Analysis Methods for Shotgun Metagenomics

Stephen Woloszynek, Zhengqiao Zhao, Gregory Ditzler, Jacob R. Price, Erin R. Reichenberger, Yemin Lan, Jian Chen, Joshua Earl, Saeed Keshani Langroodi, Garth Ehrlich, and Gail Rosen

Abstract The development of whole metagenome shotgun sequencing (WGS) has enabled the precise characterization of taxonomic diversity and functional capabilities of microbial communities in situ while obviating organism isolation and cultivation procedures. WGS created with second- and third-generation sequencing technologies will generate millions of reads and tens (or hundreds) of gigabytes of information about the organisms under investigation. Despite

Author Contributions: SW, abstract, taxonomic binning, taxonomic classification, normalization, feature selection, feature extraction, distance-based approaches, neural network approaches, statistical inference, machine learning, drafted and ordered sub-sections, coordinated co-authors; ZZ, taxonomic classification, machine learning; GD, diversity metrics, feature selection, feature extraction; JRP, abstract, diversity metrics, distance-based approaches, diversity metrics; ERR, abstract, introduction; YL, functional annotation; JC, neural network approaches; JE, feature selection, feature extraction; SKL, taxonomic binning; GR, taxonomic classification, discussion, drafted sub-sections; all authors contributed to editing and revising.

S. Woloszynek · Z. Zhao · J. Chen · G. Rosen (✉)
Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, USA
e-mail: gailr@coe.drexel.edu

G. Ditzler
Department of Electrical and Computer Engineering, University of Arizona, Tuscon, AZ, USA

J. R. Price · S. K. Langroodi
Department of Civil, Architectural, and Environmental Engineering, Drexel University, Philadelphia, PA, USA

E. R. Reichenberger
U.S. Department of Agriculture, Agricultural Research Service, Eastern Regional Research Center, Wyndmoor, PA, USA

Y. Lan
Department of Cell and Developmental Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

J. Earl · G. Ehrlich
Department of Microbiology and Immunology, Drexel University College of Medicine, Philadelphia, PA, USA

F. A. B. da Silva et al. (eds.), *Theoretical and Applied Aspects of Systems Biology*, Computational Biology 27, https://doi.org/10.1007/978-3-319-74974-7_5

containing an immense amount of information, the reads are unorganized and unlabeled, leading to a significant challenge in discerning from which genome a read originated. Thus, analysis of WGS data necessitates first determining community structure and function from the raw reads before the focus can shift to making multi-sample comparisons. A typical WGS workflow consists of read assignment (taxonomic binning and classification), preprocessing techniques (normalization, dimensionality reduction), exploratory approaches (feature selection and extraction, ordination), statistical inference (regression, constrained ordination, differential abundance analysis), and machine learning. The following chapter provides an overview of these analytical approaches (including challenges and possible pitfalls that may be encountered by researchers) as well as steps toward their solutions. Relevant software packages and resources are also discussed.

1 Introduction to Metagenomics

The term "metagenome" originated with Handelsman et al. who defined it as a collection of genomes found in the microflora of soil and described an approach used to access the organisms living in this ecosystem [1]. Their motivation was influenced by a continual decline in the discovery of new compounds from an environment that had previously provided researchers and industry with chemicals that were antimicrobial or otherwise medicinal in nature. The paucity of newly discovered compounds followed the realization that many microbes were not culturable and that microbiologists had greatly underestimated both their numbers and diversity [1–7]. The reasons behind a microbe's resistance to culturing vary; their survival may be dependent upon compounds provided by other resident organisms, and/or the conditions (e.g., temperature, atmospheric pressure, gaseous elements (along with their amounts)) may be inadequate for their survival [2]. Regardless of the cause, it became apparent that the number of organisms that could not be cultured greatly surpassed the number of microorganisms that could be cultured [1, 2, 5, 6, 8–11]. Combined, these elements drove a new and oft-interdisciplinary field known as metagenomics – the study of uncultured genetic material acquired directly from environmental communities that contain a motley population of organisms. Ensuing from these developments was the inception of numerous large-scale metagenomic studies that investigated microbial communities in water, soil, and animals [12–16]. Information acquired from these studies have exposed the intricate influence and beauty of microbes on processes as vast as the geochemical to human health.

Although specimen isolation and cultivation are not required, sophisticated computational tools are a necessity in metagenomic analysis. This analysis has been aided greatly by advances in sequencing technology, which have yielded increased accuracy in base pair identification, longer reads, and decreases in sequencing costs. The reduction in sequencing pricing as well as faster computer processors have made metagenomic analysis more accessible to institutions and laboratories looking to investigate microbial communities. As such, clinical studies and research related

to quorum sensing, antibiotic resistance, biofilms, bacteriophages, and food science along with other areas of interest have become far more common [17–23].

