Virus Research 258 (2018) 55–63

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

LVQ-KNN: Composition-based DNA/RNA binning of short nucleotide sequences utilizing a prototype-based k-nearest neighbor approach

Ariane Belka[a], Mareike Fischer[b], Anne Pohlmann[a], Martin Beer[a], Dirk Höper[a,*]

[a] Institute of Diagnostic Virology, Friedrich-Loeffler-Institut, Südufer 10, D-17493 Greifswald, Insel Riems, Germany
[b] Institute for Mathematics & Computer Science, Ernst-Moritz-Arndt University, Walther-Rathenau-Straße 47, D-17489 Greifswald, Germany

ARTICLE INFO

Keywords:
Composition-based analysis
Oligonucleotides
Metagenomics
Learning vector quantization algorithm
k-Nearest neighbor method
Cross validation

ABSTRACT

Unbiased sequencing is an upcoming method to gain information of the microbiome in a sample and for the detection of unrecognized pathogens. There are many software tools for a taxonomic classification of such metagenomics datasets available. Numerous of them have a satisfactory sensitivity and specificity for known organisms, but they fail if the sample contains unknown organisms, which cannot be detected by similarity-based classification employing available databases. However, recognition of unknowns is especially important for the detection of newly emerging pathogens, which are often RNA viruses. Here we present the composition-based analysis tool LVQ-KNN for binning unclassified nucleotide sequence reads into their provenance classes DNA or RNA. With a 5-fold cross-validation, LVQ-KNN reached correct classification rates (CCR) of up to 99.9% for the classification into DNA/RNA. Real datasets gained CCRs of up to 94.5%. Comparing the method to another composition-based analysis tool, similar or better classification results were reached. LVQ-KNN is a new tool for DNA/RNA classification of sequence reads from unbiased sequencing approaches that could be applicable for the detection of yet unknown RNA viruses in metagenomic samples. The source-code, training and test data for LVQ-KNN is available at Github (https://github.com/ab1989/LVQ-KNN).

1. Introduction

Metagenomics is the challenge of analyzing the community of organisms in a sample using unbiased genomic techniques like next-generation sequencing (NGS) bypassing the need of lab cultivation and isolation of individual species (Chen and Pachter, 2005). Because of the new, fast and cost-effective sequencing methods, a huge amount of sequence data is generated that needs to be analyzed regarding different aspects like functional characterization or taxonomic classification. For this purpose, a bunch of similarity-based and composition-based software tools were developed. Similarity-based tools like BLAST (Camacho et al., 2009), MGRast (Meyer et al., 2008), Clinical Pathoscope (Byrd et al., 2014; Hong et al., 2014), MetaPhlAn (Segata et al., 2012), Kraken (Wood and Salzberg, 2014), CLARK (Ounit and Lonardi, 2016; Ounit et al., 2015), Centrifuge (Kim et al., 2016), and RIEMS (Scheuch et al., 2015) compare nucleotide or amino acid sequences with reference sequences for the classification. BLAST is a search tool from the National Center for Biotechnology Information (NCBI, (NCBI Resource Coordinators, 2016)) which compares a query sequence with either the whole sequence database or a customized one by computing local alignments. Although the analysis of millions of sequences by BLAST is time consuming it is frequently used. To overcome this problem, several tools have been specially designed. For example, Kraken uses exact alignments of k -mers to the lowest common ancestor in a pre-computed k-mer database. In contrast, RIEMS relies on the NCBI databases (nucleotide and protein) and is a composition of several bioinformatics software tools, like the assembler NEWBLER (454/Roche), EMBOSS (Rice et al., 2000), BLAST and the statistical tool R (R Core Team, 2017) to taxonomically classify a metagenomics sequence dataset.

The aforementioned tools have in common that the classification depends solely on the nucleotide sequence identity with reference sequences. The recognition of new organisms is therefore hampered because no similar sequences are available in the databases. To circumvent these problems, new techniques were developed to classify unknown sequences. A number of tools for composition-based taxonomic classification have been published, for instance MetaCluster (Leung et al., 2011), PhyloPythia(S) (McHardy et al., 2007; Patil et al., 2011), SGSOM (Chan et al., 2008), KNNLog (Liu et al., 2012), NBC (Rosen et al., 2011), Phymm(BL) (Brady and Salzberg, 2009), TaxSOM (Weber et al., 2011), RAIphy (Nalbantoglu et al., 2011), TACOA (Diaz et al., 2009), and WSVDD (Hou et al., 2015). These tools compare elementary sequence information, e.g. oligonucleotide frequencies, 16S

* Corresponding author.
E-mail address: dirk.hoeper@fli.de (D. Höper).

https://doi.org/10.1016/j.virusres.2018.10.002
Received 29 June 2018; Received in revised form 25 September 2018; Accepted 2 October 2018
Available online 04 October 2018
0168-1702/

rRNA and relative abundance index (RAI) profiles for the classification. For example, MetaCluster applies a top-down clustering combined with a bottom-up merging of the clusters to bin sequences without a customized reference dataset. With the third version of MetaCluster an accuracy between 72.55% and 99.9% can be reached (Leung et al., 2011). TACOA classifies bacterial sequences by using oligonucleotide frequencies and a kernel based classification algorithm with the leave-one-out-method. With this approach, the program reaches a sensitivity of up to 67% for sequences of 800 bp length (Diaz et al., 2009).

While with 16S rRNA sequencing or algorithms like TACOA, recognition of bacteria and archaea appears feasible even without a suitable reference sequence, the situation is different for viruses. On the one hand, an approach like 16S rRNA analysis is impossible and on the other hand, the diversity within the virus superkingdom still needs to be elucidated. It is assumed that despite the substantial increase in our knowledge of viral species most are still to be discovered (Anthony et al., 2013). The ICTV 2017 Master Species List (MSL32; March 2018) lists 2009 RNA and 2817 DNA virus species. Taking into account the substantial genomic diversity of viruses known to date and the limited knowledge of the virosphere, it can be stated that the similarity based approaches will not be sufficient to recognize the assumingly strongly deviating sequences originating from completely unknown viruses. Therefore, recognition of sequences originating from viruses only distantly related with known viruses is hampered. Based on our current knowledge we assume that a substantial number of the viruses to be discovered will be RNA viruses. Therefore, it could help identify novel RNA viruses if it were possible to recognize a sequence is derived from a functional RNA molecule. The different physicochemical characteristics of RNA and DNA might require different compositions of the molecules to be functional in the cell, like is the case for rRNA or coding and non-coding regions of viral genomes. These different requirements might result in different oligonucleotide compositions, thus enabling the distinction between functional DNA and RNA based on the oligonucleotide frequencies. This is of special interest as metagenomics datasets derived from sequencing of total RNA might contain sequences from both RNA and DNA viruses that are not assignable to any known reference using sequence similarity alone. Classification of sequences to either RNA (viral genomes) or DNA origin (like mRNA) would be a very important step for their further investigation. Hence, here we set out to assess the suitability of oligonucleotide compositions for this purpose and if possible develop an algorithm that is suitable to classify sequences into DNA and RNA molecules based on their oligonucleotide frequencies. To this end, we combined the supervised learning vector quantization method (LVQ) and the k-Nearest Neighbor classifier (k-NN) to assign sequences to their functional classes DNA or RNA.

