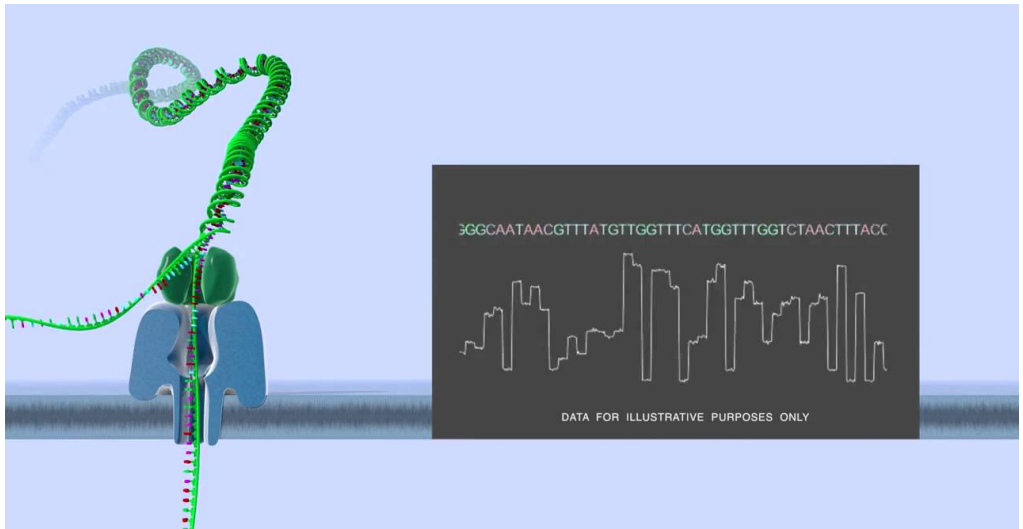
Basecalling par deep learning sur les données de séquençage par nanopores du MinION

AEROBICS

Adrien Jarretier-Yuste
Encadré par Dr Hector Hernandez-Vargas

Septembre 2019



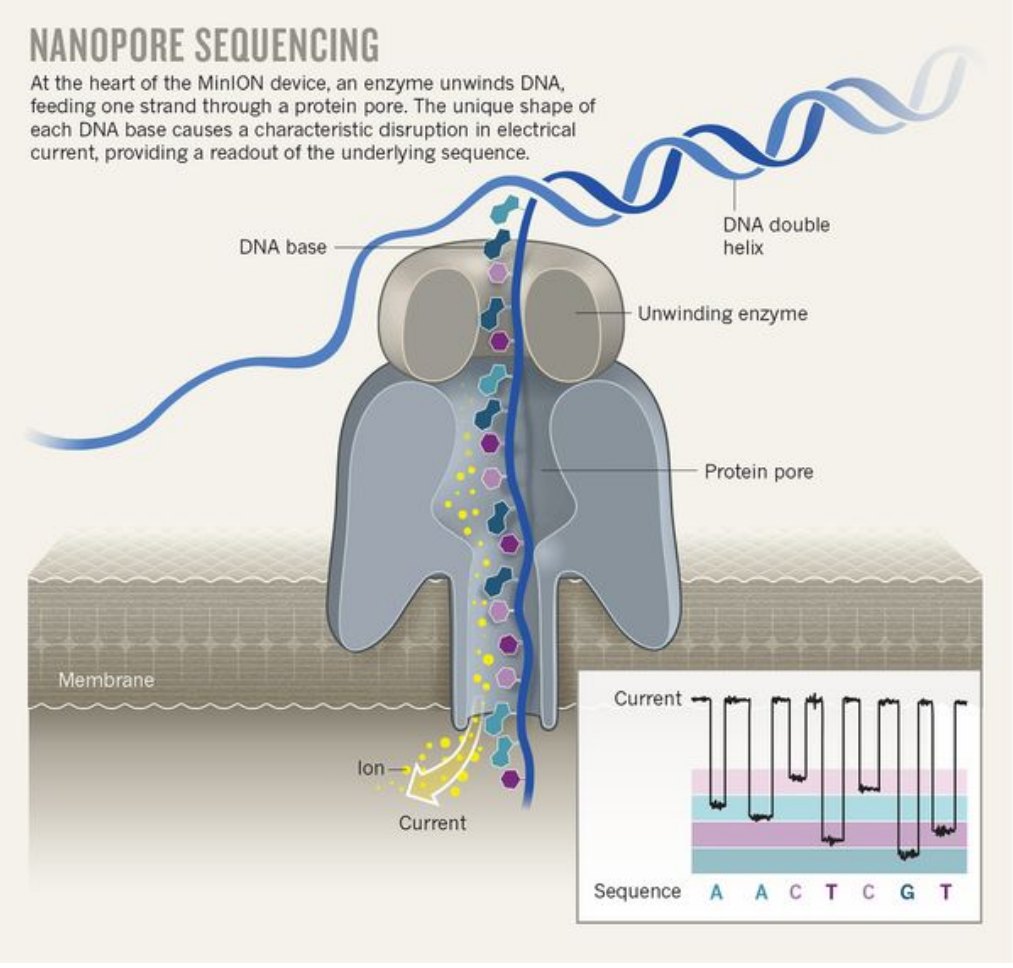
- Centre de Recherche en Cancérologie de Lyon
- Équipe TGF-beta et régulation de la réponse immunitaire
  - Hector Hernandez-Vargas
  - Chloé Goldsmith

- LIRIS
- Équipe dm2l
  - Céline Robardet
  - Stefan Duffner
  - Marc Plantevit

Le MinION

Le MinION de Oxford Nanopore Technologies.

