Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Applied Comparative Genomics

2D approach to metagenomics taxonomy classification

Victor Wang [1,], Qing Dai [2] and Harrison Huh [3,]

[1]Biomedical Engineering, Johns Hopkins University, Baltimore, 21218, United States
[2]Environmental Health and Engineering, Johns Hopkins University, Baltimore, 21218, United States
[3]Computer Science, Johns Hopkins University, Baltimore, 21218, United States

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Assembling genomes from metagenomics sequencing data is a difficult problem. One challenge is that the shotgun reads need to be labeled with their species of origin. To address this challenge, researchers have proposed several metagenomic sequence classifiers for comparative analyses, such as Phymm, Kraken, and Centrifuge. We propose using Convolution Neural Networks (CNNs) from computer vision to address the problem. Reads can be encoded as 2D images by using Gramian Angular Fields.
Results: We used genomes of 10 species from the RefSeq database as our training dataset and encoded the reads in 50-mers and converted them into 2D with Gramian Angular Field. We tested our conversion on a baseline CNN, resulting up to 71 % accuracy. As expected, the model had better accuracy under longer read lengths and suffered some performance under higher error rate in the reads.
Contact: hhuh1@jhu.edu, qingdai@jhu.edu, xwang145@jhu.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Metagenomics is one of the most useful culture independent methods to investigate complex microbial communities [4]. It provides genomic information directly from complex environmental samples. However, the interpretation of the sequencing data is limited. This is partly because it's hard to identify which sequences come from which species. Therefore, a good metagenomic taxonomy classifier would be really helpful to get more insights on the structure and function of a microbial community. There are existing algorithms for metagenomics classification such as Phymm [1], Kraken [6], and Centrifuge [3].

In computer vision, the problem of image classification has been addressed with powerful deep learning tools such as Convolutional Neural Networks (CNNs). In this paper, we explored the idea of encoding the information from 1D shotgun sequencing reads as 2D arrays or "images" and then training a CNN based classifier to assign species labels to the images. Because 1D reads can be thought of as a time series, the Gramian Angular Field [5] is an appropriate method to encode the information into a 2D representation.

Our CNN takes in the 2D data and outputs the label weights. The label with the most weight will be the prediction. For this project, we partitioned labeled data for supervised training and testing.

2 Methods

2.1 Data

The sequencing data is generated from sequencing simulators with different error rates and read lengths. We first tried the complete genome of 10 kinds of subspecies of Francisella Tularensis. Then choose 10 different species from 10 different families to train and test our model. All our raw genomic data are from the RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/). The simulation would generate sequencing reads from these 10 genomes as our metagenomic data.

2.1.1 Illumina/Short Reads Sequencing simulation

We use Mason [2] to simulate Illumina sequencing results. The package is available in bioconda (https://anaconda.org/bioconda/mason). Briefly, Mason works by generating reads with given read length from a given genome, and these reads may have mismatches, insertions, and deletions. To simulate a real Illumina sequencing result, the error rates vary

depending on the read position. We generate 10,000 150 bp reads for each species as our metagenomic data.

2.1.2 Nanopore/Long Reads Sequencing simulation

We wrote a script to simulate long reads sequencing results. Our script generates the reads with random mismatch errors without positional different error rate. Once set up, the reads we generated would have constant length and constant error rate. We generated sequencing reads with 500 bp, 800 bp, 1000 bp and 1200 bp at 1%, 2%, 5% and 10% error rate level. We use 500 bp and 1000 bp with error rate at 1% and 10% level as our metagenomic data to train and test our model. While the data at all levels of read length and error rate are used for the comparison of classification accuracy with Centrifuge.

2.2 Encoding reads as images

Reads were first converted from a string of nucleotide bases into integers (i.e. A $\rightarrow$ 0, C $\rightarrow$ 1, G $\rightarrow$ 2, T $\rightarrow$ 3). Each 50-mer in the sequence of integers were treated as a base 4 number and mapped to its corresponding base 10 number. There are 4^{50} possible 50-mers.

