Estimating the Distribution of Viral Taxa in Next-generation Sequencing data using Artificial Neural Networks

Moritz Kohls

University of Veterinary Medicine Hannover: Tierärztliche Hochschule Hannover

Magdalena Kircher

University of Veterinary Medicine Hannover: Tierärztliche Hochschule Hannover

Jessica Krepel

University of Veterinary Medicine Hannover: Tierärztliche Hochschule Hannover

Pamela Liebig

University of Veterinary Medicine Hannover: Tierärztliche Hochschule Hannover

Klaus Jung (✉ klaus.jung@tiho-hannover.de)

Tierärztliche Hochschule Hannover Institut für Tierzucht und Vererbungsforschung

<https://orcid.org/0000-0001-9955-748X>

Research article

Keywords: Artificial neural network, Virus classification, Supervised learning, Next-Generation Sequencing, Viral metagenomics

Posted Date: December 17th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-127809/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

Estimating the distribution of viral taxa in next-generation sequencing data using artificial neural networks

Moritz Kohls^{*}, Magdalena Kircher, Jessica Krepel, Pamela Liebig and Klaus Jung

^{*}Correspondence:

moritz.kohls@tiho-hannover.de

Institute for Animal Breeding and Genetics, University of Veterinary Medicine Hannover, Bünteweg 17p, 30559 Hannover, GER

Full list of author information is available at the end of the article

Abstract

Background: Estimating the taxonomic composition of viral sequences in a biological sample processed by next-generation sequencing is an important step for comparative metagenomics. For that purpose, sequencing reads are usually classified by mapping them against a database of known viral reference genomes. This fails, however, to classify reads from novel viruses and quasispecies whose reference sequences are not yet available in public databases.

Methods: In order to circumvent the problem of a mapping approach with unknown viruses, the feasibility and performance of neural networks to classify sequencing reads to taxonomic classes is studied. For that purpose, taxonomy and genome data from the NCBI database are used to sample artificial reads from known viruses with known taxonomic attribution. Based on these training data, artificial neural networks are fitted and applied to classify single viral read sequences to different taxa. Model building includes different input features derived from artificial read sequences as possible predictors which are chosen by a feature selection method. Training, validation and test data are computed from these input features. To summarise classification results, a generalised confusion matrix is proposed which lists all possible misclassification combination frequencies. Two new formulas to statistically estimate taxa frequencies are introduced for studying the overall viral composition.

Results: We found that the best taxonomic level supported by the NCBI database is that of viral orders. Prediction accuracy of the fitted models is evaluated on test data and classification results are summarised in a confusion matrix, from which diagnostic measures such as sensitivity and specificity as well as positive and negative predictive values are calculated. The prediction accuracy of the artificial neural net is considerably higher than for random classification and posterior estimation of taxa frequencies is closer to the true distribution in the training data than simple classification or mapping results.

Conclusions: Neural networks are helpful to classify sequencing reads into viral orders and can be used to complement the results of mapping approaches. The machine learning approach is not limited to already known viruses. In addition, statistical estimations of taxa frequencies can be used for subsequent comparative metagenomics.

Keywords: Artificial neural network; Virus classification; Supervised learning; Next-Generation Sequencing; Viral metagenomics

Background

Next-generation sequencing (NGS) is now regularly used to identify viral sequences in the biological sample of an infected host in order to relate the presence of a virus with symptoms of the host [1–3]. Furthermore, NGS data is regularly used to determine the taxonomic composition of pathogens for the purpose of comparative metagenomics [4] or reconstruction of networks between microbial taxa [5]. Most computational virus detection pipelines or pipelines for determining the taxonomic composition map sequencing reads or assembled contigs against viral reference sequences available in databases [6–10]. These mapping approaches have been proven to be successful in a large number of examples, however, they mostly fail to classify reads from new emerging viruses whose sequences are not yet deposited in a database. Besides such mapping approaches, machine learning approaches have been used to classify reads to taxonomic ranks [11] such as genus, family, order, class and phylum. Among the machine learning approaches listed by Peabody *et al.*, only one approach studies the performance of deep learning models for taxonomic classification (Rasheed *et al.* [12]). Rasheed *et al.* use artificial neural networks (ANN) to classify reads to different taxonomic ranks and evaluate their models on datasets representing about 500 microbial species. Some approaches have also been undertaken to use machine learning models on NGS data to identify or classify reads that belong to novel emerged viruses. Zhang *et al.* [13] as well as Bartoszewicz *et al.* [14] employ diverse machine learning models to identify viruses that can infect humans. Whilst Zhang *et al.* applied the *k*-nearest neighbour (*k*-NN) model using *k*-mer frequencies as input features for their classification, Bartoszewicz *et al.* employed a convolutional neural network (CNN) on simulated reads and contigs.

In contrast to the approaches by Zhang *et al.* as well as Bartoszewicz *et al.*, in this work we study the ability of an ANN to classify sequencing reads on the taxonomic rank of nine biological orders. While the ANN studied by Rasheed *et al.* was fitted on sequence reads from a rather small number of microbes, we train our model on the basis of several thousand viral sequences available at the NCBI database [15]. In comparison to support vector machines or linear discriminant analysis, ANNs are reported to achieve the highest accuracy rates on test data, though it depends on the specific classification problem [16, 17].

Before fitting ANNs, we link available viral reference sequences with their taxonomic description and study the distribution of different taxonomic ranks (phyla, subphyla, classes, subclasses, orders, etc.) to identify on which of these ranks a solid basis for building a training dataset would be reasonable. Next, we extract *k*-mer distributions and inter-nucleotide distances as input features from the available reference sequences. In total, we generated 120 input features.

In the methods section, we describe the building of the training data in more detail, as well as the ANN models used. Furthermore, we present the new formulas for estimating the frequency distribution of viral orders in a sample. In the results section, we present a simulation study in which we evaluate the performance of the ANNs and demonstrate how to estimate the frequency distributions. We also compare the precision of the estimations with frequency distributions derived from a mapping approach. Finally, we discuss benefits and limitations of our approach.