Knowledge obtained from these studies owe much not only to the individuals involved with undertaking the studies and improving sequencing technology but also those who have developed the algorithms and methods used to analyze next-generation sequencing (NGS) and metagenomic data [8, 24–31]. Metagenomic studies often revolve around determining what is in the sample (classification), how many organisms are in the sample (binning), and what they are doing (functional annotation). Additionally, researchers are interested in comparing samples (normalization, clustering, ordination) and determining similarities between samples (feature selection/extraction). These methods are often accomplished with machine learning techniques, and each of the aforementioned topics will be addressed in this chapter.

2 Sequence Quality and Identification

2.1 Introduction to Taxonomic Binning

High throughput whole metagenome shotgun sequencing (WGS) is a reliable technique used to characterize taxonomic diversity and function of microbial communities without cultivation of the microorganisms in a laboratory environment. After WGS, the primary goal is then to infer microbial community structure and function in the given microbiome from the millions of *unlabeled* genomic fragments (known as "reads") [32]. This is no easy task, however, since algorithmic approaches are necessary to discern taxonomic information. Extracting information from sequencing reads has accordingly been equated to simultaneously completing multiple puzzles with their pieces shuffled together [33]. While full-genome assembly is potentially an effective method for this purpose, constructing complete genomes from short reads often fails for many reasons including repetitive nucleotide patterns found within genomes, homologous regions of closely related regions, and conserved regions among different species [34, 35].

Binning is considered an alternative to full-length genome assembly [36]. Despite still relying on sequencing reads, binning is capable of approximating population composition and functional diversity of assigned genomes [37, 38]. There are two binning methods developed for disentangling metagenomic reads: "supervised" (taxa-dependent; classification) and "unsupervised" (taxa-independent; clustering). Supervised binning uses one or more phylogeny-based comparisons that involve aligning reads to reference genomes, assessing sequence composition properties such as GC content and oligonucleotide patterns (k-mers), and utilizing hybrid methods that leverage both alignment and sequence composition approaches [36, 39]. Supervised binning is often not effective for environmental samples or diverse microbial communities; however, due to bias with respect to previously sequenced

or well-studied species, many of the reference databases in which supervised approaches rely are incomplete [33], which results in many reads going unassigned or being assigned incorrectly. Also, metagenomes with high interspecies diversity fail to be accurately classified by supervised binning tools [40].

Unsupervised binning, on the other hand, relies on discriminative nucleotides, sequence composition, and taxa abundance, which is inferred in terms of contig coverage [41, 42]. Binning techniques that rely on sequence composition assume that each taxon has a unique genomic signature, which is represented as k-mer frequency vectors (Fig. 1). Example tools include 2Tbinning, LikelyBin, Metawatt, SCIMM, self-organizing maps, and VizBin (Table 1). For low-abundance taxa, composition-based techniques are prone to incorrect taxon assignments since the generated clusters for these taxa tend to be poorly described [33]. In addition, they typically require high-quality reads or contigs that are over 1000 bp in length to achieve acceptable accuracy [43]. Abundance-based techniques are much better at handling low-abundance taxa and shorter reads. For single-sample studies, limitations associated with low-abundance taxa are mitigated by enforcing distributional assumptions (e.g., the Lander-Waterman model) to the k-mer abundance coverage profile. For multi-sample studies, the taxa abundance profiles are assumed to be correlated between samples [33]. Abundance-based techniques include AbundanceBin, Canopy, and MBBC. Lastly, hybrid techniques that utilize both sequence composition and taxa abundance include COCACOLA, CompostBin, CONCOCT, differential coverage binning, GroopM, MaxBin, MetaBAT, MetaCluster, and MyCC. For a detailed review of unsupervised binning approaches, see Sedlar et al. [33].

Selection of binning methods depends on the purpose of the metagenomic study, the computational requirements, as well as the time constraints. In supervised methods, the length of metagenomic reads, which is in turn dependent upon the sequencing platform, is also a factor [44–46]. In addition, read coverage must also be considered since greater coverage may capture rare species with more accurate results. On the other hand, unsupervised binning is effective for diverse microbiomes or low-coverage datasets [36]. To improve binning results, preprocessing (e.g., quality filtering of the sequencing reads) and post-processing techniques which use different reassembly approaches (e.g., mapping reads to the bins before reassembly) are options [47–49].

2.2 Taxonomic Classification

A variety of tools are currently available that perform taxonomic classification. These include methods that rely on a subset of marker genes (MetaPhlAn [50], MetaPhyler [51], mOTU [52], MicrobeCensus [53], GOTTCHA [54]), and those that use exploit the entire set of reads, using composition-based approaches, such as alignment (MEGAN [55]) or k-mer enumeration (CLARK [56], Kraken [57], LMAT [58], MetaFlow [59], NBC [60], and PhyloSift [61]) [62, 63]. Approaches

