



Analysis of DNA Sequence Classification Using CNN and Hybrid Models.

Hemalatha Gunasekaran , K Ramalakshmi , A Rex Macedo Arokiaraj , S Deepa Kanmani , Chandran Venkatesan , C Suresh Gnana Dhas.

Journal > Publié 16 Juil 2021.

Résumé

La séquence ADN/génomique complète des virus comme la COVID. Pour éviter une épidémie comme la COVID-19 il est essentiel d'utiliser l'identification et la classification des virus, telle que plusieurs techniques d'apprentissage automatique et de deep learning (DL) ont été utilisées pour mener à bien cette tâche.

Dans cet article les auteurs fournissent une étude détaillée sur la classification des séquences d'ADN utilisant les architectures CNN, CNN-LSTM et CNN-Bidirectional LSTM et le codage Label et K-mer. Par conséquent, la classification des séquences d'ADN joue un rôle essentiel en biologie computationnelle. Lorsqu'un patient est infecté par le virus, les échantillons prélevés sur le patient et les génomes sont séquencés. Dans ce travail, trois modèles de classification différents CNN, CNN-LSTM et CNN-Bidirectional LSTM sont utilisés pour la classification des séquences d'ADN. Les techniques de codage d'étiquette et de K-mer sont utilisées pour crypter la séquence d'ADN, ce qui préserve les informations de position de chaque nucléotide dans la séquence. Les couches d'incorporation sont utilisées pour intégrer les données des deux techniques. La couche CNN est utilisée comme étape d'extraction des entités, et elle est donnée comme entrée pour LSTM et LSTM bidirectionnelle pour la classification.

La méthode proposée est comparée aux différentes techniques pour prouver la robustesse du modèle : les modèles proposés, à savoir CNN, CNN-LSTM et CNN-Bidirectionnel LSTM. Le CNN et le LSTM bidirectionnel offrent une meilleure précision de 93,16% et 93,13%, respectivement.

Mot clé

CNN, LSTM, Deep learning, K-mer, CNN-Bidirectional, ADN.