

Machine learning for metagenomics: methods and tools

Hayssam Soueidan ^{*} Macha Nikolski [†]

March 3, 2022

Abstract

Owing to the complexity and variability of metagenomic studies, modern machine learning approaches have seen increased usage to answer a variety of question encompassing the full range of metagenomic NGS data analysis. We review here the contribution of machine learning techniques for the field of metagenomics, by presenting known successful approaches in a unified framework. This review focuses on five important metagenomic problems: OTU-clustering, binning, taxonomic profiling and assignment, comparative metagenomics and gene prediction. For each of these problems, we identify the most prominent methods, summarize the machine learning approaches used and put them into perspective of similar methods. We conclude our review looking further ahead at the challenge posed by the analysis of interactions within microbial communities and different environments, in a field one could call “integrative metagenomics”.

Keywords: metagenomics, machine learning, OTU-clustering, binning, taxonomic assignment, comparative metagenomics, gene prediction

1 Introduction

While genomics is the research field relative to the study of the genome of any organism, metagenomics is the term for the research that focuses on many genomes at the same time, as typical in some sections of environmental study. Metagenomics recognizes the need to develop computational methods that enable understanding the genetic composition and activities of communities of species so complex that they can only be sampled, never completely characterized.

Analysis of microbial communities has been until recently a complicated task due to their high diversity and the fact that many of these organisms can not be cultivated. Metagenomics has essentially emerged as the only way to characterize these unculturable communities. Though “traditional” genomic studies are challenging, metagenomic studies push the requirement for highly scalable computational solutions further. Indeed, computational analysis is even more important for metagenomic analysis due not only to the large amount of metagenomic data, but also to the new questions introduced by metagenomic projects such as simultaneous assembly of multiple genomes, community-realized functions or host-microbe interactions (Wooley and Yuzhen, 2009).

Metagenomics has been popularized by the studies of bacterial communities and is still largely dominated by such applications. Indeed, the word itself is often employed in this sense. However, such a reductive view does not grant enough credit to the lively research in other areas such as viral or fungal metagenomics, and large-scale environmental studies that generate extremely heterogeneous data in terms of their origin.

Bazinet and Cummings have listed in their review 25 tools (Bazinet and Cummings, 2012) and this collection continuously expands. The specific challenges of “Big Data metagenomics” that the field is currently facing promote computational solutions originating from data science. In particular, recent years have brought a large set of work in machine learning applied to metagenomics. Indeed, machine learning is the methodology of finding patterns and making predictions from data, based on multivariate statistics, data mining and pattern recognition.

^{*}Bordeaux Bioinformatics Center, Université de Bordeaux, France

[†]Corresponding author: macha.nikolski@labri.fr; Bordeaux Bioinformatics Center, Université de Bordeaux, France and LaBRI, Université de Bordeaux, France

Machine learning currently offers some of the most promising tools for building predictive models for classification of biological data (see e.g. Libbrecht and Noble (2015)). Various biological applications cover the entire spectrum of machine learning problems including supervised learning, unsupervised learning (or clustering), and model construction. Moreover, most of biological data, such as those produced in metagenomic studies, are both unbalanced and heterogeneous, thus meeting the current challenges of machine learning in the era of Big Data.

The goal of this review is to examine the contribution of machine learning techniques for metagenomics, that is, answer the question *to what extent does machine learning contribute to the study of microbial communities and environmental samples?* We will first briefly introduce the scientific fundamentals of machine learning. In the following sections we will illustrate how these techniques are helpful in answering questions of metagenomic data analysis. We will describe a certain number of methods and tools to this end, though we will not cover them exhaustively. Finally, we will speculate on the possible future directions of this research.

2 Machine Learning Fundamentals

General principles The defining characteristic of machine learning (ML) systems is that they can improve with experience. Contrary to a system employing a classical deterministic algorithm, an ML system will exploit patterns observed in datasets to fine-tune and adapt its decisions. As defined by Mitchell (1997), a typical ML program requires at least three different components:

1. *experience*, in the form of data,
2. *task*, in the form of an output of the algorithm,
3. *objective*, in the form of performance measurement of a given output.

A program is said to “learn” if its performance for a given task improves with experience.

To design a program able to learn how to solve a given task, Flach (Flach, 2012) indicates the following requirements. First, data from the domain must be mapped to **features** constituting an (often finite) representation given as input to a **model** that will map it to an output. To build this model, some example data – often called **training data** – are fed into a **learning algorithm** that will search and try to identify an optimal model (w.r.t. the objective function) from a **hypothesis space**.

Learning algorithms are often grouped by how the model that they learn is represented in this hypothesis space. Such spaces and potential hypotheses are usually structured and reduced in order to obtain a “well-behaved” hypothesis space where the learning problem can be redefined mathematically in terms of a numerical optimization problem.

In the field of metagenomics, the most common questions that we identified as being amenable to learning are: OTU clustering, binning, taxonomic assignment, gene prediction and comparative metagenomics. These questions involve three types of machine learning tasks (see Table 1):

1. *classification*, where the task output is a discrete variable,
2. *clustering*, where the task output is a (possibly soft) partition of input data,
3. *dimensionality reduction*, where the task output is a lower-dimensional representation of the input data.

Another useful distinction is between supervised and unsupervised problems. A problem is said to be **supervised** if the output variable is provided in the training data. For example, distinguishing metagenomic samples based on known phenotypical traits is a supervised problem. The goal of supervised methods is thus to build a model from a set of labeled data points that can predict the correct category of unlabeled future data. Labels can be provided by any type of important metadata, such as the species of the microbe or of the host, but they have to be discrete and scalar. The possibility to classify unlabeled data is especially useful when alternative methods for

obtaining data labels are difficult such as it is the case, for example, for unculturable bacteria. On the other hand, **unsupervised** methods assume no labeled response but rather seek to determine hidden structure in the data. The most common application of unsupervised methods are for cluster analysis as well as dimensionality reduction.

Components of ML systems As indicated by Domingos (2012), any learning algorithm can be broken down into three components: (i) **representation**: the structure of the hypothesis space, (ii) **evaluation**: how a model will be evaluated, and (iii) **optimization**: how to navigate in the hypothesis space.

Some examples of hypothesis spaces / learning algorithm that are relevant for metagenomics are shown in Table 1.

	Representations	Learning	Applications example	Tools example
Classification/ Regression	Instances	<i>k</i> -NN	binning, OTU clustering	DOTUR
		Support Vector Machines	gene prediction, Comparative MG	MetaDistance
	Linear models	linear regression	binning	Tetra
		logistic regression	gene prediction, Comparative MG	MetaGene, MetaPhyl
		large scale linear model	taxonomic assignment	Vowpal Wabbit
	Decision trees	Random Forest	taxonomic assignment	16S Classifier
	Neural networks	neural networks	gene prediction	Orphelia
	Network	Hidden Markov Model	gene prediction	FragGeneScan
		Interpolated Markov Model	binning, gene prediction	Glimmer-MG, SCIMM
Dimension reduction	Linear combinations of features	PCA	binning	CONCOCT
		SVD	binning	LSA
Clustering	Means	<i>k</i> -means	binning	SCIMM, MetaCluster
	Medoids	<i>k</i> -medoids	binning	MultiBin
	Dendrogram	hierarchical clustering	OTU clustering	ESPRIT, MOTUR
	Mixtures / Soft partitions / Likelihoods	Gaussian mixture models	binning	CONCOCT
		Bayesian clustering / stochastic search clustering	OTU clustering, binning	CROP, BACDNAS, BEBaC, LikelyBin
		spectral clustering	binning	CompostBin
	NA	greedy heuristic clustering	binning, OTU clustering	DNAClust, USEARCH

Table 1: Examples of machine learning approaches for metagenomic applications. Methods are classified according to the three identified types of machine learning tasks: **Classification**, **Dimension reduction** and **Clustering**. These tasks can be further analysed by the type of representation of the hypothesis space (column **Representations**) and the learning algorithm that is applied to training data (column **Learning**). Finally, we mention types of metagenomic analyses to which these approaches have been applied (column **Application example**) and show examples of existing software that implements them (column **Tools example**).

In the case of supervised learning, the performance of a model shouldn't be measured on the training data. The main reason is that the real objective of inductive learning is not to perform

perfectly on data that was already seen, but to generalize well for previously unseen data. The phenomenon of poor performance on unseen data is called “overfitting”, and a simple way to ensure its absence is to leave out part of the data for evaluation purposes. It is however worth noting that different representations suffer differently from overfitting. Usually, the more complex the hypotheses are, the more care should be put in overfitting avoidance. For example, Naïve Bayes models have a simple representation (basically, just conditional probabilities / counts) and their simplicity makes them less prone to overfitting than more complex models (e.g. Random Forests), but their simplicity doesn’t allow complex decision functions to be represented. On the other hand, very complex models (such as e.g. neural networks) can approximate universal decision functions with any level of precision, but require tremendous training data and usually can not be trained to optimality.

Features Besides the learning algorithm and the model *per se*, the most important component of a learning system is how features are extracted from the domain data, a process known as **feature engineering**. As stated by Flach (Flach, 2012), “features are the workhorse of machine learning”. In our case, reads originating from an NGS project must be mapped to a domain that can be processed by the learning algorithm. Often, this domain is a vector in a multidimensional space. The combination of (i) how to map NGS reads into vector representation, (ii) learning algorithm and (iii) how a metagenomic question is formulated as a task is actually what distinguishes most of the methods described in this review.

A very natural and simple representation of DNA sequences as a numerical vector is the k -mer frequencies representation. For a given integer k , count the number of times any nucleotide subsequence of length k appears in the read. Subsequent transformation can enrich this representation (e.g. robust estimations, variance stabilization, normalization etc.). Naturally, the precision of this representation increases with k , but this comes with the cost of much higher computational and memory requirements needed to represent each read. For example, even for values of k as small as $k = 12$, each read will be represented by $4^{12} \approx 16 \times 10^6$ dimensional vector, with a memory cost that is out of reach of today’s computers, given a reasonably-sized metagenomics project. In fact, a more pervasive issue appears as k grows: generalizing a decision function becomes exponentially harder as we increase the dimensionality of the space, since only a very limited fraction of the feature space can be observed in the training data. Pushing this argument to its extreme, it is straightforward to observe that most 21-mers are actually unique to species present in public sequence databases (Kurtz et al., 2008). On one hand, this implies that we can recognize any known species with high precision, but on the other hand, this implies that novel species can not be classified and thus that such a system has close to zero generalization power. This is indeed the case for most large k -mer indexing schemes (e.g., LMAT (Ames et al., 2013)). Even for moderate values of k , it might still be the case that the feature space is too large given the training data, but that lower values of k yield too low resolution. In this case, it is possible to learn a lower-dimension manifold in which the data is projected, and instances of such techniques (e.g. PCA, SOM) are used by several of the tools reviewed here.

