Introduction to biological sequence analysis using Deep Learning in Python

Presenters : Bishnu Sarker, Sayane Shome

Date: 17-18 July, 2023

Learning Objectives of the session

Obtain understanding of high-level concepts related to sequence similarity including k-mers and then how natural language processing can be used in case of protein sequences.

Sequence similarity

when sequences are equal-length

Hamming Distance

G	G	G	A	A	T	T	T	C	C
G	G	C	A	A	T	A	A	C	C
1	1	0	1	1	1	0	0	1	1

s = 7

- If two strings are of equal length, **Hamming Distance** may score the number of mismatches.
- It counts the dissimilar letters in corresponding position of the strings.
- It does not consider the similarity from the semantic perspective.

https://www.thedatanotes.com/post/dynamic-programming-sequence-alignment/

Sequence similarity

Variable Length Strings

Levenshtein Distance

```
I N T E * N T I O N
| | | | | | | | | |
* E X E C U T I O N
d s s   i s
```

d: Deletion, s: Substitution, and i: Insertion

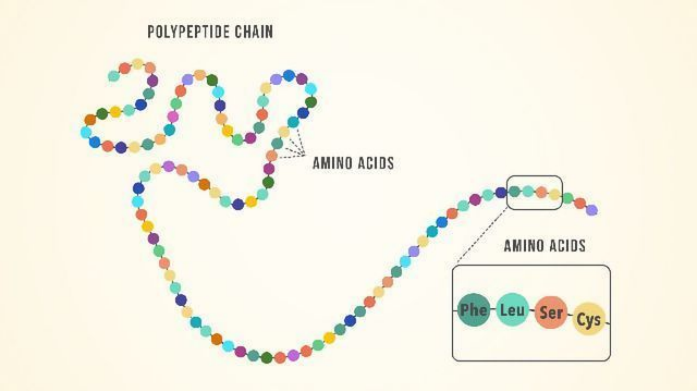
If each operation has a cost of 1, distance between the two strings is 5

If substitutions cost 2,distance between them is 8

- Edit Distance computes the minimum edits requires to transform one string to the other.
- The allowed operations are: Insertion, Deletion and Substitution.
- Each operation has a set score, that adds up to give distance or dissimilarity between two strings.
- Similar to hamming distance, it only considers the syntactical distance.
- It does not provide any insights on semantic similarity.

Jurafsky, D. and Martin, J.H. (2023) Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3/

Biological Sequences Are Strings



Given two Sequences

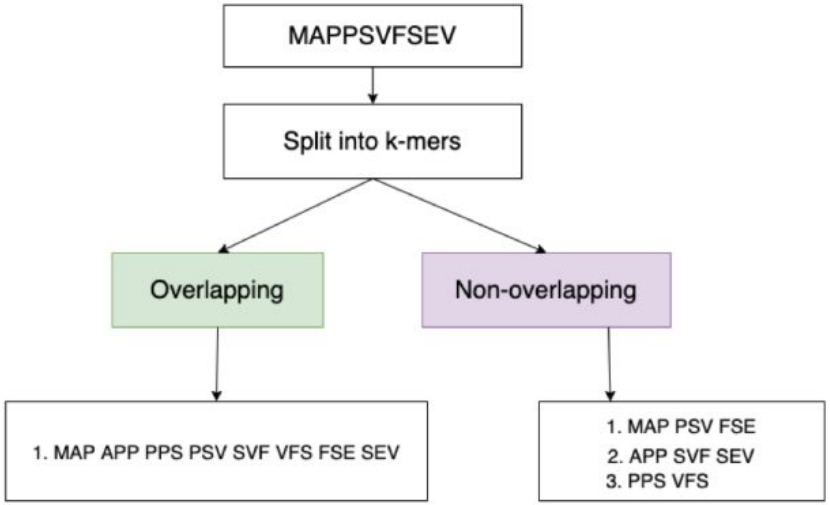
```
AGGCTATCACCTGACCTCCAGGCCGATGCCC
TAGCTATCACGACCGCGGTCGATTTGCCCGAC
```

Find the sequence Similarity or Align them..

```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
TAG-CTATCAC--GACCGC--GGTCGATTTGCCCGAC
```

