# Dinucleotide distance histograms for fast detection of rRNA in metatranscriptomic sequences

**Heiner Klingenberg[1], Robin Martinjak[1], Frank Oliver Glöckner[2,3], Rolf Daniel[4], Thomas Lingner[1], and Peter Meinicke*[1]**

1 **Department of Bioinformatics, University of Göttingen**
**Goldschmidtstr.1, Göttingen, Germany**
`{heiner,robin,thomas,peter}@gobics.de`

2 **Max Planck Institute for Marine Microbiology**
**Celsiusstrasse 1, 28359 Bremen, Germany**

3 **Jacobs University Bremen gGmbH**
**Campus Ring 1, 28759 Bremen, Germany**
`fog@mpi-bremen.de`

4 **Department of Genomic and Applied Microbiology, University of Göttingen**
**Goldschmidtstr.1, Göttingen, Germany**
`rdaniel@gwdg.de`

## Abstract

With the advent of metatranscriptomics it has now become possible to study the dynamics of microbial communities. The analysis of environmental RNA-Seq data implies several challenges for the development of efficient tools in bioinformatics. One of the first steps in the computational analysis of metatranscriptomic sequencing reads requires the separation of rRNA and mRNA fragments to ensure that only protein coding sequences are actually used in a subsequent functional analysis. In the context of the rRNA filtering task it is desirable to have a broad spectrum of different methods in order to find a suitable trade-off between speed and accuracy for a particular dataset. We introduce a machine learning approach for the detection of rRNA in metatranscriptomic sequencing reads that is based on support vector machines in combination with dinucleotide distance histograms for feature representation. The results show that our SVM-based approach is at least one order of magnitude faster than any of the existing tools with only a slight degradation of the detection performance when compared to state-of-the-art alignment-based methods.

**1998 ACM Subject Classification** J.3 Biology and genetics

**Keywords and phrases** Metatranscriptomics, metagenomics, rRNA detection, distance histograms

**Digital Object Identifier** 10.4230/OASIcs.GCB.2013.80

## 1 Introduction

Metatranscriptomics has become an essential tool for the investigation of gene expression in microbial communities [6, 16, 7, 2, 13]. Compared to metagenomics, metatranscriptomics provides a dynamic picture of the adaptation of organisms to changing environmental conditions. Depending on the particular protocol for extraction and sequencing of environmental

* corresponding author

German Conference on Bioinformatics 2013 (GCB'13).
Editors: T. Beißbarth, M. Kollmar, A. Leha, B. Morgenstern, A.-K. Schultz, S. Waack, E. Wingender; pp. 80–89
OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

RNA, a substantial amount of the resulting sequences actually correspond to ribosomal RNA (rRNA) that cannot be used for the analysis of gene expression levels. Therefore an important first step in the analysis of RNA-Seq data from metatranscriptomic experiments is to filter out the fraction of sequencing reads with significant similarity to known rRNA genes. After that the remaining messenger RNA (mRNA) reads are usually analyzed in terms of possible gene functions based on sequence similarity to known proteins from, for instance, the Pfam [17] or KEGG [10] databases. Without a prior rRNA filtering the risk is high to obtain a large number of false positive protein matches. For example, in a previous release of the Pfam database, due to misannotation, several families have been composed of spurious ORFs on the reverse strand of rRNA and therefore systematically accounted for false protein matches in metatranscriptomic RNA-Seq data [22]. Besides a time-consuming BLASTN [1] search against a comprehensive rRNA database, several recent tools can be used which all provide a computationally faster rRNA detection. The accelerating techniques in these tools include Hidden Markov Models [9, 12], the Burrows-Wheeler transformation [21] and TRIE-structures in combination with a fast bitvector matching [11]. We here propose a machine learning approach using a feature space based on oligomer distance histograms which have originally been introduced for remote homology detection in protein sequence analysis [14]. For rRNA detection we have implemented a specific feature extraction that counts the occurrences of all dinucleotide pairs over a range of possible distances (spacers) between them. Our results indicate that dinucleotide distance histograms provide a suitable representation of rRNA sequences and that SVMs are well-suitable for fast detection of rRNA in metatranscriptomic datasets.

