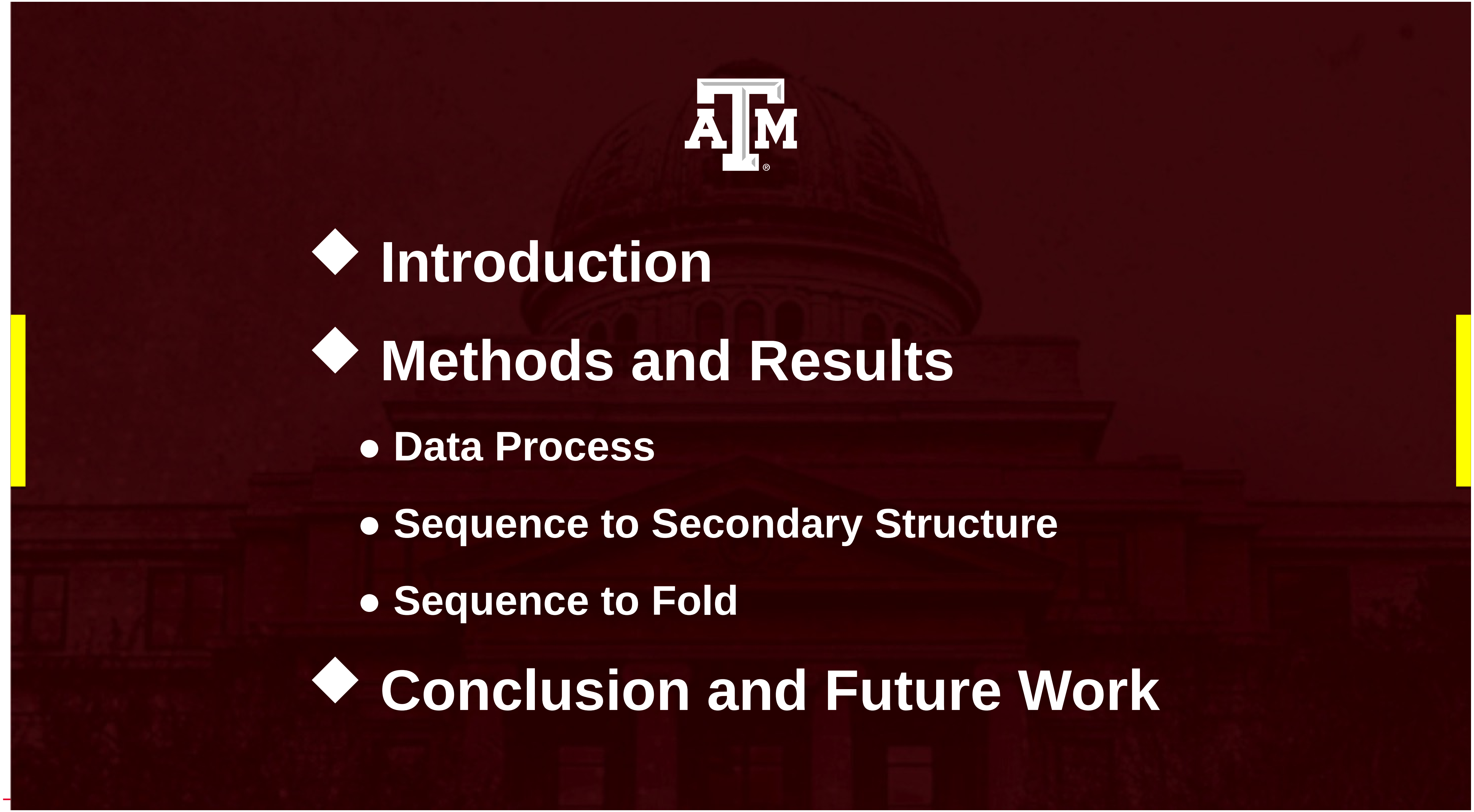
Cross Modality Learning on Proteins
(ECEN 766 Final Project)

Shaowen Zhu

April 30th, 2020



- ◆ Introduction
- ◆ Methods and Results
  - ● Data Process
  - ● Sequence to Secondary Structure
  - ● Sequence to Fold
- ◆ Conclusion and Future Work

Protein Modalities

Protein is an essential kind of nutrition materials which can be represented in different modalities.

