Higher Classification Accuracy of Short Metagenomic Reads by Discriminative Spaced *k*-mers

Rachid Ounit and Stefano Lonardi(✉)

Department of Computer Science and Engineering, University of California,
Riverside, CA 92521, USA
{rouni001,stelo}@cs.ucr.edu

Abstract. The growing number of metagenomic studies in medicine and environmental sciences is creating new computational demands in the analysis of these very large datasets. We have recently proposed a time-efficient algorithm called Clark that can accurately classify metagenomic sequences against a set of reference genomes. The competitive advantage of Clark depends on the use of discriminative *contiguous* k-mers. In default mode, Clark's speed is currently unmatched and its precision is comparable to the state-of-the-art, however, its sensitivity still does not match the level of the most sensitive (but slowest) metagenomic classifier. In this paper, we introduce an algorithmic improvement that allows Clark's classification sensitivity to match the best metagenomic classifier, without a significant loss of speed or precision compared to the original version. Finally, on real metagenomes, Clark can assign with high accuracy a much higher proportion of short reads than its closest competitor. The improved version of Clark, based on discriminative *spaced* k-mers, is freely available at http://clark.cs.ucr.edu/Spaced/.

Keywords: Metagenomics · Microbiome · Classification · Discriminative spaced k-mers · Short metagenomic reads

1 Introduction

One of the primary goals of metagenomic studies is to determine the composition of a microbial community, which typically involves the analysis of short reads obtained from sequencing a heterogenous microbial sample. The analysis can reveal the presence of unknown bacteria and viruses in a newly explored microbial habitat (e.g., in marine environment [24]), or in the case of the human body, elucidate relationships between diseases and imbalances in the microbiome (see, e.g., [7,10]).

Classification tools such as NBC [21], Kraken [25], Clark [19], among others, can be used to determine the composition of the microbial diversity from the sequenced reads for a microbial sample. We have recently proposed Clark in [19] and demonstrated that its classification speed is currently unmatched.

M. Pop and H. Touzet (Eds.): WABI 2015, LNBI 9289, pp. 286–295, 2015.
DOI: 10.1007/978-3-662-48221-6_21

Independently from us, it has been shown that Clark's classification precision is comparable or better than best state-of-the-art classifiers [15]. However, Clark's classification sensitivity is inferior compared to NBC [19].

The work presented in this manuscript describes a new approach to improve Clark's classification sensitivity. The approach exploits the concept of (discriminative) spaced k-mers. We first describe the notion of spaced k-mers as implemented in a new mode called Clark-S (S for "spaced"), then compare the performance of Clark-S against two of the most sensitive classifiers in the literature (i.e., NBC and Kraken), on several simulated/real metagenomic datasets. We show that at the phylum/genus level Clark-S outperforms both NBC and Kraken on all metrics.

2 Classification by Discriminative Spaced *k*-mers

2.1 Preliminaries

The concept and the utility of spaced seeds were initially described in context of a sequence-alignment tool called PatternHunter [17]. A *spaced seed* s is a string over the alphabet {1,*}, where '1' indicates that one should sample that position while '*' indicates that position should be ignored. The number of symbols in s is the *length* $|s|$ of s, while the number of 1s in s is the *weight* of s. A *spaced* k*-mer* is a spaced seed of length k. Let s be a spaced k-mer and weight w, and let m be a text of length k. We define $s(m)$ be the w-mer obtained from m using only the positions in s denoted by a 1. For example, if the text $m =$ `AAGTCT` and $s =$ `11*1*1` ($k = 6, w = 4$) then $s(m) =$ `AATT`. The same text processed using the spaced 6-mer $s =$ `1*11*1` would give the 4-mer $s(m) =$ `AGTT`.

The work of Ma *et al.* in [17] demonstrated that the use of single (and multiple) spaced seeds/k-mers significantly increased the chance of detecting a valid sequence alignment between the query and the target compared to contiguous seeds/k-mers, while incurring no additional computational cost. As a direct consequence of this work, spaced seeds are now used in the state-of-the-art homology search methods, such as Blast [1] or MegaBlast [26]. For more information about spaced seeds, we also refer the reader to [5,6,11–14] and references therein.