2. Methods and implementation

2.1. Data

The classification approach was based on a reference dataset X, containing 6896 bacterial and viral reference sequences. The sequences were selected from the NCBI reference database (version date: 2016-01-25), including only complete genome sequences and excluding whole genome shotgun sequences. Furthermore, for viruses it was distinguished between phages and no phages. For every sequence of X oligonucleotide frequencies were calculated using the function *compseq* within the software toolbox EMBOSS (Rice et al., 2000) with default settings. Results were stored in feature vectors

$$x = (x_1, x_2, ..., x_n, y)$$

containing the oligonucleotide frequency rate information

$$x_1, x_2, ..., x_n, \; n = 4^{\omega}, \; \omega \geq 2$$

with $\omega = |\Omega|$ for the oligonucleotide Ω and the class label $y \in C = \{1,2\}$ (class 1 - DNA and class 2 - RNA) for each sequence:

$$x \in (X, C) \subset R_{+}^{n \times m} \times \{1, 2\}, \; n = 4^{\omega}, \; \omega \geq 2, \; m = 6896$$

For example, the feature vector for the ssRNA virus *Dengue virus 1* contains following values for the oligonucleotide frequencies and the class label:

AA	**AC**	**AG**	**AT**	**CA**	**CC**	**CG**	**CT**	**GA**
1.75	1.06	1.40	1.08	1.38	0.76	0.34	0.79	1.50
GC	**GG**	**GT**	**TA**	**TC**	**TG**	**TT**	**class**	
0.72	1.11	0.69	0.65	0.73	1.18	0.79	2	

2.2. Algorithm

The reference dataset X needs to be reduced to a set of m_3 prototypes employing the *learning vector quantization 1 method* (LVQ1, (Kohonen, 2001), Chapter 6). Within this method, initial prototypes are chosen randomly from each class. The outcome of this procedure slightly differs between several runs with the same parameter settings. The learning rate, depending on the learning steps t, is a monotonically decreasing function

$$\alpha: N^{*} \longrightarrow (0, 1)$$

defined as:

$$\alpha(t) = \frac{1}{4t^2}$$

and describes the shift of the winning prototypes in one learning step. If $\alpha(t) = 0$, the winning prototype does not change anymore, and if $\alpha(t) = 1$, the prototype is replaced by the current value of x. Subsequently, the query dataset is classified using the calculated prototype sets and the *k –Nearest Neighbor method* (k -NN) (Fig. 1(a)).

The query dataset will be in both ways classified using k -NN and the prototype sets. The simplest assignment for a query q to a class y is to find a nearest neighbor p_j in the n -dimensional space.

$$y(q) = \min_{j} ||q - p_j||$$

This is called the nearest neighbor approach. To get a more precise classification it can be necessary to find more than one nearest neighbor. Therefore, the k nearest prototypes are determined by calculating the Euclidean distances between the query q and the prototypes. Then q is assigned to class y when the number of corresponding prototypes k_y is greater than or equal to l votes within the set of nearest prototypes.

2.3. Requirements

The LVQ-method was programmed in Perl and the classification with k -NN and evaluation of the results were programmed in R using the packages class (Venables and Ripley, 2002) and ggplot2 (Wickham, 2009). All calculations were either performed on a server with CentOS 6, 24 Intel® Xeon® E7450@2.40 GHz CPUs and 64GB RAM or on a personal computer with Windows 7 Enterprise, Intel® Core™ i5-6500 CPU @ 3.20 GHz and 8GB RAM.

2.4. Oligonucleotide sequences

For short-read sequencing data sets, the maximum length of a query sequence is confined to approximately 500 bp. To determine which oligonucleotides should be deployed in the analysis the following inequality was used:

$$|s| - (\omega - 1) \geq 4^{\omega}.$$

It describes the relation between the length of the query sequence and the length of the oligonucleotide. Based on the length of the current oligonucleotides Ω, $\omega = |\Omega|$, the lower bound of the minimum length of the sequences s, $|s|$, can be computed with:

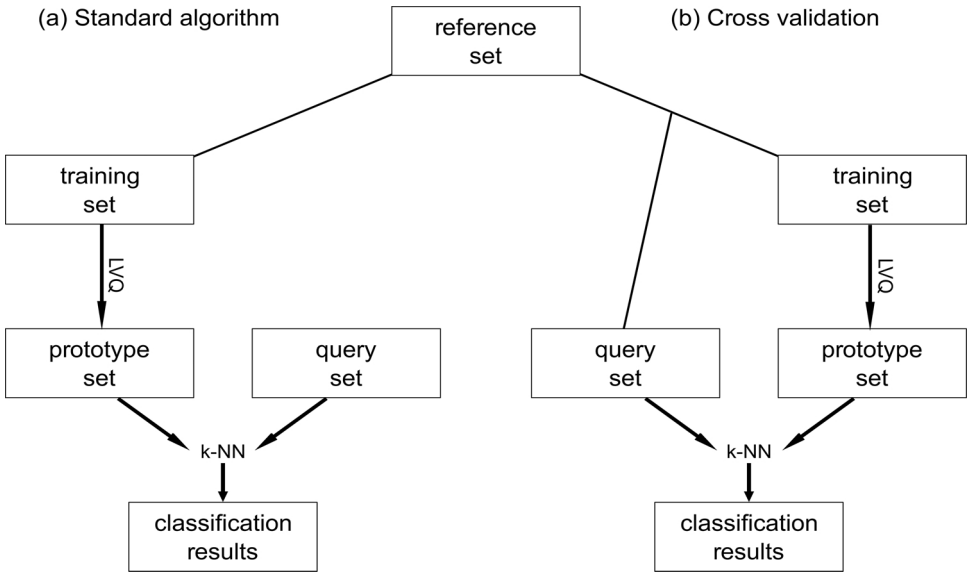

Fig. 1. Schematic display of prototype calculation and query classification. (a) The standard algorithm uses the reference dataset itself for the calculation of the prototypes. (b) The training and query datasets are created from the reference dataset for the 5-fold-crossvalidation.

