OpenVaccine: COVID-19 mRNA Vaccine Degradation Prediction

Lynda MESSAD DIT MAHTAL
Laëtitia HOLLIER
Opale RAMBAUD

M2BI / 2020

université PARIS DIDEROT PARIS 7
Université de Paris

SOMMAIRE

Introduction

- Pandémie mondiale de COVID-19 dû au SARS-CoV-2

But: **prédire les régions des molécules d'ARNm susceptibles de posséder un taux de dégradations élevé selon la position des bases et leur type**

- Méthode de Deep Learning de type Régression via un réseau récurrent GRU (Gated Recurrent Unit)

- Comparaison de 2 méthodes de traitement de données:
  - troncage des données en amont du réseau
  - troncage des données au sein du tenseur

Matériels et méthodes

Présentation des données

3 fichiers mis à disposition sur la plateforme Kaggle:

- **train.json** ⇒ data train avec 2400 lignes, 19 descripteurs
- **test.json** ⇒ data test avec 3634 lignes, 7 descripteurs
- **sample_submission.csv** ⇒ fichier de sortie au bon format attendu à la fin

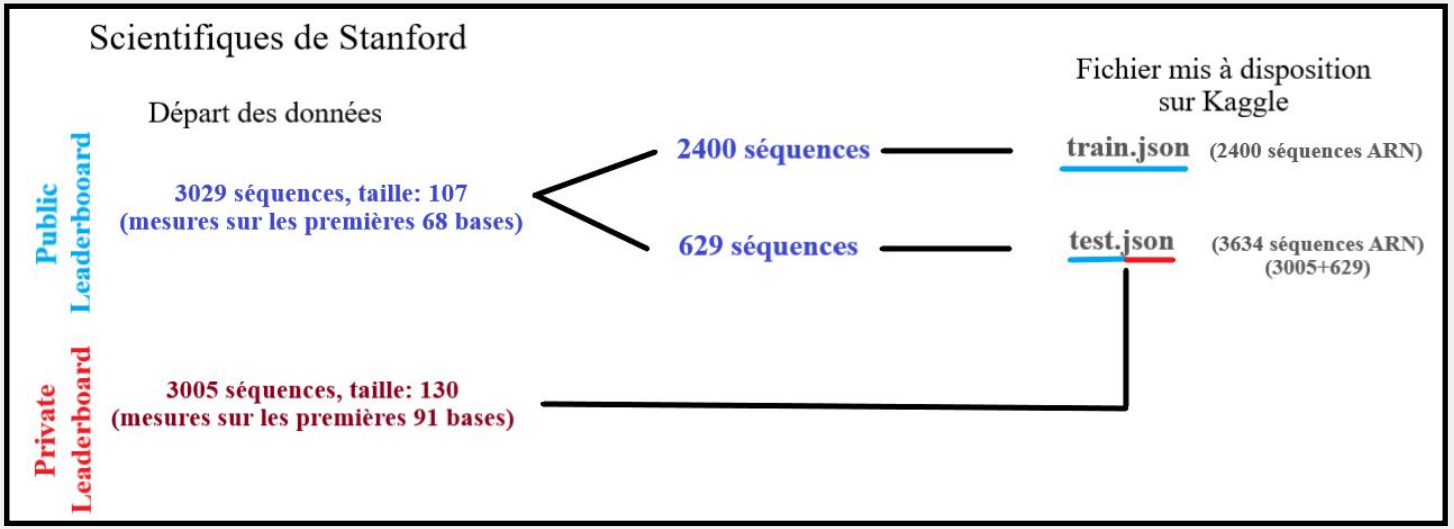


Schéma récapitulatif des données et comment elles ont été obtenues par les scientifiques de l'université de Stanford.