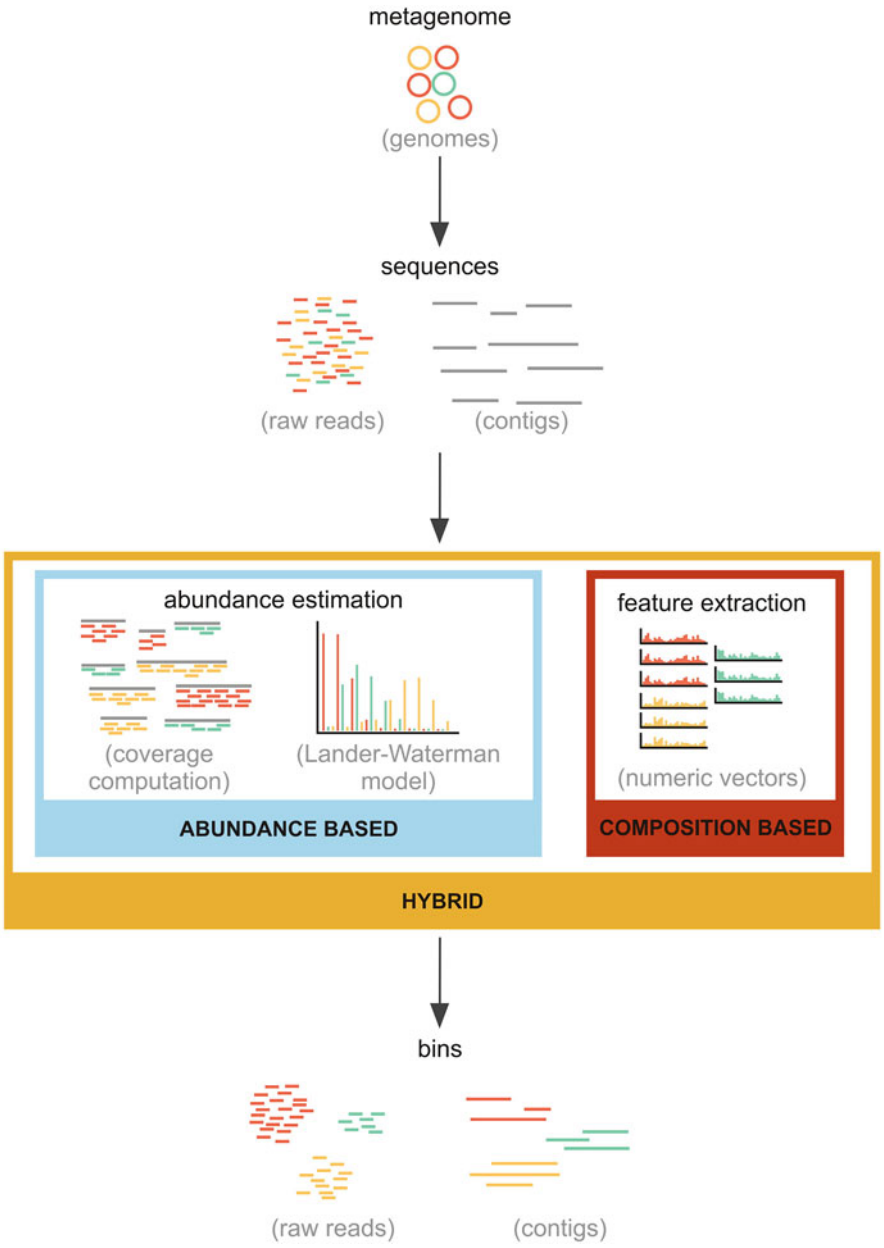

Fig. 1 Unsupervised binning workflow, originally presented in Sedlar et al. [33]

Table 1 Unsupervised binning tools. (Adapted from Sedlar et al. [33])

Method	Input data	Programming language	Type	Source
SOM	Raw reads or contigs	Perl	Composition	https://github.com/tetramerFreqs/Binning
LikelyBin	Raw reads	Perl, C	Composition	http://ecotheory.biology.gatech.edu/downloads/likelybin
SCIMM	Raw reads or contigs	Python	Composition	http://www.cbcb.umd.edu/software/scimm/
MetaWatt	Assembled contigs	Java	Composition	https://sourceforge.net/projects/metawatt/
VizBin	Contigs	Java	Composition	https://claczny.github.io/VizBin/
AbundanceBin	Raw reads	C++	Abundance	http://omics.informatics.indiana.edu/AbundanceBin/
Canopy	Gene abundance profiles	C++	Abundance	https://bitbucket.org/HeyHo/mgs-canopy-algorithm/wiki/Home
MBBC	Raw reads	Java	Abundance	http://eecs.ucf.edu/~xiaoman/MBBC/MBBC.html
CompostBin	Raw reads	C, MATLAB	Hybrid	https://sites.google.com/site/souravc/compostbin
MetaCluster	Raw reads (only pair-ends)	C++	Hybrid	http://i.cs.hku.hk/~alse/MetaCluster/index.html
DCB	Raw reads	R	Hybrid	https://github.com/MadsAlbertsen/multi-metagenome
CONCOCT	Contigs + BAM	Python	Hybrid	https://github.com/BinPro/CONCOCT
MaxBin	Contigs + (reads or abundance file)	Perl	Hybrid	https://sourceforge.net/projects/maxbin/
GroopM	Contigs + BAM	Python	Hybrid	http://ecogenomics.github.io/GroopM/
MetaBAT	Contigs + BAM	C++	Hybrid	https://bitbucket.org/berkeleylab/metabat
COCACOLA	Contigs + raw reads	MATLAB	Hybrid	https://github.com/younglululu/COCACOLA
MyCC	Contigs + BAM*optional	Python	Hybrid	https://sourceforge.net/projects/sb2nhri/files/MyCC/