Le séquençage ADN par nanopores

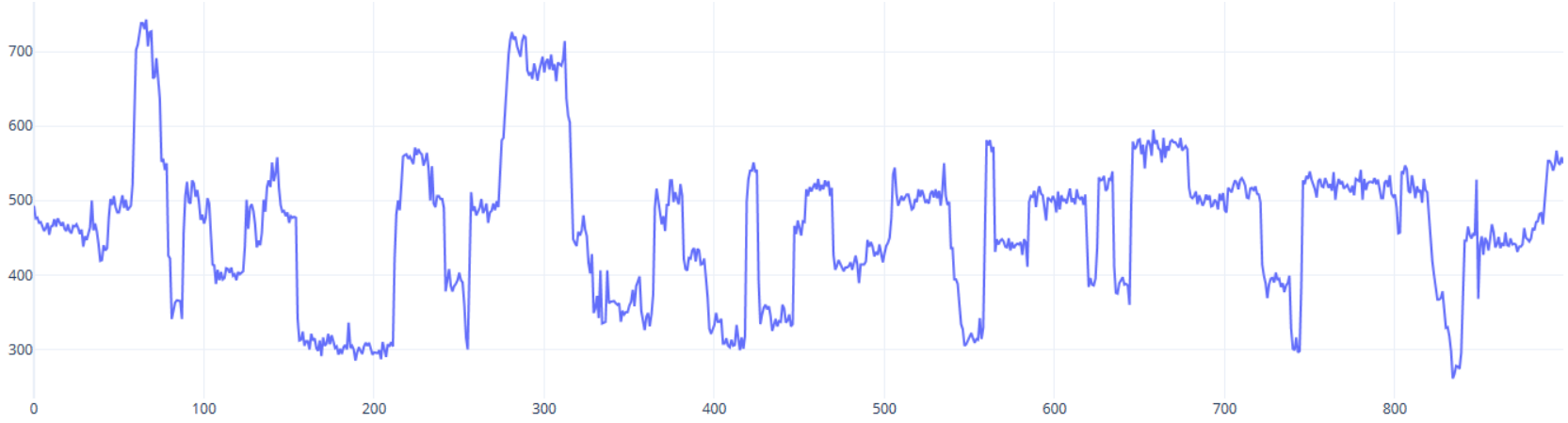


Le nanopore. Nik Spencer/Nature.

Le séquençage ADN par nanopores

Le *squiggle*

La mesure du courant est appelée *squiggle* ou "gribouillis".

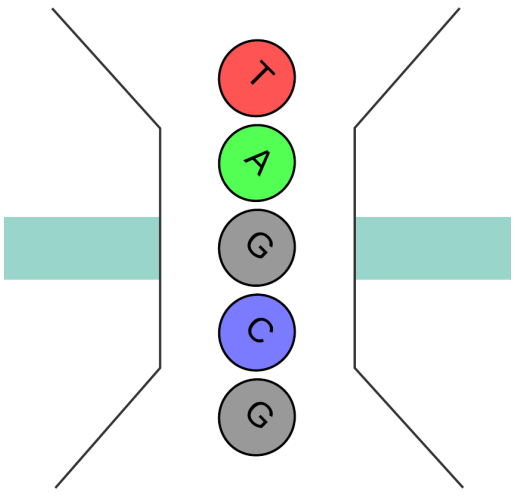


900 mesures, soit environ 100 bases.

- Échantillonage à 4kHz.
- Vitesse du brin d'ADN $\approx$ 450 bp / s, variable.
- $\approx$ 9 mesures / 5-mer.

Le *basecalling*

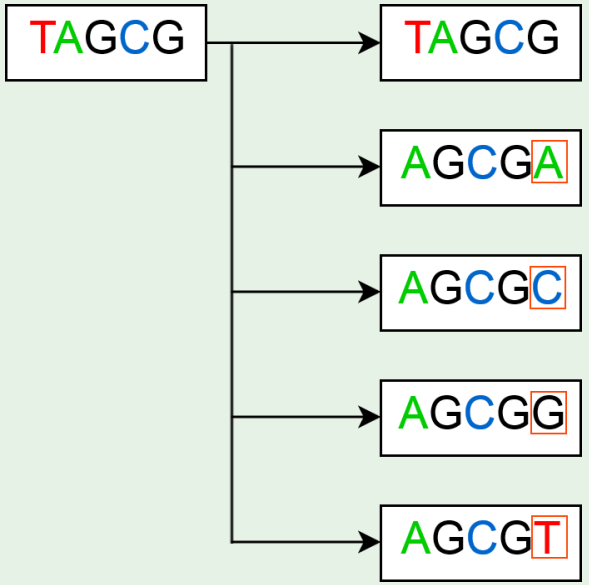
Processus de traduction du signal brut en une séquence ADN.