Gramian Angular Fields (GAF) were proposed by Z. Wang and T. Oates (2015) to encode time series data as images to enable the use of image classification techniques. The GAF is a square matrix representation of the time series in polar coordinate space. The GAF is constructed as follows:

Given a time series $X = \{x_1, x_2, ..., x_n\}$ of n observations, rescale X such that $x_i \in [-1, 1]$.

$$\tilde{x} = \frac{x_i - max(X) + x_i - min(X)}{max(X) - min(X)} \tag{1}$$

Compute the polar coordinates of the rescaled time series $\tilde{X}$. ϕ is the phase and r is the radius.

$$\begin{cases} \phi_i = \arccos(\tilde{x}_i), & -1 \leq \tilde{x}_i \leq 1, \tilde{x}_i \in \tilde{X} \\ r_i = \frac{i}{N}, & i \in [1, 2, ..., n] \end{cases} \tag{2}$$

The GAF is defined as

$$GAF = \begin{Bmatrix} \cos(\phi_1 + \phi_1) & \dots & \cos(\phi_1 + \phi_n) \\ \cos(\phi_2 + \phi_1) & \dots & \cos(\phi_2 + \phi_n) \\ \vdots & \ddots & \vdots \\ \cos(\phi_n + \phi_1) & \dots & \cos(\phi_n + \phi_n) \end{Bmatrix} \tag{3}$$

2.3 CNN architecture

Currently we use the baseline CNN, taken from a public test on CIFAR10 data. It has 2 convolutional layers, each followed by 2x2 max pool layers. Lastly, there are 3 consecutive fully connected layers. Currently, we are classifying only up to 10 labels. All code for this project is on github https://github.com/DottedGlass/MetagenomicsCNN

3 Results

3.1 Initial Experiment with 10 Subspecies

We have tested on our current implementation of the 2D encoding from the short reads of 150bp. Currently the CNN does not change in loss rate per iteration and only gives 10% accuracy (out of 10 labels). This suggests that our model is no better than randomly guessing. We suspect that this may be because of two reasons: (1) distinguishing between the 10 subspecies of *Francisella Tularensis* is too challenging and (2) short reads do not provide enough training information.

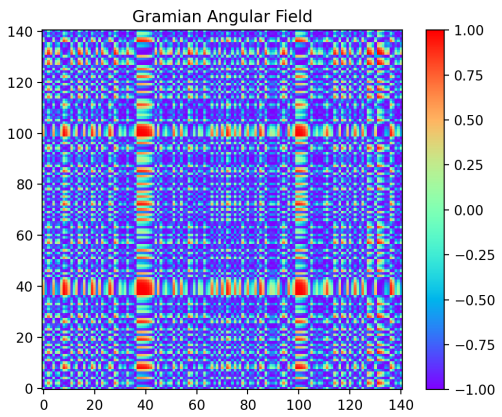

Fig. 1. This is an example of a GAF on 10-mers of a 150bp read. Since $150-10+1 = 141$, the matrix is 141 by 141. Note that the matrix is symmetric.

3.2 Second Experiment

Training and testing on 10 distinct species, we were able to get a better result of 37% accuracy. While it was not a great result, it was an improvement over the previous and showed that the convolutional neural network had predictability on the genomes. One of the reasons why it did not perform as well as we hoped was due to the small sample size in our training phase. Using only 2000 reads of 500 length for each species, it was not enough for even 1x coverage for any of the species. We hypothesized that with greater training on more coverage, we will obtain a more accurate result.

3.3 Final Experiment

The final experiment trained on 40x coverage worth of reads, experimenting on 500bp read length and 1000bp read length. We also tested on 1% and 10% error for each read length.

How does read length affect accuracy?
As hypothesized, a longer read length has improved the accuracy. On 1% error rate, 500bp read length had about 66% overall accuracy while 1000bp read length yielded 71%. On 10% error rate, 500bp read length performed at 55% overall accuracy and the 1000bp read length with 66% overall accuracy.

How does error rates on reads affect the accuracy
The error rate has an overall diminishing effect on the accuracy of the model. From 1% to 10% error rate, 500bp read length fell by 11% in accuracy from 66% to 55%. In the case of 1000bp read length, that accuracy dropped from 71% to 66%. However, it is notable that some individual genomes had improved accuracy as a result of increased error rate.

3.4 Comparison with Centrifuge

We tested how our method and Centrifuge performs on short reads and long reads.

