weekly report: amino acid seq

Jitian Zhao

University of Wisconsin Madison

December 2, 2019

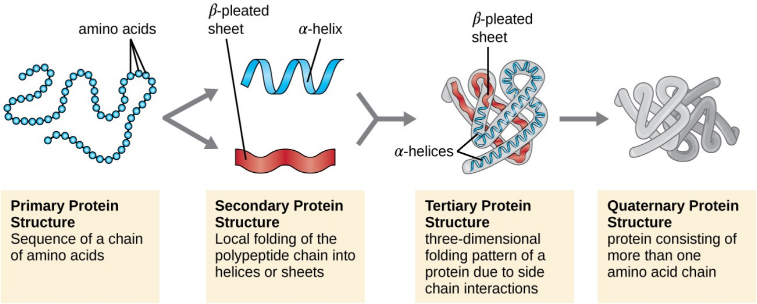
Outline

- "Unirep": protein embedding
- protein secondary structure prediction
- Transformer solution

Unirep

- **pros:** clustering distant but functionally related proteins, it has pre-trained models for different embedding dimensions
- **cons:** models are trained using protein seqs of length 200-280
- **update:** docker issue

Protein Secondary Structure Prediction



Primary Protein Structure
Sequence of a chain of amino acids

Secondary Protein Structure
Local folding of the polypeptide chain into helices or sheets

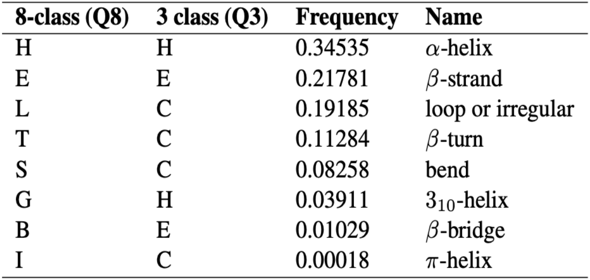
Tertiary Protein Structure
three-dimensional folding pattern of a protein due to side chain interactions

Quaternary Protein Structure
protein consisting of more than one amino acid chain

Figure: protein structure

- **target:** use amino acid sequence to predict secondary structure

Protein Secondary Structure Prediction



8-class (Q8)	3 class (Q3)	Frequency	Name
H	H	0.34535	α-helix
E	E	0.21781	β-strand
L	C	0.19185	loop or irregular
T	C	0.11284	β-turn
S	C	0.08258	bend
G	H	0.03911	3_{10}-helix
B	E	0.01029	β-bridge
I	C	0.00018	π-helix

Figure: classification

- **target:** $MDLSALRVEE \xrightarrow{predict} TTGGGSSHHH$

Protein Secondary Structure Prediction

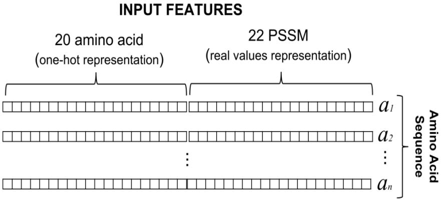


Figure: aa encoding

- Bidirectional LSTM
- S. K. Sønderby and O. Winther(2015), Hattori, Leandro Takeshi F. et al.(2017)

Protein Secondary Structure Prediction

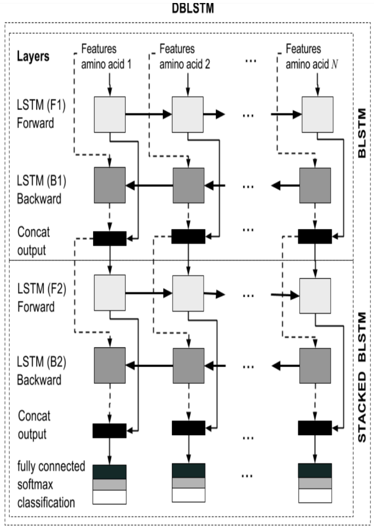


Figure: Bi-direction LSTM Network

Transformer Solutions

- semantic role labeling task
- ”Linguistically-Informed Self-Attention for Semantic Role Labeling” (2018)