$$|s| \geq 4^{\omega} + \omega - 1$$

For example, for $\omega = 2$ the minimum length of the sequences of interest is $|s| \geq 4^2 + 2 - 1 \geq 17$. If the length of the sequence s is given, the upper bound of the length of oligonucleotides ω can be calculated as follows:

$$\omega \leq \frac{-W(2^{2|s|+3}\log 2) + 2\,|s|\log 2 + 2\log 2}{2\log 2}$$

with $W(\cdot)$ the LambertW function (proof see Sup. A). For instance, if $|s| = 300$, the maximum length $\omega \leq 4.1069$ and if $|s| = 500$, the maximum length is $\omega \leq 4.4779$. Therefore, di-, tri-, and tetranucleotides were examined. As shown above, sequences derived from second generation sequencing platforms do not have enough information content to use longer oligonucleotides for the classification.

2.5. LVQ1 - prototypes

The prototype sets were optimized by changing the final number of prototypes $m_3 = \{50,\ 100,\ 500,\ 1000\}$ for every class. Moreover, the count of learning steps (ls) $t = \{5,10,15,20,30,50\}$ was optimized to get satisfying classification results.

2.6. k-Nearest neighbor

The employed function *knn* from the R package *class* is restricted to the Euclidean distance, which is used and functional in this study. For a definite decision k nearest neighbors with a minimum vote l were used for classification. This means equal or less than $k-l$ dissenting votes are allowed, otherwise the query sequence is marked as unclassified. To achieve optimal results $k = \{1,\ ...,20\}$ and $\mathrm{l} = \{1,\ ...,20\}$ were analyzed.

2.7. Optimization of parameters

Summarizing the above sections, the following parameters were tested for optimization: the number of learning steps t, the number of prototypes per class m, the number of nearest neighbors k and the vote l. The decisive factors of the classification are the *correct classification rate* (CCR), the *positive prediction value* (PPV, prediction value of class 1, DNA), the *negative prediction value* (NPV, prediction value of class 2, RNA) and the *percentage of unclassified queries*. The optimum is given by:

$$opt = (CCP,\ PPV,\ NPV,\ uncl) = (1{,}1,\ 1{,}0)$$

It means in effect the optimal parameter setting generates correct classification results for every query without producing any unclassified ones. The quasi-optimal setting of the parameters was chosen by restricting $NPV \geq 0.7$ and $uncl \leq 0.6$ and computing the minimum weighted distance to the optimum opt.

2.8. Validation

To validate LVQ-KNN the reference dataset X is divided into 5 pairwise disjoint sets

$$X = \bigcup_{i=1}^{5} Q_i$$

for the 5-fold-cross-validation (Fig. 1(b), Fig. 2). They are defined as follows:

$$T = (X_1, C) \subset R_{+}^{n \times m_1} \times \{1{,}2\}$$

$$Q = (X_2, C) \subset R_{+}^{n \times m_2} \times \{1{,}2\},\ X1 \cup X2 = X$$

with $m_1 + m_2 = m$.

The results of the cross validation were analyzed by calculating the average over the 5 disjoint query datasets.

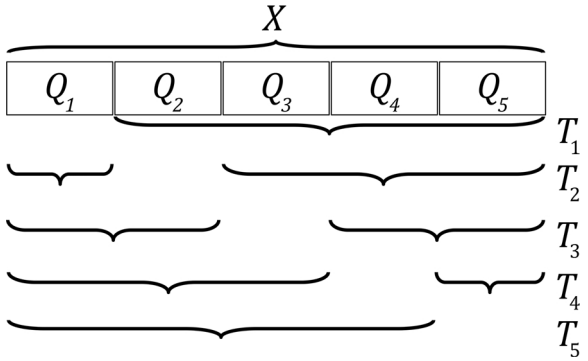

Fig. 2. Description of 5-fold-cross-valiation. 5 pairwise disjoint sets Q_i from X are chosen randomly. Every set Q_i will be one time the query dataset Q and one time part of the training dataset T.

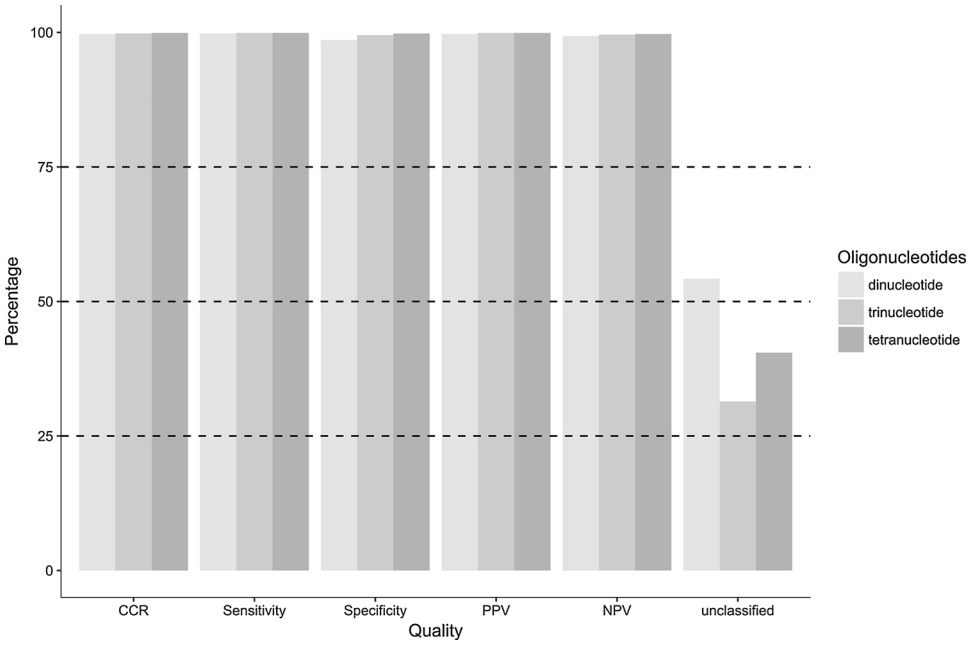

Fig. 3. Results of the DNA/RNA classification of cross-validation with di-, tri-, and tetranucleotides. Comparison of the correct classification rate (CCR), sensitivity, specificity, positive prediction value (PPV, for DNA), negative prediction value (NPV, for RNA) and percentage of unclassified reads. Dashed lines: 25%, 50% and 75% bound.