## 2 Materials & Methods

Our approach for rRNA detection in metatranscriptomic datasets is based on an RNA-specific adaption of the oligomer distance histogram (ODH) representation for biological sequences [14] and Support Vector Machines (SVM) for discrimination between rRNA and non-rRNA sequence fragments. After training of SVM classifiers using reference datasets for 16S/23S-rRNA and non-rRNA examples, we evaluate the performance of our method and several state-of-the-art rRNA detection approaches on simulated and real-world metatranscriptomics datasets. In the following sections we describe in detail the utilized datasets and the modified ODH feature space and we outline the methods used for performance comparison. The C source code for ODH-based rRNA detection is available from the authors.

### 2.1 Datasets

### Reference datasets

To construct a reference dataset for training and test of SVM classifier models, we obtained all available rRNA gene sequences from the SILVA database [18] and separated them according to their phylogenetic origin (Archaea/Bacteria) and type (16S/23S). 3' and 5' sequence overhangs were removed by trimming the sequences according to their relevant rRNA region using the ARB software package [15]. To reduce the redundancy of the dataset and avoid overlaps between training and test data, we clustered the resulting sequence sets with USEARCH (version 6.0.307) [4] using a sequence identity threshold of 95%. The final bacterial/archaeal rRNA reference datasets contained 80,832/4,217 16S-rRNA sequences and 1,930/137 23S-rRNA sequences, respectively.

For evaluation of different methods we partitioned the reference datasets into training and test sets containing 80% and 20% of the sequences, respectively. Here, we attempted to create sequence sets with similar sequence variability (for details see Appendix). By this means we obtained 64,665 (16,167) full length training (test) 16S-rRNA sequences for Bacteria and 3,373 (844) sequences for Archaea. The respective 23S datasets consist of 1,544 (386) bacterial and 109 (28) archaeal sequences.

For our discriminative learning approach we also require the training and test datasets to contain negative sequence examples, i.e. suitable non-rRNA sequences. For this purpose we masked 1,705/121 completely sequenced bacterial/archaeal genomes with respect to known rRNA and non-coding regions. For each rRNA reference dataset we extracted fragments from the remaining sequence material to yield a negative dataset of identical size and sequence length distribution.

### Simulated metatranscriptome dataset

In order to evaluate the rRNA detection performance of different methods, we generated a simulated metatranscriptome dataset with known rRNA and non-rRNA labels. For this purpose we applied MetaSim[19] to our positive and negative test sequences to produce rRNA and non-rRNA sequence reads, respectively. Here, the MetaSim default parameters for the Roche 454 sequencer model (including a relatively high error rate of 5%) were used and the average read length was set to 250 bp. Multiple (forward/reverse) reads were generated for each reference sequence until at least 80% of the original sequence was covered. Sequence fragments that had an overlap of more than 150 bp with another read were removed. In total, our simulated dataset consists of 214,270/10,535 16S/23S-rRNA and 952,215 non-rRNA reads, respectively.

### Real-world metatranscriptome datasets

In [12] two metatranscriptome datasets were used for comparative evaluation of the rRNA detection performance of the rRNASelector and Meta-RNA methods. The datasets ('Tidal salt marsh' and 'Mushroom Spring') revealed a remarkably high predicted fraction (54% and 89%, respectively) of rRNA-related sequences. The Mushroom Spring dataset (SRR106861) consists of 113,128 sequences ($\approx$30 Mbp) and an average read length of $\approx$267 bp, while the Tidal salt marsh dataset (SRR013513) comprises 238,250 sequences ($\approx$62 Mbp) and an average read length of $\approx$259 bp. Both samples have been sequenced using the Roche 454 FLX Titanium platform. We downloaded the two datasets from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA[1]), converted the sequences into the FASTA format and applied PRINSEQ [20] to the sequence sets to remove replicates and sequences exceeding 5 undefined bases. After application of PRINSEQ, 196,512 sequences ($\approx$50 Mbp) and 91,589 sequences ($\approx$24 Mbp) remained for testing in the reduced SRR013513 and SRR106861 dataset, respectively.