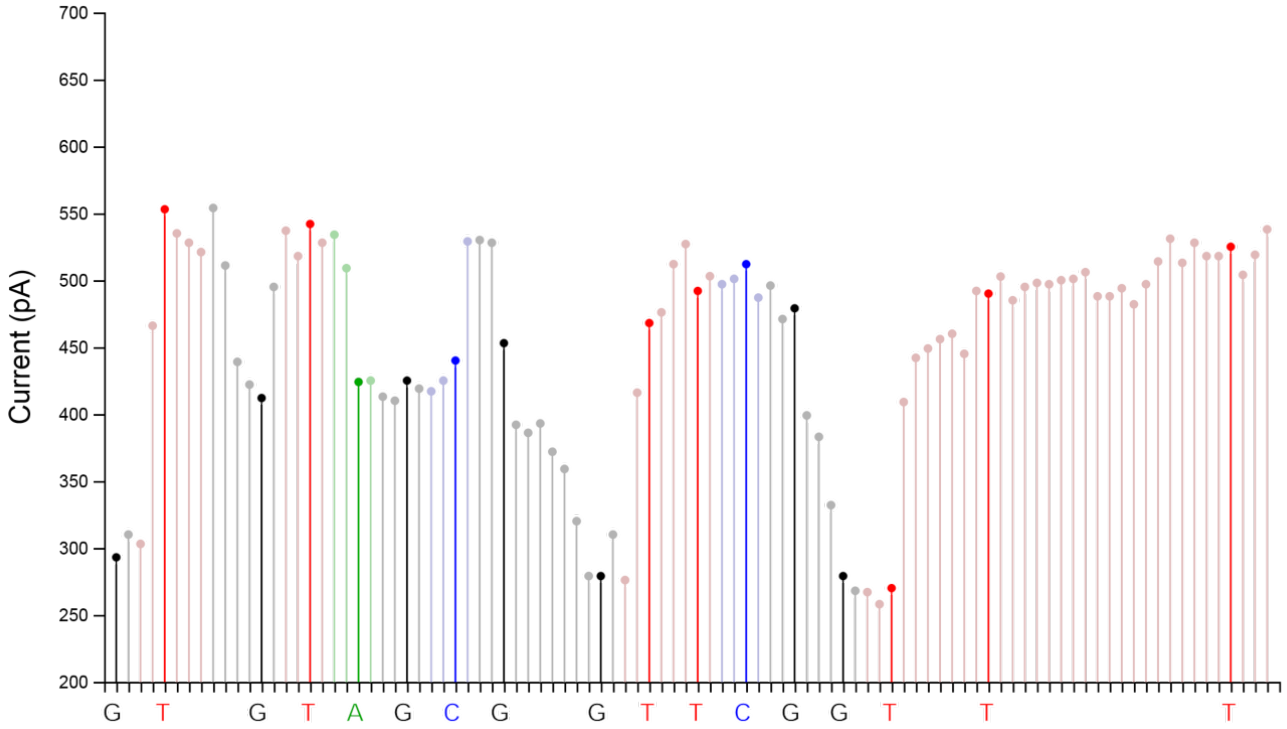
Tête du nanopore R9.

- Le courant est influencé par une sous-chaîne de longueur 5 (5-mer).
- Ici le 5-mer est TAGCG.

Transitions d'états possibles



Le *basecalling*

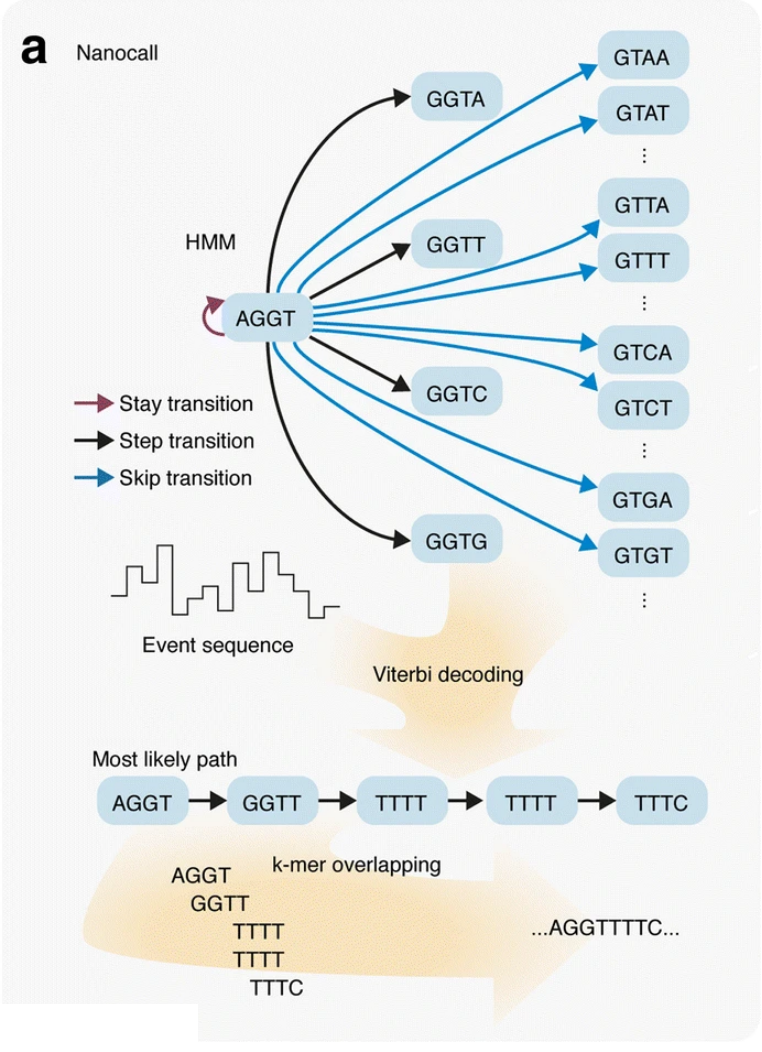


Une portion d'un *squiggle* avec la séquence ADN correspondante alignée.

Ecart entre bases Non-uniforme.

Intensité différente pour des bases identiques.

Modèles de Markov cachés



Nanocall (David et al., 2017) utilise un HMM. Illustration de (Rang et al., 2018).

Les réseaux de neurones se sont montrés meilleurs. (Rang et al., 2018), (Wick et al., 2019).

Réseaux de neurones récurrents

Réseau bidirectionnel LSTM

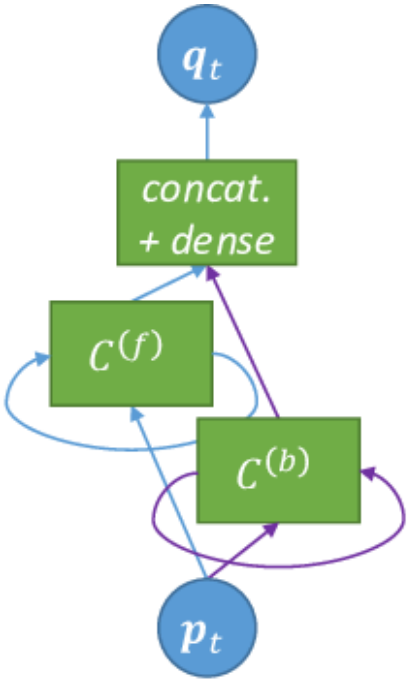


Figure 1: Illustration Pierre-Henri Conze.