3 Metagenomics: data and questions

Modern metagenomic analysis involves a variety of options: genetic diversity profiling of samples, identification of present taxa (taxonomic assignment), assembly of sequencing reads, clustering and binning, prediction of protein-coding genes and further functional analysis of community-realized functions, without forgetting the comparative aspects of multi-sample studies. Machine learning has penetrated in all these applications, even assembly, which is traditionally solved by data-structure and algorithmic approaches, has been tackled by machine learning (Sato and Sakakibara, 2015). In this review we will focus on five important questions of metagenomic studies from the perspective of machine learning: OTU clustering, binning, taxonomic assignment, comparative metagenomics and gene prediction.

Terminology The abundance of literature on metagenomic analysis has generated quite a variable terminology, which makes navigating the underlying computational concepts a complex task. For the sake of consistency, we will keep the same terminology throughout this paper. In particular, we will use the term of *assignment* to mean the assignment of a given sequence to a specific taxonomic group (see section 6). We will thoroughly distinguish this from both *OTU clustering* and *binning* which correspond to the same machine learning problem of clustering as both refer to the grouping of a data set into subgroups which are distinct from one another. Being unsupervised problems by nature, these subgroups may remain unlabeled. However, we will distinguish them by using the term *OTU clustering* exclusively for clustering of marker gene sequencing data (see section 4), while *binning* will be used for the case of whole metagenome untargeted data (see section 5).

Notice that all these different analyses perform a *classification*, the difference being that in one case it is unsupervised (OTU clustering and binning) and in another, supervised (assignment). A lot (but not all) of the methods presented below contain both some form of grouping and an assignment step. We have decided to discuss them in the sections that correspond best to the computationally innovative part of such papers.

Data

Metagenomic studies differ by a number of elements. First and foremost is the kind of data that is collected, which largely determines the downstream analysis.

Amplicon sequencing In many metagenomic studies only partial genomic information is used. DNA is extracted from all the cells of the the sample (e.g., sea, soil, gut), a set of genomic markers is chosen and then targeted and amplified by PCR. Consequently, only marker genes are sequenced. The most widely used marker gene for amplicon sequencing is the small subunit ribosomal RNA (16S) locus, that is both a taxonomically and phylogenetically informative marker for both bacteria and archaea (Pace et al., 1986; Hugenholtz and Pace, 1996). Other markers such as the small subunit 18S, the large subunit (LSU: 23S/25S/28S) and the highly variable spacers ITS1 and ITS2 of the internal transcribed spacer (ITS) region have also been used, in particular for fungi, higher plants, insects and microalgae (Lindahl et al., 2013; Koetschan et al., 2010; Cuadros-Orellana et al., 2013).

Such targeted sequencing simplifies the analysis for two reasons. First, the amount of data remains reasonable (for a high-throughput analysis) and second, known marker genes' taxonomic classification is available – in particular for bacterial communities – through reference taxonomies such as, for example, in RDP (Cole et al., 2009) or Greengenes (McDonald et al., 2012) databases.

Many synonyms for amplicon studies can be found in literature, the most frequent being **metabarcoding** and **targeted metagenomics**.

Untargeted sequencing Whole metagenome sequencing without targeted amplification provides the possibility to tap into unknown species, but generates a substantial increase in the volume of data. Here, one of the important aspects is the representativity of the extracted DNA with respect to taxa that compose the sample, and various protocols exist (Delmont et al., 2011; Burke et al., 2009). If the community of interest is associated with a host than its DNA has to be separated by either fractionation or selective lysis to ensure that minimal host DNA is present. Physical filtering (fractionation) can also be used when only part of the community is of interest to the analysis. The organisms are then separated by size, leading to untargeted sequencing of filtered samples. Various physical and molecular extraction / filtering techniques exist and can be used to separate eukaryotic, prokaryotic, or viral particles. In particular, size-based filtering is widely used for separating viruses prior to studying the composition of viral communities (John et al., 2011; Hurwitz et al., 2013).

Choice of data Notice that the choice of data to be used in a given metagenomic study is not neutral since it determines the final results of the analysis. For example, in the case of amplicon sequencing, the choice of amplicons is known to influence the resulting clusters. Regions of marker genes with high inter-species and low intra-species variability are typically chosen to be sequenced

(e.g. V1 through V9 for 16S). They are flanked by conserved sequences that enable PCR amplification of targeted sequences (amplicons) using universal or specifically designed primers. However, depending on the chosen variable regions and primers, the resulting OTUs clusters and estimations of diversity may not be the same (see for example (Jumpstart Group, 2012), (Bella et al., 2013)). Moreover, amplicon sequencing excludes unknown species or highly diverged microbes from the analysis. Analyzing whole-metagenome untargeted sequencing data is the most computationally challenging task in metagenomics for the reasons of both data volume and variety. For example, environmental samples can contain genetic material from hundreds to tens of thousands of species of different abundances, some estimations placing the number of species for certain environments in the range of millions (Gans et al., 2005). Moreover, a large fraction species collected this way may not be known, potentially leading to annotation biases since the analysis relies on the genomes present in the current databases. In this respect, identification of viruses remains especially challenging (Soueidan et al., 2014). Sequencing filtered samples alleviates the computational burden downstream, thus allowing efficient contig assembly (Coetzee et al., 2010; Minot et al., 2012). However, it introduces certain biases, such as artificially including or excluding certain species.

4 OTU clustering

A wide-spread approach to explore the genetic diversity of a microbial community is to analyse amplicon sequencing data (or barcode sequences) by grouping it into Operational Taxonomic Units (OTUs). Historically, ribosomal RNAs are the most used basis for the reconstruction of microbial phylogeny (Ludwig and Schleifer, 1994) and the definition of the corresponding taxa. Thus, rRNA approach has quite naturally become the most popular one for defining taxa in targeted bacterial studies (Blaxter et al., 2005) and has been since generalized to other marker genes. The OTUs are defined at varying levels of identity (often 97%, see e.g. (Koeppel and Wu, 2013)), and can then be used to estimate the diversity and composition in terms of taxa of a given sample. Methods for OTU clustering typically require a distance function d and a certain threshold t under which two sequences are considered to be sufficiently close. One of the characteristics of OTU-based approaches is that all sequences are clustered into OTUs, including both species for which annotation is absent in the public databases as well as novel species. Moreover, OTU clustering has the double advantage of being a light-weight approach in terms of input data (amplicon sequencing) and consequent computation, as well as of producing high-quality groups precisely because amplicon sequences are by definition taxon-specific and different between species. Such data are an ideal setting for highly sensitive and specific machine learning methods. We recapitulate OTU clustering methods presented in this review in Table 2.

While quite computationally performant, OTU-based methods also possess certain intrinsic limitations. Most importantly, the analysis and biologically meaningful interpretation of the resulting OTUs can be problematic, especially in the case of unknown species. Also, while the threshold t greatly influences the inferred clusters, it is difficult to choose *a priori* given that the evolutionary rates for marker genes can vary greatly (Patwardhan et al., 2014). Furthermore, due to sequencing errors, wrong choice of t can result in an artificial inflation of OTUs (Kunin et al., 2010). Moreover, a number of studies point out that different clustering methods provide non-equivalent clusters for the same dataset, quantitatively in terms of total number of clusters and OTU size distributions as well as qualitatively in terms of cluster composition (Huse et al., 2010; Schloss and Westcott, 2011; Sun et al., 2011; Abeln et al., 2012). This is particularly salient in the presence of sequencing errors.

4.1 Hierarchical clustering

Hierarchical clustering algorithms such as the nearest neighbor (single linkage clustering), the furthest neighbor (complete linkage clustering), the weighted neighbor, and the average neighbor (unweighted-pair group method using average linkages, UPGMA) are widely used for OTU clustering. These algorithms require the computation of the complete pairwise distance matrix between

Tool	Validated on	Feature eng.	ML model	Source code	Publication	Last update	Compared to
Dotur	16S, rpoB	multiple seq. alignment	hierarchical clustering	github.com/mothur/DOTUR	2005	2005	none
Mothur	16S	alignment against a reference DB	hierarchical clustering	github.com/mothur/Mothur	2009	2015	none
ESPRIT	16S	k -mer prefiltering, pairwise alignment	hierarchical clustering	upon request	2009	unknown	Dotur
Usearch / Uclust	16S	pairwise alignment	greedy clustering	drive5.com/usearch	2010	unknown	CD-HIT
GramCluster	16S	suffix tree to access sequences, grammar-based distance, Lempel-Ziv sequence representation, grammar-based comparisons	greedy clustering	bioinfo.unl.edu/gramcluster.php	2010	2010	CD-HIT, Uclust
ESPRIT-Tree	16S	k -mer prefiltering, probabilistic sequences, pairwise alignment	pseudo-hierarchical clustering	web access	2011	unknown	Uclust, CD-HIT, ESPRIT
DNAclust	16S	suffix trees, k -mer prefiltering, length sorting, pairwise alignment	greedy clustering	dnacust.sourceforge.net	2011	2013	Uclust, CD-HIT
CROP	16S	Gaussian mixture model to describe the data, MCMC, pairwise alignment	Baeyesian clustering	github.com/tingchenlab/CROP	2011	2014	ESPRIT, mothur
CD-HIT	16S	k -mer prefiltering, length sorting, pairwise sequence alignment	greedy clustering	cd-hit.org	2012	2015	Uclust
BEBaC	16S	k -means, pre-clustering based on k -mers, multiple seq. alignment	stochastic search clustering	available upon request	2012	unknown	ESPRIT-Tree, Uclust, CROP
Uparse	16S, ITS	maximum parsimony model, abundance sorting, pairwise alignment	greedy clustering	drive5.com/uparse	2013	unknown	QIIME, mothur
BACDNAS	16S	sequences modeled as Markov chains (no alignment!)	Bayesian Clustering with the Dirichlet process prior	www.helsinki.fi/bsg/software/BACDNAS/	2013	2013	ESPRIT-Tree, BEBaC, CROP
Swarm	16S	k -mer prefiltering, pairwise alignment	single-linkage agglomerative clustering	github.com/torognes/swarm	2014	2015	CD-HIT, DNAclust, Usearch
OTUCLUST	16S-10, ITS-10, 16S-R	abundance sorting, pairwise alignment	greedy clustering	github.com/compmetagen/micca/wiki	2015	2015	Uparse, QIIME

Table 2: Characteristics of tools for OTU clustering. Shown are types of data on which the method has been **validated**, how features have been extracted (column **Feature eng.**), the machine learning algorithm (column **ML method**), link to the **Source code**, year of **publication** as well as the last source code update (column **Last update**), and finally tools against which the performance has been evaluated (column **Compared to**).

sequences. Based on this matrix, a hierarchical tree is constructed. Given a user-specified threshold t , leaf nodes within the cutoff are assigned to an OTU. The main disadvantage of the hierarchical clustering strategy is the overestimation of the number of OTUs, especially in the presence of sequencing errors, resulting in skewed OTU abundances.

