Review of: "Deep-BGCpred: A unified deep learning genome-mining framework for biosynthetic gene cluster prediction"

Pei-Yu Lin[1], Pao-Yang Chen[1]

1 Academia Sinica

Potential competing interests: The author(s) declared that no potential competing interests exist.

Yang *et al* presented Deep-BGCpred that implemented a deep learning method for biosynthetic gene clusters (BGCs) identification within bacterial genomes. It is built based on its previous version DeepBGC[1].

Deep-BGCpred utilized a deep learning genome-mining framework, in which two customized strategies are perfomed to improve the prediction of BGCs. These two strategies, the sliding-window based and the dual-model serial screening, advance the BGC boundary identification by removing those false positive regions that frequently found with machine learning approaches.

Comparing to other machine learning-based (e.g., ClusterFinder) and the rule-based (e.g., antiSMASH) tools, the authors showed that Deep-BGCpred has a higher accuracy. It would be more helpful and informative to the readers if more explanations are offered on why the previous methods tend to identify more false positive regions. For instance, was it because that the over-fitting commonly seen in previous methods tends to result in predicting wrong BGC regions? Deep-BGCpred with the dual-models of serial screening has two steps to double-check the BGC regions using the summarised Pfam scores and BGC product classifications, therefor reduces the false positive prediction.

Deep-BGCpred is also able to detect novel BGC comparing to the rule-based methods. The results of the prediction arouse our curiosity that what are differences of the candidate BGCs predicted by different strategies, and if different tools tend to find particular types of BGCs. Additionally, as performance evaluation it would be helpful to explore those novel BGCs that cannot be found by the other methods. For example, does Deep-BGCpred tend to predict novel resistance-genes based BGCs[2] that were not easily detected by rule-based methods? Also, what about those BGC predicted only by rule-based but not by machine learning-based? Those BGC could be a clue for future improvement of machine learning-based BGC predictors. Lastly, in addition to bacteria if Deep-BGCpred can be extended to predict other species such as fungi or plants, that would make Deep-BGCpred a very useful and popular tool.

References

1. ^*Geoffrey D Hannigan, David Prihoda, Andrej Palicka, Jindrich Soukup, et al. (2019). A deep learning genome-mining strategy for biosynthetic gene cluster prediction. doi:10.1093/nar/gkz654.*
2. ^*Phuong Nguyen Tran, Ming-Ren Yen, Chen-Yu Chiang, Hsiao-Ching Lin, et al. (2019). Detecting and prioritizing biosynthetic gene clusters for bioactive compounds in bacteria and fungi. Appl Microbiol Biotechnol, vol. 103 (8), 3277-3287. doi:10.1007/s00253-019-09708-z.*