3. Results

LVQ-KNN was validated via a 5-fold cross-validation and further examined with a real and a simulated test dataset evaluating various requests. Finally, LVQ-KNN was compared with the supervised composition-based software tool RAIphy.

3.1. Cross validation results

Fig. 3 displays the comparison of the classification results using di-, tri-, or tetranucleotides, respectively. It is visible that the classification of the original reference sequences by cross validation is hyper-accurate with an average correct classification rate between 99.7% and 99.9%. The optimal classification results were reached with the trinucleotide frequencies. While similar values for CCR (99.7–99.9%), PPV (99.7–99.9%) and NPV (99.3–99.7%) were achieved with all tested oligonucleotide lengths, with 31.4% a significantly smaller amount of reads remained unclassified using trinucleotides compared with 40.5% and 54.2% unclassified reads using di- and tetranucleotides, respectively.

The optimal results using trinucleotide frequencies were achieved with t = 20 ls, m = 1000 prototypes per class, k = 11 NN and l = 11 votes.

These accurate results are a characteristic outcome for the validation of training datasets and classification methods. In order to get more realistic measures of the classification quality, both a real and simulated dataset were analyzed.

3.2. Real data results

The real dataset contained 29,106 sequences from diverse species, representing bacteria and DNA and RNA viruses (Sup. B Table S1) with read lengths between 32–1101 bp and an average length of 313 bp. The classification of this real dataset reached an optimum when tetranucleotide information was computed and evaluated (Fig. 4). Although the number of unclassified reads (57.4–59.8%) and for DNA sequences the

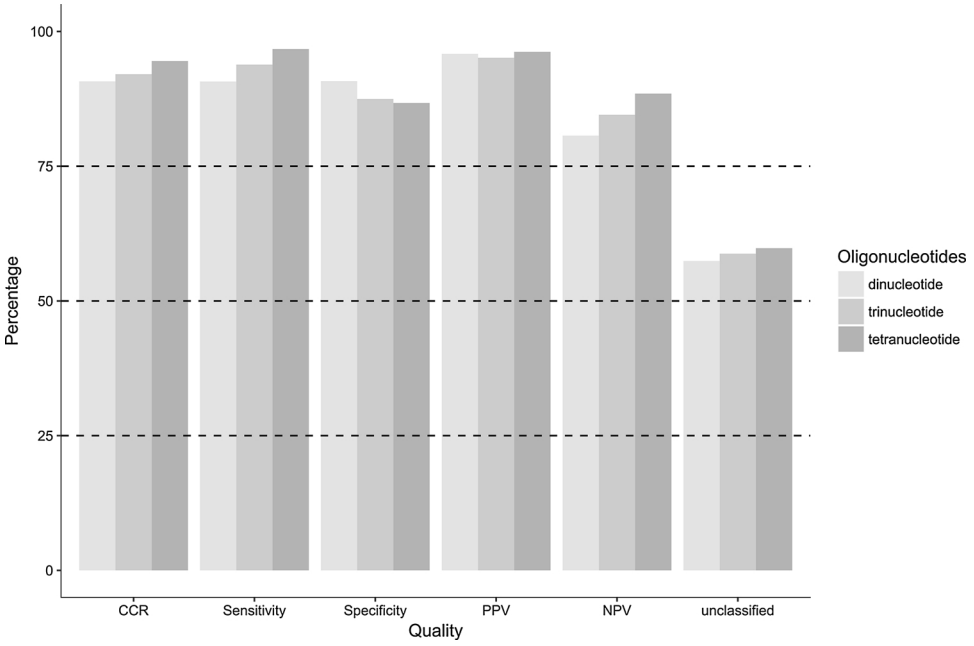

Fig. 4. Results of the DNA/RNA classification of real data with di-, tri-, and tetranucleotides. Comparison of the correct classification rate (CCR), sensitivity, specificity, positive prediction value (PPV, for DNA), negative prediction value (NPV, for RNA) and percentage of unclassified reads. Dashed lines: 25%, 50% and 75% bound.

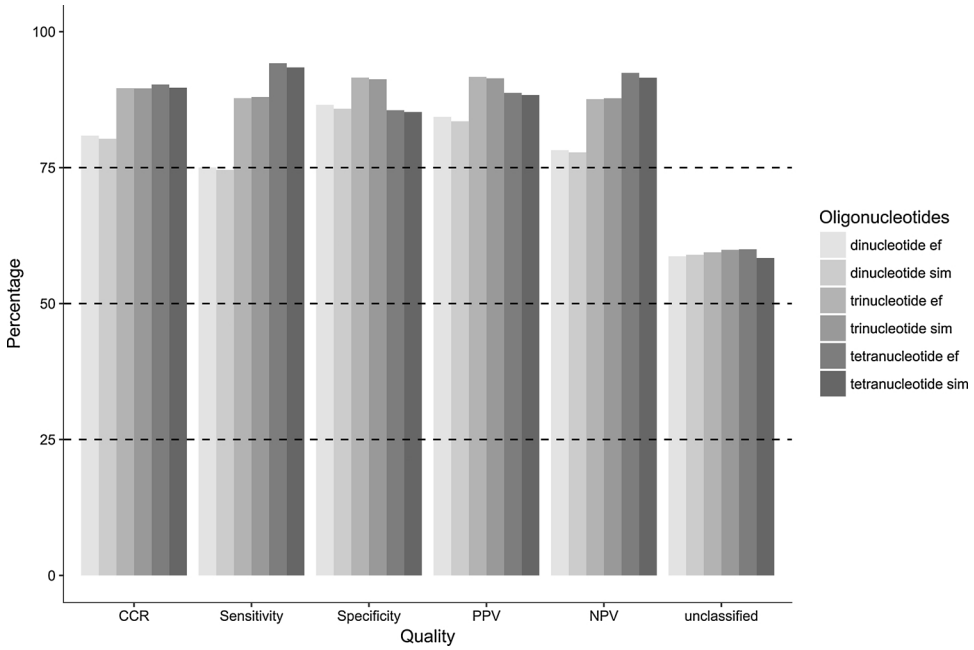

Fig. 5. Results of the DNA/RNA classification of simulated data with di-, tri-, and tetranucleotides. Comparison of the correct classification rate (CCR), sensitivity, specificity, positive prediction value (PPV, for DNA), negative prediction value (NPV, for RNA) and percentage of unclassified reads. Dashed lines: 25%, 50% and 75% bound; ef, error free sequences; sim, sequences with simulated illumina errors.