## 2.2 Dinucleotide distance histogram feature space

In [14], oligomer distance histograms (ODH) have been introduced as a vector space representation method for protein sequences. In the ODH feature space each sequence is represented as a numerical vector in which each dimension indicates the number of occurrences of a particular oligomer ($k$-mer) pair at a particular distance in the sequence. For a specific

[1] http://www.ncbi.nlm.nih.gov/sra

sequence analysis problem the ODH feature space consists of the required dimensions to consider all possible oligomer pairs for all distances (including the 'zero' distance $D = 0$) up to the longest observable distance between two oligomers.

To apply ODHs to the rRNA detection problem (in a computationally efficient manner), we here fix the length $k$ of the oligomers to $k = 2$ (dinucleotides) and introduce an upper limit $D_{max}$ for the distance between two oligomers. In contrast to [14] we here omit oligomer distances that reflect an overlap of two oligomers, i.e. the distances $D = 0$ and $D = 1$, and instead refer to an inventory of 'spacers' from $\mathcal{D}_0$ (i.e. $D = k$) to $\mathcal{D}_{max}$ ($D_{max}+k$). Therefore, our dinucleotide distance histogram feature space consists of $16^2 * \mathcal{D}_{max}$ dimensions. The maximum spacer value $\mathcal{D}_{max}$ constitutes a so-called hyperparameter whose optimal value has to be determined by evaluation.

### Discriminative classifier training

In order to distinguish between rRNA and non-rRNA sequence reads in metatranscriptomics datasets, we trained discriminative linear classifier models using Support Vector Machines (SVM) in combination with dinucleotide distance histogram feature space representatives of the reference dataset sequences. Here, we aggregated bacterial and archaeal sequences to obtain one 16S- and one 23S-rRNA classifier, respectively. For SVM training we used the LIBLINEAR implementation [5] with default parameters for the slack variables ($C = 1$) and termination tolerance ($\epsilon = 0.1$). Because the LIBLINEAR toolbox does not provide an option to account for imbalanced numbers of positive and negative training data, we 'oversampled' the positive examples to yield the same amount of rRNA and non-rRNA sequences while retaining the diversity of non-rRNA sequences. As a consequence, we used up to 30 duplicates (archaeal 23S-rRNA) of an rRNA example for model training.

## 2.3 Experimental setup

For performance evaluation we compared our approach to different state-of-the-art methods for rRNA detection. Besides the riboPicker [21] and SortMeRNA [11] methods, we used HMMER3 [3] as a representative method for HMM-based detection approaches such as Meta-RNA [9] or rRNASelector [12]. The riboPicker detection software is based on a pre-computed rRNA database[2] which does not allow the convenient removal of particular sequences. To avoid an overlap of training and test sequences, we refrained from using riboPicker for evaluation on the simulated metatranscriptome dataset and instead performed a BLASTN homology search. Here, we used the rRNA/non-rRNA assignment of the best BLAST hit up to an E-value threshold of $1e^{-3}$ for read classification. For the simulated data, the reference models/databases for all methods were built using only the training sequences. Here, separate HMMER3 models for Archaea and Bacteria were trained. For evaluation on the real-world metatranscriptome dataset we used all reference dataset sequences for training of HMMER3 models and SVM classifiers and we utilized the default databases for riboPicker and SortMeRNA.

# 3 Results

In order to evaluate our distance histogram-based rRNA detection approach, we investigated the prediction performance on simulated and real-world metatranscriptome data. First, we