Matériels et méthodes

Vérification des données

Data Train

Descriptif de la data Train

	index	signal_to_noise	SN_filter	seq_length	seq_scored
count	2400.000000	2400.000000	2400.000000	2400.0	2400.0
mean	1199.500000	4.530456	0.662083	107.0	68.0
std	692.964646	2.835142	0.473099	0.0	0.0
min	0.000000	-0.103000	0.000000	107.0	68.0
25%	599.750000	2.391000	0.000000	107.0	68.0
50%	1199.500000	4.442500	1.000000	107.0	68.0
75%	1799.250000	6.294250	1.000000	107.0	68.0
max	2399.000000	17.194000	1.000000	107.0	68.0

Data Test

Descriptif de la data Test

	index	seq_length	seq_scored
count	3634.000000	3634.000000	3634.000000
mean	1816.500000	126.018987	87.018987
std	1049.189767	8.702624	8.702624
min	0.000000	107.000000	68.000000
25%	908.250000	130.000000	91.000000
50%	1816.500000	130.000000	91.000000
75%	2724.750000	130.000000	91.000000
max	3633.000000	130.000000	91.000000

Alignement multiple Clustal Omega

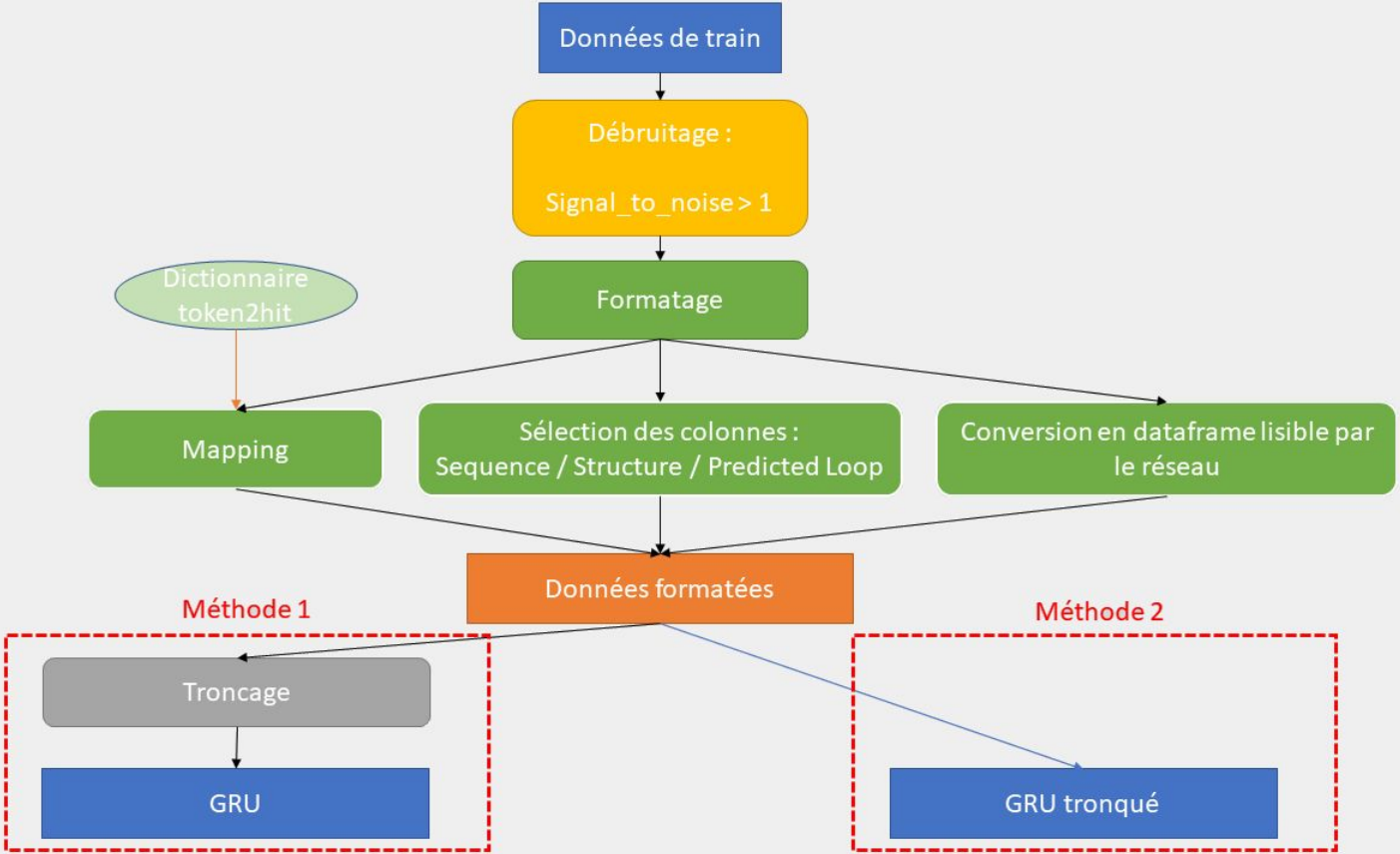
```
sequence1500    -----G-UC-G-----------------G------AG-GA-GG----AGGAA-------  45
sequence964     ----AG-CG-GA-C-G-AAU-AUA-CG-UU---CGCAU-AG-CG----AUCGA-------  58
sequence1378    ----AG-CG-AG-C-U--CG-ACG-CG-UG---CACAU-AA-UG----CACGC-------  48
sequence807     --------A-GA-G-C-ACC-CGC-CG-AU---CGUAC-AG-CC----GACGA-------  58
sequence554     ----GG-AG-AA-G-U-CAU-AGU-CA-GA---CGCGG-AG-UG----GAAGU-------  58