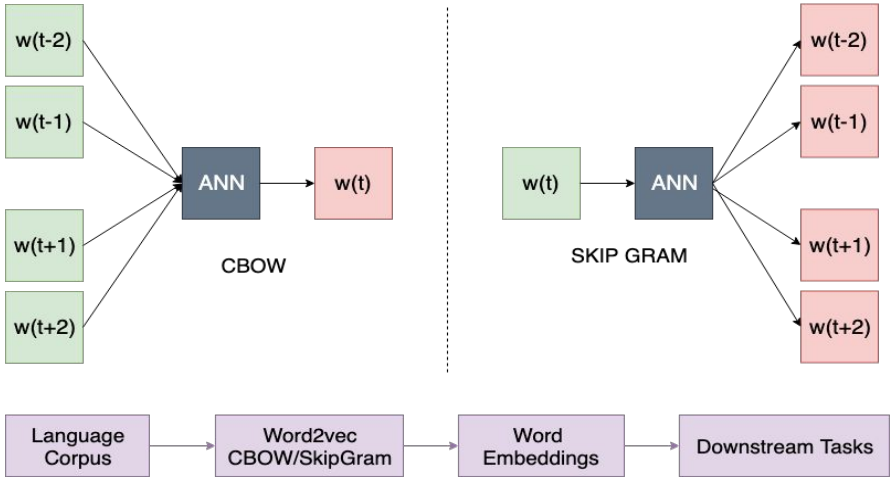
```
I N T E * N T I O N
| | | | | | | | |
* E X E C U T I O N
d s s   i s
```

K-mers based similarity



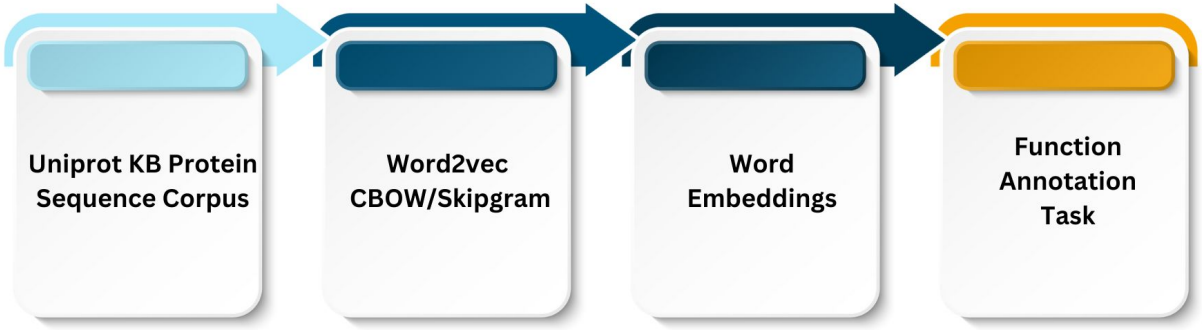
- Edit distance is computationally expensive for long sequences.
- K-mer splits the the sequence into equal K-length sub sequences.
- This provides a sentence like resemblances for biological sequences. K-mers serve as words.
- Each sequence is a set of k-mers.
- Can be encoded as fixed length vector of 1's and 0's.
- Curse of Dimensionality: (20)^3=8000 possible 3-mers. Sparse!

Word2Vec using K-mer words



- **word2vec** learns the low-rank numerical representation of words using Neural Network.
- It learns to predict the **context (surrounding) words** given a target word. The weights learned corresponding to each k-mer serve as the embedding.
- In natural language processing,sentences are composed of words. The spatial relations typically holds the meaning the full sentence.
- In case of biological sequence, **there is no notion of words.**
- K-mers such as 3-mers may serve as the words.