that utilize sequences, while slower, are enticing in their ability to leverage additional information for assembly and contamination detection, for example [62].

Marker gene approaches are faster than composition-based approaches but are limited in the number of reads they can ultimately classify [62]. They vary from method to method mostly in terms of their marker gene database construction. Composition-based approaches, on the other hand, differ more algorithmically [63]. For example, CLARK performs classification by first identifying discriminative k-mers that uniquely characterize reference sequences, which it then uses to classify query reads based on the number of shared k-mers. Kraken is similar in that it uses the number of overlapping reads between query and reference to influence the classification; however, it leverages phylogenetic information during the mapping step, building a phylogenetic tree. The reference sequence is identified by determining the lowest common ancestor that contains the k-mer from the query. Other k-mer approaches that leverage phylogenetic information include PhyloSift and LMAT. MetaFlow treats classification as a query-to-reference matching problem, using a bipartite graph. Lastly, NBC is a metagenome fragment classification tool using k-mer frequency profiles. In short, this tool trains an NBC classifier based on the frequency of k-mers. Here, $\mathbf{X} = [x_1, x_2, \cdots, x_n]$ is the set of k-mers in a sequence. In the training phase, $p(x_i|C_k)$ is estimated by the total number of k-mers x_i occurring in all the training sequences that are labeled by C_k. In the testing phase, given a query sequence, the organism containing the sequence is predicted by the class that maximizes the posterior probability $P(C_k|\mathbf{X})$.

To evaluate the performance of the tools described above, McIntyre et al. designed an analysis involving 846 species across 67 simulated and datasets [62]. The performances were evaluated by each tool's ability to (1) identify taxa in a sample at genus, species, and strain levels, (2) estimate the relative abundances of taxa in a sample, and (3) classify individual reads at the species level. For taxa identification, all tools performed optimally at the genus level, but the performance dropped noticeably at the strain level. They also determined that the performance of k-mer-based tools could be improved by introducing an abundance threshold. Read depth was another important identified factor that had an effect on performance; they found a positive relationship between the number of recovered species and read depth. BLAST-MEGAN and PhyloSift were two exceptions, but this trend could be dampened with the addition of adequate filtering. On the other hand, read depth had little impact on marker gene-based tools. The authors also showed that an ensemble classifier that combined the results from the best performing tools could produce improved results in quantifying the number of species. Combining their approach with BLAST greatly improved performance; however, because BLAST is notoriously slow, a faster ensemble showed comparable performance. For relative abundance comparisons, the authors showed that most of the tools could predict the proportion of a particular species in a sample to within a few percentage points. CLARK slightly overestimated relative abundance, but had greater precision compared to other tools. k-mer-based methods achieved the highest recall with lower sequencing depth.

Long but low quality reads generated by newer sequencing platforms are becoming more readily available. For long, lower quality reads, CLARK and Diamond-MEGAN performed more robustly than other tools. For classifying individual reads, BLAST-MEGAN gave the best precision, whereas CLARK generally gave the best recall. The last considerations were runtime and memory. The authors benchmarked all tools under the same conditions and showed that MetaPhlAn, GOTTCHA, PhyloSift, and NBC used less memory; NBC and BLAST were the slowest; and CLARK, GOTTCHA, Kraken, MetaFlow, MetaPhlAn, Diamond-Megan, and LMAT were the fastest. The authors provided a decision tree summary of usage recommendations (Fig. 2).

It should be noted that despite the large study size, 846 species is only a small subset of all species that exist. Also, the ability of a given tool to identify "unknown" organisms was never evaluated. This is highlighted by the fact that as the read depth increased, most classifiers discovered more species – leaving a perplexing open problem in metagenomic taxonomic classification. Therefore, more research should be done to determine how database size affects classification, as well as how other parameters may affect classifier performance.