sequence1713    ----GC-AG-AA-G-U-GAU-AGU-UG-GA---CGCGG-AG-UG----GAAGU-------  58


sequence1119    UA-A---AGA------CAAG----A-CA------AGA--U-AG-AAACC-AUA-AUUUCG  79
sequence976     UA-U---ACA------CAAC----A-CA------AGA--U-AG-AAACG-UGA-UAUUCG  79
sequence1293    UA-U---ACA------CAAC----A-CA------AGA--U-AG-UACCA-UAA-UGUUCG  79
sequence491     UA-U---ACA------CAAC----A-CA------AGA--U-AG-UACCU-UGA-GGUUCG  79
```

```
sequence1260    --GGA---A-AA-------------------------------------------------  6
sequence2070    --GGA---A-AA-------------------------------------------------  6
sequence1145    --GGA---A-AGCGG-A--------------------A---------------------  11
sequence819     --GGA---A-A--------------------------------------------------  5
sequence335     --GGA---A-ACUCG-A--------------------A----AU-A-A-----------  15
sequence3415    --GGA---A----------------------------------------------------  4
sequence745     --GGA---A-AAC------------------------------------------------  7
sequence2965    --GG---------------------------------------------------------  2
sequence639     --GGA---A-ACAAA----------------------------------------------  9
sequence1309    --GGA---A-ACAAC-A--------------------A----AC-A-A-----------  15
sequence2894    --GGA---A-AAACA-A--------------------A----AG-A-A-----------  15
```