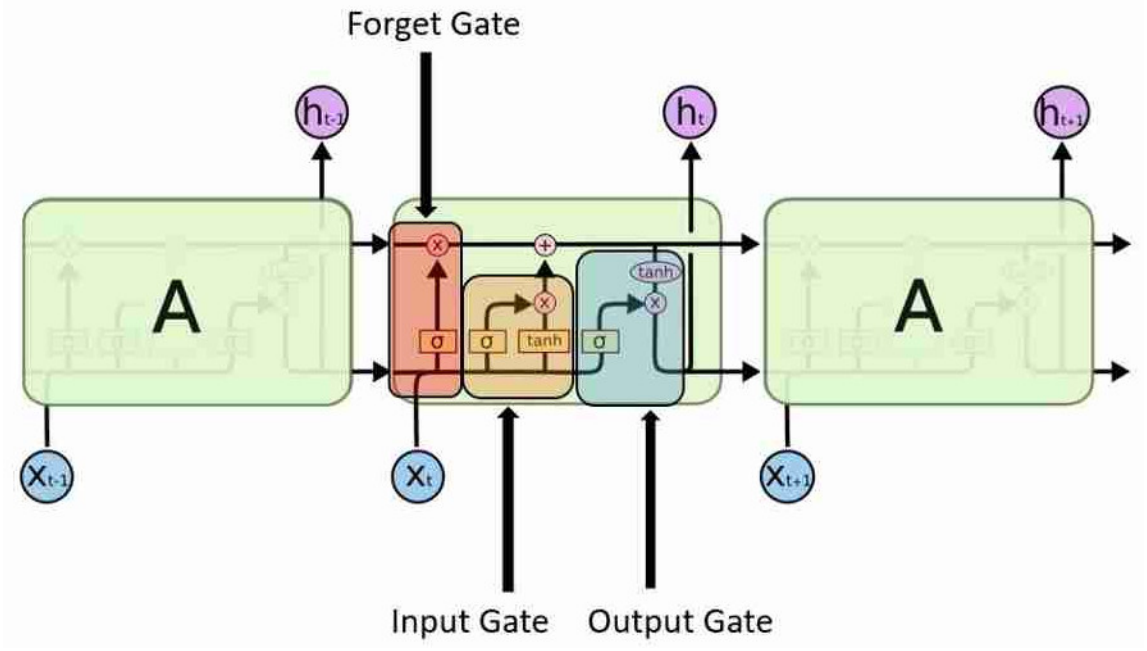


Figure 2: Unité *Long Short-Term Memory*.

1. Information à supprimer de la cellule.
2. Information à ajouter à la cellule.
3. Partie de la cellule pertinente pour la sortie.

Réseaux de neurones récurrents

Connectionist Temporal Classification

Le modèle CTC est une modification d'un réseau de neurones récurrents pour l'apprentissage de séquences sans segmentation.

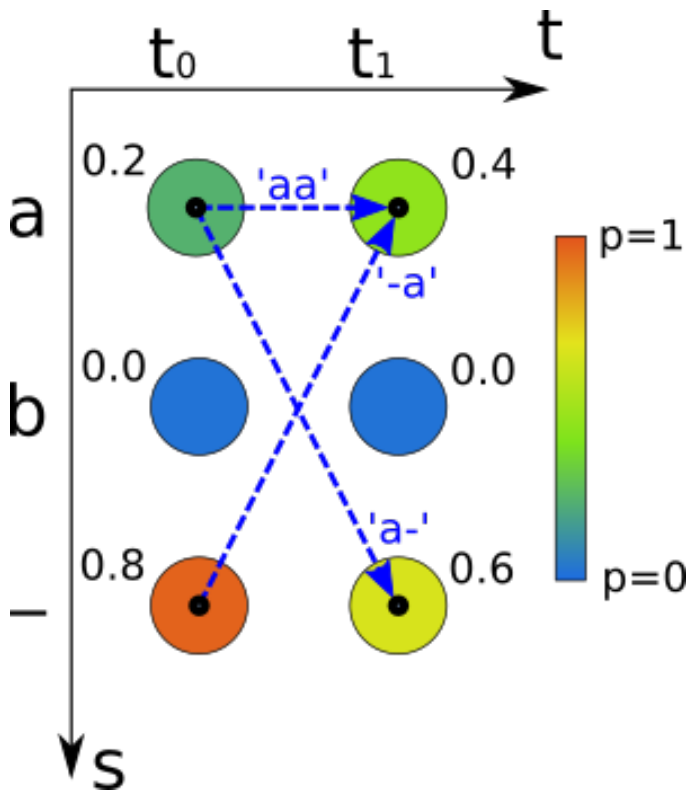
La sortie du réseaux de neurones est normalisée en matrice de probabilités sur l'ensemble de l'alphabet cible auquel on a rajouté un caractère *blank* (_).

