58.0
NetSurfp2.0 (Klausen et al, 2019)		74.4
ProtBert (from ProtTrans)	420M	72.0
ProtBert-BFD (from ProtTrans)	420M	73.0
Prot-Bert (ours; finetuned on NetSurfp2 dataset; MAX_LEN = 512)	420M	**81.7**
Prot-Bert-BFD (ours; finetuned on NetSurfp2 dataset; MAX_LEN = 512)	420M	**81.7**

Info and Thoughts on the results

- Used default hyperparameter values from the LM but trained for 3 epochs instead of 5.
- Max sequence length cut to 512
  - Recommended max length is 1024. It’s surprising I got much better result with max length of 512. The saying “Less is More” seems to apply here.
- BFD-trained LM performed same Uniref-trained LM
  - BFD dataset is about 8x bigger than Uniref. Yet it didn’t make any difference in the SSP-Q8 task. Maybe it makes a difference when learning other protein tasks.
- Most of the design decisions (e.g. no hyperparameter search; training for only 3 epochs) are due to resource constrains.

References

- Roshan Rao et al. Evaluating protein transfer learning with TAPE. In Advances in Neural Information Processing Systems, 2019.
- Neil C. Jones and Pavel A. Pevzner. An introduction to bioinformatics algorithms. MIT Press, 2004.
- Mohammed AlQuraishi. ProteinNet: a standardized data set for machine learning of protein structure. BMC Bioinformatics, 20(1):311, Jun 2019.
- Ahmed Elnaggar, et al. ProtTrans: Towards cracking the language of life’s code through self-supervised deep learning and high performance computing. bioRxiv, 2020.