DNA classification using supervised deep learning

Szuhai Iulia-Monica
Babeș-Bolyai University

WeADL 2021 Workshop

The workshop is organized under the umbrella of WeaMyL, project funded by the EEA and Norway Grants under the number RO-NO-2019-0133. Contract: No 26/2020.

Working together for a green, competitive and inclusive Europe

Problem statement

- Deoxyribonucleic acid or short DNA, is the basis of how life works
- Ancient DNA might reveal crucial information regarding past civilizations, past diseases or even extinct spcecies
- Ancient DNA is subject to contamination with modern DNA
- Our aim: classify ancient and modern DNA
- Four data representation and two learning approaches

Computational approaches to DNA analysis

Challenge: find a comprehensive and robust representation for DNA.

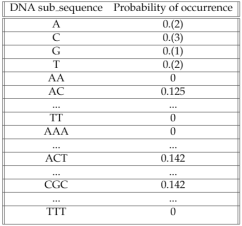
- One hot encoding
- Images
- Deep learning

Methodology

DNA sequence: [A,C,G,T]

Four different DNA representations :

- Nucleotides frequencies based representation
  - $P(s) = \frac{frequence(s)}{l - s_l + 1}$
  - 84 features



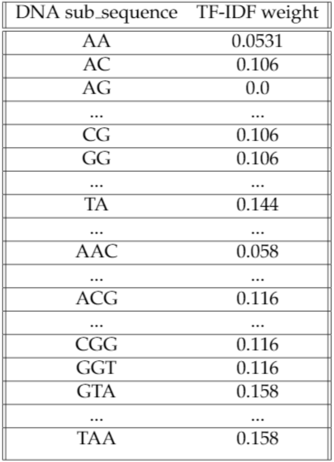
DNA sub_sequence	Probability of occurrence
A	0.(2)
C	0.(3)
G	0.(1)
T	0.(2)
AA	0
AC	0.125
...	...
TT	0
AAA	0
...	...
ACT	0.142
...	...
CGC	0.142
...	...
TTT	0

Figure: DNA representation: example of features and their values for the illustrative sequence *ACTCGCTA*.

Methodology

- TF-IDF based representation
  - $TF - IDF(s) = \frac{frequence(s)}{l - s_l + 1} * \log \frac{k}{n}$

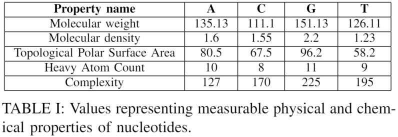
  Example of TF-IDF representation for the sequence *ACGGTAACGGTG* ,considering the coprus *ACGGTAACGGTG*, *TTGCCTGTGCATGA*, *ACCGGTTCAACGTGCAAAACGCG-CACCGC*.



DNA sub_sequence	TF-IDF weight
AA	0.0531
AC	0.106
AG	0.0
...	...
CG	0.106
GG	0.106
...	...
TA	0.144
...	...
AAC	0.058
...	...
ACG	0.116
...	...
CGG	0.116
GGT	0.116
GTA	0.158
...	...
TAA	0.158

Methodology

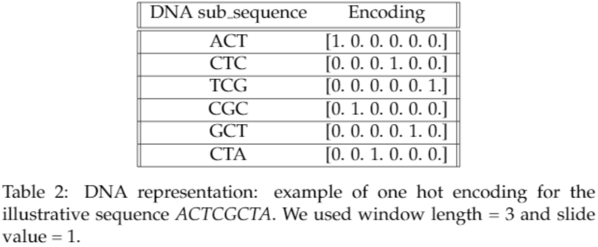
- Physical and chemical properties based representation



Property name	A	C	G	T
Molecular weight	135.13	111.1	151.13	126.11
Molecular density	1.6	1.55	2.2	1.23
Topological Polar Surface Area	80.5	67.5	96.2	58.2
Heavy Atom Count	10	8	11	9
Complexity	127	170	225	195

TABLE I: Values representing measurable physical and chemical properties of nucleotides.

- One hot encoding



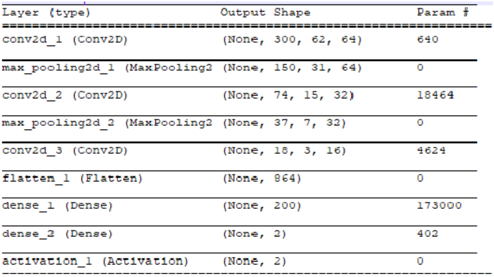
DNA sub_sequence	Encoding
ACT	[1. 0. 0. 0. 0. 0.]
CTC	[0. 0. 0. 1. 0. 0.]
TCG	[0. 0. 0. 0. 0. 1.]
CGC	[0. 1. 0. 0. 0. 0.]
GCT	[0. 0. 0. 0. 1. 0.]
CTA	[0. 0. 1. 0. 0. 0.]

Table 2: DNA representation: example of one hot encoding for the illustrative sequence *ACTCGCTA*. We used window length = 3 and slide value = 1.

Methodology

Classification problem

- Non linear models
- Multi-layer Perceptron
  - 2 hidden layers, ReLU, adaptive learning rate
- Convolutional neural network



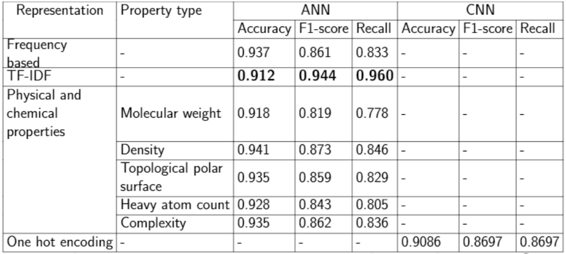
```
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 300, 62, 64)       640
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 150, 31, 64)       0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 74, 15, 32)        18464
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 37, 7, 32)         0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 18, 3, 16)         4624
_________________________________________________________________
flatten_1 (Flatten)          (None, 864)               0
_________________________________________________________________
dense_1 (Dense)              (None, 200)               173000
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 402
_________________________________________________________________
activation_1 (Activation)    (None, 2)                 0
-----------------------------------------------------------------
```

Figure: Model summary of the convolutional neural network

Dataset

- Matter collected from Capidava archaeological site
- 378.451 ancient sequences
- 115.218 modern sequences
- MLP - k- folds cross validation, k=10
- CNN - 66 % training, 33 % testing

Results



Representation	Property type	ANN			CNN		
		Accuracy	F1-score	Recall	Accuracy	F1-score	Recall
Frequency based	-	0.937	0.861	0.833	-	-	-
TF-IDF	-	**0.912**	**0.944**	**0.960**	-	-	-
Physical and chemical properties	Molecular weight	0.918	0.819	0.778	-	-	-
	Density	0.941	0.873	0.846	-	-	-
	Topological polar surface	0.935	0.859	0.829	-	-	-
	Heavy atom count	0.928	0.843	0.805	-	-	-
	Complexity	0.935	0.862	0.836	-	-	-
One hot encoding	-	-	-	-	0.9086	0.8697	0.8697

Figure: Evaluation measures for the two supervised models and the considered representations

Conclusion

- The aim was to find suitable representation and machine learning models to the goal of distinguishing between ancient and modern bacterial DNA
- Obtained results are promising
- Acquire more modern DNA sequences
- Perform these experiments on other public dataset
- Other approaches

Thank you for your attention!