Alignment-based methods suffer from the high computational cost and methods such as **Dotur** (Schloss et al., 2005) (based on multiple sequence alignment) and **Mothur** (Schloss et al., 2009) (based on Needleman-Wunsch alignments against a pre-aligned reference database) do not scale up very well despite the use of sparse matrices that represent the unique sequences only in the case of Mothur.

This gave rise to methods that attempt to achieve better scalability, such as **ESPRIT** (Sun et al., 2009) that implements a complete-linkage hierarchical clustering and minimizes the memory usage by using an on-line approach, and **ESPRIT-Tree** (Cai and Sun, 2011) that uses a probabilistic distance and a set of heuristics in order to avoid all pairwise distances. The computational speedup is also achieved by avoiding to align pairs of sequences for which it is easy to deduce from their k -mer comparison that the alignment will not be good enough.

A recent tool called **Swarm** (Mahe et al., 2014) is an agglomerative, unsupervised, single-linkage clustering algorithm that avoids the use of a global threshold t . First, it attempts to iteratively cluster highly similar amplicons using a local user-defined threshold and then refines the results by using clusters internal structure and amplicon abundances.

4.2 Heuristic clustering

Amplicon sequencing projects, especially in the case of environmental studies, produce constantly increasing datasets. For this reason, in order to compute OTUs, two important requirements are speed and scalability. However, exact clustering algorithms described in section 4.1 do not scale

well, requiring at worst the square of the number of input sequences comparisons. This has given rise to methods based on *heuristic greedy clustering*. Given a set of sequences S , a sequence $s \in S$ is chosen and it becomes the centroid of a new OTU cluster. This sequence s is then compared to all remaining sequences $S \setminus s$. Given a distance function d and a threshold t , all sequences $s' \in S$ such that $d(s, s') < t$ are added to the OTU and removed from S . Since each sequence is considered only once, heuristic greedy algorithms are usually fast. Notice, that the greedy approach provides results that are dependent on the order of input sequences.

A lot of methods in this category sort input sequences by length or abundance, such as for example **Uclust** (Edgar, 2010), **DNAclust** (Ghodsi et al., 2011) and **CD-HIT** (Suzek et al., 2007; Fu et al., 2012) and try to avoid costly all-against-all sequence alignment by a pre-filtering k -mer based step. Moreover, **GramCluster** (Russell et al., 2010) and **DNAclust** (Ghodsi et al., 2011) index the input dataset by a suffix tree for efficiency.

While remaining in the scope of greedy clustering, the **Uparse** (Edgar, 2013) and **OTUCLUST** of **MICCA** (Albanese et al., 2015) emphasize the need to rely on high quality sequences only and both include a number of steps for quality filtering, trimming and chimera filtering.

4.3 Model-based clustering

Model-based clustering aims to circumvent the overestimation of OTUs that due to the limitations of choosing an *a priori* threshold t .

For example, **CROP** (Hao et al., 2011) uses an unsupervised Bayesian clustering approach. The model relies on the notion of probability that a given sequence s belongs to a cluster. This probability is defined as function of the distance between the sequence s and the sequence that is in the center of the cluster. Moreover, CROP applies the divide-and-conquer principle by dividing the dataset into small subsets and performing Bayesian clustering on the subsets. Thus generated clusters are replaced by their consensus sequences on which a final step of Bayesian clustering is performed in order to obtain the OTUs.

In (Cheng et al., 2012) the authors introduce a probabilistic model-based **BEBaC** method. It is based on the calculation of an unnormalized posterior probability for an arbitrary partition of the reads, such that any two partitions can be compared. The best partition is then determined by a stochastic search process over the partition space.

In (Jaaskinen et al., 2013) the authors introduce the **BACDNAS** method that models sequences by Markov chains of fixed order and an expectation-maximization algorithm is used for learning a partition of these Markov chains. The prior distribution for partitions is the Dirichlet process prior.

5 Binning

The goal of *binning* is to establish the taxonomic profile of a given sample and it is often used in the case of environmental studies. We differentiate binning from OTU clustering mainly in terms of input data: while OTU clustering is used in targeted studies, binning deals with reads coming from any genomic region of any genome present in the sample. Binning is thus the most suited approach for studying complex communities. Nevertheless, almost all available methods are developed for bacterial communities, with the notable exception of **MetaVir** (Roux et al., 2013) that aims at the analysis of viromes. Analysis of other communities, such as for example fungi, are often done by ad-hoc methods or by hijacking software tools designed for bacteria (see e.g. (Lindahl et al., 2013; Cuadros-Orellana et al., 2013)).

In terms of computation, binning consists of assigning each read to a group, called *bin*, with the expected property that each bin consists of reads originating from the same taxon. Although some alignment-based methods exist (not detailed in this review), most existing computational tools for binning utilize sequence k -mer composition. Indeed, as observed by Kariin and Burge (1995), the distribution of k -mer composition is stable across a single genome and varies between genomes, even when only dinucleotides (di-mers) are considered. This observation underlies many metagenomic binning approaches.

In the existing literature, machine learning-based classification algorithms used for binning are categorized into supervised, semi-supervised, and unsupervised classes, depending on whether a training set of labeled data is used to build their models or not. We consider that the former case corresponds to solving the *taxonomic assignment* problem, that we describe in the next section. In this section, we only consider unsupervised or semi-supervised methods for the binning problem exclusively, that is methods that group sequences into bins without consideration of external reference sequences or taxonomic information in the same way it is done for OTU clustering. Indeed, it has been observed that in environmental samples, up to 99% of sampled bacteria are unknown (Eisen, 2007), hence methods relying on existing datasets will very likely miss important properties of samples. To solve the binning problem, various measures of similarity such as GC content, oligomer frequencies, the abundance of genes or contig coverage can be used during the inference. See Table 3 for the summary of binning methods presented in this review.

Tool	Validated on	Feature eng.	ML model	Source code	Publication	Last update	Compared to
TETRA	only 16s	4-mers	linear regression	www.megx.net/tetra	2004	unknown	none
CompostBin	bacterial genomes	6-mers + alignment-based weighting scheme	PCA+spectral clustering	bobcat.genomecenter.ucdavis.edu/sourav/compostbin/	2008	2012	none
LikelyBin	bacterial genomes	k -mers	partitions / stochastic search (MCMC)	ecotory.biology.gatech.edu/downloads/likelybin	2009	unknown	none
SCIMM	only 16s	preclustering	Interpolated Markov Models + k -means	www.cbcb.umd.edu/software/scimm	2010	2012	PHYSICIMM
AbundanceBin	bacterial genomes	k -mer abundances	stochastic search, expectation-maximization	omics.informatics.indiana.edu/AbundanceBin	2011	2013	MetaCluster
MetaCluster	bacterial, whole metagenome (viral included)	k -mer frequencies + Spearman footrule distance	k -means	i.cs.hku.hk/~alse/MetaCluster	2012	2014	AbundanceBin, Toss
MultiBin	bacterial genomes	pairwise alignment	similarity graph + k -medoids	none	2012	NA	AbundanceBin
Toss	bacterial genomes	unicity of k -mers	similarity graphs + MCL	www.cs.ucr.edu/~tanaseio/toss.htm	2012	unknown	CompostBin
MultiMetaGenome	bacterial genomes	scaffold coverage, 4-mers frequencies, GC content, ORF, marker proteins, taxonomic assignment	linear regression + local PCA	github.com/MadsAlbertsen/multi-metagenome	2013	2014	ESOM
CONCOCT	bacterial genomes	k -mers and coverage	Gaussian mixture models	github.com/BinPro/CONCOCT	2014	2015	MetaWatt, CompostBin, SCIMM, LikelyBin
MaxBin	bacterial genomes	4-mers, coverage, marker genes analysis	stochastic search, expectation maximization	sourceforge.net/projects/maxbin/	2014	2015	ESOM
LSA	bacterial genomes, some remarks on phage analysis	locally sensitive hashing of k -mers,	SVD + k -means	github.com/brian-cleary/LatentStrainAnalysis	2015	2015	CONCOCT, GroopM
MetaBAT	bacterial genomes	4-mers, coverage, alignment, pre-assembled contigs	custom k -medoids	bitbucket.org/berkeleylab/metabat	2015	2015	CONCOCT, Canopy, MaxBin, GroopM
DNAClust	AMD dataset	k -mers + filtering	greedy clustering	dnaplust.sourceforge.net	2011	2013	Uclust, CD-HIT

Table 3: Characteristics of binning tools. Shown are types of data on which the method has been **validated**, how features have been extracted (column **Feature eng.**), the machine learning algorithm (column **ML method**), link to the **Source code**, year of **publication** as well as the last source code update (column **Last update**), and finally tools against which the performance has been evaluated (column **Compared to**).

Binning is a computationally expensive task since the complete untagged environmental datasets are both large and heterogeneous due to the high-complexity of communities in terms of number of species, such as exemplified by ocean microbial communities (see the Science special issue on Tara Oceans project (Bork et al., 2015)) and the human microbiome project (<https://commonfund.nih.gov/hmp/index>).