PPV (95.8–96.2%) are slightly increasing with oligonucleotide lengths, the NPV for RNA sequences (80.6–88.4%), sensitivity (90.7–96.7%) and CCR (90.7–94.5%) are better when using tetranucleotides.

The optimal parameter setting is quite different in comparison to the cross-validation. For dinucleotides 50 prototypes per class, 15 ls, 9 NN and 7 votes reached the best classification result. Using trinucleotides 500 prototypes per class, 30 ls and 18 NN with a vote of 16, and using tetranucleotides 1000 prototypes per class, 15 ls and 20 NNs with a vote of 19 are necessary for a satisfactory classification. The better classification quality of DNA sequences, especially the difference between the PPV and NPV, results from the different amount of DNA (19,347) and RNA sequences (9,759) in the test dataset. This imbalance influences the prediction values of both classes and has to be considered in the interpretation of the results.

3.3. Simulated data results

The simulated datasets were created with the software tool ART (Huang et al., 2012) using 16 reference sequences from NCBI (Sup. C Table S2). The read simulation was performed with the following parameter settings:

/art_src_MountRainier_Linux/art_illumina -c noseqs -i reference.fasta -l loseq -o outputname –ef

with *noseqs* the number of generated sequences and *loseq* = 250 the length of the sequences. The parameter *ef* generated an additional samfile of the sequences without an illumina sequencing error-profile. The illumina error-profiles were produced employing default settings. Finally, two datasets were generated, one with illumina error-profiles within the sequences and one without, containing 23,998 sequences each. Hence, a direct comparison between the sequences with and without errors is possible.

The CCR increased from 80.89% for dinucleotides to up to 90.29% for tetranucleotides (Fig. 5). While the sensitivity also increased from dinucleotides (75.04%) to tetranucleotides (94.2%), the specificity reached its maximum for trinucleotides at 91.56%. In general, it was shown that no significant differences between the classification results of the error-free sequences and the sequences with illumina error-profiles were observed. In contrast, the choice of di-, tri- and tetranucleotides had an impact on the results. However, a closer consideration revealed that the source of the sequence also influenced the classification. Fig. 6 shows the same results as Fig. 5 but with a deeper look into the original source of the sequences. Bacterial DNA and viral RNA sequences reached satisfying CCRs of 85.55–98% for all analyzed oligonucleotides while viral DNA sequences reached a CCR of only 52.63–53.18% for dinucleotides. The classes DNA/RNA in this dataset are balanced; hence, no influence of the quality is given, but to some extent, the classification of viral DNA sequences tends to be less accurate than of viral RNA or bacterial DNA sequences. Finally, as Fig. 6 shows, LVQ-KNN reaches best classification results over all the classes when using tetranucleotides.

3.4. Runtime

The runtimes of all steps of the complete procedure were determined using the server as outlined in materials and methods. Fig. 7 summarizes the times necessary for oligonucleotide frequency computation for the training and test datasets, prototype calculation, and finally classification of the test dataset. It is obvious that the major determinant of the runtime of the complete procedure is prototype generation. Here, the runtime increases with the oligonucleotide length because this determines the length of the feature vectors and hence the amount of information that needs to be considered. Since prototype generation needs to be done only after addition of significant numbers of new sequences, this is not the main determinant of the general classification runtime. The runtime for classification of sequences is the sum of the times necessary for oligonucleotide frequency determination and the classification itself. For the real test dataset (see 3.2), as Fig. 7 clearly shows, the major part is the determination of the oligonucleotide frequencies. While in case of tetranucleotides the frequency computation took approx. 80 min, the classification was finished in 1 min. Noteworthy, the computation of the feature vectors was not parallelized and hence could be accelerated for larger datasets by using the available compute capacity.

3.5. Comparison with RAIphy

For the comparison, LVQ-KNN and RAIphy were tested with the same training and query datasets. To compare the tools, initially only parameter settings, which did not produce unclassified sequences were applied in finding the optimum for LVQ-KNN, because RAIphy does not have the class “unclassified”. This optimal parameter setting achieved a better sensitivity (69.2% to 77.3%), CCR (73% to 76.6%) and NPV

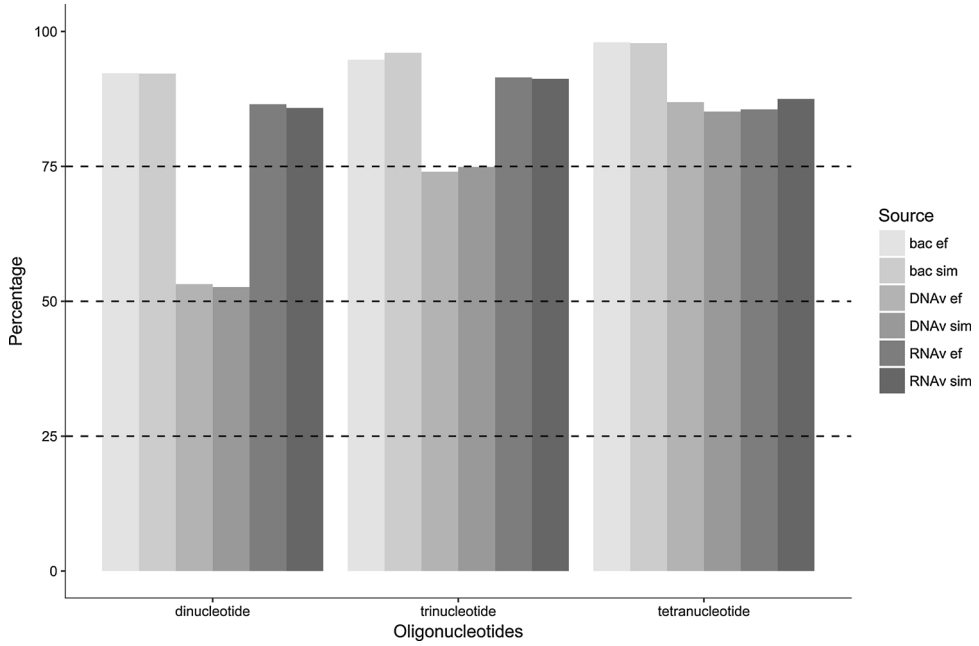

Fig. 6. Comparison of correct classification rates for di-, tri-, and tetranucleotides of simulated reads. Dashed lines indicate 25%, 50% and 75% bounds. bac, bacterial DNA; DNAv, viral DNA; RNAv, viral RNA; ef, error free sequences; sim, sequences with simulated illumina errors.