Word2Vec embeddings using Protein Domains

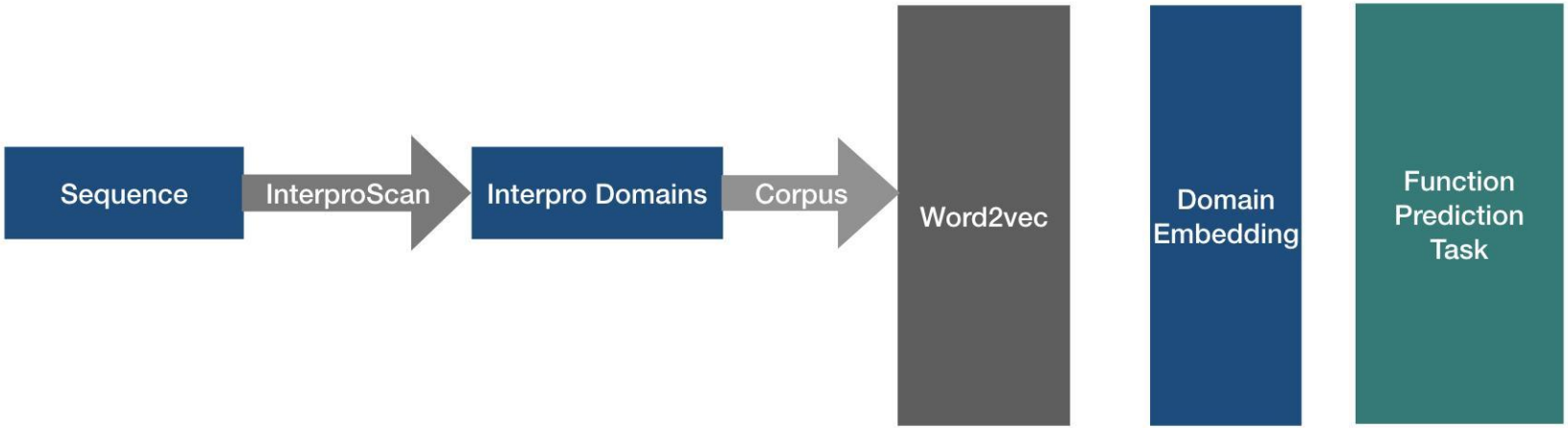


Problem 1 : No notion of word in protein sequence; instead a continuous string of amino acid symbols
Solution 1 : Split the sequence into arbitrary short words (**K-mers**)

Problem 2 : Unlike words, **K-mers** are not biologically significant unit
Solution 2 : Decompose sequence into evolutionarily conserved domains/motifs.

Word2Vec embeddings using Protein Domains

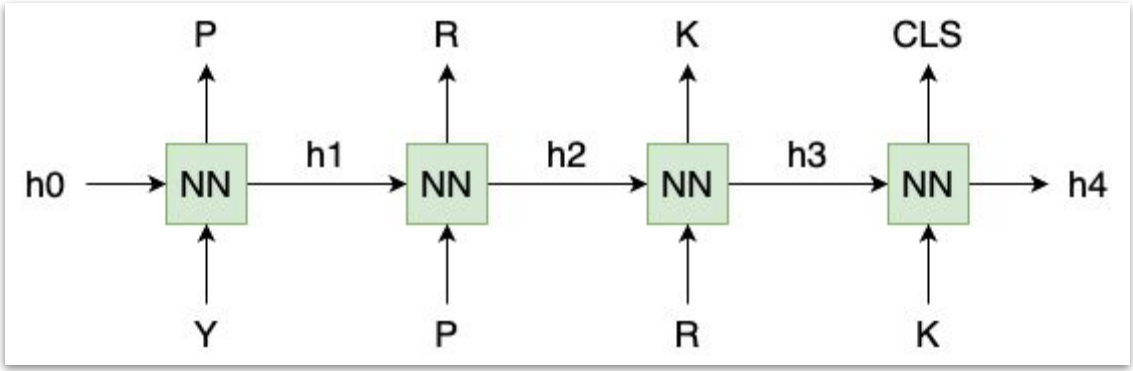
Sequence to Function prediction using Word2vec



Does not consider the ordered long range dependency among k-mers and domains.

Sarker, B., Ritchie, D. W., & Aridhi, S. (2019, September). Functional annotation of proteins using domain embedding based sequence classification. In *KDIR 2019-11th International Conference on Knowledge Discovery and Information Retrieval* (pp. 163-170).