Binning methods are known to perform better on either longer reads or on pre-assembled datasets. It has been shown that even partial assembly improves the strength of the taxonomic signal contained in individual short reads, and that even in the case of increased chimericity (Teeling and Glockner, 2012; Mende et al., 2012). Consequently, given a metagenomic dataset, a lot of methods suppose that at least partial assembly has been performed to obtain contigs of the size not less than 1000bp.

5.1 Fully unsupervised clustering

To the extent of our knowledge, the first application of k -mer distribution to the binning problem was proposed by Teeling in 2004 with a tool called **Tetra** (Teeling et al., 2004). Tetra computes a z -score for each possible 4-mer then compares two fragments by Pearson correlation over all 4-mers. However, the author indicates that this method works well for very large sequences only (in the range of 40kb).

More recently, **CompostBin** (Chatterji et al., 2008) uses an unsupervised clustering scheme based on graph cuts. First, principal component analysis extracts meaningful k -mer features and then the tool applies a normalized cuts clustering technique (spectral clustering) to group sequences into taxa-specific clusters. **MetaCluster** (Wang et al., 2012a,b) is a collection of unsupervised methods based on k -means clustering. The authors make use of three observations that are the basis of most unsupervised binning strategies.

1. k -mer frequencies from reads of a genome are linearly proportional to the genome abundance,
2. long k -mers are unique, and
3. short k -mers frequency distribution are similar for similar genomes.

Furthermore, observing that binning is sensitive to the range of abundance (for samples containing both extremely-low and high abundance), the authors propose to separate reads originating from low abundance species from reads originating from high abundance species. The group of high-abundance species' reads are then processed separately using the k -means algorithm on the 5-mer representation, while the group of low-abundance species' reads are binned using 4-mer's frequencies.

In its turn, **Toss** (Tanaseichuk et al., 2012) groups all unique k -mers into clusters and then merges these clusters based on k -mer repeat information by the Markov Cluster algorithm. In the case, where samples are expected to have highly different abundance levels of individual species, Toss relies on AbundanceBin (Wu and Ye, 2011) to pre-cluster reads. **MetaVir** (Roux et al., 2013) was developed to deal specifically with filtered viral sequencing data and is based on k -mers of size 2, 3 and 4; hierarchical clustering is performed based on the Euclidean distance between k -mer frequency vectors.

A recent method called **Latent Strain Analysis (LSA)** (Cleary et al., 2015) tries to leverage modern machine-learning techniques such as feature hashing (Weinberger et al., 2009) and online dimensionality reduction. It is a binning method based on the hyperplane clustering approach and relying on a complex-space hashing approach. They devise a k -mer hashing function that has the property of being locally sensitive: similar k -mers have a much higher probability of generating a hash collision. Such k -mers are placed in the same row within a matrix that recapitulates k -mer abundance for each sample. This matrix is then reduced using singular value decomposition (SVD) - meaning that it is performed in the space of k -mer co-variation - and reads are then assigned to the resulting k -mer clusters. The intuition behind the method is that co-variance information provided by SVD represents the relationships between k -mers found in the same genome.

Orthogonally, one of the approaches that has been fairly used in literature is self-organizing maps (SOMs), which groups sequences by oligonucleotide frequencies (Dick et al., 2009) and read coverage levels (Sharon et al., 2013). Although vastly used, application of SOMs requires a manual step, where the user selects sets of reads (or scaffolds) that should be binned together.

5.2 Model-based clustering

Another family of approaches to solve the binning problem relies on inference of a probabilistic model. These methods are unsupervised, and mostly use expectation maximization to estimate model parameter.

The main goal of **LikelyBin** (Kislyuk et al., 2009) is to bin together relatively short reads (400nt) from low complexity samples (between 2 and 10 species) that show sufficient evidence of genomic diversity. To this end, an inference scheme based on maximum likelihood principle and Markov

Chain Monte Carlo estimates a nucleotide model (a Markov model with order ranging from 3 to 5) and then assigns each read to the most likely model.

The Markov model of LikelyBin has been generalized to varying order Markov chains by (Kelley and Salzberg, 2010) in a tool called **SCIMM**. Instead of having a fixed order, varying order chains (termed Interpolated Markov Models) interpolate between all models of lower order. Furthermore, the authors propose an expectation maximization scheme similar to the k -means algorithm, but where each cluster is modeled as an Interpolated Markov chain and cluster assignment follows maximum likelihood. Cluster initialization (one of the major problems of k -means clustering) can be obtained by running LikelyBin or CompostBin. Interestingly, the authors also suggest a supervised scheme for cluster initialization based on taxonomic assignment obtained by the tool Phymm (described in section 6).

AbundanceBin (Wu and Ye, 2011) was the first method to rely on information different from similarity and composition. Indeed, it uses k -mer abundance information in order to separate reads from genomes that have different abundance levels. AbundanceBin first computes frequencies of all k -mers in a metagenomic dataset and then, assuming that these frequencies come from a mixture of Poisson distributions, models the original reads as a mixture of these distributions. An expectation-maximization algorithm is further used to find the distributions' parameters. Each read is then assigned to the bin having the highest estimated probability of this read to belong to a given bin. A drawback of this abundance-based approach is to be only able to distinguish species whose abundance levels are considerably different.

Noting that most existing method for scaffold binning (such as SOM) require the user to manually and individually select bins, the authors of **MaxBin** (Wu et al., 2014) propose a fully automated scaffold binning model-based algorithm. Using expectation maximization over 4-mers frequencies as well as scaffold coverage levels, MaxBin can estimate the probability that a sequence belongs to a particular bin. Similarly to other model-based clustering algorithms, the results of MaxBin are highly sensitive to parameter initialization, especially the number of bins. To circumvent this difficulty, MaxBin can use results from FragGeneScan to identify single-copy marker genes and thus derive estimates for the number of bins.

5.3 Multiple-sample binning

In response to the inability of abundance-based methods to separate species with similar abundances, a number of techniques have emerged that counter-balance it by considering multiple samples simultaneously. For example, algorithms proposed in **MultiBin** (Baran and Halperin, 2012) start by pooling all reads from all samples and comparing them by pairwise alignment. The algorithm then performs k -medoid clustering on vectors representing the coverage in each sample.

More extensively, **MultiMetaGenome** (Albertsen et al., 2013) improves on this idea by considering the abundance difference in any of the samples in order to separate bins. MultiMetaGenome applies a primary binning by clustering assembled genes by similarity, thus producing a non-redundant gene catalogue. Original reads are then mapped against these representative sequences and a normalized coverage for each sequence is computed. The representative sequences are further clustered into population genomes by plotting the two coverage estimates of all representative sequences against each other. In a similar way (Nielsen and et al., 2014) introduces a method for binning based on clustering of co-abundant genes, where groups of genes are binned based on their abundance across multiple samples. Seed genes are picked randomly and bins (called co-abundance gene groups CAGs) are defined by correlation in terms of abundance.

Recently, methods that combine information on both sequence abundance (as indicated by coverage) in multiple samples and sequence composition (or similarity) have received a lot of attention. For example, **Concoct** (Alneberg et al., 2014) uses Gaussian mixture models to cluster contigs based on both sequence composition and coverage across multiple samples. **MetaBat** (Kang et al., 2015) is a binning method that uses a probabilistic distance based on the tetranucleotide frequency and coverage of contigs. This pairwise distance between contigs is subsequently used in a medoid clustering algorithm to consolidate similar contigs into bins.

6 Diversity profiling and taxonomic assignment

While OTU clustering and binning provide a global view of a metagenomic sample, they do not answer the question of which species constitute it. This is the goal of **diversity profiling** and **taxonomic assignment**. Diversity profiling aims to investigate the community structure of metagenomes, by providing abundances of different taxa. In the case of taxonomic assignment, we are interested to know to which taxon belongs each read (or assembled contig). Here we have a set of known reference sequences $S = \{s_i\}$ and a set of class labels $L = \{l_i\}$ (species or any other taxonomic unit). Each sequence s is labelled by its class. A taxonomic assignment algorithm uses S in order to identify patterns in S and thus be able to classify novel (eventually unknown) metagenomic sequences M by providing a mapping $M \rightarrow L$. Notice that diversity profiling can be performed based on the results of taxonomic assignment, however dedicated methods aim at circumventing this step and rather to perform a global analysis for the reasons of computational efficiency.

Taxonomic assignment may take as input the result of OTU clustering or binning, or it can be done directly from sequencing data. Notice that many software methods provide a "two-in-one" solution integrating both clustering / binning and taxonomic assignment of the resulting clusters. As for OTU clustering and binning, taxonomic assignment methods are based either on sequence similarity provided by alignment or on genomic signatures in terms of oligonucleotide (k -mer) composition. Alignment-based methods combined with an algorithmic and data structure-driven solution (e.g. LCA) are very efficient for small-scale studies, such as amplicon sequencing. In particular, they provide state-of-the-art performance in terms of classification accuracy. Since these methods do not rely on machine learning, they will not be discussed in this review (see for review (Bazinnet and Cummings, 2012)). See Table 4 for methods that we discuss in this review.

Even if they demonstrated state-of-art accuracy, alignment-based approaches are time-consuming and highly dependent on the presence of reference genomes in databases. Consequently, whole and complex metagenomic sample can hardly be analyzed using such methods. By contrast, machine learning alignment-free methods are based on pattern recognition, and should (theoretically) be more robust to noise or missing data. Furthermore, in terms of computation, once a model has been trained using reference data, it's execution on novel sequences is generally linear and thus much less time-consuming than alignment based methods. A recent study by Vervier and co-authors (Vervier et al., 2015) that explores the possibility of a large-scale machine learning implementation for taxonomic assignment problem, has confirmed that compositional approaches achieve faster prediction times and consequently are appropriate for whole metagenome studies.

Taxonomic assignment relies on reference databases in order to assign reads to taxonomic units. Consequently, computational methods fall into the category of supervised classification and learning, classification labels being provided by the taxa present in the reference database. However, there are several competing taxonomies that differ substantially (DeSantis et al., 2006; Santamaria et al., 2012) and choosing among them is not necessarily obvious (Alonso-Alemayn et al., 2014). Moreover, for the unknown taxa, the inherent risk is to incorrectly assign their reads since they may be closely related to some of the reference sequences. Moreover, it has been long known that the assignment of short fragments (less than 1000bp) is a difficult task (McHardy and Rigoutsos, 2007), consequently as for binning, a pre-assembly step is often performed (Teeling and Glockner, 2012; Mende et al., 2012).