Methods

Taxonomy and genome data

A taxonomy database containing data that are needed for taxonomic classification can be downloaded from the ftp-server <ftp://ftp.ncbi.nih.gov/refseq/release/viral> of the National Center for Biotechnology Information [15, 18]. We downloaded these data in March 2020.

A section of the taxonomy database file `fullnamelineage.dmp` from this ftp-server is depicted in Figure 1. In this file, taxonomic information about species are listed in rows. Beginning with archaea, bacteria and fungi, viral data is included from line 2,007,414 onwards. Row entries contain species names and taxonomy ranks which are unfortunately neither labelled nor complete.

We assign these taxa to taxonomy ranks based on the International Committee on Taxonomy of Viruses [19]. To enable correct assignments, we use the taxa endings summarised in Table 1. The ending "...viridae" indicates to which family the virus belongs. The Bovine adenovirus 4 on line 2,007,418 from Figure 1, for example is a member of the family of Adenoviruses.

As the taxonomy database does not provide any unique NC identifier, the species names of the reference genome file are searched in the taxonomy file to link genome with taxonomy data.

We obtain viral reference genome FASTA files [20] again from the NCBI database and concatenate the three files into a single one. The first lines of the FASTA file (Figure 2) show its composition. Line one consists of a unique identifier together with its species name. The genome sequence consisting of the single-letter codes A, C, G and T that represent the four nucleobases follows from line two onwards. Combining genome and taxonomy data, we create a structured taxonomy table including FASTA file species which replaces the unstructured original taxonomy file. As there is data missing, we only choose species with available taxonomy information.

From more than 2 million taxonomy file lines only about 200,000 virus file lines remain. In contrast, the FASTA file contains only 12,180 viral genome sequences whereof 2,332 sequences provide taxonomy data. Hence, maximum 2,332 can be used for classification.

Taxonomy category

From a list of 14 taxonomy ranks, we need to decide which one to choose for classifying viral reads. As a necessary criterion, any classification algorithm requires at least two different taxonomy levels. To achieve a high accuracy, artificial neural networks need a big amount of training data in relation to the number of different levels one wants to predict.

As a consequence, we look for a taxonomy category that has more than one, but not too many different levels and that provides a sufficient amount of available viruses per level. This is why we choose the taxonomy category **order** from Table 2.

Table 3 depicts the total taxa frequencies. Take notice of the imbalanced count distribution. For example, in the category "order" there are only seven Bunyavirales, but 1,859 Caudovirales taxa.

Sampling of viruses and reads

We continue with the statistical sampling of viruses and reads from which the input features will be computed subsequently. In step one, from each order $j \in \{1, \dots, C\}$, we randomly sample the corresponding viruses without replacement and sort them into training, validation and test viruses. The disjunct partition is 70, 15 and 15%. For instance, from seven Bunyavirales viruses in total, five are used for training and one for validation or test data, see Table 4.

In step two, we randomly sample with replacement the same amount of viruses out of each of the $3 \cdot 9 = 27$ cell entries. From each order $j \in \{1, \dots, C\}$ and for training, validation and test data separately ($i \in \{1, 2, 3\}$), we sample $n_{i,j}$ of $N_{i,j}$ viruses which yields a total amount of n viruses:

$$\forall j \in \{1, \dots, C\} : n_{1,j} + n_{2,j} + n_{3,j} = 0.7n + 0.15n + 0.15n = n_{.,j} = n$$

The second sampling stage is illustrated in Figure 3. Let us assume, for test data we want to sample two viruses within each order. The randomly chosen viruses are surrounded by squares. However, there is the possibility that some orders only provide an insufficient amount of viruses so that viruses of these orders need to be sampled multiple times. For example, the single Bunyavirales virus must be sampled twice.

The whole sampling plan from step one and two is also known as two-stage stratified sampling with disproportionate allocation [21].

After sampling the viruses, we sample reads of the FASTA file by choosing random start positions out of a discrete uniform distribution and a read length of 150.

Input features

Possible input features to select are k -mers [22], inter-nucleotide distances [23] and DNA sequence motifs [24]. k -mers are subsequences of length k of the original read sequences such as A, TT or GTA. Inter-nucleotide distances are the distances between two identical bases. For instance, the DNA sequence TTATACTACGTGGGGGGGGTCCT exhibits T-T distances of 1, 2, 3, 4, 10 and 3 and we are interested in the count frequencies of the distances 2 to 10. A DNA sequence motif is a set of words which can have arbitrary or a subset of A, C, G and T letters on particular base positions while having fixed letters on other positions. Here, we focus on DNA sequence motifs with two or three fixed letters, therefrom two at the margins, such as C...T or G...G...A. Points represent place holders.

The number of input features sums up to

$$4^1 + 4^2 + 4^3 + 4 \cdot 9 + x + y = 4 + 16 + 64 + 36 + x + y = 120 + x + y,$$

whereof 4^k are the numbers of k -mers, 36 inter-nucleotide distances and $x = 19$ or $y = 24$ the numbers of selected DNA sequence motifs with two or three fixed letters, respectively. Since we do not select any DNA sequence motifs, which will be justified in the next subsection, we use 120 input features in total. We do not select more input features like 4-mers or DNA sequence motifs with four fixed letters because

on the one hand this will result in too many input features and on the other hand the probability of frequency zero is too high.

The decision of the selection of input features will be made on the results of Kruskal-Wallis tests and accuracy comparisons of neural network simulation runs.

Feature selection

The first step is, for each DNA sequence motif separately, to count the relative frequency in each complete nucleobase sequence of the FASTA file. In the second step, we compare the frequencies between and within the groups or orders and perform a Kruskal-Wallis test [25]. High differences of the frequencies between the orders in relation to low differences within the orders point out that this particular DNA sequence motif could be a good input feature to differentiate between the orders. On the third step, we select the DNA sequence motifs with the lowest p-value respectively highest effect size ϵ^2 [26], up to a sequence length of eight.

From all possible DNA sequence motifs with two or three fixed letters, we choose those with an effect size not less than 0.08 or 0.2, respectively. For small data, we iterate 10 simulation runs per layer in the artificial neural network and report mean and standard deviation of accuracy results (see Table 5). Each layer uses a subset of all proposed input features. If the paired t-test recognises a statistically significant and relevant accuracy increase of the subsequent layer, its corresponding input features are added to the artificial neural net model. Finally, the suggested feature selection method chooses 1-mers, 2-mers, 3-mers and inter-nucleotide distances as input features.