Recurrent Neural Networks (RNN)



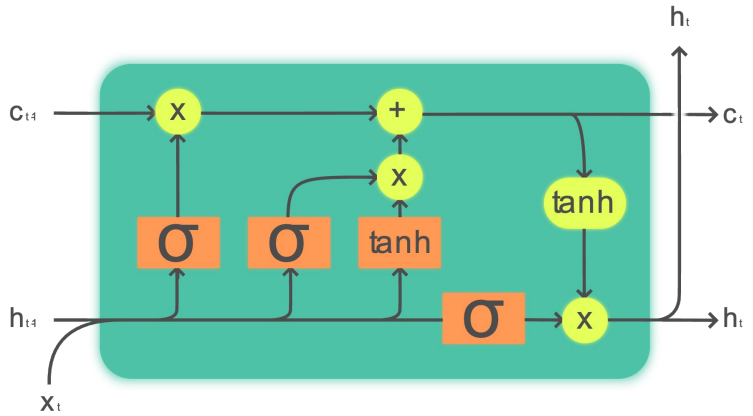
- Predicts next word/residues
- Model spatial dependency
- Take into account time signal
- Suffers from vanishing gradient problem
- Fails to model very long dependency

Hands on Tutorial on RNNs

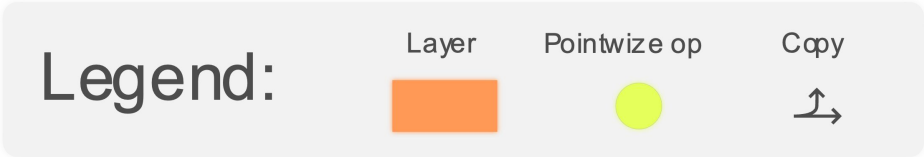
Google colab notebook

Link : Colab-Notebook-RNN

Long Short-Term Memory (LSTM)



- LSTM employs a complex flow of information to overcome the problem of Vanishing Gradient.
- One of the most popular deep learning models for sequence modelling.
- Does not support attention to support focus on different part of the input.



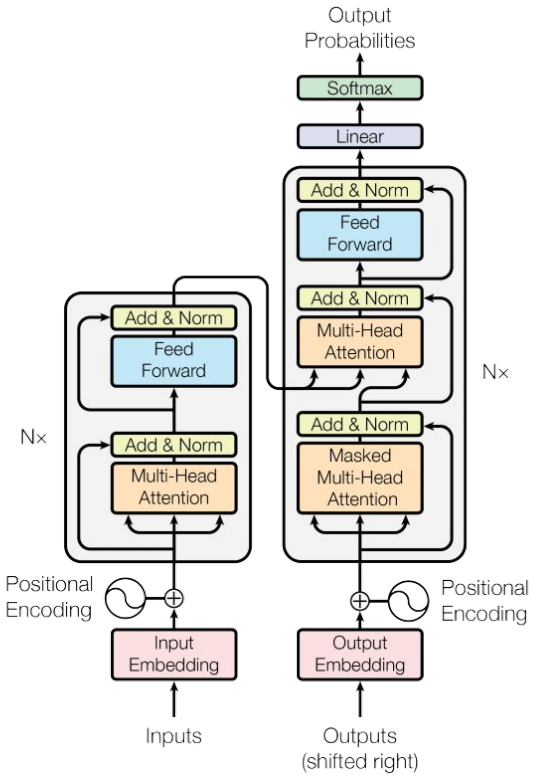
https://upload.wikimedia.org/wikipedia/commons/1/17/The_LSTM_Cell.svg

Hands on Tutorial on LSTMs

Google colab notebook

Link : Colab-Notebook-LSTM

Transformers



- The de-facto sequence model architecture includes multiple identical encoders and decoders.
- Each encoder consists of 1) **an attention layer**, 2) Feedforward layer.
- Through 8 heads, the attention layer attend different parts of the input.
- Each token is passed to individual feed forward neural network.
- The output from the encoder is passed through the top level encoders until fed to the decoders.
- The output from the top encoder is used as embeddings.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.