6.1 Naïve Bayes and Bayesian methods

The well-known ribosomal database project **RDP classifier** (Wang et al., 2007; Cole et al., 2009) relies on a reference sequence database which contains relevant species, then assigns a class label (e.g. species, genus, etc.) to each read by the naïve Bayesian algorithm based on k -mer occurrence information. Similarly, the popular **NBC** (Rosen et al., 2011) tool and **FCP** classifier Parks et al. (2011) implement a naïve Bayesian framework based on k -mer counts. Notice however, that RDP classifier is meant for 16S data, while NBC was developed for whole genome sequencing projects.

While based on Bayes' rules, Naïve Bayesian classifiers, such as learnt by RDP and NBC, are not bayesian in the formal sense, as they do not derive probability distributions over posteriors. By

Tool	Validated on	Feature eng.	ML model	Source code	Publication	Last update	Compared to
RDP classifier	16S, LSU	8-mers, with specific priors and genus-specific conditional probabilities	naive Bayes / k -NN	rdp.cme.msu.edu	2007	2015	none
NBC	16S, ITS, LSU, viruses, fungal	k -mers	naive Bayes	nbc.ece.drexel.edu	2008	2012	BLAST
Phymm	whole metagenome	variable-length k -mers	Interpolated Markov Models	www.cbcb.umd.edu/software/phymm/	2009	2012	PhyloPythia, Carma
pplacer	16S	none	likelihood based phylogenetics	matsen.fredhutch.org/pplacer/	2010	2015	RAxML
TAC-ELM	whole metagenome	GC-content, 3–4 mers	extreme learning machines, neural networks	cs.gmu.edu/~mlbio/TAC-ELM	2010	unknown	TAC-ELM, Phymm, BLAST, PhymmBL, PhyloPythia
FCP	bacterial genomes	10-mers + Laplace Smoothing	custom Naive Bayes	kiwi.cs.dal.ca/Software/FCP	2011	2015	PhyloPythiaS, TACOA, BLAST, LCA
Taxy	whole metagenome	k -mer composition of the sample	mixture modeling	gobics.de/peter/taxy	2011	unknown	Carma, Treephyler, Phymm, Galaxy
PhyloPithia / PhyloPithiaS	bacterial, whole metagenome	4, 5 and 6-mer frequencies	structural SVM	web access	2012	unknown	NBC, best BLASTN
Quikr*	16S	k -mer frequency	dimensionality reduction (compressed sensing), convex optimization	sourceforge.net/projects/quikr	2013	2014	RDP
WGSQuikr*	whole metagenome	k -mer frequency	dimensionality reduction (compressed sensing), convex optimization	sourceforge.net/projects/wgsquikr	2014	2014	RDP
SEK*	16S	k -mer composition of the sample	Kernel density estimator and mixture density models	github.com/dkoslicki/SEK	2014	2015	Quikr, Taxy, BeBAC
ARK*	16S	k -mer frequency	k -means then SEK	github.com/dkoslicki/ARK	2015	2015	Quikr, SEK, RDP
AKE	bacterial genomes	k -mers (length normalized, importance weight and over/under representation)	H2SOM classifier (neural network)	web access	2014	unknown	NBC, PhyloPythiaS, WebCarma
PhyloSift	16S	alignment against HMMs of gene families	Bayesian model	github.com/gjospin/PhyloSift	2014	2014	QIIME
16S Classifier	16S	2–6 normalized k -mers frequencies	Random Forest	metabiosys.iiserb.ac.in/16SClassifier/application.php	2015	2015	RDP, BLAST
Vowpal Wabbit	whole metagenome	4–12 mers	large scale linear models, feature hashing	unavailable	2015	NA	BWA-MEM, NBC
CSSS	viral, bacterial	k -mer based distances + alignment	1-NN	collaborators.oicr.on.ca/vferretti/borozan_css/csss.html	2015	2015	Phymm, NBC, PAIPhy, Kraken, PAUDA

Table 4: Characteristics of tools for diversity profiling (signalled by *) and taxonomic assignment. Shown are types of data on which the method has been **validated**, how features have been extracted (column **Feature eng.**), the machine learning algorithm (column **ML method**), link to the **Source code**, year of **publication** as well as the last source code update (column **Last update**), and finally tools against which the performance has been evaluated (column **Compared to**).

contrast, **pplacer** (Matsen and Armbrust, 2010) offers a full probabilistic and Bayesian framework to locate a query sequence in a reference phylogeny. From this localisation, a taxon identifier can be assigned to the query sequence. **pplacer** is used as a component of the popular **PhyloSift** (Darling et al., 2014) suite, that uses a reference database of gene families, aligns all of the input reads against this database and then calculates the posterior probability that the read diverged from particular branches of the reference tree via direct integration.

6.2 Model-based methods

Phymm (Brady and Salzberg, 2009) uses Interpolated Markov Models to learn variable-length oligonucleotides typical of a given phylogenetic grouping. In a similar fashion, **PhyloPithia** and **PhyloPithiaS** (McHardy et al., 2007; Patil et al., 2012) learn Support Vector Machine classifiers based on k -mer frequencies associated with different taxonomic groups. This classifier is then used to assign reads from a new metagenomic sample to pre-existing taxa.

The training step for such approaches can be quite slow, which gave rise to software tools that employ different optimization techniques. These approaches target the problem of diversity profiling, for which establishing a global model is particularly well-suited. Among these methods are

convex optimization strategies such as **Quikr** (Koslicki et al., 2013) (and its whole genome version **WGSQuikr** (Koslicki et al., 2014)) and **Taxy**, a mixture modeling approach of the overall oligonucleotide distribution of a metagenomic dataset (Meinicke et al., 2011). **SEK** (Chatterjee et al., 2014) aims to overcome the sensitivity of Quikr to the differences between length of sequences that are present in the database and read length, by proposing a method based on kernel density estimators and mixture density models, that enables the estimation of taxonomic units’ frequencies by solving an under-determined system of linear equations. **ARK** (Koslicki et al., 2015) is another method for diversity profiling that aims to improve the precision of convex optimisations approaches by aggregating the reads from a sample with an unsupervised machine learning approach prior to the estimation phase.

Other classical machine-learning approaches, such as Random Forests and neural networks have been advanced to perform taxonomic assignment. For example, **16S Classifier** (Chaudhary et al., 2015) uses Random Forests for the taxonomic classification of short 16S rRNA and the Greengenes taxonomy for assignment, while **TAC-ELM** (Rasheed and Rangwala, 2012) is a composition-based method (oligonucleotides and GC content) that uses a neural network-based model. Moreover, Vervier and co-authors have demonstrated that modern large scale linear classifiers with feature hashing such as **Vowpal Wabbit** (Vervier et al., 2015) can be efficiently used.

The recent **CSSS** method (Borozan et al., 2015) exploits the complementary nature of alignment-based and alignment-free similarity measures and applies the nearest neighbor clustering. In particular it aims at being able to assign taxonomic ranks to both bacterial and viral communities. However, having to compute four different similarity measures, impacts the scalability.

7 Comparative Metagenomics

The questions that we have addressed so far concern the association of DNA fragments from a given sample with a class (either an OTU, a bin or a taxon). On the other spectrum of machine learning applications for metagenomics lie methods that try to assign labels to whole samples, by analyzing features derived from all DNA fragments that compose it. Such classification aims for example at analyzing phenotypes based on metagenomic fragments and has wide applications in biomedical settings.

Following the distinction between binning and taxonomic assignment, we first review unsupervised methods that aim at clustering / comparing samples that are not labeled then we review supervised methods that can classify metagenomic samples into pre-defined classes. See Table 5 for a summary.

Tool	Validated on	Feature eng.	ML model	Source code	Publication	Last update	Compared to
R-SVM	plants, bacteria	normalized k -mer counts; feature selection integrated in R-SVM	recursive SVM	none	2014	NA	none
DectICO	whole bacterial metagenome	intrinsic correlation of oligonucleotides, feature selection by kernel partial least squares	SVM	github.com/dingxiao8715/DectICO	2015	2015	R-SVM
MetaPhyl	16S	vector of OTU frequencies + phylogenetic tree of the OTUs	multinomial logistic regression model	www.es.uer.edu/~tanaseio/metaphyl.htm	2014	unknown	Toss, Velvet
MetaDistance	16S	normalized and variance-stabilized proportion of reads assigned to a given taxon, sparse weighted distance	quadratic SVM	metadistance.igs.umaryland.edu/Welcome.html	2011	unknown	ANOVA

Table 5: Characteristics of tools for comparative metagenomics. Shown are types of data on which the method has been **validated**, how features have been extracted (column **Feature eng.**), the machine learning algorithm (column **ML method**), link to the **Source code**, year of **publication** as well as the last source code update (column **Last update**), and finally tools against which the performance has been evaluated (column **Compared to**).

7.1 Unsupervised methods

One of the first approaches to assess the distance between metagenomic samples is known as the **Directed Homogeneity test** (Mitra et al., 2009) and is implemented in the MEGAN toolbox.

This test employs Pearson’s chi-square test over all sub-nodes of a taxa, and accounts for multiple testing correction. Although this test can indicate how significant is the difference between two samples, it is not usable in a distance-based clustering approach.

Jiang and co-authors explore different sequence signature methods (14 k -mer based distances) and UPGMA clustering for the problem of clustering metagenomic samples as well as for recovering environmental gradients that affect microbial samples (Jiang et al., 2012). The authors concluded that k -mer based descriptors are well-suited for this task and have singled out the d_2^S dissimilarity measure, (corresponding to the uncentered correlation between the number of counts of k -mers for two sequences normalized by standard deviation) that outperformed others for both application scenarios. This measure has been recently employed in the DectICO method (Ding et al., 2015) (see section 7.2).

The **DSM Framework** of Seth et al. (2014) aims at computing a dissimilarity measure between samples based on their k -mer frequencies. Such a measure can be used for initial exploratory analysis of a large set of metagenomic samples. The authors propose to compare samples based on the Jaccard dissimilarity between normalized relative frequencies of informative k -mers. By using an efficient distributed implementation, informative k -mers can be identified *over all values of k simultaneously*. Although this is an unsupervised method in the sense that samples are *a priori* not labeled, the authors demonstrate that their method can discriminate samples more efficiently than supervised methods based on protein families. In our opinion, such feature estimation and feature selection can be used for other machine learning approaches aiming at sample classification.