Artificial neural net

For computation of the artificial neural network (ANN) [27], we apply the R-package `keras` [28]. The sequential model contains 100,000 training reads per order. The input layer contains 120 input nodes, the hidden layer 64 nodes and the output layer nine nodes for the nine different orders. The respective activation functions used are Rectified Linear Unit [29] for the hidden layer and Softmax [30] for multinomial classification of the output layer.

As optimisation algorithm we choose stochastic gradient descent. To quantify the model performance, sparse categorical crossentropy and accuracy are suitable metrics.

While training the model, the data are shuffled before each epoch to allow faster convergence and to prevent memorizing of the order of the training samples. [31]. The batch size is 2^{10} and maximum epoch count is 500 with an early stopping method to prevent overfitting [32].

The whole simulation procedure is iterated N_{sim} times on independently generated sets of data samples each, see Figure 5.

Generalised confusion matrix

A generalised confusion matrix [33] provides an appropriate summary of classification result metrics. Cell entry $(j, k + 2)$ of Table 6 counts the reads of order $j - 1$ which were classified to order $k - 1$. The test data provide 21,429 reads per order as row-sums, in total $9 \cdot 21429 = 192861$. The diagonal from top left to bottom right provides the correct classification frequencies.

As additional metrics, true positive and negative rates

$$TPR = \frac{TP}{TP + FN} \quad \text{and} \quad TNR = \frac{TN}{TN + FP}$$

as well as positive and negative predictive values

$$PPV = \frac{TP}{TP + FP} \quad \text{and} \quad NPV = \frac{TN}{TN + FN}$$

are reported in the special confusion matrix in Table 7. Note that in both Tables 6 and 7 all numbers are mean values on 10 simulation runs of which each utilises a different partition of training, validation and test data.

Estimation of taxa frequencies

The artificial neural network classifies single reads into taxonomy ranks. However, we do not only want to predict to which order a particular read belongs, but also we want to estimate the global taxa count distribution of the nine orders. In this way we gain a summary of species abundance on taxonomy ranks. Prior and posterior taxa count distributions are calculated as follows.

There are $N_{sim} = 10$ simulation runs and $N_{cat} = 9$ categories, in this case orders.

$$C_s \in \mathbb{N}_0^{N_{cat} \times N_{cat}}, s = 1, \dots, N_{sim},$$

is a finite sequence of generalised confusion matrices, without first two columns, which contain absolute read count frequencies of the simulation runs.

$$P_s \in [0, 1]^{N_{cat} \times N_{cat}}, s = 1, \dots, N_{sim},$$

is a finite sequence of generalised positive predictive values, that is the probability that a read classified as taxon k^* actually belongs to taxon j :

$$p_{s;j^*,k} = \frac{c_{s;j^*,k}}{\sum_{j=1}^{N_{cat}} c_{s;j,k}}, s \in \{1, \dots, N_{sim}\}, j^*, k \in \{1, \dots, N_{cat}\} \quad (1)$$

For each simulation run with $s \in \{1, \dots, N_{sim}\}$, we compute a leave-one-out mean matrix

$$\hat{P}_s, s = 1, \dots, N_{sim}.$$

The mean values are calculated over all P_s matrices except the one corresponding to the particular simulation run:

$$\bar{p}_{s^*;j,k} = \sum_{s \in \{1, \dots, N_{sim}\} \setminus \{s^*\}} \frac{p_{s^*;j,k}}{N_{sim} - 1}, s^* \in \{1, \dots, N_{sim}\}, j, k \in \{1, \dots, N_{cat}\} \quad (2)$$

Note that all training, validation and test samples of all simulation runs are generated artificially from the genome and taxonomy database and thus their real taxonomy ranks are known to the analyst, with the exception of test data of simulation

run s . Correct classifications of test samples of this simulation run are unknown to the analyst in real-world scenarios, see Figure 5, and hence to be predicted. This is why those data are excluded from the computation of leave-one-out mean matrix $\hat{P}_s$. The column sums of

$$C_s \in \mathbb{N}_0^{N_{cat} \times N_{cat}}, s = 1, \dots, N_{sim},$$

compose the prior estimation

$$\hat{A} \in \mathbb{N}_0^{N_{cat} \times N_{sim}}$$

of the read count distribution:

$$\hat{a}_{s;k} = \sum_{j=1}^{N_{cat}} c_{s;j,k}, s \in \{1, \dots, N_{sim}\}, k \in \{1, \dots, N_{cat}\} \quad (3)$$

By applying the law of the total probability and multiplying a leave-one-out mean $\hat{P}_s$ matrix by A , we receive the posterior estimation of read count distribution:

$$\hat{B} \in \mathbb{R}_{\geq 0}^{N_{cat} \times N_{sim}}; \hat{B} = \bar{P} \cdot \hat{A} \quad (4)$$

By using information about misclassification rates and generalised positive predictive values, the estimation of taxa frequencies should be improved. These, however, includes the necessity to perform multiple simulation runs on independently generated artificial data to estimate misclassification rates.

Apart from estimating taxa frequencies based on the classification results of artificial neural net simulation runs, we implement a supplementary mapping approach where we generate artificial 150 base pair read sequences with constant quality values and map these versus our virus reference genomes represented in the training data using Bowtie2 [34]. As we want to classify new, yet unknown, viruses from each order we randomly sample 15% of the viruses and use these to sample 21,429 artificial read sequences which are stored in a FASTQ file. The remaining viruses from each order are stored in a FASTA file so that virus sets of both files are disjunct. This procedure is conform to sampling viruses and reads of the ANN approach which ensures comparability. Finally, from the mapping results the taxonomy ranks are determined and taxa frequencies are computed.

Whether the posterior estimation is actually superior to prior estimation and mapping estimation will be examined in the results section.