Décodage d'une séquence

1. _AAGG_CC_C
2. Répétitions supprimées : _AG_C_C
3. *Blanks* supprimés : AGCC

Réseaux de neurones récurrents

Connectionist Temporal Classification

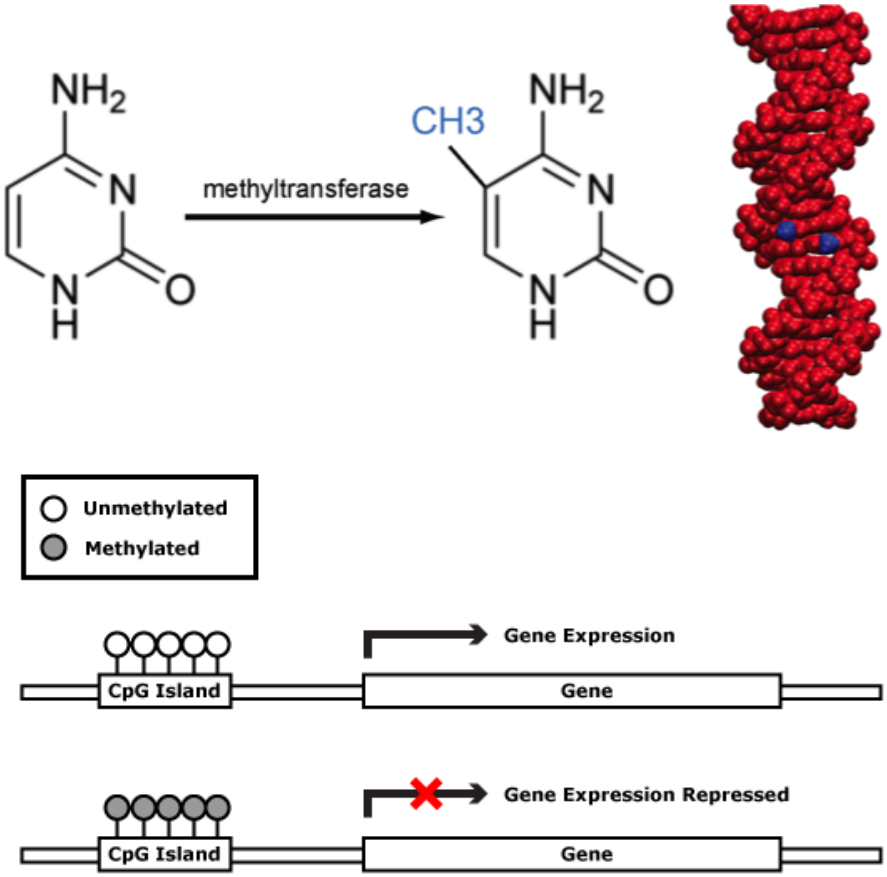


- 3 chemins donnent "a" après décodage, "aa", "_a" et "a_".
- p("aa") = 0.08, p("_a") = 0.32, p("a_") = 0.12.
- on les additionne et obtient 0.52.

- 1 chemin donne "" après décodage, "__".
- p("__") = 0.48.

La séquence "a" est alors plus probable que la séquence vide.

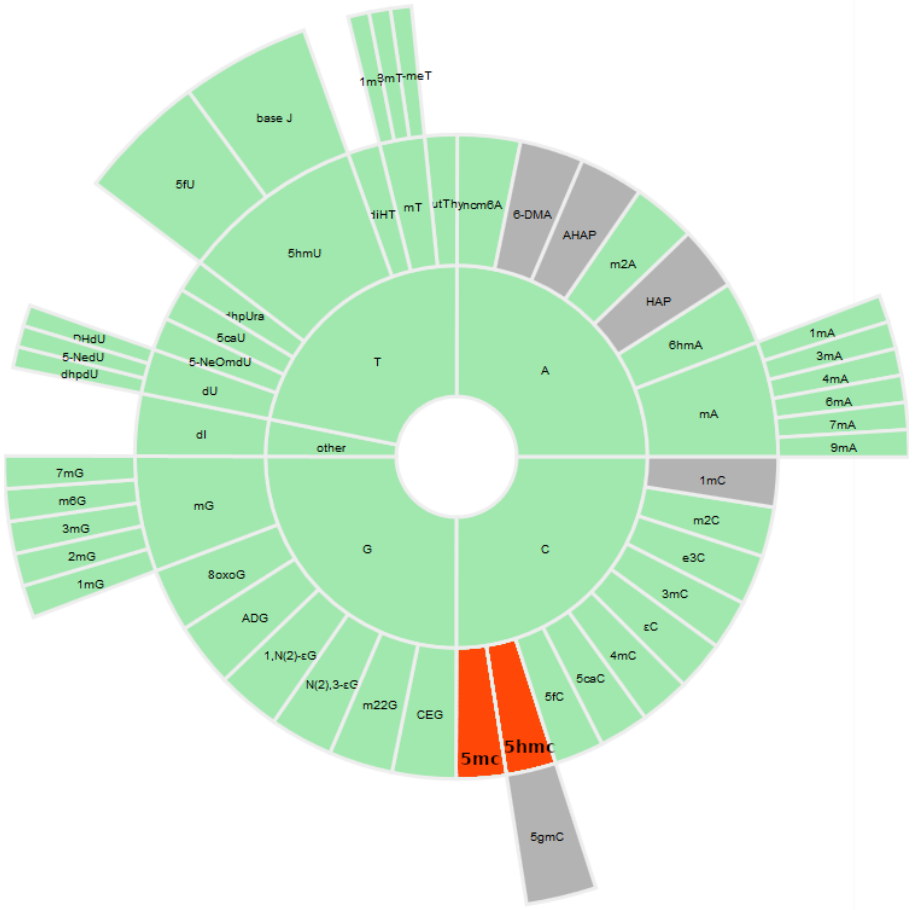
La méthylation de l'ADN



(Schémas de Mikhail Dozmorov)

1. Un groupe méthyle est ajouté sur le 5ème carbone d'une cytosine la convertissant en 5-méthylcytosine (5mC).
2. La Méthylation dans le promoteur est associée à l'inactivation du gène.

Les autres modifications chimiques



Sood, Viner, et Hoffman, ≪ DNAmod ≫.

Le groupe méthyle peut être oxydé pour donner une 5-hydroxyméthylcytosine (5hmC).

5hmC a un rôle déterminant dans la différentiation des cellules.[1]

[1] Ecsedi, Rodríguez-Aguilera, et Hernandez-Vargas, ≪ 5-Hydroxymethylcytosine (5hmC), or How to Identify Your Favorite Cell ≫.

AEROBICS

Analysis and Epigenetic Recognition Of dysBalanced Immune Cell plaSticity

- Le système immunitaire est flexible, (Ziegler and Buckner, 2009)
- Un déséquilibre de certains types de cellules du système immunitaire est un indicateur de la présence d'un cancer.