7.2 Supervised methods

The unsupervised approaches are powerful for exploratory data analysis, but are not adequate when samples are known to belong to different groups or phenotypes. In this setup, supervised machine learning aims at inferring the relationship between the sample’s content and the label of interest.

To the best of our knowledge, the first attempt at supervised learning for comparative metagenomics was performed by Yang et al. (2006) to distinguish between soil and sediment samples obtained via amplicon sequencing. They applied both Random Forest and k -nearest neighbors algorithms over all identified hyper-variable regions. Specifically for metagenomics samples, the **MetaDistance** (Z. Liu et al., 2011) method was devised for multi-class classification of amplicon sequenced metagenomic datasets. MetaDistance represents samples as a normalized and variance-stabilized proportion of reads assigned to a given taxon. This matrix is then used to learn a sparse weighted distance by using a quadratic SVM formulation that maximises intra-class distance and minimized inter-class distance. This sparse distance encodes feature selection and is then used in a k -NN algorithm.

Owing to the observation that OTUs can be organized as a tree, Tanaseichuk et al. (2014) proposed a multi-class metagenomic classifier called **MetaPhyl** that accounts for OTUs relationships. Each sample is represented by a vector of OTU frequencies, and a multinomial logistic regression model with a tree penalty is then trained. The authors report greatly improved performance over MetaDistance and over generic classifiers (logistic regression and Random Forest).

Another approach specifically devised for metagenomic sample classification is the **R-SVM** of Cui and Zhang (2013) where the authors generalized recursive SVMs to perform both feature selection and discrimination of human whole metagenome samples for both control and inflammatory bowel disease patients.

More recently, this approach has been generalized in the **DectICO** tool (Ding et al., 2015), which combines (i) kernel partial least squares as a strategy for feature selection, (ii) *intrinsic correlation of oligonucleotides* (ICO) that generalizes the k -mer frequencies to describe samples and (iii) SVM as a learning algorithm. The authors claim that DectICO method outperforms other approaches when long k -mers are considered.

8 Gene prediction

Another broad question that metagenomics tries to answer is the **functional analysis** for metagenomic studies, that is determining what are the functional and metabolic repertoires of a given community, what are the differences between communities and what are the members that enable them to exert different effects. Ideally, a functional profiling of a community would establish a complete geneset for each species present in the sample and perform further functional annotation and network analysis. This is still not possible.

Considering a microbial community as a whole (and thus ignoring the exact assignment of a gene to a specific species) gives rise to community network analysis. It has even been argued that a functional view should better characterize a microbial community than its taxonomic composition since different organisms are able to perform similar biological functions, giving rise to the concept of community-realized functions (Liu, 2012; Fuhrman, 2009; Zhou et al., 2010). To the best of our knowledge, neither functional annotation nor network reconstruction have been tackled by machine learning approaches. Consequently, this section is dedicated to gene prediction in metagenomic samples, see Table 6 for an overview.

Gene prediction methods attempt to identify patterns within DNA sequences that correspond to those recognized by transcription and translation machinery. Typically, gene prediction tools predict the start and stop sites of protein-coding genes and produce *in silico* translations of these genes.

Standard tools for gene finding were designed to work on single genome sequencing data (e.g. Glimmer (Delcher et al., 1999)). Typically, gene prediction is based on Hidden Markov models that are trained on the gene structure of known organisms similar to the one that is being studied or alternatively on generalized models of prokaryotic or eukaryotic genes. Often, genes present in a metagenomic sample come from different organisms, this is why such an approach can not produce predictions with high confidence. Indeed, true positive rates of gene identification in metagenomic samples are at best around 70% of true positives for reads, missing a significant subset of genes (Filippo et al., 2012; Trimble et al., 2012). Since the accuracy in complete genomes is very high (> 95%), an assembly step can greatly improve gene prediction for metagenomic samples and some methods explicitly aim at longer sequences.

It has also to be noted that most of the proposed computational methods target bacterial communities. Thus, gene prediction for mixed metagenomic datasets remains a largely unresolved question.

Tool	Validated on	Feature eng.	ML model	Source code	Publication	Last update	Compared to
MetaGene	bacterial	GC%, ORF lengths, distance from left start codons, distance between neighboring ORFs	Hidden Markov Model	omics.informatics.indiana.edu/FragGeneScan	2006	2015	Glimmer, MetaGene
MetaGeneMark	bacteria, archaea	3, 4, 5 and 6-mer and nucleotide frequencies	three-periodic Markov chain model	web access	2010	unknown	GeneMarkS, MetaGeneAnnotator, MetaGene
FragGeneScan	bacteria, whole metagenome	codon usage bias, sequencing error models and start/stop codon patterns	Hidden Markov Model	omics.informatics.indiana.edu/FragGeneScan	2010	2015	BLASTX, MetaGene
Orphelia	bacterial	codon, dicodon usage, orf length, tis, gc content	neural network	orphelia.gobics.de	2008	unknown	MetaGene
MetaGeneAnnotator	bacteria, archaea, prophage	as MetaGene + models for prophage genes and RBS	logistic regression	metagene.cb.k.u-tokyo.ac.jp	2008	unknown	GeneMarkS, Glimmer, MetaGene
MetaProdigal	bacterial	start site information, translation table, hexamer statistics, RBS motifs and upstream base composition	log-likelihood function on subsets of training data	github.com/hyattpd/Prodigal	2012	2015	MetaGeneAnnotator, MetaGeneMark
Glimmer-MG	whole metagenome	gene length, start/stop codon presence, TIS, rbs, start codon usage	Interpolated Markov Models	www.cbc.umd.edu/software/glimmer-mg	2012	2014	MetaGeneAnnotator, MetaGeneMark, FragGeneScan

Table 6: Characteristics of gene detection tools. Shown are types of data on which the method has been **validated**, how features have been extracted (column **Feature eng.**), the machine learning algorithm (column **ML method**), link to the **Source code**, year of **publication** as well as the last source code update (column **Last update**), and finally tools against which the performance has been evaluated (column **Compared to**).

8.1 Statistical methods

One of the first gene prediction methods for metagenomic datasets, **MetaGene** (Noguchi et al., 2006), uses logistic regression models whose features are the GC content and the di-codon frequencies in order to distinguish between gene-coding and non-gene coding ORFs. **MetaGeneAnnotator** (Nogucki et al., 2008) builds on the MetaGene approach by integrating more thorough statistical models for bacterial, archaeal and prophage genes and incorporating species-specific patterns of ribosome binding sites in order to increase the confidence of the translation starts site prediction.

MetaProdigal (Hyatt et al., 2012) is a method that improves on the Prodigal software Hyatt et al. (2010) and is based on the same idea of gene prediction. The coding regions are distinguished from the background by using a log-likelihood function that incorporates information such as translation table, hexamer statistics, ribosomal binding site motifs and upstream base composition. The main difference with Prodigal resides in the fact that training data is pre-clustered using a complete-linkage clustering and the log-likelihood function is defined on these clusters.

8.2 Model-based methods

MetaGeneMark (Zhu et al., 2010) is built on the same principle as the popular HMM-based gene prediction tool GeneMark, but adapts it to process short DNA fragments. First, training genomes are used to estimate a polynomial and logistic approximations of oligonucleotide frequencies as a function of the GC-content. These estimates are then used to set the parameters of the MetaGeneMark gene finder HMM. **FragGeneScan** (Rho et al., 2010) is another HMM-based approach that integrates codon usage bias, start/stop codon patterns as well as an explicit modeling of sequencing errors; it seems to outperform other solutions and is currently used by the EBI Metagenomics pipeline.

Glimmer-MG (Kelley et al., 2012) relies on Interpolated Markov Models to identify the coding regions and distinguish them from noncoding DNA. First, binning is performed using SCIMM and the IMM models are trained within each cluster for gene predictions, which makes this method quite computationally expensive.

Orphelia (Hoff et al., 2008, 2009) is a two-stage machine learning approach. First, linear discriminants for monocodon usage, dicodon usage and translation initiation sites are used to extract features from sequences. Second, ORFs are classified as being protein-coding or not by a neural network that combines these features with ORF length and GC-content. The neural network is trained on random sub-sequences of a given length of genomes in the reference database.

In addition to the novel computational methods presented in this review a number of integrated pipelines has been developed such as the EBI Metagenomics Webserver (Hunter et al., 2014) based on FragGeneScan, RAMMCAP (Li, 2009) based on MetaGene and SmashCommunity based on GeneMark and MetaGene (Arumugam et al., 2010).

9 Discussion

Applications of machine learning in metagenomics have penetrated all major questions asked in this field of research. Moreover, this idiom of data analysis has been a major driver in all areas of metagenomics: OTU clustering, binning, taxonomic assignment, comparative metagenomics and functional analysis. In turn, these biological applications have triggered many original developments in machine learning: in the way data are represented, how relevant features are selected and how prior knowledge is introduced in the algorithm.

Looking at the tables that recapitulate the examined tools, we can notice a clear progression towards the sophistication of machine learning approaches. Even better, today certain metagenomic questions can be considered to be essentially solved, in the sense that they can be answered in a computationally efficient way and tend to produce high quality results. The most noteworthy example is the taxonomic assignment for bacterial communities in the case of amplicon sequencing (for the cases where the resolution provided by reference taxonomies is considered to be sufficient for the study).

Despite the spectacular technological and methodological advances, the science of metagenomics still has a large number of questions for which no definitive answer has been provided. Here we enumerate some of these questions, which in our opinion constitute the major avenues for current and future research in metagenomics. These questions pertain to both biological and computational aspects of metagenomics.

- *Novel species.* Dealing with unknown species complicates both computation and interpretation. For example, OTU clustering and binning can be efficiently performed; however, interpretation of the resulting groups is difficult. At the same time, taxonomic assignment is unable to assign sequences from such species if no sufficiently close taxa exist in the reference databases.
- *Whole metagenomic samples.* As we have seen, many computational approaches have been validated on bacterial data. Being able to efficiently and accurately deal with samples containing, for example, host traces, archaea, viruses, fungi, protists and small algae, is still an open question.
- *Functional analysis.* Detection of protein-coding genes and consequently their downstream analysis remain challenging. In particular this involves developing methods able to efficiently analyse heterogeneous and fragmented data in the presence of sequencing errors.
- *Big Data challenge.* Machine learning is still seldom deployed in the “Big Data” fashion. Many authors remark that current methods cannot be applied to datasets with sizes over a Terabyte. This constitutes a bottle-neck for large-scale environmental studies. An effort in porting existing methods to modern Big Data machine learning frameworks (e.g., Spark etc.) or in developing novel ones has to be undertaken.