Results

Accuracy performance history

As a first result, the performance history of four different simulation runs is exemplarily depicted in Figure 4. Primarily, the smooth curves stand out which is caused by a large batch size. The training time differs greatly, from 500 epochs on plot A to only 50 on plot D. Training loss decreases negatively exponentially and training accuracy increases logarithmically. Validation accuracy is considerably higher than

11% which means that our neural net classifications are clearly more precise than simple random classifications. Plot A shows an exceptional validation performance behaviour. Validation loss decreases slightly until epoch 150, then increases to a value higher than on start, while validation accuracy is still increasing. This indicates that the uncertainty for predicting class probabilities increases, whereas the classification results are becoming more precise. Plots B to D partly show better loss and accuracy performances on validation data in comparison to training data.

Generalised confusion matrix

While according to Table 7 the relative majority of reads from most orders is correctly classified, the misclassification rate is rather high regarding other orders. Due to the close relationship between Bunyavirales order and Paramyxoviridae family, which belong to Mononegavirales order, most of Bunyavirales reads are falsely classified into Mononegavirales order, but not the other way around. In detail, 7,606 Bunyavirales reads are classified to Mononegavirales and only 2,728 Mononegavirales reads are classified to Bunyavirales. This is justified by basic principles of set theory. Since Bunyavirales and Mononegavirales, to which the Paramyxoviridae family belongs, share the same ancestor, a relative majority of Bunyavirales reads is classified to the subset of Mononegavirales viruses that belong to the Paramyxoviridae family and thus is classified correctly. Contrariwise, only Mononegavirales reads selected from Paramyxoviridae family are majoritarian classified to Bunyavirales order and this does not apply for reads from other Mononegavirales families. Nidovirales, Picornavirales and Tymovirales are biologically related and Herpesvirales descend from Caudovirales which explains high misclassification rates between these orders (see Table 6). For instance, 3,816 or 4,458 Picornavirales reads are mapped to Nidovirales or Tymovirales, respectively.

Estimation of taxa frequencies

The boxplots on Figure 6 show the distributions of mapping, prior and posterior estimation of taxa frequencies of the nine orders. Indeed, both variance and bias and thus mean squared error of posterior estimation are considerably smaller for every order in comparison to prior or mapping estimation.

The mapping estimation performed worst by far. Mapping rate was low with a mean value of 14.5% and a standard deviation of 3.3% which is caused by the disjoint partition of FASTQ and FASTA viruses in the simulation runs. From the remaining mapped reads, not a single read was mapped to a Bunyavirales or Tymovirales virus and thus taxa frequencies of these orders were underestimated. In contrast to that, Caudovirales and Ligamenvirales taxa ranks were overestimated considerably. In median, more than 40% of all reads were mapped to Caudovirales although only $100/9 \approx 11\%$ was expected. Since the difference in the statistical estimation error of the mapping approach in comparison to both other approaches is graphically obvious, no statistical estimation was necessary to prove its inferiority.

Besides the graphical comparison, we apply Mood's median test to compare the median absolute deviation of estimated frequencies from the true value of 21,429 test reads between the groups "Prior" and "Posterior". The p-value of 0.00000002197 is highly significant. A paired alternative to the median test is the

sign test. Indeed both prior and posterior estimation are computed on same data for each simulation run which allows the usage of the sign test, yet we used the more conservative median test to confirm the MAD (mean absolute deviation) error differences of both estimators. Median absolute deviations of both groups to the true value are 8,872.5 or 2,741.1 which results in a difference of 6,131.4. Thus, median posterior estimation of taxa frequencies is more than 6,000 reads more precise than median prior estimation. That implies, using generalised positive predictive values contributes vastly to more precise taxa frequency distribution and even prior estimation is superior to a mapping approach based estimation, at least if the probe is composed of new virus reads.

Discussion

Performance comparison between machine learning and mapping approach

In this work, we classified viral read sequences to nine different viral taxonomic orders by using ANNs. Furthermore, we developed two formulas to statistically estimate species abundance on taxonomy ranks. After building and training, the performance history of the model was evaluated graphically on validation data. Prediction accuracy of the derived methods was evaluated on test data and classification results were summarised. From this confusion matrix, diagnostic measures such as true and false positive rates as well as positive and negative predictive values were calculated. The prediction accuracy of the artificial neural net was considerably higher than for random classification and posterior estimation of taxa frequencies was relatively precise in comparison to prior estimation or to order distributions obtained by the mapping approach, at least for yet unknown, new viruses.

Evaluation of our new approach shows that an ANN model can be helpful to classify mapped viral read sequences into viral taxa. The results of this classification could also be helpful to support or complement the findings of a mapping approach by classifying reads from novel viruses, too. The calculation of misclassification rates and training of the ANN can be done independently of the test data the analyst wants to classify and thus can involve large datasets and require a long runtime which could help to improve accuracy and estimation performance. As a benefit in comparison to mapping, our new approach is not limited to already known viruses. In addition, statistical estimations of taxa frequencies provide an insight into taxa abundance of viral species.

For fitting the ANN we draw training reads from only 70% of viruses from each order. With this, we took into account scenarios with reads of novel viruses in the data of a test sample. If an ANN model was trained on reads from all available viruses in a database instead, we would expect that a mapping based approach would excel the ANN performance. Nevertheless, the ANN model could then be used to validate the mapping results. Independently of which of those two approaches, mapping or machine learning, are used for taxa classification, in both cases taxa abundance estimations could be enhanced by our posterior estimation formulas applying the law of the total probability. While these formulas were evaluated for our ANN model, here, as a future perspective we want to also implement an algorithm to enhance classification accuracy of mapping results based on estimated misclassification rates.

Phylogenetic explanations for misclassifications

In our evaluation of the neural net for classification of viral orders, several misclassifications were observed. A reason for the false classification of reads from Bunyavirales to the order Mononegavirales could be explained by the close phylogenetic relation between these two orders (cf. Table 1 from Wolf et al. [35]).

Furthermore, the phylogenetic relationship between Picornavirales as an ancestor and Nidovirales as a descendant on branch two [35], which both belong to positive-sense RNA viruses, is somewhat close which results in a high one-way misclassification rate from Picornavirales to Nidovirales.