[2] http://edwards.sdsu.edu/ribopicker/rrnadb/rrnadb_2012-01-17.tar.gz

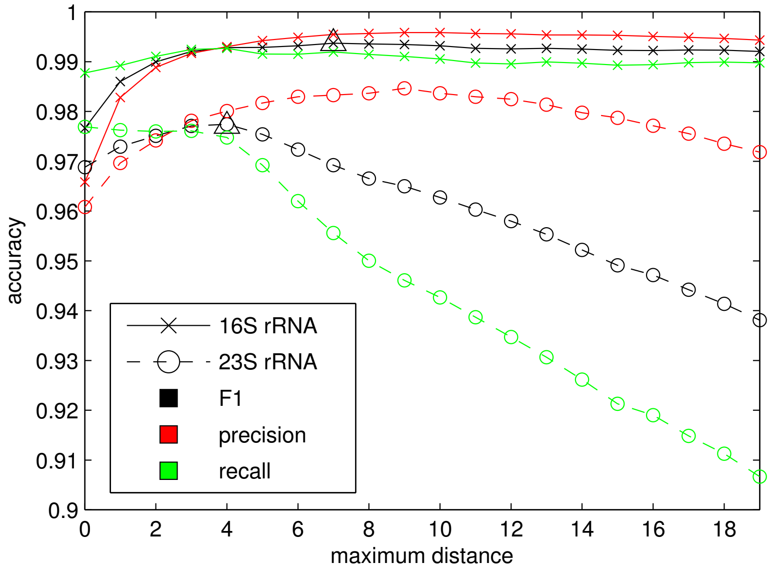

**Figure 1** Dependency of the 16S-/23S-rRNA detection performance on the maximum distance parameter. Maximum $F_1$ performance values are indicated by triangles.

determined optimal values for the maximum distance parameter of our method for 16S- and 23S-rRNA classifiers, respectively. Then we compared the rRNA detection performance and the runtime of our approach to those of state-of-the-art methods.

## Selection of optimal values for the maximum distance parameter

Our distance histogram-based feature space for nucleotide sequences (DDH) requires the definition of a maximum distance (spacer) $\mathcal{D}_{max}$ between two dimers (see also section 2.2). While small values for $\mathcal{D}_{max}$ lead to a memory-efficient feature space representation of DNA/RNA sequences, higher values allow to model conserved long-range correlations between particular residues in the sequence. To determine optimal values for the maximum distance parameter, we performed a 5-fold cross-validation on simulated 16S- and 23S-rRNA training datasets using different values for $\mathcal{D}_{max} = [0, .., 19]$. To yield meaningful performance measure values, we balanced the number of positive and negative test examples by oversampling the rRNA example sequences analogously to the classifier training procedure (see section 2.2).

Figure 1 shows the dependency of the rRNA detection performance on the maximum distance value in terms of precision ($\frac{\#TP}{\#TP+\#FP}$) and recall ($\frac{\#TP}{\#TP+\#FN}$) curves. While the recall values already start to decrease for medium values of $\mathcal{D}_{max}$, the precision increases until a plateau is reached. The $F_1$ measure, which combines precision and recall, shows different specific local performance maxima for 16S ($\mathcal{D}_{max} = 7$) and 23S ($\mathcal{D}_{max} = 4$) data. In the following, we use these optimal values for all evaluations.

## Performance on simulated data

Our simulated dataset allows to evaluate the rRNA detection performance of different methods independently for 16S- and 23S-rRNA sequence fragments based on a known classification of the reads. Table 1 shows the performance values of four different methods for our simulated 16S- and 23S-rRNA datasets, respectively. For 16S-rRNA data, the HMMER3 method

**Table 1** rRNA detection performance on simulated metatranscriptome data for different methods. All values represent percentages.

| | | DDH | HMMER3 | BLASTN | SortMeRNA | # reads |
|---|---|---|---|---|---|---|
| 16s | recall | 98.79 | 99.94 | 99.06 | 99.90 | 428540 |
| | precision | 99.60 | 100.0 | 100.0 | 100.0 | |
| | $F_1$ | 99.19 | 99.97 | 99.53 | 99.95 | |
| 23s | recall | 97.17 | 99.28 | 91.40 | 98.92 | 21070 |
| | precision | 98.00 | 100.0 | 100.0 | 100.0 | |
| | $F_1$ | 97.58 | 99.63 | 95.51 | 99.46 | |

achieves almost perfect classification of the reads, closely followed by SortMeRNA. BLASTN and our DDH approach show a slightly lower detection performance in terms of the $F_1$ measure due to a higher fraction of overlooked 16S-rRNA sequences. Remarkably, HMMER3, SortMeRNA and BLASTN only classify very few non-rRNA reads as ribosomal RNA and thus yield a (rounded) precision of 100%.

