A data-centric pipeline using convolutional neural network to select better multiple sequence alignment method

Mengmeng Kuang
Department of Computer Science
University of Hong Kong
mmkuang@cs.hku.hk

Hing-fung Ting
Department of Computer Science
University of Hong Kong
hfting@cs.hku.hk

ABSTRACT

Multiple sequence alignment (MSA) is widely used to find out the evolutionary relationship of every input sequence as well as the functional or structural roles of every aligned residue. Traditionally, the MSA problem was tackled by algorithm-centric approaches which had applied many classical computer algorithms (such as dynamic programming, divide-and-conquer algorithm and so on) and proven strategies (such as progressive strategy, non-progressive strategy, consistency-based strategy, iterative refinement etc.). Different single-algorithm MSA methods have different accuracies on different similarity protein families. Therefore, to integrate the advantages of different MSA methods, we present a brand-new data-centric pipeline using the convolutional neural network (CNN) [3] to choose better MSA method for different similarity protein families.

An MSA is very similar to a 2D picture, which has a good hierarchical structure. The conserved regions and the corresponding conserved columns in an MSA could be seen as boxes and lines in a picture. CNN is known to be very good at recognizing imperfect pictures which containing existed noises, which means it may perform well for recognizing draft MSAs. Briefly, the method first using a quick MSA method to construct large-scale draft MSAs from the simulated protein families produced by protein simulation tool INDELible [1]. The main point is training a classifier by CNN which employing the draft MSAs as input and giving the better MSA method as output.

In our research, we simulated more than 640,000 protein families with sequence number range from 3 to 64. The fastest (but not accurate) mode of a famous MSA tool, Mafft(FFT-NS-1) [2], with default parameters used for constructing draft MSAs from those families. We regard these MSAs as two-color images, one color for the aligned residues, and the other color for the gaps. Two layers of CNN and a fully connected layer with 0.5 of dropout were used for training the decision model.

The preliminary results suggest more than 85% accuracy in classification of choosing better alignment solution between the newest versions of Mafft(L-INS-i) and Mafft(G-INS-i). Currently, we are improving the performance of this pipeline by selecting better categories for protein families and fine-tuning the decision model.

BCB '20, September 21–24, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7964-9/20/09.
https://doi.org/10.1145/3388440.3414909

CCS CONCEPTS

• **Applied computing** → *Molecular sequence analysis.*

KEYWORDS

Multiple sequence alignment, classification, convolutional neural network, data-centric, decision model

ACM Reference Format:
Mengmeng Kuang and Hing-fung Ting. 2020. A data-centric pipeline using convolutional neural network to select better multiple sequence alignment method. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '20), September 21–24, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3388440.3414909

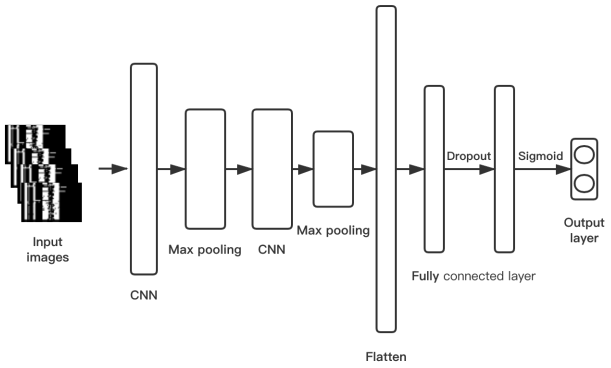

Figure 1: The structure of decision model

ACKNOWLEDGMENT

Ting is partially supported by the GRF Grant HKU17208019.

REFERENCES

[1] William Fletcher and Ziheng Yang. 2009. INDELible: a flexible simulator of biological sequence evolution. *Molecular biology and evolution* 26, 8 (2009), 1879–1888.
[2] Kazutaka Katoh and Daron M Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* 30, 4 (2013), 772–780.
[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems.* 1097–1105.