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Species	Read 500, Error 1%	Read 500, Error 10%	Read 1000, Error 1%	Read 1000, Error 10%
Thermosynechococcus elongatus BP-1 (NC_004113)	82%	40%	85%	72%
Thermofilum pendens (NC_008698)	52%	67%	69%	50%
Methanobrevibacter smithii ATCC 35061 (NC_009515)	84%	46%	85%	78%
Coxiella burnetii (NC_010117)	50%	75%	80%	46%
Acidilobus saccharovorans 345-15 (NC_014374)	66%	46%	47%	69%
Streptococcus pneumoniae AP200 (NC_014494)	55%	63%	81%	64%
Chlamydia psittaci VS225 (NC_018621)	29%	54%	33%	54%
Sulfolobus acidocaldarius N8 (NC_020246)	70%	59%	74%	70%
Haloarcula hispanica N601 (NC_023013)	91%	45%	94%	87%
Escherichia coli K-12 (NZ_LN832404)	80%	59%	58%	73%
Overall	**66%**	**55%**	**71%**	**66%**

Fig. 2. Accuracy of our CNN model trained on four sets of simulated reads. Read length 500 and 1000, with substitution error rates 1% and 10%.

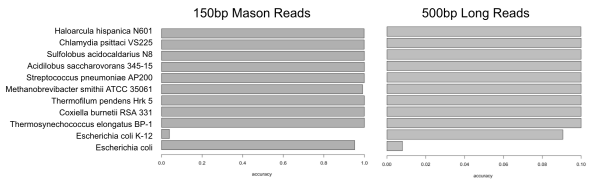

Fig. 3. Confusion Matrix for 500bp read length with 1% error rate.

According to the result in Fig. 3, we can see the read length can help centrifuge differentiate E. coli K-12 strain from the E. coli. In the 150bp mason simulator data, only 3.87% of the reads of E. coli K-12 strain are identified from E. coli. While about 90.1% of the reads of E. coli K-12 strain are identified in 500 bp data, which indicates that the read length may have a great influence on the accuracy for subspecies identification. This might be due to the higher probability of the long reads containing unique fragments of E. coli K-12 strain.

The error rate is another factor that influences the accuracy of classification. We compared the classification accuracy controlling the read length equal to 500 bp. According to Fig. 4 and Fig. 5, the centrifuge classification accuracy would change from about 100% to 97% with error rate from 1% to 10%. This accuracy decrease can be observed in all species we choose. Besides, the error rate also changes the identification of E. coli K-12 strain from E. coli. The identified reads are 90.1%, 88.5%, 83.5% and 68.0% with error rate at 1%, 2%, 5% and 10% level respectively.

Besides, the centrifuge identification of subspecies is more sensitive to error rate. The accuracy for the 10 different species classification did not change much with the error rate from 1% to 5%, while the accuracy for E. coli K-12 strain identification changed a lot at these error rate levels. Since the genomic difference between different species is much greater than the genomic difference between subspecies, the classification of a higher phylogenetic level would have a greater fault tolerance.

This fault tolerance would vary with different read length. In the Centrifuge result for 1000 bp data, the classification accuracy with 10% error rate can reach over 99.8% for all 10 species. For real Nanopore sequencing data, where the read length would be a few thousand or longer, the benefit of long read length could account for the downside of higher error rate. Centrifuge results for 500 bp, 800 bp, 1000 bp and 1200 bp at 1%, 2%, 5% and 10% error rate levels can be found here (https://docs.google.com/spreadsheets/d/1ZCnp55spOz-9Fcjck2udYWpihs4MnfSvFsdnmAOmYIQ/edit#gid=0)