For the simulated 23S-rRNA data HMMER3 and SortMeRNA also achieve very high detection performance values. The precision and recall values of the DDH method slightly decrease as compared to the 16S dataset, which has presumably to be attributed to the much smaller number of training examples for the 23S dataset. The detection performance of BLASTN in terms of the recall value substantially decreases for 23S-rRNA reads due to a considerably reduced number of significant hits to the database sequences. Additional experiments with longer sequence reads (400bp) and a lower simulated read error rate (2.5%) indicated that the detection performance of all methods increases, with BLASTN showing the biggest improvement (data not shown).

## 3.1 Performance on metatranscriptome data

In contrast to the simulated datasets, the classification of reads from real-world metatranscriptome data is not known and thus no ground truth exists. Because the HMMER3 method outperformed the other approaches on the simulated data, we first measured the overlap with the other methods in terms of a hypothetical ground truth on the real-world datasets provided by HMMER3 predictions. Table 2 shows the results of the overlap analysis in terms of hypothetical recall, precision and $F_1$ estimates for different methods on the two real-world metatranscriptome datasets. For the Mushroom spring dataset, SortMeRNA achieved the highest agreement with the HMMER3 prediction followed by riboPicker and our distance histogram approach. While the overlap recall ranged from 89% to almost 100%, the overlap precision of all methods was very high. This can be attributed to the high fraction of predicted rRNA ($\approx$89%) in this dataset (see also section 2.1). In contrast, the Tidal salt marsh dataset showed a substantially lower predicted fraction of rRNA reads ($\approx$54%). Here, the precision value for riboPicker and our DDH approach decreased to $\sim$92%, while the SortMeRNA method still showed a high value (99%). However, because of a substantially diminished recall value for SortMeRNA on this dataset, riboPicker exhibited the highest agreement with the HMMER3 predictions in terms of the combined $F_1$ measure.

Figure 2 shows Venn diagrams representing the overlap of predicted rRNA fractions as obtained from the three abovementioned methods without HMMER3. For the Mushroom Spring datasets the three methods agreed on $\approx$86% of the sequence fragments, while no method exclusively classified more than 1% of the reads as rRNA. The Venn diagram associated with

**Table 2** rRNA detection overlap of different methods with HMMER3 predictions on real-world metatranscriptome data. All values represent percentages.

| | | riboPicker | SortMeRNA | DDH |
|---|---|---|---|---|
| Mushroom Spring | recall | 97.60 | 99.78 | 89.39 |
| | precision | 99.75 | 99.78 | 98.93 |
| | $F_1$ | 98.66 | 99.78 | 93.92 |
| Tidal salt marsh | recall | 97.54 | 81.81 | 82.72 |
| | precision | 91.57 | 99.00 | 92.80 |
| | $F_1$ | 94.46 | 89.60 | 87.47 |

the Tidal marsh dataset shows a substantially smaller consensus of the method predictions. Here, riboPicker and our DDH approach filtered ≈ 15% and 5% of the reads exclusively. The classification overlap between the DDH method and riboPicker/SortMeRNA was ≈7.3%/0.9%, respectively.

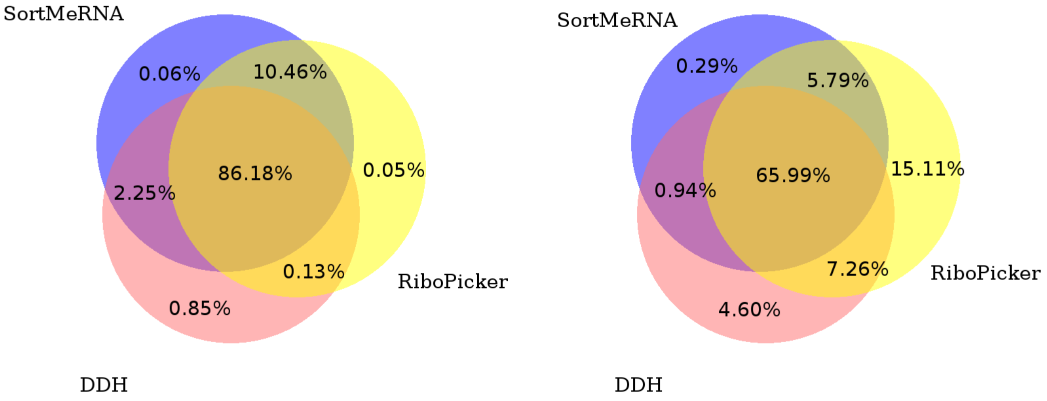

**Figure 2** Venn diagrams showing the overlap of rRNA classification results for different methods on real-world metatranscriptome data. Left-hand side: Mushroom spring dataset, right-hand side: Tidal salt marsh dataset.