(71.4% to 77%) than RAIphy (Fig. 8). Comparing the results from RAIphy with the optimal results from LVQ-KNN including the possibility of "unclassified" as a status over all parameter settings (Fig. 9), LVQ-KNN is more reliable than RAIphy, independent of the applied oligonucleotide or optimal parameter setting.

4. Discussion

The classification of nucleotide sequences based on their oligonucleotide frequency composition is not a new but an underexplored field of metagenomics analyses. In comparison to TACOA, WSVDD and other software tools, which try to assign bacterial and archaeal sequences correctly to taxonomic classes, we have been successful in differentiating viral and bacterial sequences by assigning them not into their taxonomic classes but to their original nucleic acid classes DNA or RNA. The focus of our work was on the correct classification of viral RNA sequences, particularly those that could not be classified by sequence similarity based tools. Due to the different analytical purposes and prerequisites, a comprehensive comparison between these methods is not feasible. Despite this, a comparison of our method and the tool RAIphy was performed. LVQ-KNN produces significantly better results for the tested parameter settings. Since the computation of the prototypes depends on the training dataset, the selection of data is crucial. By selection of more suitable training datasets (e.g. by excluding rRNA sequences that are here assigned as DNA) the classification output can be optimized. Another important fact that influences the classification quality is the length and quality of the query sequences.

In general, it can be assumed that longer sequences contain more information and will therefore allow better classification results. This can be confirmed by comparing the cross validation results with real and simulated data. The cross validation data contains whole genome sequences with an average length of 70,241 bp. The simulated and real

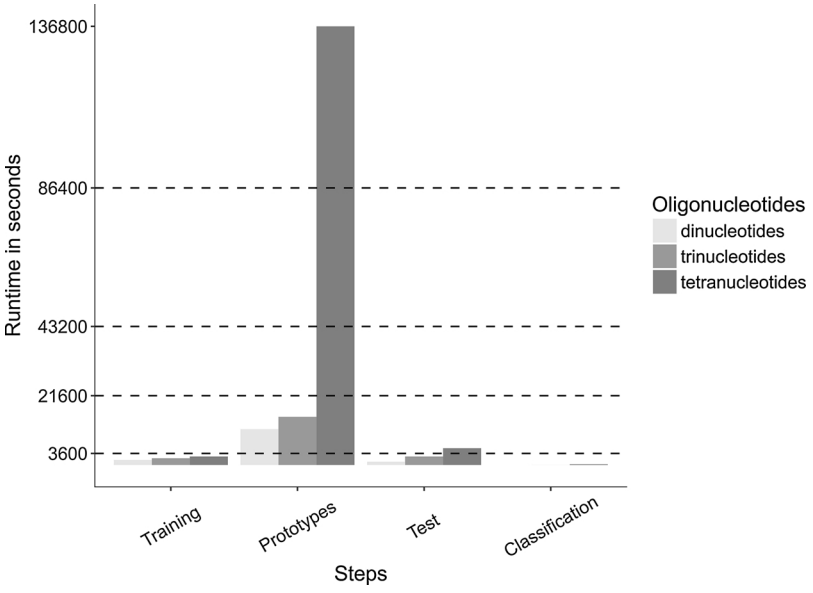

Fig. 7. Runtime of the different steps necessary for prototype generation and sequence classification. The time necessary for the respective step was determined using the training dataset for prototype computation and the real dataset for classification. The dependence of the necessary time on the length of the used oligonucleotides is shown in comparison. The first two steps "Training" and "Prototypes" refer to oligonucleotide computation from the reference dataset and subsequent calculation of the optimal prototype set. The next two steps are the oligonucleotide computation for the test dataset ("Test") and the subsequent classification of the sequences ("Classification").

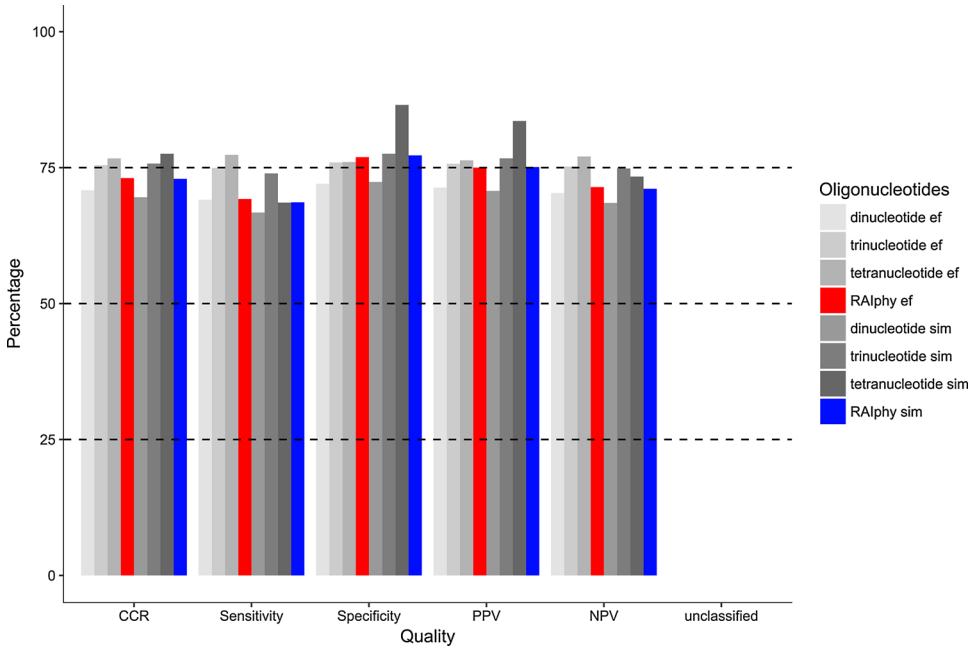

Fig. 8. Comparison of classification results from RAIphy and LVQ-KNN without status "unclassified". Comparison of the correct classification rate (CCR), sensitivity, specificity, positive prediction value (PPV, for DNA), and negative prediction value (NPV, for RNA). For LVQ-KNN only parameter settings which not produced unclassified reads were applied. ef: error free sequences; sim: sequences with simulated illumina errors. Dashed lines: 25%, 50% and 75% bound.