AEROBICS

Analysis and Epigenetic Recognition Of dysBalanced Immune Cell plaSticity

- Le système immunitaire est flexible, (Ziegler and Buckner, 2009)
- Un déséquilibre de certains types de cellules du système immunitaire est un indicateur de la présence d'un cancer.
- Étudier et mesurer ce déséquilibre.

AEROBICS

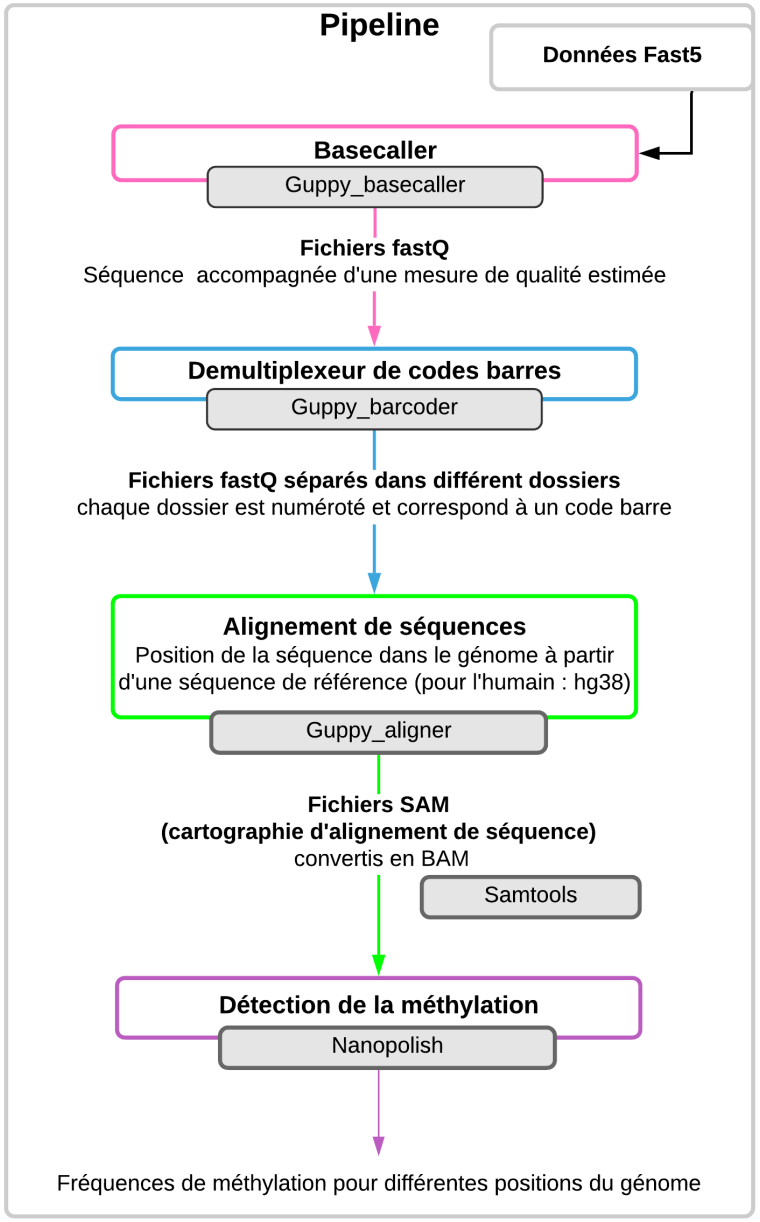
Analysis and Epigenetic Recognition Of dysBalanced Immune Cell plaSticity

- Le système immunitaire est flexible, (Ziegler and Buckner, 2009)
- Un déséquilibre de certains types de cellules du système immunitaire est un indicateur de la présence d'un cancer.

- Étudier et mesurer ce déséquilibre.
- Détecter correctement la méthylation dans l'ADN.
- Identifier les sous-types de cellules immunitaires en se basant sur ces modifications (5mC, 5hmC).

1. Définitions
   - Le MinION
   - Le séquençage ADN par nanopores
   - Le basecalling
2. Les algorithmes de machine learning
   - Modèles de Markov cachés
   - Réseaux de neurones récurrents
3. Contexte biologique
   - La méthylation de l'ADN
   - AEROBICS
4. Missions du stage
   - Développement d'un pipeline d'analyse du squiggle
   - Développement d'un basecaller personnalisé

Pipeline d'analyse du *squiggle*



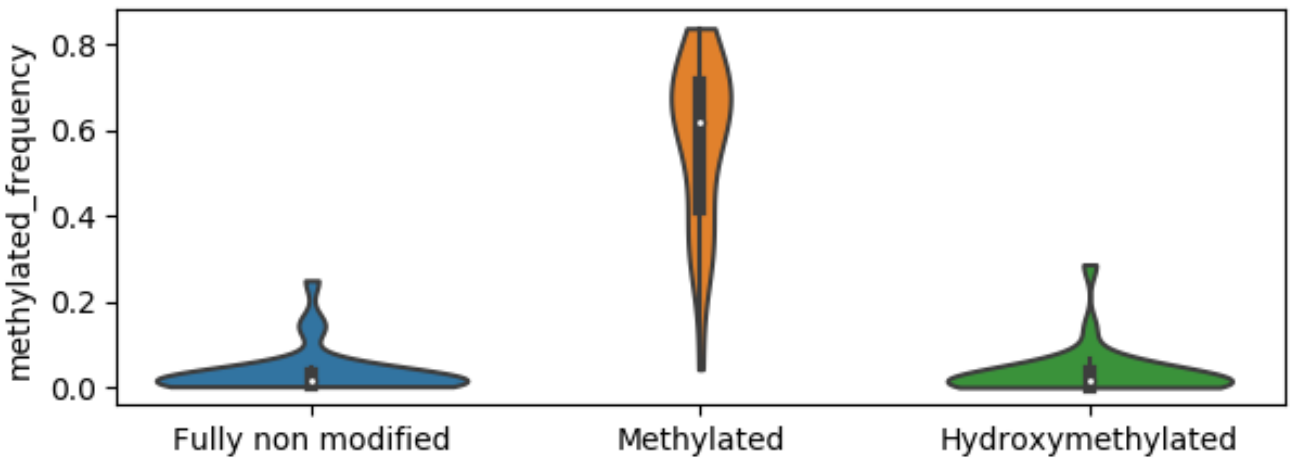
1. Basecaller (Guppy, RNN GRU).
2. Démultiplexeur, sépare les échantillons.
3. Alignement sur la référence.
4. Détection de 5mC avec Nanopolish (HMM).