In comparison with our performance analysis on simulated data the results on the real-world datasets indicate a much larger disagreement of different methods than expected. This discrepancy, in principle, could already be seen in the evaluation of the SortMeRNA tool [11]. On one hand this indicates that the simulation setup that we used in a similar way like other researchers have done before, is too simple to capture the complexity of real metatranscriptomic data. On the other hand this shows the difficulty of measuring the rRNA detection performance in general and we have to admit that the assumption of a putative best method (HMMER3) is possibly not appropriate to tackle this problem.

### 3.1.1 Runtimes

Current next-generation sequencing methods yield a large number of sequencing reads with rapidly growing sizes of the resulting datasets. Therefore, the speed of an rRNA detection

**Table 3** Runtimes of different methods on real-world metatranscriptome datasets.

| dataset | method | time in sec |
|---|---|---|
| Mushroom Spring | DDH | 5.7 |
| | SortMeRNA | 81.2 |
| | riboPicker | 363 |
| | HMM | 3016 |
| Tidal salt marsh | DDH | 12.1 |
| | SortMeRNA | 156 |
| | riboPicker | 591 |
| | HMM | 4487 |

method is an important aspect for the timely downstream analysis of the functionally relevant metatranscriptome data. To compare the runtimes of different detection methods, we performed all classification analyses for the real-world metatranscriptome datasets in single core mode on an Intel(R) Core(TM) i5-3470 CPU @ 3.20GHz with 32GB RAM. Here, the runtimes for HMMER3 and our DDH approach are aggregated over all (16S, 23S, Archaea, Bacteria) model/classifier evaluations.

As shown in table 3, our distance histogram approach is ≈13 times faster than the second-fastest method, SortMeRNA. The speed-up factor of the DDH method over riboPicker and HMMER3 ranged between ≈47 to 64 and 370 to 530, respectively. These numbers indicate the suitability of our approach for fast rRNA detection in very large datasets.

## 4 Discussion

We introduced a machine learning approach to the detection of ribosomal RNA in metatranscriptomic sequences. The utilized feature space is composed of frequencies of spacer lengths between pairs of dinucleotides. The corresponding dinculeotide distance histograms (DDH) provide a natural representation of mismatches and therefore can cope with a relatively high rate of sequencing errors. In our experiments we found that a maximum distance of approximately 10 nt, i.e. a feature space with at most 2500 dimensions, is sufficient to provide a good discrimination of rRNA from coding regions. In comparison with alignment-based methods the DDH implies a position independent analysis and therefore neglects some conserved position information that is present in rRNA sequences. As a consequence, our results indicate a slightly lower detection sensitivity and specificity. The advantage over the existing methods is the computational speed, which is more than 10 times higher than for the previously fastest method (SortMeRNA). In future work we will address some of the limitations that result from our current training setup where we utilize the LIBLINEAR SVM implementation. With this library we have to keep all training vectors in memory and therefore the number of examples is restricted to about 150,000 DDH feature vectors for 32GB RAM. Using regularized least squares training (see e.g. [8]), we will be able to substantially increase training sets and therefore more realistic training examples may be obtained from a large number of simulated sequencing reads. In particular, we expect a better representation of 23S-rRNA and a better coverage of the negative examples in terms of a more comprehensive sampling of coding regions.

**Acknowledgements** This work was supported by the Deutsche Forschungsgemeinschaft [grant numbers Me3138 (“Computational models for metatranscriptome analysis”) and Li2050 (“Machine learning methods for functional characterization of the peroxisome”)].