Matériels et méthodes

Pré-traitement des données



Matériels et méthodes

Gated Recurrent Unit (GRU)

- **Réseau de neurones récurrents** : constitué de neurones interconnectées interagissant non-linéairement et pour lequel il existe au moins un cycle dans la structure
- **Reset Gate** : sert à contrôler combien d'information passée le réseau doit oublier.
- **Update Gate** : décide des informations à conserver et de celles à oublier

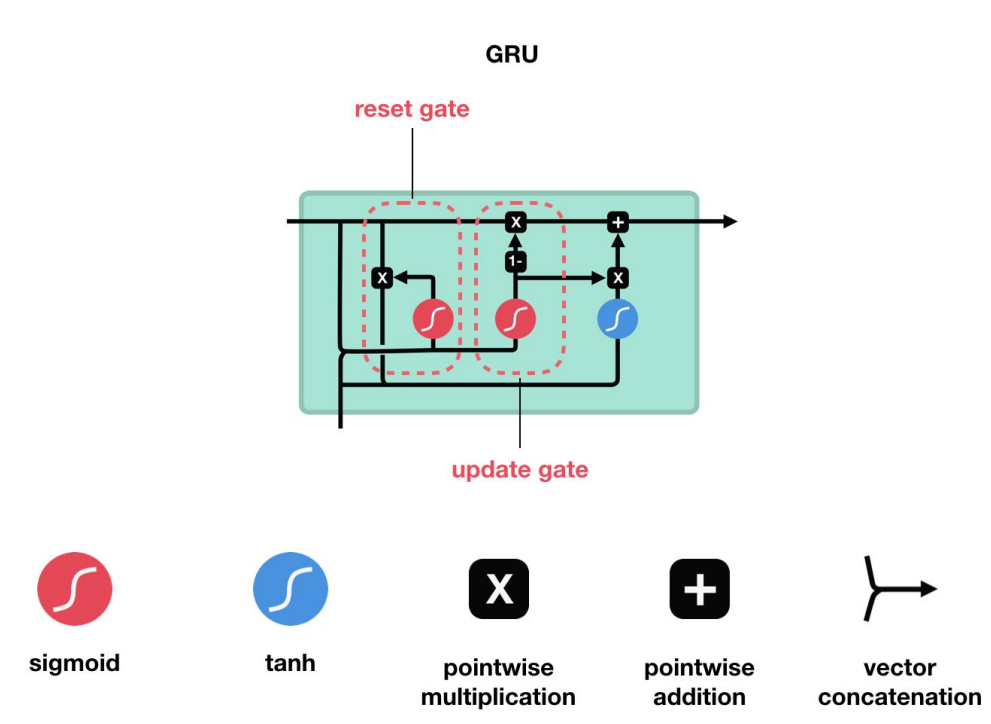


Schéma d'un GRU (source : towardsdatascience.com)

Matériels et méthodes

Design du réseau

Réseau normal

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 68, 3)]	0
embedding (Embedding)	(None, 68, 3, 100)	1400
tf_op_layer_Reshape (TensorF	[(None, 68, 300)]	0
spatial_dropout1d (SpatialDr	(None, 68, 300)	0
bidirectional (Bidirectional	(None, 68, 300)	406800
bidirectional_1 (Bidirection	(None, 68, 300)	406800
bidirectional_2 (Bidirection	(None, 68, 300)	406800
tf_op_layer_strided_slice (T	[(None, 68, 300)]	0
dense_3 (Dense)	(None, 68, 5)	1505

Total params: 1,223,305
Trainable params: 1,223,305
Non-trainable params: 0

Réseau tronqué

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 107, 3)]	0
embedding_1 (Embedding)	(None, 107, 3, 100)	1400
tf_op_layer_Reshape_1 (Tenso	[(None, 107, 300)]	0
spatial_dropout1d_1 (Spatial	(None, 107, 300)	0
bidirectional (Bidirectional	(None, 107, 300)	406800
bidirectional_1 (Bidirection	(None, 107, 300)	406800
bidirectional_2 (Bidirection	(None, 107, 300)	406800
tf_op_layer_strided_slice (T	[(None, 68, 300)]	0
dense (Dense)	(None, 68, 5)	1505

Total params: 1,223,305
Trainable params: 1,223,305
Non-trainable params: 0

- Loss function : **MCRMSE**
- Metric : **Accuracy**

Résultats

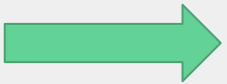
Prédictions

	reactivity	deg_Mg_pH10	deg_Mg_50C	deg_pH10	deg_50C	id_seqpos
0	0.714691	0.723065	0.636777	2.171737	0.835025	id_00073f8be_0
1	2.397619	3.301251	3.539931	4.499967	3.084261	id_00073f8be_1
2	1.509842	0.562209	0.624255	0.602935	0.686217	id_00073f8be_2
3	1.421088	1.264342	1.809805	1.328547	1.812573	id_00073f8be_3
4	0.892619	0.693451	0.907562	0.570246	0.904023	id_00073f8be_4