Herpesvirales originate from Caudovirales [36, 37] which explains the high two-way misclassification rates between these orders.

Generally, the creation of phylogenetic classification of RNA viruses is a hard task due to the higher mutation rates of RNA viruses in comparison to DNA viruses, or, as Ryabov and Eugene [38] stated: *"Another reason why it might not be possible to reconstruct the "evolutionary trees" of RNA viruses is the crucial role of horizontal gene transfer, between viral groups and even between viruses and their hosts, in the emergence of novel virus genomes. Individual genome segments, gene blocks and single genes may have independent evolutionary histories, making virus genomic evolution "network-like", rather than "tree-like" [39]."*

Acknowledgements

Not applicable.

Author's contributions

KJ formulated the scientific problem. MKo, KJ, JK and MKi designed the artificial neural net simulations. MKo performed the data analysis and the simulation study and drafted the manuscript. PL contributed to the biological interpretation of the results. All authors read and approved the final manuscript.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 398066876/GRK 2485/1.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

References

- Piewbang, C., Jo, W.K., Puff, C., van der Vries, E., Kesdangakonwut, S., Rungsipipat, A., Kruppa, J., Jung, K., Baumgärtner, W., Techangamsuwan, S., et al.: Novel canine circovirus strains from thailand: Evidence for genetic recombination. *Scientific reports* **8**(1), 1–9 (2018)
- Piewbang, C., Jo, W.K., Puff, C., Ludlow, M., van der Vries, E., Banlunara, W., Rungsipipat, A., Kruppa, J., Jung, K., Techangamsuwan, S., et al.: Canine bocavirus type 2 infection associated with intestinal lesions. *Veterinary pathology* **55**(3), 434–441 (2018)
- Bratman, S.V., Bruce, J.P., O'Sullivan, B., Pugh, T.J., Xu, W., Yip, K.W., Liu, F.-F.: Human papillomavirus genotype association with survival in head and neck squamous cell carcinoma. *JAMA oncology* **2**(6), 823–826 (2016)
- Jünemann, S., Kleinbölting, N., Jaenicke, S., Henke, C., Hassa, J., Nelkner, J., Stolze, Y., Albaum, S.P., Schlüter, A., Goesmann, A., et al.: Bioinformatics for ngs-based metagenomics and the application to biogas research. *Journal of biotechnology* **261**, 10–23 (2017)
- Peschel, S., Mülller, C.L., von Mutius, E., Boulesteix, A.-L., Depner, M.: Netcomi: Network construction and comparison for microbiome data in R. *bioRxiv* (2020)
- Wang, Q., Jia, P., Zhao, Z.: Virusfinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PloS one* **8**(5), 64465 (2013)

7. Scheuch, M., Höper, D., Beer, M.: Riems: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. *BMC bioinformatics* **16**(1), 69 (2015) 380
8. Alawi, M., Burkhardt, L., Indenbirken, D., Reumann, K., Christopeit, M., Kröger, N., Lütgehetmann, M., Aepfelbacher, M., Fischer, N., Grundhoff, A.: Damian: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples. *Scientific reports* **9**(1), 1–17 (2019)
9. Saremi, B., Kohls, M., Liebig, P., Siebert, U., Jung, K.: Measuring reproducibility of virus meta-genomics analyses using bootstrap samples from fastq-files. *Bioinformatics* (2020) 385
10. Kruppa, J., Jo, W.K., van der Vries, E., Ludlow, M., Osterhaus, A., Baumgaertner, W., Jung, K.: Virus detection in high-throughput sequencing data without a reference genome of the host. *Infection, Genetics and Evolution* **66**, 180–187 (2018)
11. Peabody, M.A., Van Rossum, T., Lo, R., Brinkman, F.S.: Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC bioinformatics* **16**(1), 362 (2015) 390
12. Rasheed, Z., Rangwala, H.: Metagenomic taxonomic classification using extreme learning machines. *Journal of bioinformatics and computational biology* **10**(05), 1250015 (2012)
13. Zhang, Z., Cai, Z., Tan, Z., Lu, C., Jiang, T., Zhang, G., Peng, Y.: Rapid identification of human-infecting viruses. *Transboundary and emerging diseases* **66**(6), 2517–2522 (2019) 395
14. Bartoszewicz, J.M., Seidel, A., Renard, B.Y.: Interpretable detection of novel human viruses from genome sequencing data. *BioRxiv* (2020)
15. Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., McVeigh, R., O'Neill, K., Robbertse, B., et al.: Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020** (2020) 400
16. Raczko, E., Zagajewski, B.: Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral apex images. *European Journal of Remote Sensing* **50**(1), 144–154 (2017)
17. Ren, J.: Ann vs. svm: Which one performs better in classification of mcs in mammogram imaging. *Knowledge-Based Systems* **26**, 144–153 (2012) 405
18. Sayers, E.W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K.D., Karsch-Mizrachi, I.: Genbank. *Nucleic acids research* **47**(D1), 94–99 (2019)
19. Walker, P.J., Siddell, S.G., Lefkowitz, E.J., Mushegian, A.R., Dempsey, D.M., Dutilh, B.E., Harrach, B., Harrison, R.L., Hendrickson, R.C., Junglen, S., et al.: Changes to virus taxonomy and the international code of virus classification and nomenclature ratified by the international committee on taxonomy of viruses (2019). *Archives of virology* **164**(9), 2417–2429 (2019) 410
20. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al.: Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**(D1), 733–745 (2016)
21. Shahrokhi Esfahani, M., Dougherty, E.R.: Effect of separate sampling on classification accuracy. *Bioinformatics* **30**(2), 242–250 (2014) 415
22. Perry, S.C., Beiko, R.G.: Distinguishing microbial genome fragments based on their composition: evolutionary and comparative genomic perspectives. *Genome biology and evolution* **2**, 117–131 (2010)
23. Afreixo, V., Bastos, C.A., Pinho, A.J., Garcia, S.P., Ferreira, P.J.: Genome analysis with inter-nucleotide distances. *Bioinformatics* **25**(23), 3064–3070 (2009) 420
24. D'haeseleer, P.: What are dna sequence motifs? *Nature biotechnology* **24**(4), 423–425 (2006)
25. Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* **47**(260), 583–621 (1952)
26. Mangiafico, S.: Summary and analysis of extension program evaluation in r, version 1.15. 0. Rutgers Cooperative Extension, New Brunswick, NJ: <https://rcompanion.org/handbook/>. [Google Scholar] (2016) 425
27. Haykin, S.: *Neural Networks: a Comprehensive Foundation*. Prentice-Hall, Inc., ??? (2007)
28. Chollet, F.: Allaire, J. keras: R Interface to 'Keras'. R package version 2.2. 4.1 (2019)
29. Hara, K., Saito, D., Shouno, H.: Analysis of function of rectified linear unit used in deep learning. In: 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2015). IEEE
30. Goodfellow, I., Bengio, Y., Courville, A.: 6.2. 2.3 softmax units for multinoulli output distributions. In: *Deep Learning*, pp. 180–184. MIT Press, ??? (2016) 430
31. Bengio, Y.: Practical recommendations for gradient-based training of deep architectures. In: *Neural Networks: Tricks of the Trade*, pp. 437–478. Springer, ??? (2012)
32. Prechelt, L.: Early stopping-but when? In: *Neural Networks: Tricks of the Trade*, pp. 55–69. Springer, ??? (1998) 435
33. Stehman, S.V.: Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment* **62**(1), 77–89 (1997)
34. Langdon, W.B.: Performance of genetic programming optimised bowtie2 on genome comparison and analytic testing (gcat) benchmarks. *BioData mining* **8**(1), 1 (2015)
35. Wolf, Y.I., Kazlauskas, D., Iranzo, J., Lucia-Sanz, A., Kuhn, J.H., Krupovic, M., Dolja, V.V., Koonin, E.V.: Origins and evolution of the global rna virome. *MBio* **9**(6) (2018) 440
36. Koonin, E.V., Dolja, V.V., Krupovic, M.: Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology* **479**, 2–25 (2015)
37. Koonin, E.V., Yutin, N.: Evolution of the large nucleocytoplasmic dna viruses of eukaryotes and convergent origins of viral gigantism. In: *Advances in Virus Research* vol. 103, pp. 167–202. Elsevier, ??? (2019) 445
38. Ryabov, E.V.: Invertebrate rna virus diversity from a taxonomic point of view. *Journal of invertebrate pathology* **147**, 37–50 (2017)
39. Koonin, E.V., Dolja, V.V.: Expanding networks of rna virus evolution. *BMC biology* **10**(1), 54 (2012)