NGS has catalysed the research in metagenomics, and paved the way for scientists to build fundamental knowledge on various communities over the past decade. The technical ability to routinely generate large datasets of sequences from extremely variable communities has enabled the study of different environments in a less biased way. Part of the solution to the challenges identified above lies in the technical advances for data generation, especially longer reads. The field is finally mature enough to think “beyond the metagenomics” and advance our understanding of microbial communities and different environments by including such aspects as, for example, host-pathogen interactions, quorum sensing and cell-to-cell interactions. This would require collecting additional experimental measurements and metadata, and in turn call for yet another generation of computational methods for what we can call *integrative metagenomics*.

Acknowledgements The authors would like to thank the anonymous reviewers for their helpful comments. This work was partly supported by Région Aquitaine grant 2014.

References

- Abeln, M. B. S., Zaura, E., and Brandt, B. (2012). Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics*, 28:2891–2897.
- Albanese, D., Fontana, P., Filippo, C. D., Cavalieri, D., and Donati, C. (2015). Micca: a complete and accurate software for taxonomic profiling of metagenomic data. *Scientific Reports*, 5:e9743.
- Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K., Tyson, G., and Nielsen, P. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, 31:533–538.
- Alneberg, J., Bjarnason, B., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U., Lahti, L., Loman, N., Andersson, A., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11:1144–1146.

- Alonso-Aleman, D., Barre, A., Beretta, S., Bonizzoni, P., Nikolski, M., and Valiente, G. (2014). Further steps in TANGO: improved taxonomic assignment in metagenomics. *Bioinformatics*, 30(1):17–23.
- Ames, S., Hysom, D., Gardner, S., Lloyd, G., Gokhale, M., and Allen, J. (2013). Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics*, 29(18):2253–2260.
- Arumugam, M., Harrington, E., Foerstner, K., Raes, J., and Bork, P. (2010). Smashcommunity: a metagenomic annotation and analysis tool. *Bioinformatics*, 26(23):2977–2978.
- Baran, Y. and Halperin, E. (2012). Joint analysis of multiple metagenomic samples. *PLoS Computational Biology*, 8(2):e1002373.
- Bazinet, A. and Cummings, M. (2012). A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, 13:1–13.
- Bella, J., Bao, Y., Gloor, G., Burton, J., and Reid, G. (2013). High throughput sequencing methods and analysis for microbiome research. *Journal of Microbiological Methods*, 95(3):401–414.
- Blaxter, M., Mann, J., Chapman, T., Thomas, F., and Abebe, C. W. R. F. E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462):1935–1943.
- Bork, P., Bowler, C., de Vargas, C., Gorsky, G., Karsenti, E., and Wincker, P. (2015). Tara oceans studies plankton at planetary scale. *Science*, 348(6237):873–873.
- Borozan, I., Watt, S., and Ferretti, V. (2015). Integrating alignment-based and alignment-free sequence similarity measures for biological sequence classification. *Bioinformatics*, 31(9):e1396.
- Brady, A. and Salzberg, S. (2009). Phymm and PhymmBL: Metagenomic phylogenetic classification with Interpolated Markov Models. *Nature Methods*, 6(9):673–676.
- Burke, C., Kjelleberg, S., and Thomas, T. (2009). Selective extraction of bacterial DNA from the surfaces of macroalgae. *Applied and Environmental Microbiology*, 75(1):252–256.
- Cai, Y. and Sun, Y. (2011). ESPRIT-Tree: hierarchical clustering analysis of millions of 16s rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Research*, 39(14):e95.
- Chatterjee, S., Koslicki, D., Dong, S., Innocenti, N., Cheng, L., Lan, Y., Vehkaperä, M., Skoglund, M., Rasmussen, L., Aurell, E., and Corander, J. (2014). SEK: sparsity exploiting k -mer-based estimation of bacterial community composition. *Bioinformatics*, 30(17):2423–2431.
- Chatterji, S., Yamazaki, I., Bai, Z., and Eisen, J. (2008). CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. In *Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology*, Lecture Notes in Computer Science, pages 17–28.
- Chaudhary, N., Sharma, A., Agarwal, P., Gupta, A., and Sharma, V. (2015). 16s classifier: A tool for fast and accurate taxonomic classification of 16s rRNA hypervariable regions in metagenomic datasets. *PLoS ONE*, 10(2):e0116106.
- Cheng, L., Walker, A., and Corander, J. (2012). Bayesian estimation of bacterial community composition from 454 sequencing data. *Nucleic Acids Research*, 40(12):5240–5249.
- Cleary, B., Brito, I., Huang, K., Gevers, D., Shea, T., Young, S., and Alm, E. (2015). Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nature Biotechnology*, 33(10):1053–1060.

- Coetzee, B., Freeborough, M., Maree, H., Celton, J.-M., Rees, D., and Burger, J. (2010). Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard. *Virology*, 400(2):157–163.
- Cole, J., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R., Kulam-Syed-Mohideen, A., McGarrell, D., Marsh, T., Garrity, G., and Tiedje, J. M. (2009). The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37(Database issue):D141–D145.
- Cuadros-Orellana, S., Leite, L., Smith, A., Medeiros, J., Badotti, F., Fonseca, P., Vaz, A., Oliveira, G., and Gooes-Neto, A. (2013). Assessment of fungal diversity in the environment using metagenomics: a decade in review. *Fungal Genomics and Biology*, 3:e110.
- Cui, H. and Zhang, X. (2013). Alignment-free supervised classification of metagenomes by recursive SVM. *BMC Genomics*, 14:e641.
- Darling, A., Jospin, G., Lowe, E., Bik, F. M. H., and Eisen, J. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, 2:e243.
- Delcher, A., Harmon, D., Kasif, S., White, O., and Salzberg, S. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 27(23):4636–4641.
- Delmont, T., Robe, P., Clark, I., Simonet, P., and Vogel, T. (2011). Metagenomic comparison of direct and indirect soil DNA extraction approaches. *Journal of Microbiological Methods*, 86(3):397–400.
- DeSantis, T., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. (2006). Greengenes, a chimera-checked 16s rRNA gene database and workbench compatible with arb. *Applied and Environmental Microbiology*, 72(7):5069–5072.
- Dick, G., Andersson, A., Baker, B., Simmons, S., Yelton, A., and Banfield, J. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biology*, 10(8):1–16.
- Ding, X., Cheng, F., Cao, C., and Sun, X. (2015). DectICO: an alignment-free supervised metagenomic classification method based on feature extraction and dynamic selection. *BMC Bioinformatics*, 16:e323.
- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.
- Edgar, R. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461.
- Edgar, R. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, 10(10):996–998.
- Eisen, J. (2007). Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biology*, 5(3):e82.
- Filippo, C. D., Ramazzotti, M., Fontana, P., and Cavalieri, D. (2012). Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Briefings in Bioinformatics*, 13(6):696–710.
- Flach, P. (2012). *Machine Learning. The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press.
- Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152.

- Fuhrman, J. (2009). Microbial community structure and its functional implications. *Nature*, 459:193–199.
- Gans, J., Wolinsky, M., and Dunbar, J. (2005). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science*, 309:1387–90.
- Ghodsi, M., Liu, B., and Pop, M. (2011). DNACLUSt: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics*, 12:e271.
- Hao, X., Jiang, R., and Chen, T. (2011). Clustering 16s rRNA for OTU prediction: a method of unsupervised bayesian clustering. *Bioinformatics*, 27(5):611–618.
- Hoff, K., Lingner, T., Meinicke, P., and Tech, M. (2009). Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Research*, 37(suppl 2):W101–W105.
- Hoff, K., Tech, M., Lingner, T., Daniel, R., Morgenstern, B., and Meinicke, P. (2008). Gene prediction in metagenomic fragments: A large scale machine learning approach. *BMC Bioinformatics*, 9:e217.
- Hugenholtz, P. and Pace, N. (1996). Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends Biotechnol.*, 14:190–197.
- Hunter, S., Corbett, M., Denise, H., Fraser, M., Gonzalez-Beltran, A., Hunter, C., Jones, P., Leinonen, R., McAnulla, C., Maguire, E., Maslen, J., Mitchell, A., Nuka, G., Oisel, A., Radhakrishnan, S. P. R., Rocca-Serra, P., Scheremetjew, M., Sterk, P., Vaughan, D., Cochrane, G., and Sansone, D. F. S.-A. (2014). Ebi metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research*, 42(D1):D600–D606.
- Hurwitz, B., Deng, L., Poulos, B., and Sullivan, M. (2013). Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environmental Microbiology*, 15:1428–1440.
- Huse, S., Welch, D., Morrison, H., and Sogin, M. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol*, 12(7):1889–1898.
- Hyatt, D., Chen, G., LoCascio, P., Land, M., Larimer, F., and Hauser, H. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:e119.
- Hyatt, D., LoCascio, P., Hauser, L., and Uberbacher, E. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, 28(17):2223–2230.
- Jaaskinen, V., Parkkinen, V., Cheng, L., and Corander, J. (2013). Bayesian clustering of DNA sequences using Markov chains and a stochastic partition model. *Statistical applications in genetics and molecular biology*, 13:105–121.
- Jiang, B., Song, K., Ren, J., Deng, M., Sun, F., and Zhang, X. (2012). Comparison of metagenomic samples using sequence signatures. *BMC Genomics*, 13(1):e730.
- John, S., Mendez, C., Deng, L., Poulos, B., Kauffman, A., Kern, S., Brum, J., Polz, M., Boyle, E., and Sullivan, M. (2011). A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environmental Microbiology Reports*, 3:195–202.
- Jumpstart Consortium Human Microbiome Project Data Generation Working Group (2012). Evaluation of 16s rDNA-based community profiling for human microbiome research. *PLoS ONE*, 7:e39315.
- Kang, D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165.