2.3 Functional Annotation

Unraveling the functional composition of metagenomes is crucial to understanding the microbe's metabolic dynamics and how they shape the environment or adapt to environmental changes. From either assembled individual genomes or the metagenome as an entity, protein-coding genes can be predicted by scanning the sequences for start/stop codons. However, gene prediction and the following functional profiling do not depend on full gene sequences. Functional profiling can be achieved using short reads directly, as they may be highly similar to gene sequence fragments or contain characteristic protein domains for recognition. As easy as it sounds, functional profiling of metagenomes remains challenging. One of the fundamental difficulties is that metagenomic sequences can be highly divergent in comparison to genes and proteins currently identified [64]. Therefore, profiling tools that rely on sequence similarity are subject to a tough dilemma between sensitivity and specificity. Another difficulty is that short sequencing reads may not contain sufficient information for us to accurately infer their functions. Therefore, increasing the number of annotated reads and improving the annotation accuracy remain top challenges for tools in development for functional profiling [65].

Recently, a lot of effort has been devoted to creating an accurate knowledge base of metagenomic functions and developing reliable and scalable profiling tools. These two types of efforts are tightly coupled, and in most cases, the choice of which database to profile against also decides which profiling software/tool should be used.

As of now, various databases have been constructed, and they represent different resolutions of metagenomic function. For example, NCBIs RefSeq database [66]

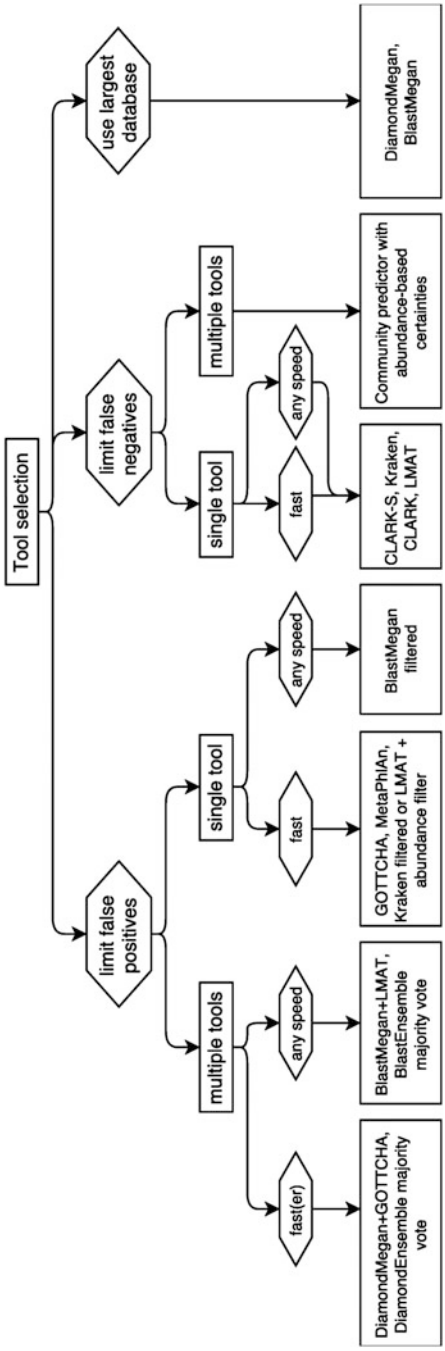

Fig. 2 Decision tree summary of usage recommendations, originally presented in McIntyre et al. [62]

and the UniProt database [67] are two of the largest reference sequence collections, both containing over 100 million annotated protein sequences. When provided with a reference database as comprehensive as these two, it is more likely to find annotated proteins that share high sequence similarity to an unknown read. However, one downside of big databases is that most of the "annotated" proteins in these databases were annotated automatically, i.e., the reference itself is subject to error. In many cases, these reference databases are adequate for profiling metagenomes and discovering significant changes between profiles. In other cases, however, one can opt for a reduced database with higher credibility, such as the Swiss-Prot database [68] which contains only a half-million annotated proteins but is manually curated and reviewed. There are also other reduced databases focusing on specific metagenomes (such as the UniMES database that hosts proteins inferred from environmental metagenomes) or datasets generated from specific metagenomic studies that can be used as reference databases of related metagenomes (such as the functional profiles generated from the Tara oceans project [69] of global ocean microbiomes and profiles generated from the Human Microbiome Project [70]).

As previously mentioned, the largest databases now contain up to 100 million annotated proteins. Although we may be able to annotate metagenomic reads with these proteins, it is not easy to interpret and understand a metagenomic profile without summarizing similar or relevant protein functions into groups. The gene ontology database is one of the many databases that strive to address this problem [71]. It annotates reference proteins with a carefully standardized vocabulary (called GO terms) and constructs a comprehensive relationship network between GO terms from the molecular level to larger pathways, as well as cellular and organismal-level systems. Therefore, we can use GO terms to profile metagenomes at molecular, pathway, or cellular levels. Besides gene ontology, several databases also summarize protein annotation into groups or hierarchical groups, such as the COG/EggNOG categorizations [72, 73], the KEGG pathways [74, 75], the MetaCyc pathways, and the SEED subsystems [76, 77]. The COG/EggNOG was generated by grouping orthologous proteins from numerous organisms into clusters, whereas KEGG, MetaCyc, and SEED group (or related) proteins are based on their related metabolic roles.