sequences have an average length of 284 bp, which is much shorter compared with the training sequences. Hence the procedure reached much better results with the cross validation than with the test datasets. Moreover, the reference sequences of the simulated dataset were split into subsequences of different lengths (250b, 500b, 2,000b, 5,000b, 10,000b) and the classification results were then compared between the groups (Fig. 10). The figure clearly shows that the longer the sequences are, the better the classification is; in addition, the percentage of unclassified reads decreases when the length of the sequences increases. Noteworthy, the sequence composition also greatly influences the sensitivity. In case of low complexity sequences like Poly-A or sequences with extremely high or low G + C content, the information content can be low, regardless of the sequence length. Lastly, the percentage of unclassified sequences is also determined by the parameter settings for number of k nearest neighbors and the l votes. The presented results were obtained with parameter settings that were optimized for a highly specific rather than a sensitive classification, with good positive and negative prediction values. The amount of unclassified sequences can be reduced by using different settings but this leads to a lower correct classification rate.

Regarding the choice of the used oligonucleotide lengths, we confined the analyses to di-, tri, and tetranucleotides for two reasons: (i) as shown (see 2.4), the upper bound of the length of oligonucleotides that are useful for the analysis of short-read sequences is 4 and (ii) CCR and percentage of unclassified reads apparently approach the optimum with tetranucleotides. Taking in addition the increase of runtime (Fig. 7) depending on the chosen oligonucleotide into account, we decided not to consider oligonucleotides longer than 4. Even when analyzing reads generated from third-generation sequencers which can be substantially longer (up to kilobases), use of longer oligonucleotides will not significantly improve CCR (see Fig. 10). Only read-length will improve the percentage of the unclassified sequences.

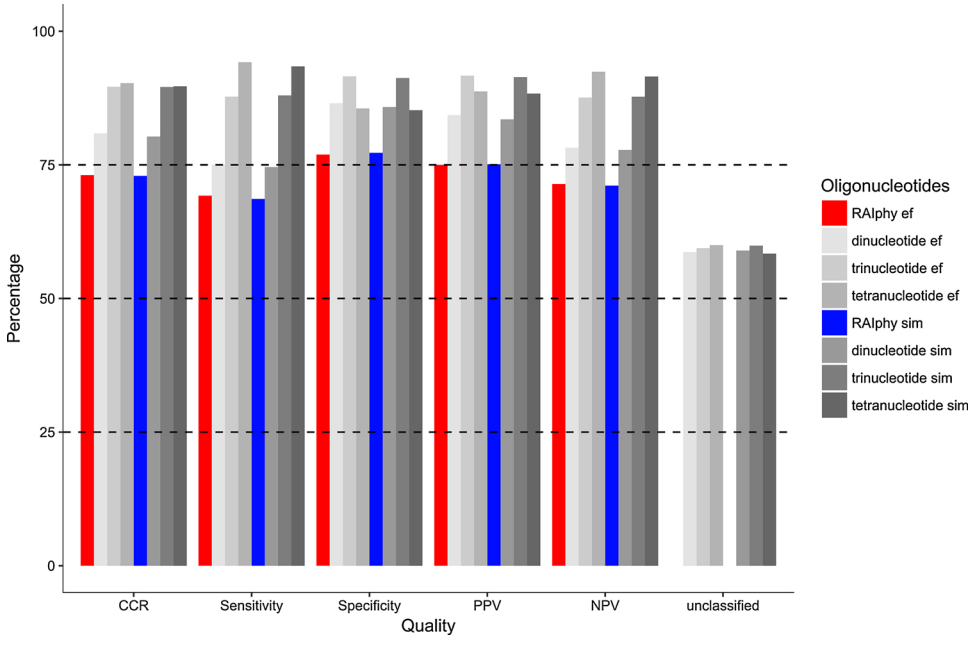

Fig. 9. Comparison of classification results from RAIphy and LVQ-KNN including status "unclassified". Comparison of the correct classification rate (CCR), sensitivity, specificity, positive prediction value (PPV, for DNA), negative prediction value (NPV, for RNA) and percentage of unclassified reads. ef: error free sequences; sim: sequences with simulated illumina error. Dashed lines: 25%, 50% and 75% bound.

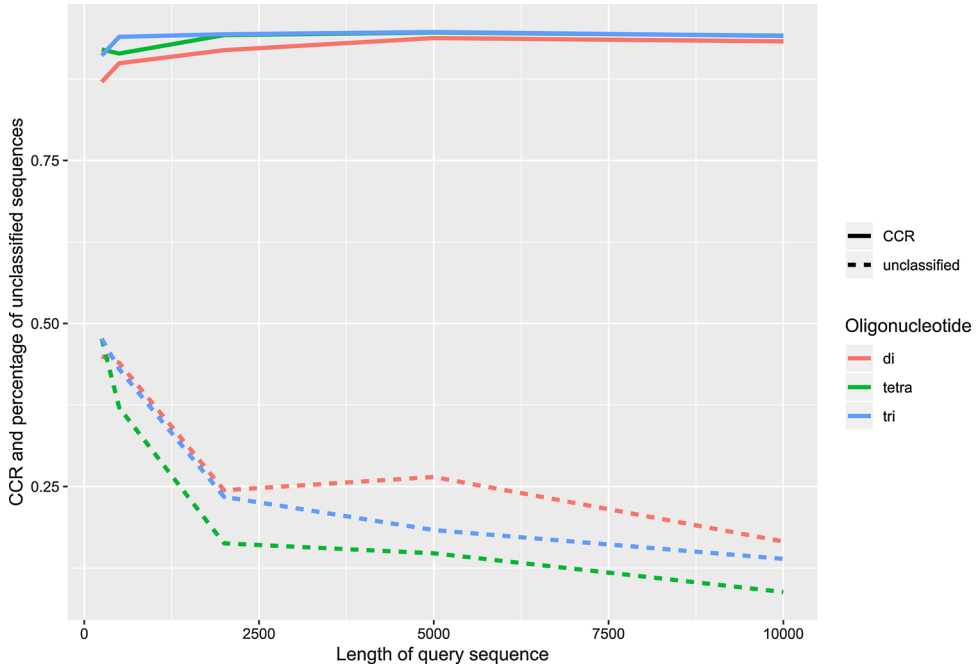

Fig. 10. Comparison of the classification results from sequences of different lengths. Depicted are the percentage of unclassified reads (dashed lines) and the CCR.

k-NN compares the query features with those from the prototypes by computing their distances and chooses the class with the majority of the nearest prototypes. While all feature vectors have the same value range, the best metric choice is the Euclidean distance. None of the features, in this case the oligonucleotide frequencies, are dominating the distance calculation and therefore the classification results. Furthermore, no standardization approach is necessary.