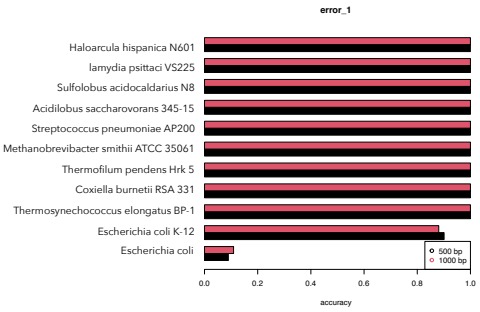

Fig. 4. Centrifuge classification under 1% error

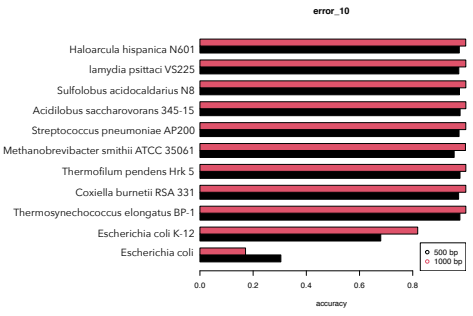

Fig. 5. Centrifuge classification under 10% error

4 Discussion

4.1 Analysis of Early Results

Our initial test with CNN on our current image encoding does not yield any predictability. We determined that attempting to classify 10 subspecies was too difficult that we decided to tone down our goal to 10 microbial species

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

each from a different family. In the second attempt, the model performed better with 37%, proving that the 10 distinct species were somewhat easier to predict. However we merely trained on the data of 2000 samples of 500bp reads which obviously is not enough for 1x coverage for most genomes. Thus we had to figure out a better way of generating our 2D images without straining too much on the GPU. Using reads of 500 and 1000 base pairs and training on at least 40x coverage, we were able to yield a much better result.

4.2 Overfitting and K-mers

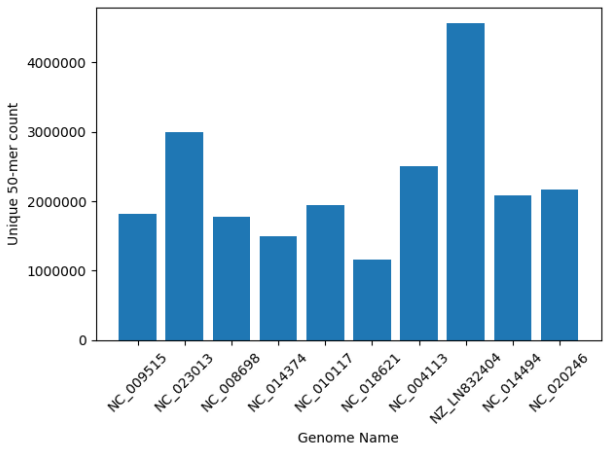

Fig. 6. Unique 50-mer count. This notes the number of 50-mers for each genome which are not present in other genomes

One concern with convolutional neural networks is that the model may have been trained to overfit. Our image was made using 50-mers which is beyond the standard use of k-mers in most genomics research. Analyzing the genomic data of our selected species, we learned that the majority of the 50-mers were found only once in each genome. Moreover, most of the 50-mers found in each genome were unique to that species, approximately 99.999% across the board. In order to generalize our model more with other microbial species, we may have to use smaller k-mers.

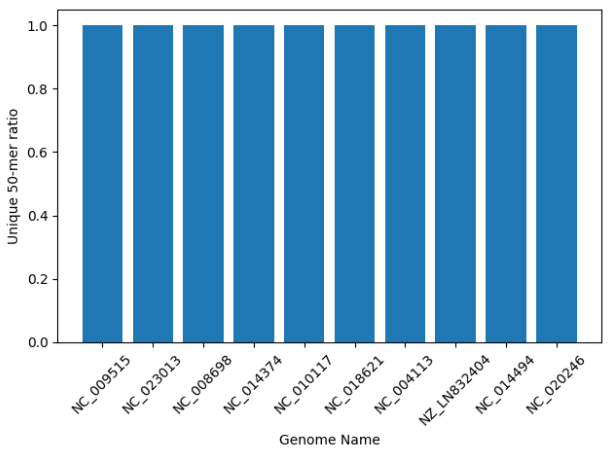

Fig. 7. Unique 50-mer ratio. The percentage of 50-mers that are not present in other genomes.

4.3 Reflection on the 2D Approach

Lastly, we examine the effectiveness of our 2D imaging approach to classifying the reads.While the model seems to show some predictability, we are unsure if it is due to the way we trained the model, limiting to 10 species due to the tendency of CNNs to overfit. Furthermore, we had to train on 40x coverage, increasing the training time by a significant amount. Compared to Centrifuge which gave very quick results, our model took several hours to complete just one epoch of training, transforming the 1D data into 2D along its way as storing the hundreds of thousands of transformed images was not viable.