Pipeline d'analyse du *squiggle*

Résultats

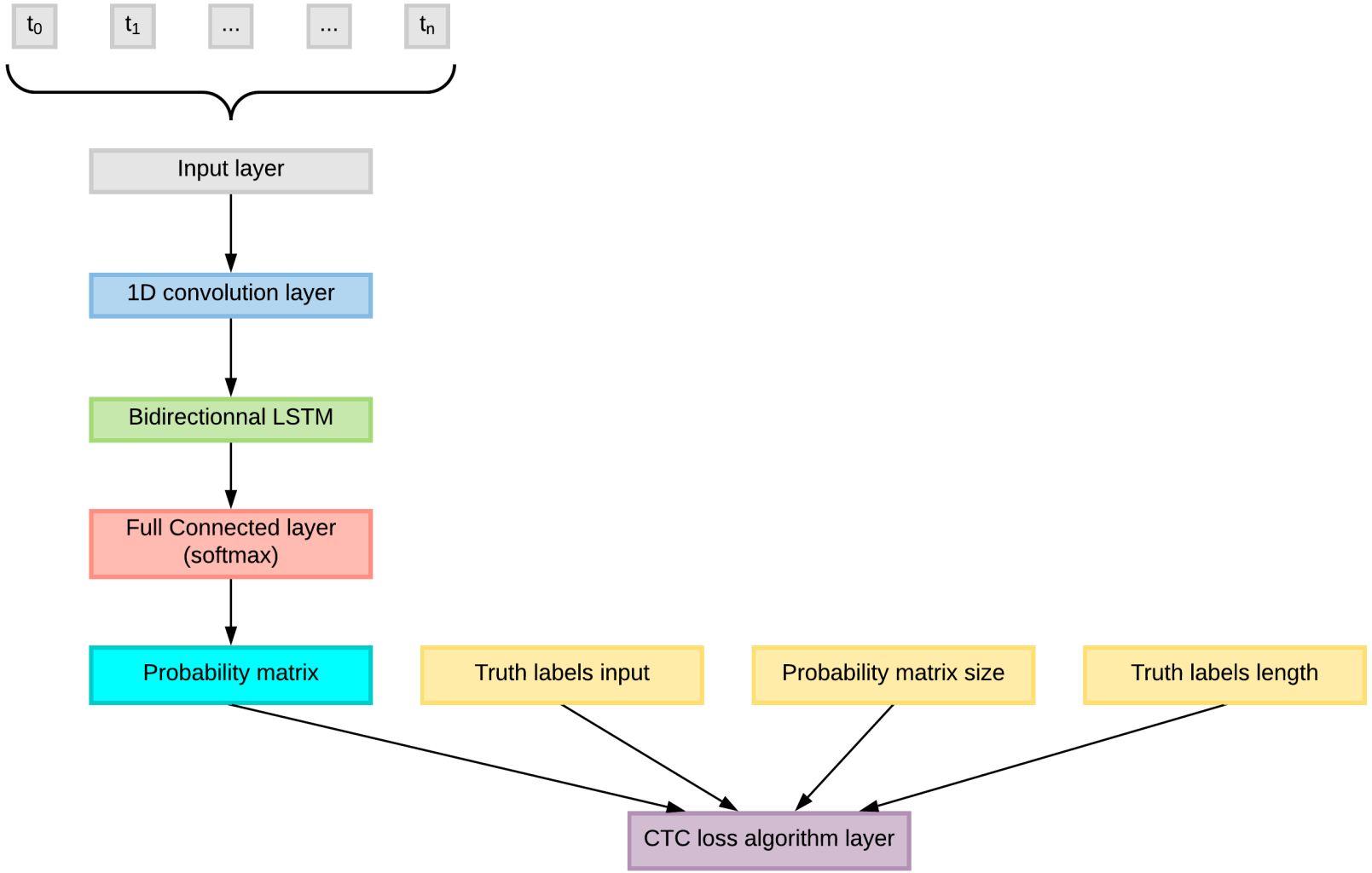
Des échantillons d'ADN synthétique nous servent de contrôles

1. Complètement non modifié.
2. Toutes les cytosines sont des 5mC.
3. Uniquement des 5hmC.



Nombre de sites CpG	2391329	10521	3485
Distribution du nombre de sites	99%	0.4%	0.1%
Méthylation moyenne	.014	.660	.019