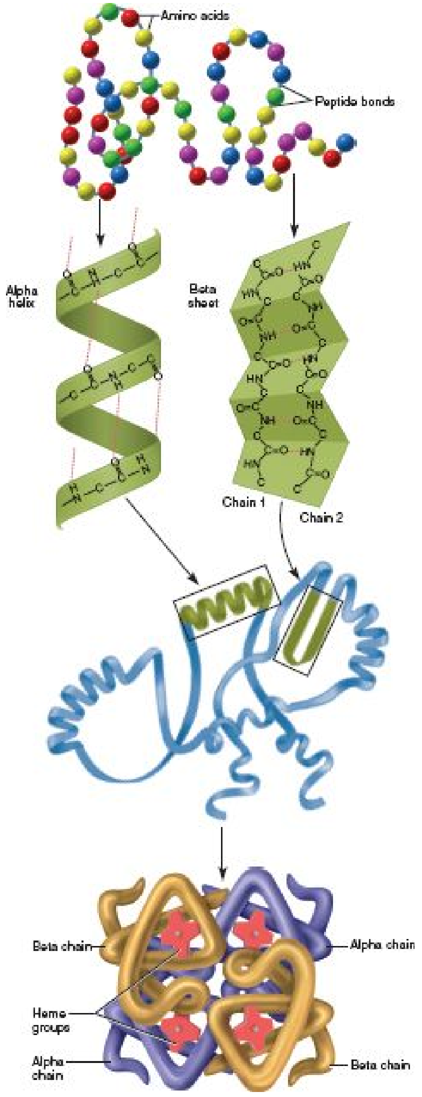
1-D: AA (amino acid) sequence,

SS (secondary structure)

2-D: contact map, distance matrix

3-D: structure

What is the relationship between the modalities and can one modality help to learn another?