Annotating metagenomic reads using these databases – either large databases or reduced ones – relies on sequence similarity with reference proteins. Therefore, alignment-based methods such as BLAST search are often used for the functional profiling [78]. Additionally, numerous software tools were developed to make protein alignment and hence functional profiling computationally efficient. Besides individual tools, several large-scale pipelines have also been developed to annotate metagenomic data against multiple databases at once, such as IMG/M [79, 80], MG-RAST [81], MEGAN [82], and HUMAnN [83]. These pipelines stitch together multiple bioinformatics steps from raw metagenomic reads to functional profiling, making it easier for the user to interpret and compare the functional potential of different microbial communities.

Although proteins with similar functions may have evolved and become highly divergent in terms of nucleotide sequences, the protein domains they contain are

more conserved and may function independently from the rest of the protein. Therefore, grouping proteins based upon their functional protein domains is yet another way of summarizing different protein sequences into a manageable profile. One example of such an approach is Pfam [84], which is a collection of protein sequence alignments and hidden Markov models (HMMs) and provides a good repository for identifying protein families, domains, and repeats. SMART is another example of a protein domain database, which is abundant with domains in signaling, extracellular, and chromatin-associated proteins. Other protein domain databases commonly used are PROSITE [85], PANTHER [86], HAMAP [87], ProDom [88], etc. Although each of these databases can be acquired separately and many of them have specific software that can profile metagenomes against it (such as HMMER for Pfam, ScanProsite for PROSITE, and HAMAP-Scan for HAMAP), it is worth mentioning that InterPro [89] has combined signatures from all of the aforementioned domain databases, as well as several others, into a single searchable resource for functional profiling. Therefore, InterProScan (developed for InterPro) can be a very handy software package to scan metagenomic reads against most domain databases.

2.4 Normalization

After sequencing DNA from microbial communities of interest and determining the abundances of genomes or genes present in the community, the next step is to perform comparisons between samples. However, to make these comparisons requires that the abundances first be normalized because raw metagenomic abundances fail to accurately represent the true configuration of the taxonomic community. Simply put, in a given sample from an environment of interest, the total number of sequenced reads does not accurately reflect the true amount of DNA present in the environment. This is primarily due to study-level variation in sample collection, DNA extraction, library preparation, and sequencing depth [90]. Obtaining true "absolute abundance" cannot be achieved with sequencing data alone; for example, quantitative PCR would have to be performed in tandem [90]. Thus, differences in library size is often mitigated by calculating relative abundances where each count is divided by the total abundance from its corresponding sample [91]. Specifically, given a vector of J raw abundances from sample i:

$$\left[x_{i,1}, x_{i,2}, \ldots, x_{i,J}\right] \tag{1}$$

the relative abundance for raw abundance j in sample i is given by

$$x^*_{i,j} = \frac{x_{i,j}}{\sum_j x_{i,j}} \tag{2}$$

However, this approach is considered inappropriate [92]: dividing a raw abundance by its sample's sum constrains it to the unit simplex (all values for that sample must sum to 1), thereby rendering the data compositional [93]. Increases in any gene or genomes abundance is coupled with a corresponding decrease in the relative abundance of all other genes or genomes. In other words, while the absolute abundance of a particular taxon may be constant between two communities, their relative abundances will be different if the abundances of other taxa differ. This complicates interpretability and introduces spurious correlations. Thus, techniques such as linear regression and Pearson's correlation are no longer appropriate [94, 95]. A better alternative is to perform a centered log-ratio transformation (CLR) [96]:

$$x^*_{i,j} = \log \frac{x_{i,j}}{g_i} \tag{3}$$

where g_i is the geometric mean for sample i given by

$$g_i = \sqrt[J]{x_{i,1} x_{i,2} \cdots x_{i,J}} \tag{4}$$

The CLR is free of the compositional artifacts described above, but is limited by a singular covariance matrix, which may limit its use in downstream modeling approaches [93]. In addition, sparse abundance data further complicates calculating the CLR due to a zero denominator and the calculation being done in log space. This necessitates stringent filtering or, more commonly, the addition of small non-zero values (pseudo-counts), which may introduce bias [97]. Also, if the pseudo-count is set to 1 and the dataset is very sparse, then each raw abundance will be divided by a geometric mean close to 1, drastically dampening any normalization effect, and use of smaller pseudo-counts does not remedy the situation [98]. Recent work has suggested using values based on percentiles in place of the geometric mean, but whether this approach is robust to highly sparse datasets is currently unknown [92, 98].

Silverman et al. [93] has introduced a phylogeny-based normalization approach (PhILR) that utilizes the isometric log-ratio transformation (ILR), which, unlike the CLR, returns an invertible covariance matrix. The ILR scales CLR transformed abundances by taxa-level weights $\boldsymbol{p}$ and a weight matrix ψ given by the binary partitioning of the phylogenetic tree:

$$x^*_{i,j} = \mathrm{CLR}(x_{i,j})\mathrm{diag}(\boldsymbol{p})\psi^T \tag{5}$$

The taxa-level weighting allows for soft-thresholding of low-abundance taxa and may dampen the bias resulting from use of a pseudo-count.