Figures

```
2007412 1301311 | zooplankton environmental sample | cellular organisms; Eukaryota; environmental samples; |
2007413 148771 | unidentified amoeba from potato | cellular organisms; Eukaryota; environmental samples; |
2007414 10239 | Viruses |
2007415 10508 | Adenoviridae | Viruses; |
2007416 100953 | Adenovirus | Viruses; Adenoviridae; |
2007417 130499 | Bovine atadenovirus D | Viruses; Adenoviridae; Atadenovirus; |
2007418 70333 | Bovine atadenovirus 4 | Viruses; Adenoviridae; Atadenovirus; Bovine atadenovirus D; |
2007419 111166 | Bovine atadenovirus 5 | Viruses; Adenoviridae; Atadenovirus; Bovine atadenovirus D; |
2007420 120509 | Bovine atadenovirus 8 | Viruses; Adenoviridae; Atadenovirus; Bovine atadenovirus D; |
2007421 2169706 | Deer atadenovirus A | Viruses; Adenoviridae; Atadenovirus; |
```

Figure 1 Taxonomy database. Taxonomy database for taxa classification, downloaded from National Center for Biotechnology Information [15, 18] on March 2020 from the ftp-server <ftp://ftp.ncbi.nih.gov/refseq/release/viral>. A section of the taxonomy database file `fullnamelineage.dmp` containing taxonomic information about viral species is depicted.

```
>NC_001798.2 Human herpesvirus 2 strain HG52, complete genome
AGTCCCCGTCTTCCGCGCGGGGCGGGCGGGGAAAAAGCCGCGGGGGCGCCGCGGGAAGGCAGCCCCGCGGCGC
GCGGGGGAGGGGCGGCGCCGCGGGGGAGCGGCGCGCTCCGGGGAGGGACGGGGAAGGGGGCGCGGGGGCTGCCCTG
CCGCCGCCCGCCGCGCGCCGCGCTTCGCGCCCCCCCCCAAAAAACACCCCCCGGGGTTGACTCCCCGGGGGAAA
AGAGGCGGGGCGGGAGTCCCGTCTCGCCGCGCCCCCTTAAGAGGGCCCGCAACACGGCCCGGGCTGCGCACGCCAGCCG
GGACGGGTGAGTTGCTAGGCAAGCACGGAAGTGGCGGTTACACGTGCATGCGTGCCGAGTGAAGTCTCCCGCCCCGACGC
GCTCCGGCTCCGGGCTACGCCGAGCCAGCCGCGCCGCCATGTCCCGCCGCGGGGTCCCCGCCCGGGGTCCCCGGCG
CCGGCCGCGCCCGGGCTCCAGCGGTGCGCGCCCGGGCGCTCCAGCCGTGCGCGCCCGGGCGGCTCCCAACCGCAG
ACTCCCAATGGTCCCTGCGTACGACTCGGGAACCGGGTCGAGAGCGCGCGCGCGCTCTCGCTCTGCGGCGCTGG
...
```

Figure 2 FASTA file. Viral reference genome FASTA file from NCBI [20]. The first line is composed of a unique identifier and the corresponding species name. Consecutively, the genome sequence consisting of the nucleobases A, C, G and T is listed, beginning on line 2.

Figure 3 Virus sampling step 2. We sample a fixed number of viruses from each order, with varying sample sizes depending on the selection of training, validation or test data. Let the number of viruses to sample test data be two within each order. For five orders, two different randomly sampled viruses can be chosen and surrounded by squares. Nonetheless, the remaining four orders only provide one and thus too little viruses so that in this case the single available virus must be sampled twice.