- Kariin, S. and Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends in genetics*, 11(7):283–290.
- Kelley, D., Liu, B., Delcher, A., Pop, M., and Salzberg, S. (2012). Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Research*, 40(1):e9.
- Kelley, D. and Salzberg, S. (2010). Clustering metagenomic sequences with interpolated markov models. *BMC Bioinformatics*, 11:e544.
- Kislyuk, A., Bhatnagar, S., Dushoff, J., and Weitz, J. (2009). Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinformatics*, 10:e316.
- Koeppel, A. and Wu, M. (2013). Surprisingly extensive mixed phylogenetic and ecological signals among bacterial operational taxonomic units. *Nucleic Acids Research*, 41(10):5175–5188.
- Koetschan, C., Forster, F., Keller, A., Schleicher, T., Ruderisch, B., Schwarz, R., Muller, T., Wolf, M., and Schultz, J. (2010). The ITS2 database iii—sequences and structures for phylogeny. *Nucleic Acids Research*, 38(suppl 1):D275–D279.
- Koslicki, D., Chatterjee, S., Shahrivar, D., Walker, A., Francis, S., Fraser, L., Vehkaperä, M., Lan, Y., and Corander, J. (2015). Ark: Aggregation of reads by k-means for estimation of bacterial community composition. *PloS one*, 10(10):e0140644.
- Koslicki, D., Foucart, S., and G. Rosen, G. (2013). Quikr: a method for rapid reconstruction of bacterial communities via compressive sensing. *Bioinformatics*, 29(17):2096–2102.
- Koslicki, D., Foucart, S., and Rosen, G. (2014). WGSQuikr: Fast whole-genome shotgun metagenomic classification. *PLoS ONE*, 9(3):1–5.
- Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental microbiology*, 12(1):118–123.
- Kurtz, S., Narechania, A., Stein, J., and Ware, D. (2008). A new method to compute k-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics*, 9:e517.
- Li, W. (2009). Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics*, 10:e359.
- Libbrecht, M. and Noble, W. (2015). Machine learning applications in genetics and genomics. *Nat Rev Genet*, 16(6):321–332.
- Lindahl, B., Nilsson, R., Tedersoo, L., Abarenkov, K., Carlsen, T., Kjäller, R., Kõljalg, U., Pennanen, T., Rosendahl, S., Stenlid, J., and Kauserud, H. (2013). Fungal community analysis by high-throughput sequencing of amplified markers—a user’s guide. *The New Phytologist*, 199(1):288–299.
- Liu, B. (2012). Computational metagenomics: Network, classification and assembly. PhD thesis, University of Maryland, USA.
- Ludwig, W. and Schleifer, K. (1994). Bacterial phylogeny based on 16s and sequence analysis. *FEMS Microbiology Reviews*, 15:155–173.
- Mahe, F., Rognes, T., and C. de Vargas, C. Q., and Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2:e593.
- Matsen, F. and Armbrust, R. K. E. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11(1):e538.

- McDonald, D., Price, M., Goodrich, J., Nawrocki, E., DeSantis, T., Probst, A., Andersen, G., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.*, 6(3):610–618.
- McHardy, A., Martín, H., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, 4:63–72.
- McHardy, A. and Rigoutsos, I. (2007). What’s in the mix: phylogenetic classification of metagenome sequence samples. *Current opinion in microbiology*, 10(5):499–503.
- Meinicke, P., Aßhauer, K., and Lingner, T. (2011). Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics*, 27(12):1618–1624.
- Mende, D., Waller, A., Sunagawa, S., Jarvelin, A., Chan, M., Arumugam, M., Raes, J., and Bork, P. (2012). Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS One*, 7(2):e31386.
- Minot, S., Grunberg, S., Wu, G., Lewis, J., and Bushman, F. (2012). Hypervariable loci in the human gut virome. *Proceedings of the National Academy of Sciences*, 109(10):3962–3966.
- Mitchell, T. (1997). *Machine Learning*. McGraw-Hill, Inc. New York, NY, USA.
- Mitra, S., Klar, B., and Huson, D. (2009). Visual and statistical comparison of metagenomes. *Bioinformatics*, 25(15):1849–1855.
- Nielsen, H. and et al., M. A. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, 32:822–828.
- Noguchi, H., Park, J., and Itoh, T. (2006). Metagene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Research*, 34(19):5623–5630.
- Nogucki, H., Tamiguchi, T., and Itoh, T. (2008). MetaGeneAnnotator: Detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 15(6):387–396.
- Pace, N., Stahl, D., Lane, D., and Olsen, G. (1986). The analysis of natural microbial populations by ribosomal rna sequences. *Advances in Microbial Ecology*, 9:1–55.
- Parks, D., MacDonald, N., and Beiko, R. (2011). Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics*, 12:328–328.
- Patil, K., Rouné, L., and McHardy, A. (2012). The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. *PLoS ONE*, 7(6):e38581.
- Patwardhan, A., Ray, S., and Roy, A. (2014). Molecular markers in phylogenetic studies - a review. *Journal of Phylogenetics & Evolutionary Biology*, 2:e131.
- Rasheed, Z. and Rangwala, H. (2012). Metagenomic taxonomic classification using extreme learning machines. *Journal of bioinformatics and computational biology*, 10(05):e1250015.
- Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research*, 38(20):e191–e191.
- Rosen, G., Reichenberger, E., and Rosenfeld, A. (2011). NBC: the Nave Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, 27:127–129.
- Roux, S., Krupovic, M., Debroas, D., Forterre, P., and Enault, F. (2013). Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biology*, 3(12):e130160.

- Russell, D., Way, S., Benson, A., and Sayood, K. (2010). A grammar-based distance metric enables fast and accurate clustering of large sets of 16s sequences. *BMC Bioinformatics*, 11:e601.
- Santamaria, M., Fosso, B., Consiglio, A., Caro, G. D., Grillo, G., Licciulli, F., S. Liuni, S., Marzano, M., Alonso-Aleman, D., Valiente, G., and Pesole, G. (2012). Reference databases for taxonomic assignment in metagenomics. *Briefings in Bioinformatics*, 13(6):682–695.
- Sato, K. and Sakakibara, Y. (2015). MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. *DNA research*, 22(1):69–77.
- Schloss, P., Patrick, D., and Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology*, 71(3):1501–1506.
- Schloss, P. and Westcott, S. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16s rRNA gene sequence analysis. *Applied and Environmental Microbiology*, 77:3219–3226.
- Schloss, P., Westcott, S., Ryabin, T., Hall, J., Hartmann, M., Hollister, E., Lesniewski, R., Oakley, B., Parks, D., Robinson, C., Sahl, J., Stres, B., Thallinger, G., Horn, D. V., and Weber, C. (2009). Introducing Mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–7541.
- Seth, S., Valimaki, N., Kaski, S., and Honkela, A. (2014). Exploration and retrieval of whole-metagenome sequencing samples. *Bioinformatics*, 30(17):2471–2479.
- Sharon, I., Morowitz, M., Thomas, B., Costello, E., Relman, D., and Banfield, J. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Research*, 23(1):111–120.
- Soueidan, H., Schmitt, L., Candresse, T., and Nikolski, M. (2014). Finding and identifying the viral needle in the metagenomic haystack: Trends and challenges. *Frontiers in Microbiology*, 5:e739.
- Sun, Y., Cai, Y., Huse, S., Knight, R., Farmerie, W., Wang, X., and Mai, V. (2011). A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Briefings in bioinformatics*, 13:107–121.
- Sun, Y., Cai, Y., Liu, L., Farrell, F. Y. M., McKendree, W., and Farmerie, W. (2009). ESPRIT: estimating species richness using large collections of 16s rRNA pyrosequences. *Nucleic Acids Research*, 37(10):e96.
- Suzek, B., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. (2007). UniRef: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288.
- Tanaseichuk, O., Borneman, J., and Jiang, T. (2012). Separating metagenomic short reads into genomes via clustering. *Algorithms for Molecular Biology*, 7:e27.
- Tanaseichuk, O., Borneman, J., and Jiang, T. (2014). Phylogeny-based classification of microbial communities. *Bioinformatics*, 30:449–456.
- Teeling, H. and Glockner, F. (2012). Current opportunities and challenges in microbial metagenome analysis - bioinformatic perspective. *Briefings in Bioinformatics*, 13(6):728–742.
- Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glockner, F. (2004). Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatics*, 5:1471–2105.

- Trimble, W., Keegan, K., D'Souza, M., Wilke, A., Wilkening, J., Gilbert, J., and Meyer, F. (2012). Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. *BMC Bioinformatics*, 13:183–183.
- Vervier, K., Mahé, P., Tournoud, M., Veyrieras, J.-P., and Vert, J.-P. (2015). Large-scale machine learning for metagenomics sequence classification. *Bioinformatics*, page e683.
- Wang, Q., Garrity, G., Tiedje, J., and Cole, J. (2007). Naive Bayesian Classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–5267.
- Wang, Y., Leung, H., Yiu, S., and Chin, F. (2012a). MetaCluster 4.0: A novel binning algorithm for ngs reads and huge number of species. *Journal of Computational Biology*, 19(2):241–249.
- Wang, Y., Leung, H., Yiu, S., and Chin, F. (2012b). MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics*, 28(18):i356–i362.
- Weinberger, K., Dasgupta, A., Langford, J., Smola, A., and Attenberg, J. (2009). Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1113–1120. ACM.
- Wooley, J. and Yuzhen, Y. (2009). Metagenomics: Facts and artifacts, and computational challenges. *Journal of Computer Science and Technology*, 25(1):71–81.
- Wu, Y., Tang, Y., Tringe, S., Simmons, B., and Singer, S. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, 2(26):e26.
- Wu, Y. and Ye, Y. (2011). A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *Journal of Computational Biology*, 18(3):523–534.
- Yang, C., Mills, D., Mathee, K., Wang, Y., Jayachandran, K., Sikaroodi, M., Gillevet, P., Patrick, Entry, J., and Narasimhan, G. (2006). An ecoinformatics tool for microbial community studies: Supervised classification of amplicon length heterogeneity (ALH) profiles of 16s rRNA. *Journal of Microbiological Methods*, 65(1):49–62.
- Z. Liu, Z., Hsiao, W., Cantarel, B., Drábek, E., and Fraser-Liggett, C. (2011). Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics*, 27(23):3242–3249.
- Zhou, J., Deng, Y., Luo, F., He, Z., Tu, Q., and Zhi, X. (2010). Functional molecular ecological networks. *mBio*, 1(4):e00169–10.
- Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research*, 38(12):e132–e132.