5 premières lignes du fichier de prédictions avec la méthode 1

	reactivity	deg_Mg_pH10	deg_Mg_50C	deg_pH10	deg_50C	id_seqpos
0	0.588973	0.625430	0.539917	1.986531	0.792241	id_00073f8be_0
1	2.160224	3.196975	3.410743	4.401863	3.056409	id_00073f8be_1
2	1.367732	0.595218	0.638190	0.572630	0.758136	id_00073f8be_2
3	1.266218	1.225548	1.709989	1.134518	1.741791	id_00073f8be_3
4	0.830383	0.563159	0.879607	0.518978	0.931565	id_00073f8be_4

5 premières lignes du fichier de prédictions avec la méthode 2



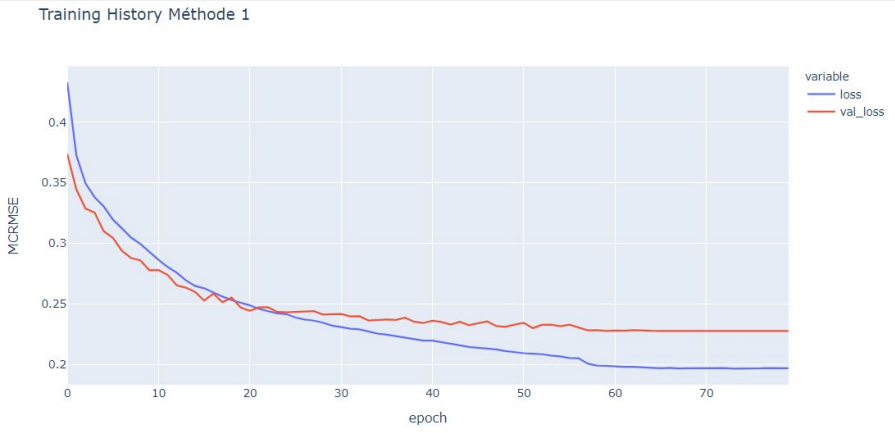
Soumission Kaggle

Scoring

Résultats

Comparaison des méthodes

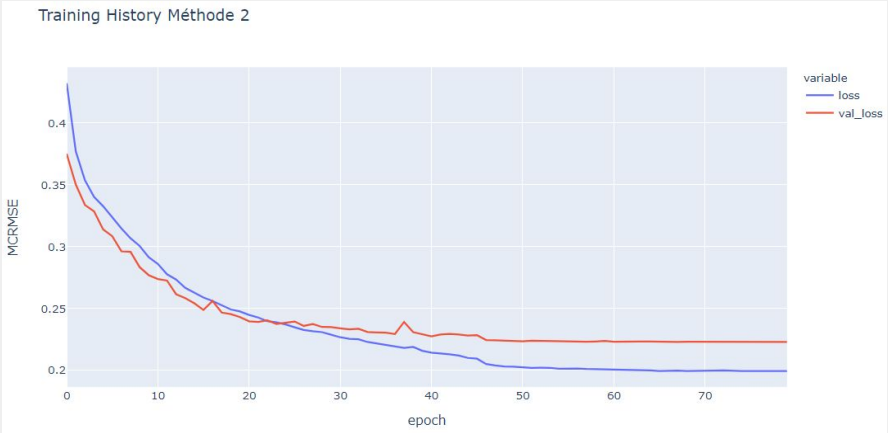
Méthode 1



loss = 0.182 acc = 0.550

Private Score	Public Score
0.38704	0.27812

Méthode 2



loss = 0.184 acc = 0.549

Private Score	Public Score
0.38772	0.27014

Conclusion

- Les 2 méthodes possèdent des résultats similaires

- Le fait de tronquer les données en amont ou au sein du tenseur ne change pas les performances du réseaux

- Intéressant de comparer ces résultats avec une 3ème méthode qui crée une fenêtre glissante de taille 68 sur la séquence afin de ne prédire que le nucléotide central ⇒ amélioration des résultats de prédictions ?

Merci de votre attention