The initial appeal of transforming a one-dimensional streaming data into a two dimensional image data was to utilize the convolutional neural network for a problem otherwise inapplicable. However, whether or not the transformation provides any additional information not present in the 1-D data is an unanswered question. The most evident caveat to the method is that it quadratically multiplies the input data, significantly consuming the computational space and increasing the train time of the models.

Acknowledgements

We would like to thank Dr. Michael Schatz and Melanie Kirsche for sponsoring our project. Computation was done using Maryland Advanced Research Computing Center (MARCC).

References

[1]Arthur Brady and Steven L Salzberg. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated markov models. *Nat. Methods*, 6(9):673–676, September 2009.

[2]Manuel Holtgrewe. Mason: a read simulator for second generation sequencing data. 2010.

[3]Daehwan Kim, Li Song, Florian P Breitwieser, and Steven L Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.*, 26(12):1721–1729, December 2016.

[4]Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata. Corrigendum: Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.*, 35(12):1211, December 2017.

[5]Zhiguang Wang and Tim Oates. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[6]Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, 15(3):R46, March 2014.

Supplement Figures

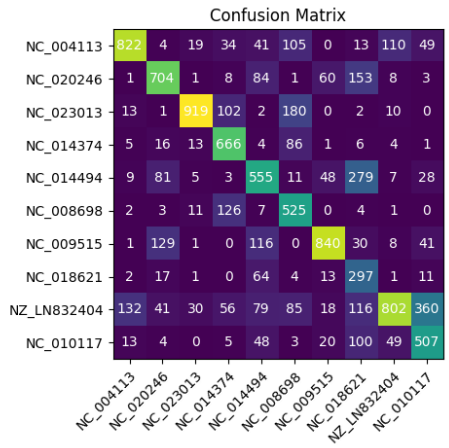

Fig. 8. Confusion Matrix for 500bp read length with 1% error rate.

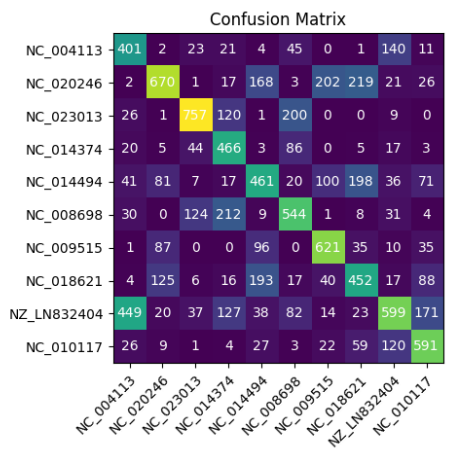

Fig. 9. Confusion Matrix for 500bp read length with 10% error rate.

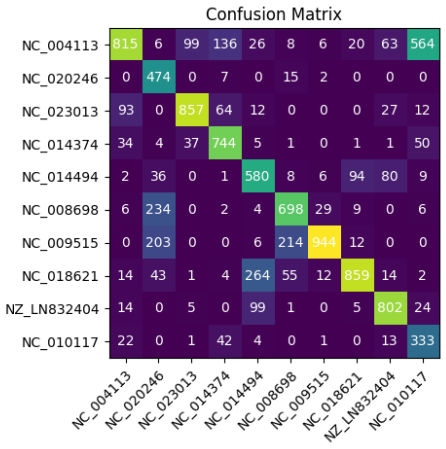

Fig. 10. Confusion Matrix for 1000bp read length with 1% error rate.

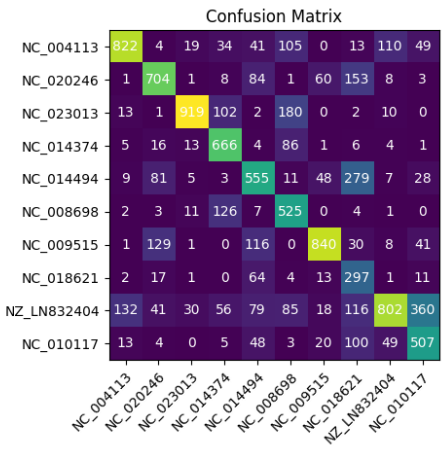

Fig. 11. Confusion Matrix for 1000bp read length with 10% error rate.