Consider now the following problem: we are given a read r and two target sequences g_1 and g_2, and we want to classify r to g_1 or g_2, *i.e.*, we want to know whether r is more likely to originate from g_1 or from g_2. As it is done in homology search methods, we can use seeds/k-mers as "witnesses" of possible valid alignments. A time-efficient solution is to count the number shared k-mers between r and targets g_1 and g_2, and assign r to the target that has the highest count. As said, spaced seeds/k-mers increases the probability of detecting a valid alignment compared to contiguous seeds/k-mers. It is always possible, however, that a shared seed/k-mer (whether it is spaced or not) may be a false positive. In order to compensate for false positives, we use discriminative spaced k-mers, as described next.

2.2 Discriminative Spaced *k*-mers

Given a set of reference sequences (or *targets*) $\{g_1, g_2, \ldots, g_p\}, i \in \{1, 2, \ldots, p\}$, the set D_i of discriminative k-mers for target g_i is the set of all k-mers in g_i that do not appear in any other reference sequences [19]. Given a spaced seed s of length k and weight w, we define $D_{i,s}$ to be the set of all w-mers obtained via s from k-mers in D_i. We then define the set $E_{i,s}$ of discriminative spaced k-mers as the set of all w-mers of $D_{i,s}$ that do not appear in any set $D_{j,s}$ where $j \neq i$. Thus, any w-mer in $E_{i,s}$ is a spaced k-mer of weight w that can be found in one and only one target.

As stated earlier, the concept of spaced k-mers is not new. Several popular metagenome classifiers, such as MetaPhyler [16], PhymmBL [4] or MEGAN [9], as BLAST-based methods, have been implicitly using spaced seeds. In addition, other similarity-based methods that analyze genomic and metagenomic sequences use spaced k-mers, such as Seed [2]. However, to the best of our knowledge, the concept of discriminative spaced k-mers is novel and introduced for the first time in this manuscript.

2.3 Selection of Optimal Spaced Seeds and Index Creation

The selection of specific spaced seed is critical to achieve high precision and sensitivity (see, e.g., [5,6,11–14,17]). For contiguous k-mers, the classification precision increases as we increase k. However, the highest sensitivity occurs with somewhat shorter k-mers. Clark is more precise for long contiguous k-mers (e.g., $k = 31$), but the highest sensitivity occurs for k-mers of length 19–22 [19]. As a consequence, we considered here spaced seeds of length $k = 31$ and weight $w = 22$. The choice of selecting a length of 31 is also motivated by a fair comparison against Clark and Kraken, which achieve high accuracy thanks to long 31-mers in their default mode. However, we realize that a more exhaustive analysis of k and w would be necessary, but (i) the intent of this work is to show the advantage of replacing discriminative contiguous seed with discriminative spaced seed, (ii) an analysis of other choices of w will be reported in the journal version of this paper.

Given k and w, the second step is to determine the structure of the spaced seed. In order to determine the optimal structure we proceeded to model sequence similarly as it is done in alignments-based method (see, e.g., [17]). We considered that the succession of matches/mismatches follows a Bernoulli distribution with parameter p, where p represents the similarity level between the read and the reference sequence. If a short read belongs to a known reference sequence, then the similarity level should be high since the amount of mismatches dues to genomic variations or sequencing errors are low. This is why we assumed a high similarity level, and chose $p = 95\,\%$.

We searched exhaustively through all the spaced seeds of length $k = 31$ and weight $w = 22$ (starting/ending with '1') using a similarity level of 95 %, and a random region of length 100 bp, by using the dynamic programming approach from [17] and implemented in [12]. The spaced seed with the highest hit probability [17], 0.998113, is `1111*111*111**1*111**1*11*11111`. In addition, we have also selected two additional spaced seeds with the highest hit probability namely `11111*1**111*1*11*11**111*11111` (0.998099) and finally `11111*1*111**1*11*111**11*11111` (0.998093).

Before a read can be classified, Clark-*S* builds a database of discriminative spaced k-mers for each target. Clark-*S* can take advantage of multiple spaced seeds, thus multiple databases can be created. For each spaced seed, discriminative spaced k-mers were built from contiguous discriminative 31-mers. Once the three databases of discriminative spaced k-mers were computed, they are stored in disk so they can be loaded for classification.

The classification algorithm of the "Spaced" mode is identical to that of the "full" mode (extensively described in [19]), except for two differences, namely (i) Clark-*S* queries against discriminative spaced k-mers instead of discriminative k-mers and (ii) Clark-*S* does three queries for each k-mer in a read, because there are three different databases. Finally, as done in the full and other modes, the read is assigned to the target that has the highest amount of successful queries, and several statistics (such as the confidence score and gamma score, see [19]) are computed as well.

3 Results

3.1 Datasets

To evaluate numerically the performance of the classifiers we used simulated datasets. From the available literature, we have selected the following three simulated metagenomes, which we made available at http://clark.cs.ucr.edu/Spaced/. The first dataset is "A1.10.1000" which was derived from "A1", the first group of paired-end reads in the dataset "A" from [15]. According to authors, this dataset closely mimics the complexities, size and characterization of real metagenomes. The A1 dataset contains about 28.9M reads, 80 % of which correspond to known sequenced genomes (from bacterial, archaeal and eukaryotes genomes), and 20 % of which are randomized reads (from real genomes) that should not be assigned to any taxa. We have extracted 10,000 reads from A1 as follows. We have arbitrarily taken nine different genomes from the list of genomes used to build "A1" (see Supplementary Table 1 in [15]). Then, we took the first 1,000 reads for each selected genome, and also 1,000 "random" reads. The resulting dataset, called "A1.10.1000", contains 10,000 reads (each 100 bp long) and can be considered as medium/high complexity.

The second dataset is "B1.20.500" which was derived from "B1", the first group of reads in the dataset "B", from [15]. Similarly as done for A1.10.1000, we have extracted 10,000 reads from B1 as follows. We have arbitrarily taken 19 different genomes from the list of genomes used to build "B1" (see Supplementary Table 2 in [15]). Note that these 19 selected genomes are different from those selected in A1. Then we took the first 500 reads for each selected genome, and also 500 "random" reads. The resulting dataset, called "B1.20.500", contains 10,000 reads (each 100 bp long) and can be considered as medium/high complexity.

The third dataset "simBA-5" comes from the Kraken paper and is described in it. According to the authors, it was created using bacterial and archaeal genomes, and with an error rate five times higher than the default. It contains 10,000 reads, each read is 100 bp long, and can be considered as high complexity.

To classify these metagenomic datasets, we use the entire set of bacterial/archaeal genomes from NCBI/RefSeq as reference genomes. At the time of writing, they represent 2,644 genomes and distributed in 36 phyla. The cumulative length of these genomes is 9.1 billion base pairs, where the average genome length is 3.4 million base pairs.

3.2 Comparison with Other Tools

A large set of metagenomic classifiers exists in the literature. However, a comparison between Clark and all existing classifiers is not necessary. An independent comprehensive evaluation of a wide range of metagenomics classifiers has been carried out recently using six large datasets of short paired-end reads [15]. On the data tested, Kraken is among the most accurate methods at the phylum level compared to other popular and used methods, such as mOTU [23], MetaPhlAn [22], MetaPhyler or MEGAN. However, the experimental results in [25] shows that NBC is more sensitive than Kraken, Megablast and PhymmBL at the genus level. In our study [19], we have also shown that NBC is more sensitive than Kraken at the genus level. In addition, NBC is more sensitive than Clark, at the genus level, even when the latter is run in its most sensitive settings (*i.e.*, "full" mode and $k = 20$) [19]. Note that the study [3] also shows the high sensitivity of NBC. As a consequence of this analysis, it appears sufficient to compare Clark against NBC and Kraken, as they are the two most accurate classifiers among current published methods, at the phylum and genus level.

3.3 Classification Accuracy

In this section, we present the performance of Clark (v1.2.1-beta), NBC (v1.1) and Kraken (v0.10.5-beta) on the three simulated datasets described above. Consistently with other published studies (e.g., [19,25] or [3]), the sensitivity is defined as the ratio between the number of correct assignments at a given taxonomy rank (e.g., phylum or genus) and the number of reads defined for that rank. The precision is defined as the ratio between the number of correct assignments at a given taxonomy rank (e.g., phylum or genus) and the number of assigned reads.

We present below results for the phylum and genus level. In Tables 1 and 2, the first three rows report results from Kraken Clark, and NBC, all run in their default/recommended parameters. We ran Kraken and Clark in the default mode, with $k = 31$, and NBC, with $k = 15$. The last two rows in these tables report the performance of Clark-S. In the last row we report the precision and sensitivity when filtering only high confidence (HC) assignments (*i.e.*, assignment with confidence score ≥ 0.75 and gamma score ≥ 0.03).

Table 1. Phylum-level accuracy (%) of Kraken, NBC, Clark, Clark-*S* and Clark-*S* (HC) on A1.10.1000, B1.20.500 and simBA-5

	A1.10.1000		B1.20.500		simBA-5	
	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity
Kraken	99.91	77.59	99.98	90.91	99.98	94.49
Clark	**99.93**	76.87	**100.00**	90.12	99.99	93.46
NBC	79.86	79.86	94.91	94.91	99.89	**99.89**
Clark-*S*	94.50	**79.99**	98.95	**94.98**	99.87	99.70
Clark-*S* (HC)	99.63	79.97	99.99	94.93	**100.00**	99.29

Table 2. Genus-level accuracy (%) of Kraken, NBC, Clark, Clark-*S* and Clark-*S* (HC) on A1.10.1000, B1.20.500 and simBA-5

	A1.10.1000		B1.20.500		simBA-5	
	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity
Kraken	**99.80**	70.61	99.94	90.55	**99.85**	91.97
Clark	**99.80**	69.98	**99.95**	89.69	99.82	90.77
NBC	77.94	77.94	94.76	**94.76**	98.97	**98.97**
Clark-*S*	92.71	**78.38**	98.76	94.74	98.58	98.22
Clark-*S* (HC)	99.35	76.41	**99.95**	94.52	99.61	97.24

Observe in Table 1 that (i) Clark-*S* (HC) and NBC achieve very high sensitivity, (ii) Kraken's sensitivity is lower than NBC or Clark-*S* for all datasets, (iii) Clark-*S* outperforms NBC's sensitivity in A1.10.1000 and B1.20.500, (iv) both Clark and Kraken have high precision and achieve more than 99.9 % in all datasets (even though A1.10.1000 and B1.20.500 contain reads that do not belong to any bacterial/archaeal genomes), but (v) Clark-*S* (HC) is as precise as them and outperforms NBC in all datasets.

Table 2 shows that (i) Clark's sensitivity is lower than NBC, (ii) Clark-*S* (HC) and NBC achieve the highest sensitivity and outperforms Kraken, (iii) Clark-*S* is more NBC in A1.10.1000, (iv) Kraken and CLARK show high precision and achieve both more than 99.8 % in our datasets, (v) Clark-*S* (HC) is as precise as Kraken and Clark, it outperforms NBC in all datasets, especially for A1.10.100 or B1.20.500. For simBA-5, NBC achieves the best sensitivity with 98.97, less than 2 % more than the level performed by Clark-*S* (HC).

3.4 Real Metagenomic Samples

In this section, we evaluate the performance of Clark-*S* (HC) on a large real metagenomic dataset. We have selected the dataset from [18], which is a recently published study on the population dynamics in microbial communities present in surface seawater in Monterey Bay, CA.

This dataset contains 42M reads, and the average read length is 510 bp. We pre-processed the dataset of raw reads using the following trimming steps: (i) we removed the first five bases and kept the following 100 bases using FastQ Trimmer[1], (ii) we removed reads containing sequencing adapters using Scythe[2], (iii) we trimmed the read ends if contained bases with a quality score below 30 and discarded reads containing any Ns using Sickle[3]. The resulting dataset contained 37M short reads.

We classified these 37M short reads using Kraken (default) and Clark-*S*, using the bacterial/archaeal genomes from NCBI/RefSeq. Kraken was able to classify only 1,1M reads (or 3 % of the total). Clark in its default mode also classifies about 1,1M reads. However, Clark-*S* classifies 20M reads (or 54 % of the total), about 20 times more than Kraken. Among these 20M classified reads, there are 7M high confidence assignments (or 19 % of the total), which is about 6 times more than Kraken.

The fact that Kraken assigns only 3 % of the reads can be explained by the fact that (i) Kraken relies on matching exact *k*-mer, and (ii) the current database of bacterial/archaeal likely contains only a limited fraction of the bacterial/archaeal diversity in seawater. Seawater metagenomes are likely to contain a high proportion of organisms that are missing in NBCI/RefSeq database because while the marine environment is one of the most biologically diverse on the planet [8], the culture in laboratory of bacteria from seawater is difficult [20]. Since Clark-*S* allows mismatches on the *k*-mers, it can identify at least the phylum/genus of unknown organisms.

Kraken identified, as dominant phyla, *Proteobacteria* (57 %) and *Bacteroides* (27 %). This is consistent with results reported in [18], as well as phyla in low-abundance such as *Actinobacteria* (1 %) or *Thaumarchaeota* (2 %). Within high confidence assignments of Clark-*S*, the two dominant phyla are, as expected by estimations from [18], *Proteobacteria* (56 %) and *Bacteroides* (32 %). Consistently with [18], phyla in low-abundance were correctly identified, for example, *Actinobacteria* (1 %) and *Thaumarchaeota* (2 %).

Experimental results from Kraken and Clark-*S* (HC) indicate the expected dominant phyla in the dataset (with the expected abundance for each). While Kraken and Clark-*S* (HC) are consistent for this dataset, we do notice one significant disagreement. The expected abundance of *Cyanobacteria* is 0–2 %, according to [18], but Kraken reports 9 % and Clark-*S* (HC) reports 3 %. Such discrepancies can be explained by our pre-processing to create this dataset, however, the estimation by Clark-*S* (HC) is more accurate than Kraken. As a consequence, Clark-*S* was able to assign about 20 times more short reads than Kraken, and its high confidence assignments show stronger consistency with expected results than Kraken's results.

[1] http://hannonlab.cshl.edu/fastx_toolkit/index.html.

[2] https://github.com/ucdavis-bioinformatics/scythe.

[3] https://github.com/ucdavis-bioinformatics/sickle.

3.5 Time and Space Complexity

All experiments presented in this study were run on a Dell PowerEdge T710 server (dual Intel Xeon X5660 2.8 Ghz, 12 cores, 192 GB of RAM). NBC's speed is the slowest at 8–9 reads per minute, Kraken's speed is 1.8–2M reads per minute, while Clark (default mode) runs the fastest, at 2.8–3M reads per minute. However, Clark-*S* runs slower than Clark, and classifies about 150–200 thousand reads per minute. While Clark is the fastest in the default mode, it does not provide the same classification accuracy of NBC or Clark-*S*. The fact that Clark-*S* computes spaced *k*-mers and uses several spaced seeds explains this difference of speed. However, Clark-*S* is still several thousand of times faster than NBC.

NBC consumed less than 500 MB of RAM, while Clark and Kraken used 70 and 77 GB respectively. Finally, Clark-*S* used 110 GB. This larger RAM usage is due to the multiple databases corresponding to the three spaced seeds. However, this amount remains significantly lower than 160 GB, which is the amount needed to build/construct the database of discriminative *k*-mers.

4 Discussion

We have introduced for the first time the use of discriminative spaced *k*-mers for the classification problem of short metagenomic reads. To the best of our knowledge, Clark is the first metagenome classifier using (multiple) discriminative spaced *k*-mers. We have tested Clark-*S* against Clark, Kraken and NBC.

Our results on several realistic metagenomic samples show that (i) Clark/Kraken achieves high precision while being less sensitive than NBC at the phylum/genus level, (ii) NBC achieves high sensitivity while being less precise than the other tools, however, (iii) Clark-*S* (HC) can be *both* as precise as (or more precise than) Kraken and as sensitive as NBC. While Clark-*S* is slower than Clark because its uses mutiple spaced seeds, it is still faster than NBC by several order of magnitude. Finally, in the context of real metagenomic data, we proved that Clark-*S* (HC) can classify with high accuracy a much higher proportion of short reads than Clark/Kraken.

We are currently improving the speed and the RAM usage of Clark-*S*. A public release of Clark-*S* is available at http://clark.cs.ucr.edu/Spaced/.

Acknowledgments. This work was supported in part by the U.S. National Science Foundation [IIS-1302134]. We are thankful to the anonymous reviewers for their constructive feedback.

References

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. **215**(3), 403–410 (1990)
2. Bao, E., Jiang, T., Kaloshian, I., Girke, T.: Seed: efficient clustering of next-generation sequences. Bioinformatics **27**(18), 2502–2509 (2011)

3. Bazinet, A.L., Cummings, M.P.: A comparative evaluation of sequence classification programs. BMC Bioinformatics **13**(1), 92 (2012)
4. Brady, A., Salzberg, S.: PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nat. Methods **8**(5), 367–367 (2011)
5. Brown, D.G., Li, M., Ma, B.: A tutorial of recent developments in the seeding of local alignment. J. Bioinform. Comput. Biol. **2**(04), 819–842 (2004)
6. Choi, K.P., Zeng, F., Zhang, L.: Good spaced seeds for homology search. In: Proceedings of Fourth IEEE Symposium on Bioinformatics and Bioengineering, BIBE 2004, pp. 379–386. IEEE (2004)
7. Human Microbiome Project Consortium: A framework for human microbiome research. Nature **486**(7402), 215–221 (2012)
8. Felczykowska, A., Bloch, S.K., Nejman-Falenczyk, B., Baranska, S.: Metagenomic approach in the investigation of new bioactive compounds in the marine environment. Acta Biochim. Pol. **59**, 501–505 (2012)
9. Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C.: MEGAN analysis of metagenomic data. Genome Res. **17**(3), 377–386 (2007)
10. Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J., Chinwalla, A., et al.: Structure, function and diversity of the healthy human microbiome. Nature **486**(7402), 207–214 (2012)
11. Ilie, L., Ilie, S.: Multiple spaced seeds for homology search. Bioinformatics **23**(22), 2969–2977 (2007)
12. Ilie, L., Ilie, S., Bigvand, A.M.: Speed: fast computation of sensitive spaced seeds. Bioinformatics **27**(17), 2433–2434 (2011)
13. Li, M., Ma, B., Kisman, D., Tromp, J.: Patternhunter ii: highly sensitive and fast homology search. J. Bioinform. Comput. Biol. **2**(03), 417–439 (2004)
14. Li, M., Ma, B., Zhang, L.: Superiority and complexity of the spaced seeds. In: Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm. Society for Industrial and Applied Mathematics, pp. 444–453 (2006)
15. Lindgreen, S., Adair, K.L., Gardner, P.: An Evaluation of the Accuracy and Speed of Metagenome Analysis Tools. Cold Spring Harbor Laboratory Press (2015). doi:10.1101/017830
16. Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., Pop, M.: Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics **12**(Suppl 2), S4 (2011)
17. Ma, B., Tromp, J., Li, M.: Patternhunter: faster and more sensitive homology search. Bioinformatics **18**(3), 440–445 (2002)
18. Mueller, R.S., Bryson, S., Kieft, B., Li, Z., Pett-Ridge, J., Chavez, F., Hettich, R.L., Pan, C., Mayali, X.: Metagenome sequencing of a coastal marine microbial community from Monterey Bay, California. Genome Announc. **3**(2), e00341-15 (2015)
19. Ounit, R., Wanamaker, S., Close, T.J., Lonardi, S.: Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics **16**(1), 236 (2015)
20. Pace, N.R.: Mapping the tree of life: progress and prospects. Microbiol. Mol. Biol. Rev. **73**(4), 565–576 (2009)
21. Rosen, G.L., Reichenberger, E.R., Rosenfeld, A.M.: NBC: the naive bayes classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics **27**(1), 127–129 (2011)
22. Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., Huttenhower, C.: Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods **9**(8), 811–814 (2012)

23. Sunagawa, S., Mende, D.R., Zeller, G., Izquierdo-Carrasco, F., Berger, S.A., Kultima, J.R., Coelho, L.P., Arumugam, M., Tap, J., Nielsen, H.B., et al.: Metagenomic species profiling using universal phylogenetic marker genes. Nat. Methods **10**(12), 1196–1199 (2013)
24. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., et al.: Environmental genome shotgun sequencing of the Sargasso Sea. Science **304**(5667), 66–74 (2004)
25. Wood, D., Salzberg, S.: Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. **15**(3), R46 (2014)
26. Zhang, Z., Schwartz, S., Wagner, L., Miller, W.: A greedy algorithm for aligning DNA sequences. J. Comput. Biol. **7**(1–2), 203–214 (2000)