Modalities Translation

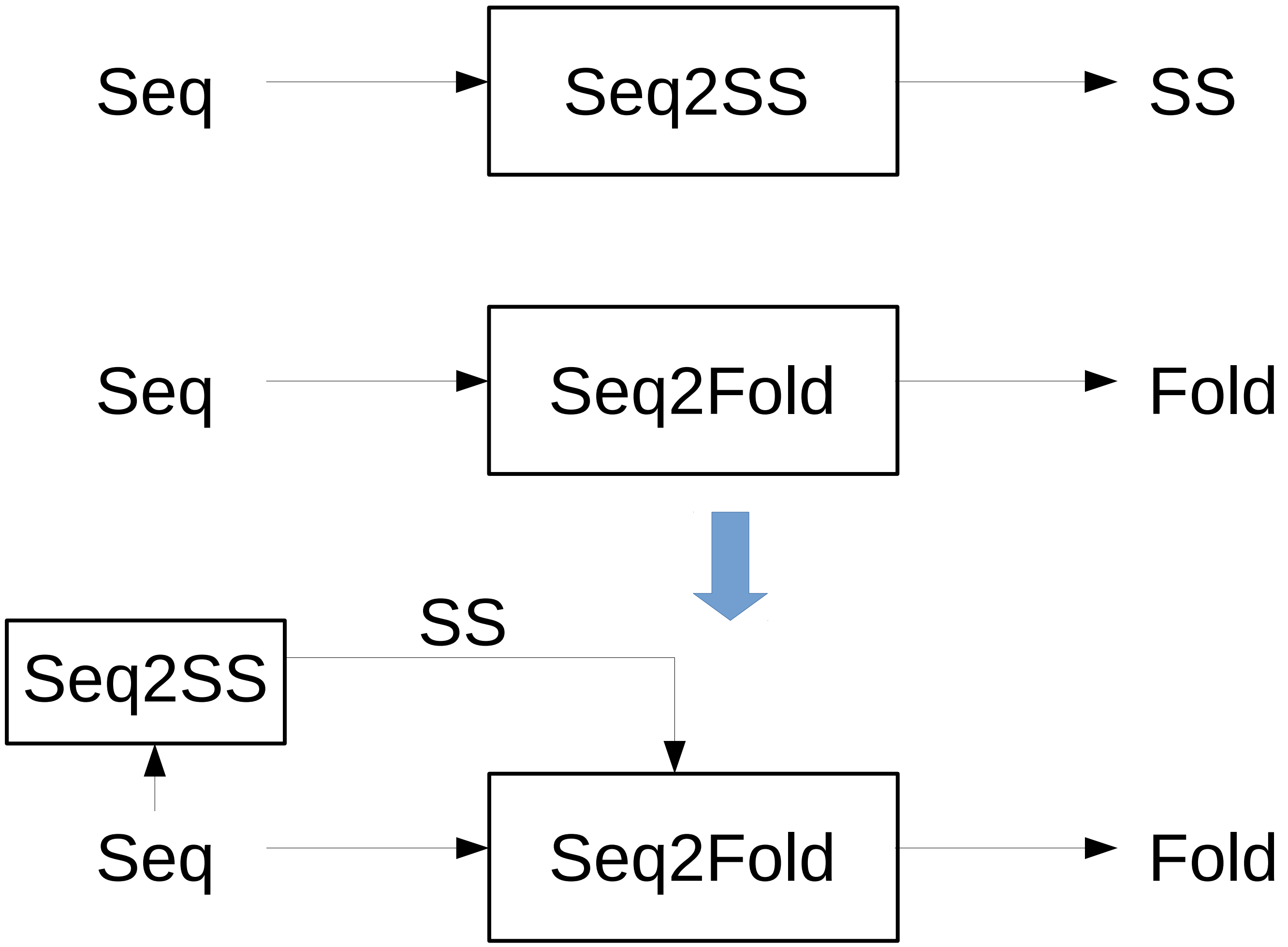
Structure to others: DSSP

Sequence to SS: SCRATCH, TAPE Transformer

Seq & SS & PSSM & SA to Structure (Fold): DeepSF

Most SOTA methods are based on MSA (multiple sequence alignment) and can be rather time-consuming (minutes for just one sequence). How can we model the relationship and make the the prediction directly on the sequence?

Project Idea



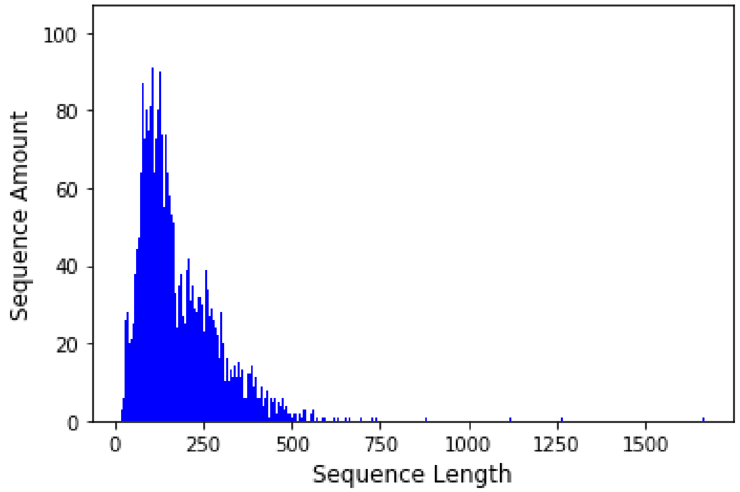
- Introduction
- Methods and Results
  - Data Process
  - Sequence to Secondary Structure
  - Sequence to Fold
- Conclusion and Future Work

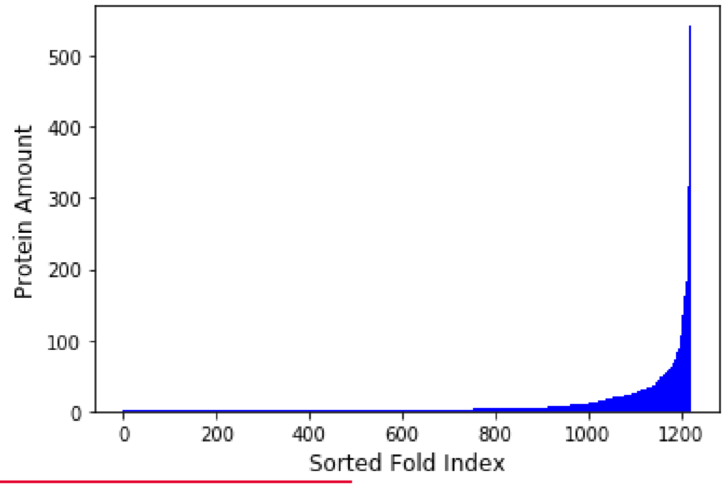
Data Process

This project is based on the ASTRAL SCOPe 2.07 Dataset with less than 40% identity.

- Remove the sequences with missing residues or abnormal residues.
- Remove the sequence longer than 512.
- Only consider single-chain proteins.
- For Seq2Fold, only consider the folds with at least 3 sequences.