## A Construction of training and test sets

To create training and test sets with similar sequence variability, we analyzed the original SILVA alignments regarding the variability of the alignment columns and assigned a score to each sequence that reflects its distance from the consensus sequence/profile. Given an alphabet $\mathcal{A} = \{$A,C,G,T,N,-$\}$ and a sequence $\mathcal{S}$, the score $\mathcal{T}_{k,j}$ for a symbol $k \in \mathcal{A}$ and an alignment position $j$ is calculated by

$$\mathcal{T}_{k,j} = \begin{cases} 0, & \text{if } \frac{\sum_{i=1}^{m} \mathcal{S}_{i,j}='\text{-}'}{m} > 0.5 \\ log_{10}(\frac{\sum_{i=1}^{m} \mathcal{S}_{i,j}=k}{m}), & \text{else} \end{cases}$$

whereby $m$ represents the number of sequences. The sequence score is then calculated as $score(\mathcal{S}) = \sum_{j=1}^{L} \mathcal{T}_{\mathcal{S}_{k,j}}$. As a result, a lower sequence score indicates a higher deviation from the consensus sequence. Sequences for the final training and test datasets were then sampled from the reference datasets subject to similarly distributed sequence scores.

## References

1 S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

2 L. C. Carvalhais, P. G. Dennis, G. W. Tyson, and P. M. Schenk. Application of metatranscriptomics to soil environments. *J. Microbiol. Methods*, 91(2):246–251, Nov 2012.

3 S. R. Eddy. Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, 7(10):e1002195, Oct 2011.

4 R. C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, Oct 2010.

5 R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.

6 J. A. Gilbert, D. Field, Y. Huang, R. Edwards, W. Li, P. Gilna, and I. Joint. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE*, 3(8):e3042, 2008.

7 J. A. Gilbert and M. Hughes. Gene expression profiling: metatranscriptomics. *Methods Mol. Biol.*, 733:195–205, 2011.

8 K. J. Hoff, M. Tech, T. Lingner, R. Daniel, B. Morgenstern, and P. Meinicke. Gene prediction in metagenomic fragments: a large scale machine learning approach. *BMC Bioinformatics*, 9:217, 2008.

9 Y. Huang, P. Gilna, and W. Li. Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics*, 25(10):1338–1340, May 2009.

10 M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The KEGG databases at GenomeNet. *Nucleic Acids Res.*, 30(1):42–46, Jan 2002.

11 E. Kopylova, L. Noe, and H. Touzet. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, 28(24):3211–3217, Dec 2012.

12 J. H. Lee, H. Yi, and J. Chun. rRNASelector: a computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *J. Microbiol.*, 49(4):689–691, Aug 2011.

13 B. Leis, A. Angelov, and W. Liebl. Screening and expression of genes from metagenomes. *Adv. Appl. Microbiol.*, 83:1–68, 2013.

14 T. Lingner and P. Meinicke. Remote homology detection based on oligomer distances. *Bioinformatics*, 22(18):2224–2231, Sep 2006.

15 W. Ludwig, O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Förster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. König, T. Liss, R. Lüssmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K. H. Schleifer. ARB: a software environment for sequence data. *Nucleic Acids Res. 25;32(4):1363-71*, 32(4):1363–1371, Feb 2004.

16 R. S. Poretsky, S. Gifford, J. Rinta-Kanto, M. Vila-Costa, and M. A. Moran. Analyzing gene expression from marine microbial communities using environmental transcriptomics. *J Vis Exp*, (24), 2009.

17 M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. The Pfam protein families database. *Nucleic Acids Res*, 40:D290–D301, Jan 2012.

18 C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, 41(Database issue):D590–596, Jan 2013.

19 D. C. Richter, F. Ott, A. F. Auch, R. Schmid, and D. H. Huson. MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS ONE*, 3(10):e3373, 2008.

20 R. Schmieder and R. Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, Mar 2011.

21 R. Schmieder, Y. W. Lim, and R. Edwards. Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics*, 28(3):433–435, Feb 2012.

22 H. J. Tripp, I. Hewson, S. Boyarsky, J. M. Stuart, and J. P. Zehr. Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. *Nucleic Acids Res.*, 39(20):8792–8802, Nov 2011.