In addition to differences in sample read depth, there remain other potential biases – most notably from biological sources. These include a gene or genome's mappability and length. First, relative abundances are often overestimated since metagenomes are represented as the proportion of mapped reads present in the

sample, and this ignores the variability of unmapped reads stemming from novel taxa or genes. Second, the probability of sequencing a read is a function of the length of the gene or genome from which the read originated. Correcting for gene-length permits gene-to-gene comparisons and is possible for well-described genes where the gene length is available; however, performing a genome-length correction is impractical due to the degree of diversity within a metagenome and the variability in genome lengths [90].

Accurately estimating the relative abundances of taxa in a metagenome can be accomplished via marker gene approaches, which circumvent issues with genome size since the marker genes themselves are well-characterized [90]. Marker gene approaches include MetaPhlAn [50], MetaPhyler [51], and MicrobeCensus [53]. Recent work has focused on calculating average genomic copy number, which corrects for the biases stemming from average genome size, genome mappability, and species richness. One approach, called MUSiCC, utilizes the median abundance of universal single-copy genes to normalize gene relative abundances. It is currently, however, only applicable to KEGG annotated data [91].

3 Comparative Analysis

3.1 Diversity Metrics and Distances

β-diversity allows us to examine the similarities and dissimilarities between multiple samples in a metagenomic study. Microbial ecologists begin by first computing a pair-wise distance matrix, $\mathbf{D} \in \mathbb{R}_{+}^{n \times n}$, where entry (i, j) is the distance between sample i and j with $i, j \in [n]$. One of the most important steps in this part of the analysis is the selection of the distance matrix. In general, microbial ecologists rarely, if ever, use the standard Euclidean distance to compare samples; rather, they use distances that are based on set theory or a distance between distributions.

The Jaccard index is a simple measure to determine the dissimilarity based solely on the presence or absence of a taxon in two samples. The index is given by

$$D_{\text{JAC}}(X_i, X_j) = 1 - \frac{|X_i \cap X_j|}{|X_i \cup X_j|} \tag{6}$$

where X_i and X_j represent a set of metagenomic features in sample i and j, respectively. One of the drawbacks to the Jaccard index is that it does not account for the magnitude of taxa presence, rather it only identifies whether the taxa were present in a sample. Bray-Curtis is another metric which, unlike the Jaccard index, has the abundances incorporated into the calculation. Formally, the Bray-Curtis dissimilarity is given by

$$D_{\mathrm{BC}}(X_i, X_j) = \frac{2C_{ij}}{S_i + S_j} \tag{7}$$

where S_i and S_j are the total number of taxa counted at both sites and C_{ij} is the sum of the lesser value for only those species in common between samples. Note that because the triangle inequality does not hold, Bray-Curtis is a dissimilarity metric and not a distance metric.

The Hellinger distance is the distance between two probability distributions, and it has been used occasionally in microbial ecology. Also, similar to Jaccard and Bray-Curtis, the Hellinger distance is bounded. Let $\mathcal{P} := \{p_j : j \in [n]\}$ and $\mathcal{Q} := \{q_j : j \in [n]\}$ be the probability distributions over two different samples that are represented by n taxa. The Hellinger distance is defined as

$$D_{\mathrm{HEL}}(\mathcal{P}, \mathcal{Q}) = \frac{1}{\sqrt{2}} \| \sqrt{\mathcal{Q}} - \sqrt{\mathcal{P}} \|_2 \tag{8}$$

where $\| \cdot \|_2$ is the ℓ^2-norm.

The aforementioned distances can all be found in traditional mathematical literature; however, given that microbial ecologists are using β-diversity in their studies, it should be the case that the distance measure being used in the analysis has some biological connection. The unique fraction metric (UniFrac) is perhaps the most widely used measure of distance in microbial ecology [99, 100]. UniFrac was proposed to measure the phylogenetic difference between microbial communities, as other measures such as Bray-Curtis, Hellinger, and Jaccard do not. The unweighted version of UniFrac, like the Jaccard index, only deals with the presence or absence of taxa. Unweighted UniFrac is implemented as follows: consider that you are provided two samples A and B, which are made up of metagenomic sequences, and build a phylogenetic tree using all available reads (see Zvelebi and Baum [101]). Color all the branches of the tree red where a path between two sequences in A exists, and perform the same operation for B but using a different color (e.g., blue). If a branch is colored both red and blue, then it is marked gray. The UniFrac distance is the ratio of the number of branches in the tree that are unique to either A or B to the total number of branches in the tree. Weighted UniFrac takes the concept of using this ratio to incorporate the frequency of the reads in the calculation.