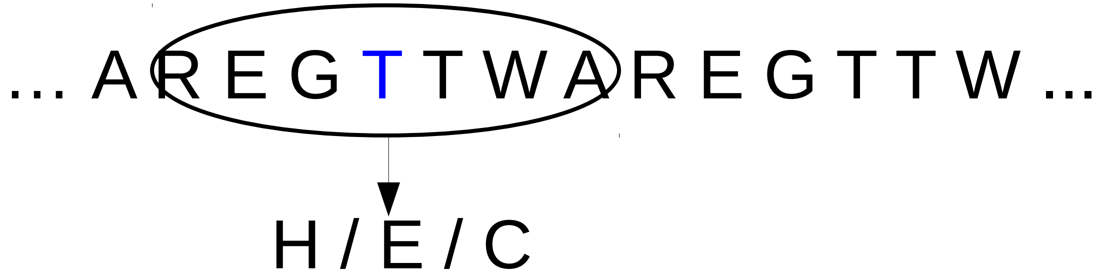
	a	b	c	d	e	f	g
S2S	2497	2793	3924	3405	233	253	709
S2F	2276	2681	3868	3141	178	206	632





Sequence to SS

Focus only on protein segments to predict the local structure.



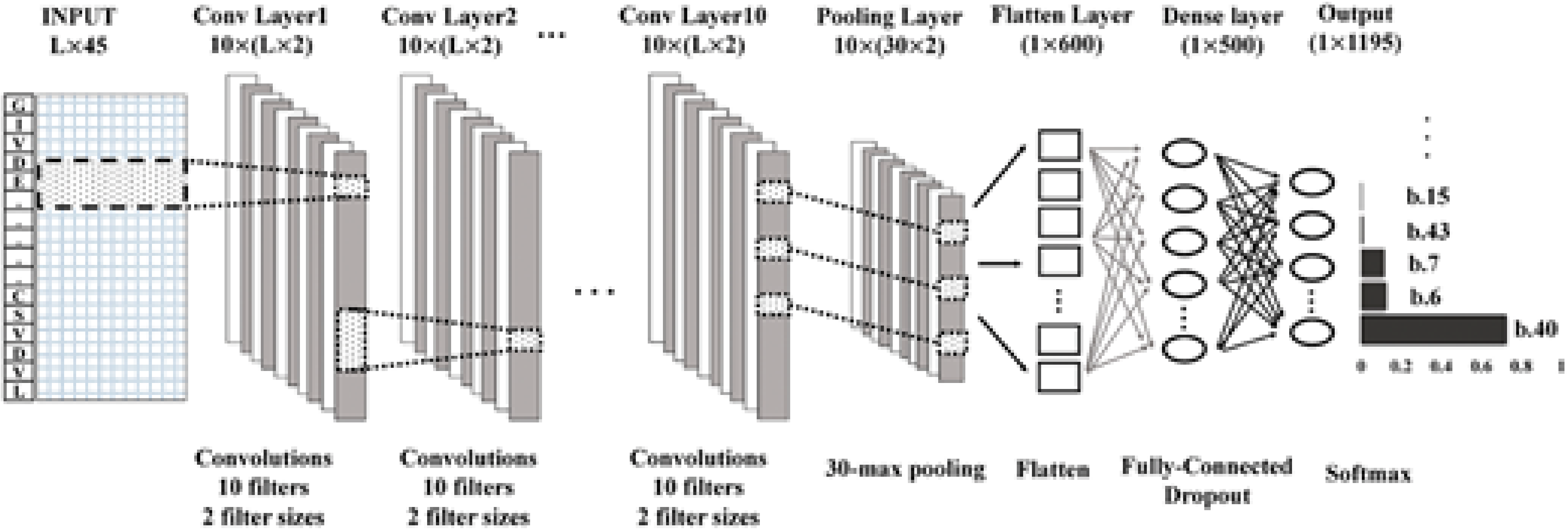
... A R E G T T W A R E G T T W ...