Basecaller CNN-RNN-CTC



Basecaller CNN-RNN-CTC

Fonction objectif

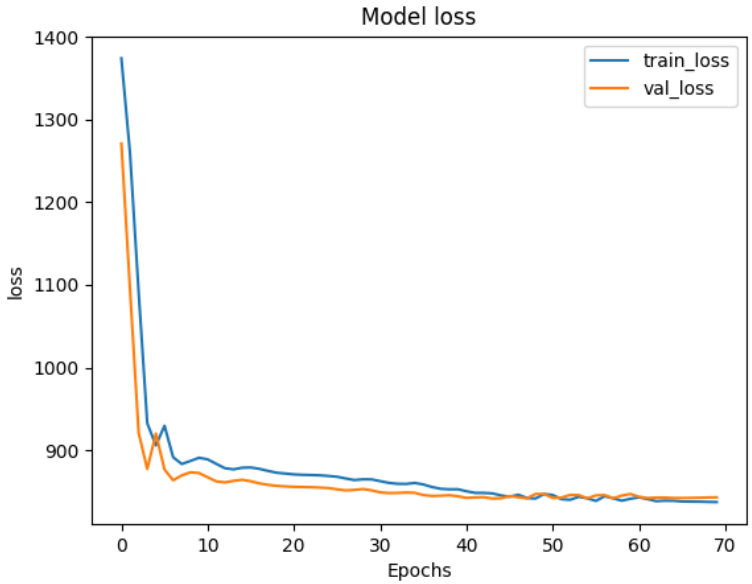
L'entraînement d'un CTC vise à maximiser le logarithme des probabilités de classifications correctes sur le jeu de données d'entraînement, formellement défini comme suit :

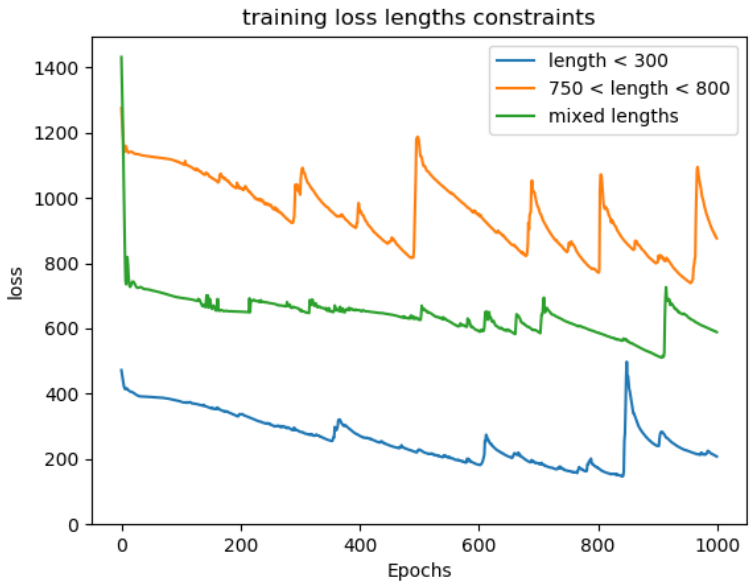
Avec S l'ensemble des données d'entraînement, $(x, z) \in S$, x les données d'entrée, z les séquences attendues :

$$-\sum_{(x,z)\in S} ln(p(z|x)) \tag{1}$$

Basecaller CNN-RNN-CTC

Résultats d'entraînements





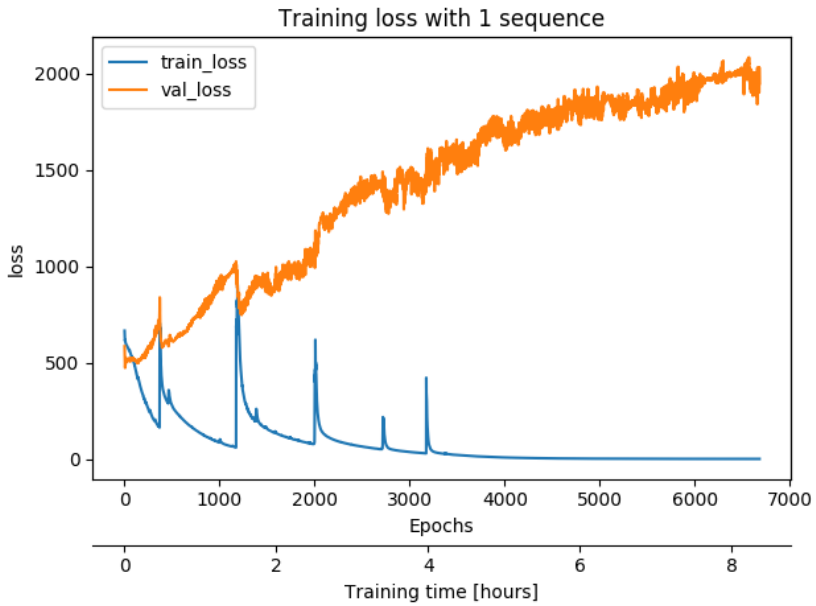
40 séquences d'entraînement et 10 de validation.

3 jeux de 4 séquences de longueurs spécifiques.

Le calcul de l'erreur est biaisé par la longueur des séquences.

Basecaller CNN-RNN-CTC

Résultats d'entraînements



8 heures d'entraînement sur une seule séquence.

- Erreur finale : 0.2748.
- Distance entre prédite attendue : 973.
- En supprimant les 973 dernier labels, distance = 0.

- Le modèle a bien appris par coeur.
- Des labels supplémentaires sont rajoutés durant la prédiction.

Pour aller plus loin

Pour aller plus loin

- Le modèle CTC implémenté reste un prototype.
- L'entraînement est long, et la prédiction faussée par un bug.

Pour aller plus loin

- Le modèle CTC implémenté reste un prototype.
- L'entraînement est long, et la prédiction faussée par un bug.
- Approfondir en modifiant la boucle d'entraînement, et en testant divers paramètres de décodage CTC, Graves et Jaitly, ≪ Towards End-to-End Speech Recognitionwith Recurrent Neural Networks ≫..

Pour aller plus loin

- Le modèle CTC implémenté reste un prototype.
- L'entraînement est long, et la prédiction faussée par un bug.
- Approfondir en modifiant la boucle d'entraînement, et en testant divers paramètres de décodage CTC, Graves et Jaitly, « Towards End-to-End Speech Recognitionwith Recurrent Neural Networks »..
- Implémenter une normalisation par batch pour accélérer l'entraînement, Ioffe et Szegedy, « Batch Normalization ».


Tank
yu!