5. Conclusion

This work is one of the first that enables classification of not only bacterial but also viral sequences into their correct original classes DNA or RNA without sequence similarity comparisons. It combines a supervised learning algorithm with a standard classification method to overcome the difficulty of a slow classification caused by too many comparison computations. LVQ-KNN is suitable for de novo classification with satisfactory results.

Acknowledgements

We thank Florian Pfaff and Claudia Wylezich for helpful comments on the manuscript. We are grateful to Susanne Fechtner for her support in comparing various methods and settings. The study has been funded in part by the European Union Horizon 2020 programme (European Commission Grant Agreement No. 643476 'COMPARE') and the BMBF project DetektiVir (Grant No. 13N13783).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi: https://doi.org/10.1016/j.virusres.2018.10.002.

References

Anthony, S.J., Epstein, J.H., Murray, K.A., Navarrete-Macias, I., Zambrana-Torrelio, C.M., Solovyov, A., Ojeda-Flores, R., Arrigo, N.C., Islam, A., Ali Khan, S., Hosseini, P., Bogich, T.L., Olival, K.J., Sanchez-Leon, M.D., Karesh, W.B., Goldstein, T., Luby, S.P., Morse, S.S., Mazet, J.A., Daszak, P., Lipkin, W.I., 2013. A strategy to estimate unknown viral diversity in mammals. mBio 4 (5) e00598-e00513.

Brady, A., Salzberg, S.L., 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat. Methods 6 (9), 673–676.

Byrd, A.L., Perez-Rogers, J.F., Manimaran, S., Castro-Nallar, E., Toma, I., McCaffrey, T., Siegel, M., Benson, G., Crandall, K.A., Johnson, W.E., 2014. Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. BMC Bioinformatics 15, 262.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Chan, C.-K.K., Hsu, A.L., Halgamuge, S.K., Tang, S.-L., 2008. Binning sequences using very sparse labels within a metagenome. BMC Bioinformatics 9, 215.

Chen, K., Pachter, L., 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput. Biol. 1 (2), e24.

Diaz, N.N., Krause, L., Goesmann, A., Niehaus, K., Nattkemper, T.W., 2009. TACOA - taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics 10, 56.

Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J.F., Byrd, A.L., Castro-Nallar, E., Crandall, K.A., Johnson, W.E., 2014. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome 2, 33.

Hou, T., Liu, F., Liu, Y., Zou, Q.Y., Zhang, X., Wang, K., 2015. Classification of metagenomics data at lower taxonomic level using a robust supervised classifier. Evol. Bioinform. Online 11, 3–10.

Huang, W., Li, L., Myers, J.R., Marth, G.T., 2012. ART: a next-generation sequencing read simulator. Bioinformatics 28 (4), 593–594.

Kim, D., Song, L., Breitwieser, F.P., Salzberg, S.L., 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 26 (12), 1721–1729.

Kohonen, T., 2001. Self-Organizing Maps. Springer Series in Information Sciences, vol. 30 Springer-Verlag, Berlin Heidelberg.

Leung, H.C., Yiu, S.M., Yang, B., Peng, Y., Wang, Y., Liu, Z., Chen, J., Qin, J., Li, R., Chin, F.Y.L., 2011. A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. Bioinformatics 27 (11), 1489–1495.

Liu, Z., Bensmail, H., Tan, M., 2012. Efficient feature selection and multiclass classification with integrated instance and model based learning. Evol. Bioinform. Online 8, 197–205.

McHardy, A.C., Martín, H.G., Tsirigos, A., Hugenholtz, P., Rigoutsos, I., 2007. Accurate phylogenetic classification of variable-length DNA fragments. Nat. Methods 4 (1), 63–72.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., Edwards, R.A., 2008. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9, 386.

Nalbantoglu, O.U., Way, S.F., Hinrichs, S.H., Sayood, K., 2011. RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. BMC Bioinformatics 12, 41.

NCBI Resource Coordinators, 2016. Database resources of the national center for biotechnology information. Nucleic Acids Res. 44 (D1), D7–D19.

Ounit, R., Lonardi, S., 2016. Higher classification sensitivity of short metagenomic reads with CLARK-S. Bioinformatics 32 (24), 3823–3825.

Ounit, R., Wanamaker, S., Close, T.J., Lonardi, S., 2015. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative *k*-mers. BMC Genomics 16, 236.
Patil, K.R., Haider, P., Pope, P.B., Turnbaugh, P.J., Morrison, M., Scheffer, T., McHardy, A.C., 2011. Taxonomic metagenome sequence assignment with structured output models. Nat. Methods 8 (3), 191–192.
R Core Team, 2017. R: A Language and Environment for Statistical Computing, 3.4.1 ed. R Foundation for Statistical Computing, Vienna, Austria.
Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16 (6), 276–277.
Rosen, G.L., Reichenberger, E.R., Rosenfeld, A.M., 2011. NBC: the Naïve Bayes classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics 27 (1), 127–129.
Scheuch, M., Höper, D., Beer, M., 2015. RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. BMC Bioinformatics 16 (1), 69.
Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., Huttenhower, C., 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods 9 (8), 811–814.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics With S. Statistics and Computing. Springer-Verlag, New York.
Weber, M., Teeling, H., Huang, S., Waldmann, J., Kassabgy, M., Fuchs, B.M., Klindworth, A., Klockow, C., Wichels, A., Gerdts, G., Amann, R., Glöckner, F.O., 2011. Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics. ISME J. 5 (5), 918–928.
Wickham, H., 2009. ggplot2 - Elegant Graphics for Data Analysis. Use R!. Springer-Verlag, New York.
Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15 (3), R46.