Figure 4 Accuracy performance history. The performance history of four chosen simulation runs is depicted. The smoothness of the curves could be reached by a large batch size. The training time varies vastly, exceeding 500 epochs on plot A and falling below 50 on plot D. Whilst training loss decreases subject to exponential decay, training accuracy increase is logarithmic. Validation accuracy exceeds simple random classification probability of 11% considerably. Plot A holds a remarkable behaviour of validation performance. From epoch 150 on, validation loss and accuracy both increase simultaneously. This portends that predictions of class probabilities incline to become increasingly imprecise, whereas classification results are increasing in accuracy. In contrary to plot A, on plots B to D loss and accuracy performance on validation data is superior to that in training data.

Table 1 Taxa endings according to International Committee on Taxonomy of Viruses [19].

Realm	Subrealm	Kingdom	Subkingdom	Phylum	Subphylum	Class	Subclass
...viria	...vira	...virae	...virites	...viricota	...viricotina	...viricetes	...viricetidae
Order	Suborder	Family	Subfamily	Genus	Subgenus	Species	
...virales	...virineae	...viridae	...virinae	...virus	...virus	...virus	

Table 2 Available taxonomy data. There are 14 possible taxonomy ranks to choose for classification of viral reads. We aim to select a taxonomy factor that provides a moderate number of different levels and a fair amount of available viruses per level. Thus, the taxonomy rank order appears to be a natural decision.

Taxonomy factor	Levels	Available viruses	Available viruses / levels	Not available viruses
Domain (= Viruses)	1	3747	3747	0
Realm	1	883	883	2864
Subrealm	0	0		3747
Kingdom	0	0		3747
Subkingdom	0	0		3747
Phylum	1	77	77	3670
Subphylum	2	77	38.5	3670
Class	2	77	38.5	3670
Subclass	0	0		3747
Order	9	2332	259.1	1415
Suborder	5	47	9.4	3700
Family	100	3482	34.8	265
Subfamily	57	921	16.2	2826
Genus	559	2511	4.49	1236

Figure 5 Data samples for simulation runs. We perform 10 simulations runs on 10 different partitions of training, validation and test samples. For simulation run s , for instance, the true taxonomy ranks of test data samples s are unknown to the ANN, whilst taxonomy ranks of all training, validation and test samples of the remaining simulation runs are considered as known and thus used for posterior statistical estimation of taxa abundances.

Figure 6 Statistical estimation of taxa frequencies. The comparison of mapping, prior and posterior estimations of taxa frequencies on 10 simulation runs show that the posterior estimations are much closer to the true value of 21,429 (or 11%) in terms of mean absolute deviation (MAD) and thus more precise in relation to the prior estimations. MAD values are 8872.5 or 2741.1 for prior or posterior estimation, respectively. The difference of 6131.4 is statistically significant, too.

Table 3 Levels of taxonomy factor order. There is an imbalance of the count distribution.

Bunyavirales	Caudovirales	Herpesvirales	Ligamenvirales	Mononegavirales
7	1859	22	21	70
Nidovirales	Ortervirales	Picornavirales	Tymovirales	
47	81	107	118	

Table 4 Sample sizes for taxonomy category order. Step one of the sampling procedure is to randomly sample from each order viruses without replacement and allocate them to training, validation and test viruses. 70, 15 or 15% of sampled viruses are used for training, validation or test samples, respectively.

	Training	Validation	Test	Σ
Bunyavirales	5	1	1	7
Caudovirales	1301	279	279	1859
Herpesvirales	16	3	3	22
Ligamenvirales	15	3	3	21
Mononegavirales	50	10	10	70
Nidovirales	33	7	7	47
Ortervirales	57	12	12	81
Picornavirales	75	16	16	107
Tymovirales	82	18	18	118
Σ	1634	349	349	2332

Table 5 Feature selection. We choose DNA sequence motifs with an effect size not less than 0.08 or 0.2, respectively. This yields 19 or 24 DNA sequence motifs. Using short-runtime artificial neural networks on small data, we iterate the simulation runs 10 times per depicted layer and report the accuracy results. In this step-wise procedure, we select only those sets of features that contribute a significant and relevant accuracy increase. The paired t-test is statistically not significant on layer four and accuracy increase is not relevant on layer four, which is why DNA sequence motifs are excluded as features and thus not selected. Only 1-mers, 2-mers, 3-mers and inter-nucleotide distances remain for the actual simulation procedure on big data.

Feature selection	Input features	Mean	Standard deviation
1-mers	4	0.303	0.032
1-mers, 2-mers	20	0.380	0.041
1-mers, 2-mers, inter-nucl. dist.	56	0.425	0.049
1-mers, 2-mers, inter-nucl. dist., seq. motifs 2	75	0.424	0.040
1-mers, 2-mers, inter-nucl. dist., seq. motifs 3	80	0.436	0.050
1-mers, 2-mers, inter-nucl. dist., 3-mers	120	0.501	0.025

Table 6 Generalised confusion matrix. All cell entries except those in first two columns count the absolute frequencies of how many of the 21429 reads per order were classified to the orders listed in the columns. In total, there are $9 \cdot 21429 = 192861$ test samples for each of the 10 simulation runs and these test samples as well as the classified read frequencies differ between the simulation runs, hence mean absolute frequencies are noted. All reads on the diagonal were correctly classified. The misclassification rate is high for Bunyavirales and Picornavirales and low for the remaining viruses.

Order	Order ID	Ord. 0	Ord. 1	Ord. 2	Ord. 3	Ord. 4	Ord. 5	Ord. 6	Ord. 7	Ord. 8
Bunyavirales	0	5321	190	184	2665	7606	2471	1293	596	1104
Caudovirales	1	778	8782	2673	2108	1270	1960	1350	512	1996
Herpesvirales	2	480	4950	6193	1808	1706	1682	1407	752	2453
Ligamenvirales	3	1271	1635	488	12071	1451	2099	1233	491	691
Mononegavirales	4	2728	497	258	839	9669	1615	2668	949	2206
Nidovirales	5	1068	855	414	740	1636	12134	681	1164	2737
Ortervirales	6	1510	999	1015	1391	3859	1013	8198	673	2771
Picornavirales	7	1622	1157	710	1046	3067	3816	2047	3504	4458
Tymovirales	8	1295	1231	405	315	2770	1730	1731	1176	10775

Table 7 Special confusion matrix. True positive and negative rates (TPR and TNR) as well as positive and negative predictive values (PPV and NPV) are reported as mean values from 10 simulation runs. TNR and NPV are consistently high and PPV is moderate, yet TPR is remarkably low for Bunyavirales and Picornavirales.

