Meetings International SciTechnol

Ruilin Li et al., J Appl Bioinforma Comput Biol 2019, Volume: 8
DOI: 10.4172/2329-9533-C1-008

International Conference on
BIOINFORMATICS & SYSTEM BIOLOGY
&
3rd International Conference on
SURGERY & ANAESTHESIA

March 20-21, 2019 | Singapore City, Singapore

Gene prediction algorithm based on feature selection of open reading frame for metagenomic sequences

Ruilin Li[1,2], Xinyin Han[1,2], Zhipeng He[1,2], Xiaoyu He[1,2], Shanyu Chen[1,2], Haidong Zhu[1,2], Yu Zhang[1,2], Dan Zhao[1,2], Weizhong Li[3], and Beifang Niu[1]

[1] Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, China
[3] J. Craig Venter Institute, USA

Gene prediction is an important approach to improve the annotation of metagenomic genes. A variety of gene prediction models based on different principles had been implemented, with emphasis on statistical models, Markov or improved Markov models, deep learning models, and so on. The current gene prediction algorithms, such as FragGeneScan, Prodigal, MetaGeneAnnotator, Orphelia, Glimmer3, GeneMarkS-2, were specially designed for short fragments or whole genomes; however, the former will result in the identified genes being incomplete and the latter is not suitable for unknown species. Meanwhile, according to our previous benchmark results of these algorithms, the prediction error rate was relatively high (27.10%~54.70%), especially for datasets with low coverage (staggered dataset). In this study, we proposed an algorithm based on feature selection of ORFs named as Consensus, which combined the ORFs generated from known models, extracted the ORFs' feature matrix and the corresponding label matrix. Finally, the optimal solution was obtained by the least square's solution of the feature and label matrixes. The overall indicator of gene prediction via Consensus was better than that of single software (F-score was 82.94% on staggered dataset). Even more remarkably, we compared the results of models using two longer assembled scaffolds datasets of the real mock metagenomic samples containing 20 bacterial strains from NCBI (National Center for Biotechnology Information) instead of simulated reads, which would truly reflect the predictive power of the models. We believe our findings will improve the study of novel genes and annotation pipelines in unknown metagenomic species.

Biography

Ruilin Li is PhD candidate and her research interests include high performance computing and bioinformatics.

lirl@sccas.cn

Notes: