BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 6. 614–625
Advance Access published on 18 April 2011
doi:10.1093/bib/bbr019

Computational challenges of sequence classification in microbiomic data

Paolo Ribeca and Gabriel Valiente

Submitted: 12th December 2010; Received (in revised form): 15th March 2011

Abstract

Next-generation sequencing technologies have opened up an unprecedented opportunity for microbiology by enabling the culture-independent genetic study of complex microbial communities, which were so far largely unknown. The analysis of metagenomic data is challenging: potentially, one is faced with a sample containing a mixture of many different bacterial species, whose genome has not necessarily been sequenced beforehand. In the simpler case of the analysis of 16S ribosomal RNA metagenomic data, for which databases of reference sequences are known, we survey the computational challenges to be solved in order to be able to characterize and quantify a sample. In particular, we examine two aspects: how the necessary adoption of new tools geared towards high-throughput analysis impacts the quality of the results, and how good is the performance of various established methods to assign sequence reads to microbial species, with and without taking taxonomic information into account.

Keywords: *metagenomics; next-generation sequencing*

INTRODUCTION

To solve the core problem of metagenomics is to determine and quantify the species composition of a sample containing material from a mixture of different (and possibly previously unknown) microorganisms [1]. Direct sequencing techniques are fundamental to this end, since they give access to microbial species that otherwise could not be isolated and grown in the laboratory.

Although some pioneering studies in metagenomics date back to several years ago [2], the recent advent of high-throughput sequencing (HTS) has paved the way to a variety of projects trying to capture the diversity of microbial ecosystems [3, 4, 5, 6, 7]. In particular, cheap sequencing coupled to large-scale analysis techniques promises to play a major role in the exploration and understanding of multi-organism interactions relevant to human health, which remain so far mostly unknown [8]. It is not hard to envision for the close future scenarios where the bacterial composition of an individual's gastrointestinal tract is constantly monitored in clinical trials, or environmental samples from public places are periodically checked for the emergence of new pathogens. Applications to functional metagenomics, in particular finding genes encoding for novel biocatalysts and drugs [9] in environmental samples from aquatic, soil, animal and plant habitats, also promise to have an enormous impact on biotechnology [10].

However, metagenomics stands out by itself as a daunting challenge: in addition to the usual

Corresponding author. Gabriel Valiente, Complexity and Formal Methods Research Group, Technical University of Catalonia, E-08014 Barcelona, Spain. Tel: +34 934 134 017; Fax: +34 934 137 833; E-mail: valiente@lsi.upc.edu

Paolo Ribeca is leader of the Algorithm Development Group at the Spanish National Center for Genomic Analysis (CNAG) in Barcelona, Spain. His research interests focus on the application of various disciplines of computer science, physics and mathematics to high-performance scientific computing. Since the inception of high-throughput sequencing techniques, he has specialized on algorithms for short-read processing.

Gabriel Valiente is an accredited Full Professor at the Department of Software and a member of the Algorithms, Bioinformatics, Complexity and Formal Methods Research Group of the Technical University of Catalonia in Barcelona, Spain. He is also a member of the Computational Biology and Bioinformatics Research Group at the University of the Balearic Islands and a permanent visiting professor at the Centre for Genomic Regulation in the Barcelona Biomedical Research Park. His research interests fall in the general areas of computational and systems biology, with emphasis on algorithms in bioinformatics and mathematical models in computational and systems biology.

© The Author 2011. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/article/12/6/614/222008 by guest on 13 March 2024

difficulties inherited from genome sequencing (in particular those coming from the short read length accessible to present-day machines), many other problems need to be solved if a precise characterization of a metagenomic sample is desired.

To keep things simple and due to lack of space, in this article we will remove from the picture an important source of uncertainty: we will assume that (a good approximation of) the genomic reference for the biological sample has already been established, thus essentially reducing the metagenomic problem to the precise quantitative determination of the composition of a mixture of species already known in advance. In particular, we will not examine the complications that arise when an unknown number of unknown genomic components need to be assembled out of HTS data.

In general, the sources of uncertainty contributing to the problem of metagenomics in an HTS setup can be modeled after the successive stages needed to analyze a typical dataset.

(i) Two metagenomic samples can present a very high variability, showing wildly varying complexities (number of species present in the mixture) and distributions of relative abundances (one or more dominant species); furthermore, many extreme possibilities are all represented by corresponding realistic biological situations.
(ii) The sequencing technology employed influences the species accessibility and resolution that can be achieved in a study (different read lengths and yield lead to different discriminative power in the detection of both the number of species and their relative abundances).
(iii) The way we understand and model taxonomies (multiple alignments, trees, etc.) constrains the extent up to which we are able to distinguish samples of genomes coming from more or less related species.
(iv) The way we assign HTS reads to the taxonomies (by using different alignment algorithms, *k*-mer analysis, Bayesian models and so on) defines complex trade-offs between the sensitivity of the method, the amount of data produced, and the effectiveness of the assignment. In particular, due to the high yield offered by modern sequencing technologies, some methods established in the field no longer offer adequate performance.

In the rest of this article, we will try to survey how, in our understanding, the aforementioned sources of uncertainty contribute to make the chosen metagenomic setup a difficult problem. In particular, we will consider each stage separately, examining various alternative computational methods, and critically assessing the respective performance and limitations.

SIMULATING A METAGENOMIC SAMPLE

The signal obtained when sequencing metagenomic data can be extremely complex. Due to this reason, the availability of standardized simulated data as a starting point is fundamental, especially when trying to evaluate the influence of different analysis pipelines on the quality of the results.

Sequence reads should be simulated in a way that reflects the diverse taxonomic composition of a metagenomic dataset. The scientific community has been aware of this problem since several years. One of the first in-silico metagenomic datasets [11] was designed with the goal of simulating microbial communities of varying complexity: low-complexity communities with one dominant population, medium-complexity communities with more than one dominant population flanked by low-abundance populations and high-complexity communities with no dominant population.

Subsequently a more general tool, MetaSim [12], was developed. It allows not only to choose a metagenomic scenario, but also to take the sequencing biases introduced by different technologies into account: due to such capabilities, it appears to be an adequate tool for the scope of this study. In particular, in order to simulate a metagenomic dataset with MetaSim, the number of genomes present at each level of the NCBI taxonomy [13], the sequencer error model and the read length distribution have to be specified.

General considerations

As previously mentioned, it is possible to envisage many different experimental biological setups (targeted or random sequencing, one or many dominant species) giving rise to different categories of metagenomic datasets, each one showing a different degree of redundancy. To keep things simple, we will focus in this article on the problem of distinguishing and quantifying the relative abundance of sequences all

Downloaded from https://academic.oup.com/bib/article/12/6/614/222008 by guest on 13 March 2024

coming from the ribosomal 16S subunit of bacterial species [1]. Although this is a classical problem in metagenomics, we will examine it in the new context of HTS techniques, which impose new analysis protocols.

Due to historical reasons in the development of HTS techniques, most of the metagenomic 16S ribosomal databases available today have been generated using Roche/454 technology [14]. For several years, this platform has been capable of providing long reads (200–500 nt, with the perspective of becoming 1000 nt in the near future) at the price of many insertions/deletions, particularly when long homopolymeric stretches are present in the nucleic acid being sequenced. This has motivated the development of sequencing error correction methods for Roche/454 data [15, 16]. The yield is moderate.

In recent years, however, other technologies have been introduced. Among them, the sequencing-by-synthesis by Illumina/Solexa, which produces shorter reads (36 nt as of 2008, 150 nt as of today) but offers a very high fidelity (notably, a very low rate of insertions/deletions) and much higher yields. In general, longer reads are more effective in highlighting differences between closely related sequences; however, the read length provided today by Illumina/Solexa sequencing is enough to achieve a good resolution, with the additional advantage of a much lower cost. These considerations suggest that technologies like this one will become more and more relevant to metagenomics in the near future.

Such considerations explain why we decided to simulate both Roche/454 and Illumina/Solexa reads for this survey.

Generating Roche/454 and Illumina/Solexa datasets with MetaSim

We took a reference bacterial taxonomy of 5165 near-full-length cultures of high quality obtained from the TOBA database [17, 18, 19] with a uniform scheme of seven taxonomic ranks (domain, phylum, class, order, family, genus, species) [18]. These 16S ribosomal RNA sequences range from 1202 to 1780 nt and cover the whole spectrum of known bacteria, the dominant phyla being Proteobacteria, Actinobacteria, Firmicutes, Bacteroidetes and Tenericutes, with 1925, 1285, 1178, 355 and 160 species, respectively. Despite their high quality, 2169 of the 5165 reference sequences have ambiguous base calls.

We then sampled metagenomic reads from the reference sequences with MetaSim, using empirical error models for both Roche/454 (pyrosequencing) and Illumina/Solexa (sequencing-by-synthesis) technologies. Our four datasets contain Roche/454 reads of variable length and 100 and 250 nt on average, and Illumina/Solexa reads of 50 and 75 nt, with ∼100 000 reads for each error model and length.

The size of our controlled experiment is very small if compared to the typical yield provided by a modern HTS machine, and thus admittedly unrealistic—for instance, one single lane of Illumina/Solexa technology currently contains ∼20 million reads. However, datasets of the size considered in this article still allow for precise statistical conclusions to be drawn; they also make possible accurate comparisons at the single-read level, which would otherwise be very difficult due to the excessive amount of data.

It should be noted that after running MetaSim it was necessary to filter its results due to the presence of various simulation errors, which were producing results not completely consistent with what one would expect from existing technologies. In particular, we have filtered out sampled Roche/454 reads which were too short (that is, for the two datasets of 100 and 250 nt, reads having length at most 20 and 50 nt, respectively); as well, we have excluded simulated Illumina/Solexa reads not corresponding to a full-length match at the given position in the reference sequence. We also mention that our Illumina/Solexa datasets of length 50 and 75 have been obtained by truncation from datasets of 62 and 80 nt, respectively. In fact, the latter ones were the hard-coded read lengths provided by MetaSim, but they do not correspond to the most typical values seen when using Illumina/Solexa instruments. The resulting total number of reads and the average read lengths for each empirical error model are shown in Table 1.

In addition, we should mention that at the moment MetaSim does not offer the possibility of simulating paired-end reads, which is why we had to limit our study to single-end reads. In a similar way, the absence of simulated quality scores kept us from studying the influence of sequencing qualities on the accuracy of analysis protocols for metagenomic data.

Downloaded from https://academic.oup.com/bib/article/12/6/614/222008 by guest on 13 March 2024

Table 1: Features of the simulated datasets

Simulation run	Total reads	Mean length	Diversity	Richness	Taxonomic diversity
454–100	99 784	107.021	8.5200	5148	11.0149
454–250	99 387	264.767	8.5190	5148	11.0146
Solexa-50	95 874	50.000	8.5191	5148	11.0168
Solexa-75	94 567	75.000	8.5186	5148	11.0172

Evaluation methods

Several indicators have been proposed and used in the literature to quantify variability across metagenomic samples.

Together with the number of species in a sample [20], one of the simplest and most widely used measures of species diversity in microbial ecology is the Shannon–Wiener index [21, 22], which measures the information entropy of the distribution of a sample taken from a population of species. Along with the Clarke–Warwick index [23], which measures the average distance in a taxonomic reference between the sampled species, other widely accepted notions are those of α-diversity (species diversity within an ecosystem) and β-diversity (change in species diversity between two ecosystems) [20].

In our controlled experiment, given each simulated read we know exactly which is the sequence in the reference originating it: owing to this fact, for our artificial datasets we can directly compute the correct values of any indicator. The exact values of diversity (Shannon–Wiener index), richness (total number of species) and taxonomic diversity (Clarke–Warwick index) for each empirical error model and read length are also shown in Table 1. Ideally, the various read mapping and assignment methods discussed below should reproduce precisely such values: this requirement will be the core of our evaluation strategy.

MAPPING SEQUENCE READS

As explained in the previous section, our study will be focused on the analysis of an HTS sample of mixed ribosomal RNA 16S subunits of bacterial species. In such a situation, the first computational challenge arises when trying to align (*map*, in HTS parlance) a very large number of sequence reads to the metagenomic reference—which is, in this case, a database of 16S ribosomal RNA sequences coming from a large set of different organisms.

The high yield of HTS technologies produces impressive amounts of data. It is nowadays common to obtain several hundred million of reads during a single experiment: those figures require extremely efficient alignment programs. In fact, typical HTS mappers are able to align tens of million of reads per hour on a single-core processor. As a matter of fact, such a high-performance requirement rules out traditional alignment programs like BLAST [24], which are slower by about three orders of magnitude.

Speed, however, comes at the price of accuracy. While programs like BLAST are easily able to highlight big differences between the HTS read and the genome it is being aligned to, typical mappers usually explore a much more limited number of possibilities. For instance, by default BWA [25] would allow only up to six mismatches for a read of 150 nt, corresponding to 96% of sequence similarity. Another problem is even more relevant. Since most HTS analysis protocols discard ambiguous reads potentially coming from more than one single location in the genome, HTS mappers are usually geared towards reporting one single 'best' match, and not towards performing exhaustive alignment. In other words, either due to their algorithmic limitations they are unable to report all of the—possibly many—matches found in a reference, or they have as a default behavior that of reporting one match, becoming much slower if asked to find them all.

In general, it is an open question—which is neither clearly stated nor clearly answered in the literature—whether modern mappers are adequate to address metagenomic problems, in particular when the reference is composed by many sequences very close to each other. In the latter case, exhaustive mapping would seem to become paramount, since given only one match it is impossible to say whether the read has many hits in the database (making its attribution in terms of a taxonomic tree ambiguous) or it actually corresponds to a single species.

Downloaded from https://academic.oup.com/bib/article/12/6/614/222008 by guest on 13 March 2024

In this section, we try to assess and quantify this problem. In particular, we compare the results obtained when analyzing our simulated Roche/454 and Illumina/Solexa datasets with different HTS mappers.

Mapping Roche/454 reads

For the mapping of simulated Roche/454 reads (which following MetaSim's sequencing error models—as explained in 'Simulating a Metagenomic Sample'—contain no mismatch and up to 22 indels, although over 97% of the reads have at most 13 indels; see Table 2) we have compared two programs suitable for the alignment of long reads, BWA/SW [26] and BLAT [27]. BWA/SW is a tool based on the Burrows–Wheeler transform (BWT) employing a heuristic Smith-Waterman algorithm specifically adapted to long HTS reads, while BLAT is a BLAST-like generic alignment tool. Although BLAT is not able to do exhaustive mapping, it usually shows both good sensitivity and a reasonable speed, hence appearing to be a good standard for our comparison (BLAST is typically orders of magnitude slower than BLAT, and we had to rule it out for this study).

Some important differences should be noted. Due to its algorithmic design, BWA/SW can only report one hit, thus producing a relatively small output (∼20 MB and ∼40 MB for the 100 and 250-nt datasets, respectively). On the other hand, BLAT finds thousands of hits per read, most of them being partial—that is, such that not all the bases in the query have been successfully aligned: as a consequence, BLAT produces a comparatively much bigger output (of the order of 10 GB for both datasets) and is comparatively slower. To be able to perform a meaningful assessment of the results we have then filtered the BLAT output, keeping only the hits belonging to the best stratum—that is, the ones having the smallest number of edit operations. The size of the filtered BLAT output is similar to that of BWA/SW (Table 3). Apart from these choices, we ran both tools with their default settings: as both programs—and BLAT in particular—accept many parameters that define complicated speed-versus-accuracy trade-offs, we felt that fine-tuning them was outside the scope of this article.

Table 2: Number of simulated reads classified in terms of their number *k* of mismatches (for Illumina/Solexa reads) and indels (for Roche/454 reads)

k	454–100	454–250	Solexa-50		Solexa-75	
0	5255	237	44 287	(41 416)	48 861	(44 752)
1	15 206	946	34 463	(34 109)	32 540	(32 035)
2	22 528	2271	13 170	(14 156)	10 493	(12 059)
3	22 363	4636	3292	(4326)	2254	(3473)
4	16 634	7776	577	(1128)	367	(1149)
5	9737	11 184	77	(363)	44	(481)
6	4834	13 305	8	(124)	8	(217)
7	2073	13 994		(66)		(108)
8	813	13 180		(42)		(80)
9	252	10 851		(23)		(38)
10	67	8129		(24)		(26)
11	17	5531		(17)		(25)
12	5	3349		(11)		(10)
13		1990		(8)		(19)
14		1051		(6)		(13)
15		518		(6)		(6)
16		258		(2)		(8)
17		102		(3)		(5)
18		42		(3)		
19		24		(3)		(6)
20		8		(3)		(4)
21		4		(2)		(1)
22		1		(1)		
≥23				(32)		(52)

In parentheses: numbers obtained classifying the reads when the IUPAC ambiguities coming from the sequences in the reference and present in the reads due to a bug in MetaSim (see text) are also taken into account.

Comparing BWA/SW and BLAT

It is interesting to note that on one hand, the HTS-tuned BWA/SW algorithm is much faster and less space-consuming than that of BLAT; on the other hand, the sensitivity shown by BLAT (both tools taken with their default parameter values) is always higher for the read lengths considered (74 versus 39%—almost twice as much—of correctly recovered hits in the 100-nt dataset, 62 versus 53% in the 250-nt dataset; Table 3).

Table 3: Mapping simulated Roche/454 reads with BWA/SW and BLAT

Dataset	Tool	Output size	All reads	
			Mapped	Unmapped
454–100	BWA/SW	18 M	35 999 (38.57%)	57 329
	BLAT	11 G/13 M	69 189 (74.14%)	24 139
454–250	BWA/SW	39 M	54 050 (52.73%)	48 444
	BLAT	14 G/4 M	63 094 (61.56%)	39 400

Numbers in parentheses indicate sensitivity: the number of reads for which the simulated read has been correctly aligned to the originating sequence in the database.

Downloaded from https://academic.oup.com/bib/article/12/6/614/222008 by guest on 13 March 2024

Another very important point is that from the output of BWA/SW it is not possible to estimate the redundancy (number of equivalent matches) of each read, since only one hit is returned by the aligner. In addition, any further computation in order to precisely assign the read to some branch of the taxonomic tree becomes problematic, since equivalent matches are not listed.

As a final remark, it should be emphasized that in spite of its better sensitivity, it is unclear to us if BLAT could really be used in a production environment: the daunting size of the produced output and the much longer running times it implies would probably prevent BLAT from providing a yield adequate for a realistic metagenomic HTS setup.

Mapping Illumina/Solexa reads

For the mapping of simulated Illumina/Solexa reads (which following MetaSim's sequencing error models—as explained in 'Simulating a Metagenomic Sample'—contain up to eight mismatches and no indel, although over 97% of the reads have at most four mismatches; see Table 2) we have compared two programs suitable for the alignment of short reads, BWA [25] and GEM [28]. Both are tools based on the BWT. However, one relevant difference is that insofar as only nucleotide substitutions and no indels are involved, GEM implements an exhaustive search algorithm: all the existing matches in the reference up to the specified number of substitutions are always counted, and the user can specify how many of them should be output.

Following BWA's default policy, we looked for matches in the reference having at least 96% of sequence similarity. Hence we ran both programs with at most two allowed substitutions in the case of the 50-nt dataset, and at most three substitutions in the case of the 75-nt dataset. In addition, to make the results fully comparable and ready for subsequent taxonomic assignment, we asked both programs to report all matches found. Such a choice actually corresponds to all existing matches within the specified edit distance in the case of GEM, and to a subset of the latter in the case of BWA, selected by BWA following a complicated and difficult-to-describe criterion of 'equivalent best alignment' proprietary to the BWA algorithm itself.

As illustrated in 'Simulating a Metagenomic Sample', our simulated Illumina/Solexa reads have no indels. In consequence, the subset of reads alignable with at most two substitutions in the 50-nt dataset (and that of reads alignable with at most three substitutions in the 75-nt dataset) can be considered a gold standard to quantify exhaustiveness: GEM is bound to find all the existing matches for each of the reads belonging to it.

Apart from the aforementioned choices for the number of mismatches, both programs were run with default parameters. Notably, it was necessary to filter out some non-existent mappings that BWA had incorrectly reported close to sequence boundaries in the genomic reference.

Comparing BWA and GEM

The sensitivities of the two programs appear to be similar (respectively 93 and 94% for BWA and GEM in the case of the 50-nt dataset, and 97% for BWA and 98% for GEM in the case of the 75-nt dataset; Table 4), GEM performing slightly better than BWA when considering the complete datasets. As one would expect, in particular, GEM is able to correctly recover all the hits if one considers the subsets of alignable artificial reads with at most two substitutions in the case of the 50-nt dataset, and at most

Table 4: Mapping simulated Illumina/Solexa reads with BWA and GEM

Dataset	Tool	Output size	All reads		Alignable reads	
			Mapped	Unmapped	Mapped	Unmapped
Solexa-50	BWA	231 M	89 421 (93.27%)	6453	89 416 (99.70%)	265
	GEM	511 M	89 681 (93.54%)	6193	89 681 (100.00%)	0
Solexa-75	BWA	87 M	92 042 (97.33%)	2525	92 039 (97.33%)	280
	GEM	284 M	92 319 (97.62%)	2248	92 319 (100.00%)	0

Numbers in parentheses indicate sensitivity: the fraction of reads for which the simulated read has been correctly aligned to the originating sequence in the database. Alignable reads are the ones having an edit distance from the original sequence of at most two substitutions in the case of the 50-nt dataset, and at most three substitutions in the case of the 75-nt dataset.

Downloaded from https://academic.oup.com/bib/article/12/6/614/222008 by guest on 13 March 2024

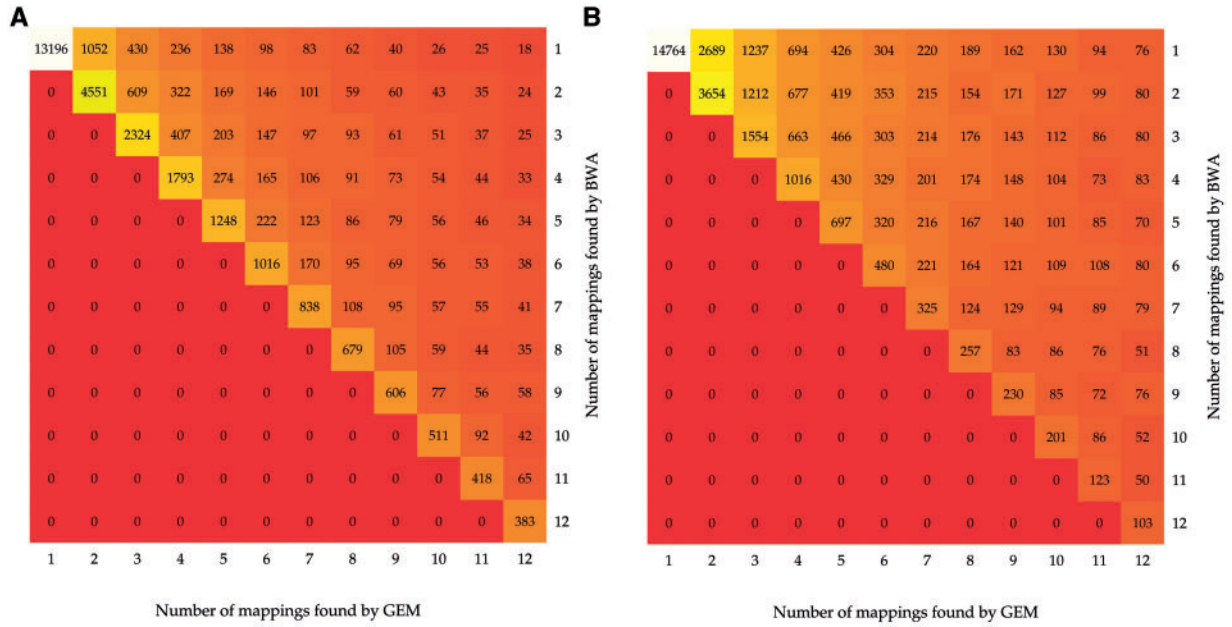

Figure 1: Fidelity of BWA and GEM mappers measured on 50-nt reads aligned with at most two mismatches (**A**) and 75-nt reads aligned with at most three mismatches (**B**). The heatmap shows how reads can be classified in terms of the number of times they can be aligned to the reference with GEM (on the *x*-axis) and with BWA (on the *y*-axis); to keep the figure readable, only the distribution for the first twelve matches is shown (the complete range of matches found is between 1 and ∼4500 for the 50-nt dataset and between 1 and ∼2900 for the 75-nt dataset).

three substitutions in the case of the 75-nt dataset (see the last column of Table 4).

However, the ability of recovering the genomic location that originates the read is not the only important one as far as metagenomics is concerned. As mentioned before, it is also paramount to be able to correctly estimate and output the exact number of hits, in order to make the successive attribution of the read to the correct part of the taxonomy tree as precise as possible. In Figure 1, we quantitatively examine this aspect: the contingency tables show the alignable reads separated in groups after their different classification in terms of number of matches found by BWA (*y*-axis) and GEM (*x*-axis). The reads on the diagonal have been attributed the same number of matches by both aligners; for the off-diagonal reads in the upper triangle BWA, being not an exhaustive mapper, incorrectly found less matches than it should have. In particular, the figures clearly show how BWA misclassifies a relevant fraction of reads (25% for the 50-nt dataset and 30% for the 75-nt dataset) as uniquely mapping when they are not (see the first row of each panel). In addition, reads that map into many locations are often classified by BWA as having a much lower number of hits (see for instance the rightmost columns in each panel).

We believe that this result clearly illustrates how large the systematic bias produced by non-exhaustive mapping schemes can be when trying to assign a read to a database of redundant sequences, as in the metagenomic example considered in this article.

ASSIGNMENT OF SEQUENCE READS

After sequence reads have been mapped to a reference, one has to assign each of them to species (albeit some integrated methods also exist which do not require a preliminary alignment to be performed externally). This is the second computationally challenging step when analyzing 16S ribosomal RNA data. The assignment can be done in two possible ways.

(i) Non-taxonomic assignment. Possibly using knowledge of the genomic reference for the dataset, but assuming no taxonomic reference for the metagenomic dataset. When it is known, the genomic reference serves as a template for a multiple alignment of the sequence

Downloaded from https://academic.oup.com/bib/article/12/6/614/222008 by guest on 13 March 2024

reads; the alignment in turn defines pairwise similarities that can be used to group reads into clusters of related species. In the absence of a genomic reference, pairwise similarities among reads can still be used to group them into clusters of related species, although the genomic reference allows for a better assessment of their similarity.

(ii) Taxonomic assignment. Reads are attributed to species at the closest possible taxonomic rank, using both mapping information and knowledge of the genomic and taxonomic reference for the metagenomic dataset. Ambiguities may arise while mapping sequence reads to the genomic reference, when a single read is mapped to more than one reference sequence. These ambiguities are usually solved by assigning ambiguous reads to either the consensus (lowest common ancestor) of all matching sequences in the taxonomic reference [29], or to a sequence in the taxonomic reference that provides optimal sensitivity and specificity [30, 31].

In the rest of this section, we examine the performance of various methods for non-taxonomic assignment (MOTHUR/non-taxonomic and QIIME) as well as some methods for taxonomic assignment (RDP classifier, MOTHUR/taxonomic and TANGO).

Non-taxonomic assignment

MOTHUR/non-taxonomic

To perform the non-taxonomic assignment of our simulated metagenomic dataset, making use of a genomic reference of 4938 near-full-length, aligned 16S ribosomal RNA sequences of high quality obtained from the SILVA database [32], we used MOTHUR [33] version 1.14.0. The multiple alignment of the sequence reads was performed with MOTHUR's default parameter values, which correspond to *k*-mer searching with 8-mers [34] followed by pairwise alignment [35] with a reward of 1 for a match and penalties of 1 and 2 for a mismatch and a gap, respectively.

Usually, 16S ribosomal RNA gene sequences are grouped into operational taxonomic units with a sequence identity threshold of 97%: this produces phylotypes that are reasonably close to species [36, 37, 38]. We thus specified a cutoff value of 0.03 to cluster the sequence reads with MOTHUR.

The multiple alignment was then used to group the reads into clusters of related species, again using MOTHUR's default parameter values. The resulting values of diversity (Shannon–Wiener index) and richness (total number of species) for each of the clustering methods available in MOTHUR (complete linkage or nearest neighbor, single linkage or furthest neighbor, and UPGMA or average neighbor) are shown in Table 5. Comparison with the exact values in Table 1 shows that non-taxonomic assignment of reads results in an overestimation of both diversity and richness of the metagenomic dataset.

QIIME

For the non-taxonomic assignment of our simulated metagenomic dataset, we also used QIIME (Quantitative Insights into Microbial Ecology) version 1.2.0 [39], which does not require any genomic reference. In particular, we selected the CD-HIT [40] method of QIIME, where the sequence reads are grouped into operational taxonomic units at different sequence identity thresholds. The default sequence identity threshold of 97% produces phylotypes that are close to species; in addition to this, we have also used a threshold of 94%, which produces phylotypes that are close to genus, and a threshold of 90%, which produces phylotypes that are close to family [41]. The sequence reads are pairwise aligned unless a statistical analysis of their *k*-mer

Table 5: Non-taxonomic assignment with MOTHUR

Dataset	Complete linkage (Nearest neighbor)		Single linkage (Furthest neighbor)		UPGMA (Average neighbor)	
	D	R	D	R	D	R
454–100	10.9427	56 542	10.9692	58 058	11.5127	99 974
454–250	11.4562	94 487	11.4563	94 494	11.5129	99 996
Solexa-50	11.4718	95 973	11.4719	95 976	11.5129	100 000
Solexa-75	11.1665	70 723	11.1669	70 748	11.3920	88 612

Diversity (D) and richness (R) of the datasets for various clustering methods (nearest, furthest and average neighbor).

Downloaded from https://academic.oup.com/bib/article/12/6/614/222008 by guest on 13 March 2024

Table 6: Non-taxonomic assignment with QIIME and CD-HIT

Dataset	97% (species)		94% (genus)		90% (family)	
	D	R	D	R	D	R
454–100	11.4396	95 146	11.2416	83 945	10.7552	62 743
454–250	11.4675	96 642	11.2252	81 916	10.6435	56 121
Solexa-50	11.3237	89 893	10.9031	72 527	10.3348	54 088
Solexa-75	11.2916	87 505	10.9590	72 863	10.3600	52 948

Diversity (D) and richness (R) of the datasets for various sequence similarity threshold values.

Table 7: Taxonomic assignment with RDP classifier

Dataset	T = 100%		T = 90%		T = 80%		T = 70%	
	D	R	D	R	D	R	D	R
454–100	5.0099	1449	4.4206	1422	3.4922	1373	1.5970	952
454–250	6.2334	1444	6.0476	1426	5.7182	1407	4.1619	1357
Solexa-50	3.9299	1375	3.2198	1302	2.3419	1146	1.2636	534
Solexa-75	5.2719	1446	4.8026	1406	4.0484	1341	2.1618	970

Diversity (D) and richness (R) of the datasets as a function of the confidence threshold (T).

frequencies shows that their sequence identity falls below the threshold value; in the latter case, the longest read in a cluster becomes the representative of the operational taxonomic unit.

The resulting values of diversity (Shannon–Wiener index) and richness (total number of species) for a sequence identity threshold of 97% (species), 94% (genus) and 90% (family) are shown in Table 6. Again, comparison with the exact values in Table 1 shows that non-taxonomic assignment of sequence reads results in an overestimation of both diversity and richness of the metagenomic dataset.

Taxonomic assignment

RDP classifier

We used the RDP (Ribosomal Database Project) Classifier version 2.2 [42] for the taxonomic assignment of the simulated datasets, with the genomic reference of 5165 near-full-length type cultures of high quality described in 'Generating Roche/454 and Illumina/Solexa Datasets with MetaSim' and default parameter values. The RDP Classifier reported the top seven hits for each read in the simulated dataset, ranked by confidence estimate.

We then extracted the best hit (having the lowest taxonomic rank) for each read, given a fixed confidence threshold. The resulting values of diversity (Shannon-Wiener index) and richness (total number of species) for a confidence threshold between 70 and 100% are shown in Table 7. Comparison with the exact values in Table 1 shows that taxonomic assignment of sequence reads results in an underestimation of both diversity and richness of the metagenomic dataset.

MOTHUR/taxonomic

For the taxonomic assignment of the simulated dataset we also used MOTHUR version 1.14.0 [33], with the genomic reference of 4938 near-full-length, aligned 16S ribosomal RNA sequences of high quality obtained from the SILVA database [32] and default parameter values, which correspond to k-mer searching, where we have chosen $6 \leq k \leq 10$. MOTHUR reported the top hit for each read in the simulated dataset.

The resulting values of diversity (Shannon–Wiener index) and richness (total number of species) for a confidence threshold between 70 and 100% are shown in Table 8. Again, comparison with the exact values in Table 1 shows that taxonomic assignment of sequence reads results in an underestimation of both diversity and richness of the metagenomic dataset.

Downloaded from https://academic.oup.com/bib/article/12/6/614/222008 by guest on 13 March 2024

Table 8: Taxonomic assignment with MOTHUR

Dataset	$k=6$		$k=7$		$k=8$		$k=9$		$k=10$	
	D	R	D	R	D	R	D	R	D	R
454–100	2.8589	635	3.0214	649	3.0566	666	3.0695	653	3.0465	649
454–250	3.3534	653	3.4070	662	3.4428	664	3.4538	678	3.4505	680
Solexa-50	2.4020	626	2.6129	642	2.6448	633	2.6440	637	2.6304	637
Solexa-75	2.7633	634	2.9011	628	2.9195	642	2.9286	643	2.9204	653

Diversity (D) and richness (R) of the datasets as a function of the search method (k-mer with $6 \leq k \leq 10$).

Table 9: Taxonomic assignment with TANGO

Dataset	BWA			GEM (first stratum)			GEM (second stratum)			GEM (first/second stratum)		
	D	R	TD	D	R	TD	D	R	TD	D	R	TD
Solexa-50	6.2114	3,595	8.0441	6.5149	3,739	8.3293	6.0989	3,570	7.6383	6.5189	3,879	8.1438
Solexa-75	6.4927	3,625	8.6793	6.7536	3,623	8.9289	6.4596	3,606	8.4800	6.7753	3,864	8.7930

Diversity (D), richness (R), and taxonomic diversity (TD) of the datasets as a function of the mapping algorithm.

TANGO

Finally, for the taxonomic assignment of the simulated Illumina/Solexa datasets we used TANGO (Taxonomic Assignment in Metagenomics) version 1.2.0 [30, 31], with default parameter values. As input, we employed the mappings obtained for our datasets in 'Comparing BWA/SW and BLAT' using BWA and GEM against the genomic reference of 'Generating Roche/454 and Illumina/Solexa Datasets with MetaSim'. Since the assignment process is sensitive to noise, we filtered the GEM output stratum-wise—that is, among all possible matches we kept only those being within some small edit distance from the best one. In detail, we considered three sets of matches: first stratum, the set of matches having the minimum possible number of mismatches; second stratum, the set of matches having at most one mismatch more than the best match and first/second stratum, an intermediate set (consisting of the set of best matches, plus those having one mismatch more than the best ones, but the latter being included only if they are not more than twice as many as the matches belonging to the first stratum). We did not perform such a filtering for BWA, since it does not find all the matches, and because it already provides the set of alignments that are the best ones after its own algorithmic criterion. TANGO reported the top hit for each read in the simulated dataset.

The values of diversity (Shannon–Wiener index), richness (total number of species) and taxonomic diversity (Clarke–Warwick index) for the various sets of matches considered can be found in Table 9. The best result is obtained using the GEM matches filtered to keep only the first stratum, thus confirming the importance of exhaustive mapping for taxonomic attribution. These values are closer to the exact values in Table 1 than the ones obtained with other taxonomic assignment tools; however, despite the more accurate taxonomic assignment produced by TANGO, they still represent an underestimation of our indicators.

CONCLUSIONS

We have surveyed some of the computational problems that are involved in the analysis of a metagenomic 16S ribosomal RNA dataset, along with the performance of several methods suitable for coping with such a situation.

(i) Mapping of HTS reads to the genomic reference (database of more or less related sequences coming from a set of different species) for the metagenomic dataset. The high yield of HTS technologies requires time and space-efficient alignment programs, and current programs sacrifice accuracy for efficiency, exploring a limited

Downloaded from https://academic.oup.com/bib/article/12/6/614/222008 by guest on 13 March 2024

space of parameters and/or reporting the best matches only. Exhaustive alignment becomes paramount here, but most current mapping programs—typically based either on seeding strategies or on a more or less arbitrary choice of a 'best' alignment out of many possible ones—are not exhaustive. The quantification of this effect for contiguous mapping (no indels allowed) shows that the use of non-exhaustive alignment schemes can lead to a substantial misrepresentation of the correct number of existing matches, which is likely to result in a bias during the subsequent attribution of the read to a database of redundant sequences.

(ii) Assignment of HTS reads to species, using knowledge of the genomic reference but assuming no taxonomic reference (classification tree of the sequences in the genomic reference) for the metagenomic dataset. Pairwise similarities among reads can be used to group them into clusters of related species (operational taxonomic units). Alternatively, the genomic reference serves as a template for a multiple alignment of the reads, which in turn defines pairwise similarities that can also be used to group the reads and assign them to clusters of related species. Both solutions lead to an overestimation of diversity and richness in the metagenomic dataset.

(iii) Assignment of HTS reads to species at the closest possible taxonomic rank, using mapping information and also knowledge of the genomic and taxonomic reference for the metagenomic dataset. The ambiguities that arise when a single read is mapped to more than one sequence in the genomic reference can be solved in two possible ways: by assigning ambiguous reads to either the lowest common ancestor of all matching sequences in the taxonomic reference, or to a sequence in the taxonomic reference that provides optimal sensitivity and specificity. Both solutions lead to an underestimation of diversity and richness in the metagenomic dataset. However, they get closer to the actual diversity and richness values the more accurate the taxonomic assignment is.

This long list of problems is still far from describing all the open computational challenges in metagenomics. In particular, it goes without saying that many of the scenarios we did not consider in this article (one for all, the case of a metagenomic sample for which a genomic reference is not known and should be assembled from the sample itself) will likely require the careful invention of even more sophisticated analysis protocols.

Key Points

- Computational challenges in metagenomics can be surveyed by means of appropriate analysis protocols like the one we have devised for this article.
- Most current tools for the mapping of HTS reads to the genomic reference sacrifice accuracy for efficiency, and non-exhaustive alignment schemes lead to a misrepresentation of the correct number of matches and to a bias in the subsequent attribution of the read to a database of redundant sequences.
- Current tools for the assignment of HTS reads to species, using knowledge of the genomic reference but assuming no taxonomic reference, lead to an overestimation of diversity and richness in the metagenomic dataset.
- Current tools for the assignment of HTS reads to species at the closest possible taxonomic rank, using mapping information and also knowledge of the genomic and taxonomic reference, lead to an underestimation of diversity and richness in the metagenomic dataset, getting closer to the actual values the more accurate the taxonomic assignment is.

Acknowledgements

We would like to thank Simon Heath for carefully reading the manuscript. We also thank the anonymous reviewers for their helpful comments and suggestions, which have lead to a significant improvement of the original article.

References

1. National Research Council. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. Washington, DC: The National Academies Press, 2007.
2. Venter JC, Remington K, Heidelberg JF, *et al*. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004;**304**(5667):66–74.
3. Dinsdale EA, Edwards RA, Hall D, *et al*. Functional metagenomics profiling of nine biomes. *Nature* 2008;**452**(7187): 629–32.
4. Pond SK, Wadhawan S, Chiaromonte F, *et al*. Windshield splatter analysis with the Galaxy metagenomic pipeline. *Genome Res* 2009;**19**(11):2144–53.
5. Sogin ML, Morrison HG, Huber JA, *et al*. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci USA* 2006;**103**(32):12115–20.
6. Turnbaugh PJ, Hamady M, Yatsunenko T, *et al*. A core gut microbiome in obese and lean twins. *Nature* 2009; **457**(7228):480–4.
7. Turnbaugh PJ, Ley RE, Hamady M, *et al*. The Human Microbiome Project. *Nature* 2007;**449**(7164):804–10.
8. Vijay-Kumar M, Aitken JD, Carvalho FA, *et al*. Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science* 2010;**328**(5975):228–31.

Downloaded from https://academic.oup.com/bib/article/12/6/614/222008 by guest on 13 March 2024

9. Simon C, Daniel R. Achievements and new knowledge unraveled by metagenomic approaches. *Appl Microbiol Biotechnol* 2009;**85**(2):265–76.
10. Lorenz P, Eck J. Metagenomics and industrial applications. *Nat Rev Microbiol* 2005;**3**(6):510–6.
11. Mavromatis K, Ivanova N, Barry K, *et al*. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* 2007;**4**(6):495–500.
12. Richter DC, Ott F, Auch AF, *et al*. MetaSim: A sequencing simulator for genomics and metagenomics. *PLoS ONE* 2009;**3**(10):e3373.
13. Wheeler DJ, Chappey C, Lash AE, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2000;**28**(1):10–4.
14. Rothberg JM, Leamon JH. The development and impact of 454 sequencing. *Nature Biotechnol* 2008;**26**(10):1117–24.
15. Quince C, Lanzén A, Curtis TP, *et al*. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 2009;**6**(9):639–41.
16. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 2011;**12**:38.
17. Cole JR, Wang Q, Cardenas E, *et al*. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* 2009;**37**(D):141–5.
18. Garrity GM, Lilburn TG, Cole JR, *et al*. *The Taxonomic Outline of Bacteria and Archaea. TOBA release 7.7*. Michigan State University Board of Trustees. http://www.taxonomicoutline.org/, 2007 (29 March 2011, date last accessed).
19. Lilburn TG, Harrison SH, Cole JR, *et al*. Computational aspects of systematic biology. *Brief Bioinform* 2006;**7**(2):186–95.
20. Whittaker RH. Evolution and measurement of species diversity. *Taxon* 1972;**21**(2–3):213–51.
21. Begon M, Harper JL, Townsend CR. *Ecology: Individuals, Populations and Communities*. 2nd edn. Oxford: Blackwell Science, 1996.
22. Krebs CJ. *Ecological Methodology*. 2nd edn. Menlo Park, CA: Benjamin Cummings, 1998.
23. Clarke KR, Warwick RM. A taxonomic distinctness index and its statistical properties. *J Appl Ecol* 1998;**35**(4):523–31.
24. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**(3):403–10.
25. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**(14):1754–60.
26. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;**26**(5):589–95.
27. Kent WJ. BLAT: the BLAST-like alignment tool. *Genome Res* 2002;**12**(4):656–64.
28. Ribeca P. GEM: GEnomic Multi-tool. http://gemlibrary.sourceforge.net/, 2009 (29 March, 2011 date last accessed).
29. Huson DH, Auch AF, Qi J, *et al*. MEGAN analysis of metagenomic data. *Genome Res* 2007;**17**(3):377–86.
30. Clemente JC, Jansson J, Valiente G. Accurate taxonomic assignment of short pyrosequencing reads. In: *Proc 15th Pacific Symp. Biocomputing* 2010;**15**:3–9.
31. Clemente JC, Jansson J, Valiente G. Flexible taxonomic assignment of ambiguous sequencing reads. *BMC Bioinformatics* 2011;**12**:8.
32. Pruesse E, Quast C, Knittel K, *et al*. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 2007;**35**(21):7188–96.
33. Schloss PD, Westcott SL, Ryabin T, *et al*. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009;**75**(23):7537–41.
34. Vinga S, Almeida J. Alignment-free sequence comparison: A review. *Bioinformatics* 2003;**19**(4):513–23.
35. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970;**48**(3):443–53.
36. Lozupone CA, Knight R. Species divergence and the measurement of microbial diversity. *FEMS Microbiol Rev* 2008;**32**(4):557–8.
37. Martin AP. Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl Environ Microbiol* 2002;**68**(8):3673–82.
38. Quince C, Curtis TP, Sloan WT. The rational exploration of microbial diversity. *The ISME J* 2008;**2**(10):997–1006.
39. Caporaso JG, Kuczynski J, Stombaugh J, *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010;**7**:335–6.
40. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**(13):1658–9.
41. Vandamme P, Pot B, Gillis M, de Vos P, Kersters K, Swings J. Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev* 1996;**60**(2):407–38.
42. Wang Q, Garrity GM, Tiedje JM, *et al*. Naïve bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 2007;**73**(16):5261–7.

Downloaded from https://academic.oup.com/bib/article/12/6/614/222008 by guest on 13 March 2024