Order	Order ID	TPR (%)	TNR (%)	PPV (%)	NPV (%)
Bunyavirales	0	24.8	93.7	33.1	90.9
Caudovirales	1	41.0	93.3	48.7	92.8
Herpesvirales	2	28.9	96.4	51.0	91.6
Ligamenvirales	3	56.3	93.6	53.6	94.5
Mononegavirales	4	45.1	86.3	31.2	92.8
Nidovirales	5	56.6	90.4	46.1	94.4
Ortervirales	6	38.3	92.8	41.6	92.3
Picornavirales	7	16.4	96.3	36.0	90.2
Tymovirales	8	50.3	89.3	40.1	93.6

Figures

2007412	1301311	zooplankton environmental sample	cellular organisms; Eukaryota; environmental samples;
2007413	148771	unidentified amoeba from potato	cellular organisms; Eukaryota; environmental samples;
2007414	10239	Viruses	
2007415	10508	Adenoviridae	Viruses;
2007416	100953	Atadenovirus	Viruses; Adenoviridae;
2007417	130499	Bovine atadenovirus D	Viruses; Adenoviridae; Atadenovirus;
2007418	70333	Bovine adenovirus 4	Viruses; Adenoviridae; Atadenovirus; Bovine atadenovirus D;
2007419	111166	Bovine adenovirus 5	Viruses; Adenoviridae; Atadenovirus; Bovine atadenovirus D;
2007420	120509	Bovine adenovirus 8	Viruses; Adenoviridae; Atadenovirus; Bovine atadenovirus D;
2007421	2169706	Deer atadenovirus A	Viruses; Adenoviridae; Atadenovirus;

Figure 1

Taxonomy database. Taxonomy database for taxa classification, downloaded from National Center for Biotechnology Information [15, 18] on March 2020 from the ftp-server <ftp://ftp.ncbi.nih.gov/refseq/release/viral>. A section of the taxonomy database file fullnamelineage.dmp containing taxonomic information about viral species is depicted.

```
>NC_001798.2 Human herpesvirus 2 strain HG52, complete genome
AGTCCCCGTCTTGCCGCGCGGGGGCGGGCGCGGGAAGCCGCGCGGGGGCGCCCGCGGGAAGGCAGCCCCGCGGCGC
GCGGGGGGAGGGGCGGCGCCCGCGGGGGAGCGGCCGGCTCCGGGGGAGGGACGGGGAAGGGGGCGCGGGGGCTGCCCTG
CCGCCCGCCCGCGCCGCGCGCCCGCCTTCGCGCCCCCCCCCAAAAAACACCCCCCGGGGGTTGACTCCCCGGGGGAAA
AGAGGCGGGGCGGGAGTCCCCGTCTTGCCGCGCGCCCTTAAGAGGGCCCGCAACACGGCCCGGGCTGCGCACGCCAGCCG
GGACGGGTGAGTTCGCTAGGCAAGCACGGAAGTGGCGGTTACACGTGCATGCGTGCCGAGTGAAGTCTCCCGCCCCGACGC
GCTCCGGCTCCGGGCTACGCCGAGCCAGCCGCCCGCCATGTCCCGCCGCGGGGTCCCCGCCGCGGGGTCCCCGGCG
CCGGCCGCGCCCCGGCGCTCCAGCCGTGCCGCGCCCCGGCGCTCCAGCCGTGCCGCGCCCCGGCGCGCTCCCAACCGCAG
ACTCCCAAATGGTCCCTGCGTACGACTCGGGAACCGCGGTGAGAGCGCGCGGGCCGCGTCTCGCTCCTGCGGGCGCTGG
...
```

Figure 2

FASTA file. Viral reference genome FASTA file from NCBI [20]. The first line is composed of a unique identifier and the corresponding species name. Consecutively, the genome sequence consisting of the nucleobases A, C, G and T is listed, beginning on line 2.

Figure 3

Virus sampling step 2. We sample a fixed number of viruses from each order, with varying sample sizes depending on the selection of training, validation or test data. Let the number of viruses to sample test data be two within each order. For five orders, two different randomly sampled viruses can be chosen and surrounded by squares. Nonetheless, the remaining four orders only provide one and thus too little viruses so that in this case the single available virus must be sampled twice.

Figure 4

Accuracy performance history. The performance history of four chosen simulation runs is depicted. The smoothness of the curves could be reached by a large batch size. The training time varies vastly, exceeding 500 epochs on plot A and falling below 50 on plot D. Whilst training loss decreases subject to exponential decay, training accuracy increase is logarithmic. Validation accuracy exceeds simple random classification probability of 11% considerably. Plot A holds a remarkable behaviour of validation performance. From epoch 150 on, validation loss and accuracy both increase simultaneously. This portends that predictions of class probabilities incline to become increasingly imprecise, whereas classification results are increasing in accuracy. In contrary to plot A, on plots B to D loss and accuracy performance on validation data is superior to that in training data.

Figure 5

Data samples for simulation runs. We perform 10 simulation runs on 10 different partitions of training, validation and test samples. For simulation run s , for instance, the true taxonomy ranks of test data samples s are unknown to the ANN, whilst taxonomy ranks of all training, validation and test samples of the remaining simulation runs are considered as known and thus used for posterior statistical estimation of taxa abundances.

Figure 6

Statistical estimation of taxa frequencies. The comparison of mapping, prior and posterior estimations of taxa frequencies on 10 simulation runs show that the posterior estimations are much closer to the true value of 21,429 (or 11%) in terms of mean absolute deviation (MAD) and thus more precise in relation to the prior estimations. MAD values are 8872.5 or 2741.1 for prior or posterior estimation, respectively. The difference of 6131.4 is statistically significant, too.