H / E / C

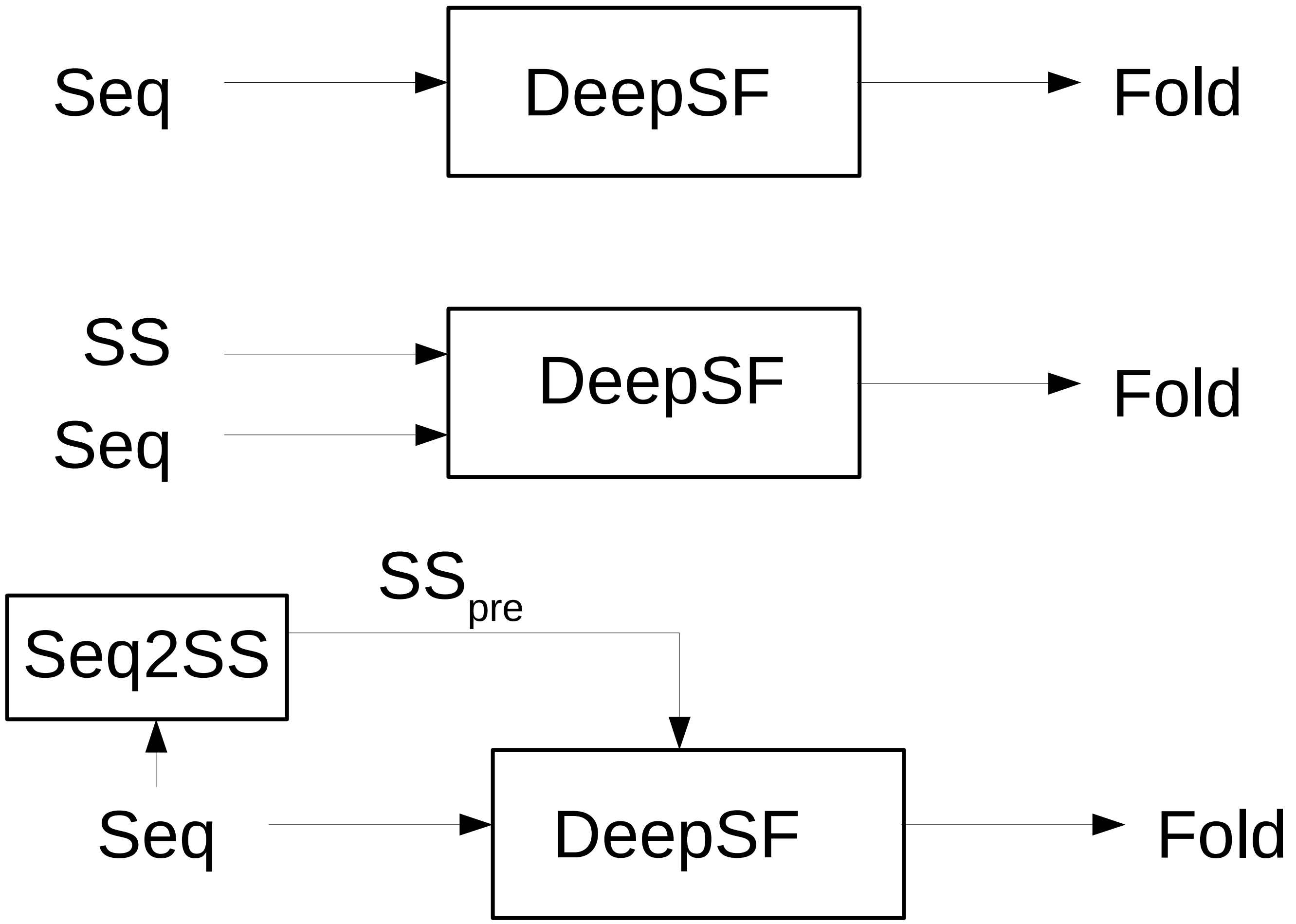
Window Size	SVM		FCNN
	linear	rbf	
3	0.539 / 0.722 1246.3 / 0.772	0.539 / 0.722 1511.9 / 1.043	0.553 / 0.706 221.5 / 0.045
5	0.585 / 0.762 2950.0 / 1.326	0.587 / 0.770 1806.1 / 1.984	0.591 / 0.738 226.0 / 0.032
7	0.616 / 0.794 5812.0 / 1.639	0.616 / 0.802 2248.3 / 2.010	**0.611 / 0.833** 230.1 / 0.057

DeepSF

DeepSF is an 1-D CNN that can predict the fold of given sequences.



Sequence to Fold



Sequence to Fold

TEXAS A&M UNIVERSITY

Accuracy	DeepSF (AA)	DeepSF (AA + SS_{pre})	DeepSF (AA + SS)
Top 1	0.122	0.131	0.502
Top 5	0.307	0.317	0.757
Top 10	0.422	0.438	0.830
Top 15	0.503	0.507	0.879
Top 20	0.558	0.569	0.903

- Introduction
- Methods and Results
  - Data Process
  - Sequence to Secondary Structure
  - Sequence to Fold
- Conclusion and Future Work

Conclusion Future Work

This project provided a quick and efficient method to predict the protein SS, and showed that SS can help to improve the SF model performance.

Future Work:

- Other sequence-based models for SS (Transformer, Seq2Seq, …)
- Test whether some other modalities can be efficiently predicted and applied to help the fold prediction.

Thank you !