3.2 *Feature Representation and Dimensionality Reduction*

Metagenomic datasets are often made up of thousands of features that represent abundances (i.e., the relative proportion of a protein family), and these datasets frequently have more features than the number of samples. A dataset with more features than samples is a challenging problem because the system is underdetermined. Furthermore, many of these features are often uncorrelated with sample data or

even redundant with each other. For example, consider a metagenomic dataset that is being used to find the taxa that favor a high saline environment. This dataset has 50 samples from both a high and low saline environment, and each sample is made up of 5000 taxa. We refer to the high and low saline environments as the data that describe the sample classes. Feature selection and dimensionality reduction allow us to (1) represent these samples by either the bacteria that are relevant to differentiate between high and low environments and (2) visualize the 100 samples in a 2-D or 3-D space, respectively.

Feature selection is the process of identifying a subset of features that are relevant and possibly non-redundant with the class. This feature subset allows metagenomic research to identify the informative variables in a dataset, such that there is still significant predictive power in the reduced set. It is important to know that the feature subset has variables that still have physical meaning (e.g., bacteria, protein family, etc.). In contrast, feature extraction is an approach to transform the data into a new (lower) dimensional space, and the new features are typically combinations of all the other features; however, these new features no longer have physical meaning.

3.2.1 Feature Selection

Feature selection plays a central role in nearly all tasks of the data analysis; however, many popular feature selection algorithms do not scale well with a large metagenomic dataset. Therefore, computationally cheap methods are used to remove so-called low zero variance metagenomic features. This low zero variance is not the best method to use in every situation; however, it is a good place to start to remove complexity when faced with a high degree of dimensionality in the data. Related methods exist in information theory (i.e., measuring features for the amount of mutual information between a metagenomic feature and the class [102]). The objective is to eliminate metagenomic features with low mutual information and not redundant with the other features.

More sophisticated methods exist for performing feature selection, including ones utilizing more than just variance, in addition to other probabilistic quantities. For example, the Relief algorithm examines paired samples (based on Euclidean distance) and weights features based on the samples' proximity in Euclidean space. It updates a weight matrix by determining if the features belong to the same or different classes [103]. Correlation-based feature selection (CFS) identifies features that have a high correlation with the supplied class of the sample but low correlation with other features while being less computationally intensive than Relief [104]. Both of these approaches are known as filter-based feature selection since they are classifier independent. Brown et al. provide a comprehensive review of information-theoretic filter feature selection algorithms [105].

In addition to filter-based approaches, embedded feature selection algorithms jointly optimize the feature selector and classifier. The least absolute shrinkage and selection operator (Lasso) is an approach to feature selection that optimizes a model

for linear regression [106]. However, unlike standard linear regression, Lasso adds a penalty on the ℓ_1-norm of the linear model. This penalty, which is shown in (9), forces the solution to be sparse (i.e., many entries in θ are zero), thus performing feature selection for the linear model $\theta^{\mathsf{T}}\mathbf{x}$. Lasso is formally given by

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \theta^{\mathsf{T}}\mathbf{x}_i)^2 + \lambda \sum_{j=1}^{p} |\theta_j| \tag{9}$$

where θ are the parameters of the linear model, $\mathbf{x}_i$ is the metagenomic feature vector, y_i is the class (± 1) or dependent variable, n is the number of samples, and p is the total number of metagenomic features. Bates and Tibshirani have recently adapted Lasso for compositional data, using log ratios as described above for CLR normalization [107]:

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \mu - \sum_{1 \le j < k \le p} \theta_{j,k} \log \frac{x_{i,j}}{x_{i,k}} \right)^2 + \lambda \sum_{j=1}^{p} |\theta_j| \tag{10}$$

The log-ratio Lasso differs from (9) in that it aims to detect models composed of a sparse subset of ratios as opposed to models composed of a sparse subset of regression coefficients. Ditzler et al. have implemented an open-source feature selection software tool for analyzing metagenomic and 16S datasets [108]. Lasso and other sparse regression techniques are easily implemented in glmnet, available in R, MATLAB, and Python.

3.2.2 Feature Extraction

Feature selection reduces the set of metagenomic features to a subset that is informative – potentially non-redundant – and still maintains a physical interpretation. Feature extraction is a technique for dimensionality reduction that embeds the original set of features in a lower-dimensional space (e.g., apply a linear projection of the metagenomic data vectors from $\mathbb{R}^p$ to $\mathbb{R}^2$ where $p \gg 2$). Principal component analysis (PCA) is one of the more popular projections for feature extraction. In PCA, we seek to represent the p-dimensional data in a lower-dimensional space that maximizes the variance of the projections. It turns out these projections are the eigenvectors of the covariance matrix of the data that correspond to the largest eigenvalues. Note that PCA does not take the class into account when the projections are calculated. Sparse PCA can also be performed for feature extraction [109], where the difference between PCA is that the projection is made by adding a sparsity constraint on the input metagenomic features. Note that this form of feature extraction will result in a new set features that have a high variance; however, these features do not have any biological meaning because the new feature set is made up of linear combinations of all other features. There is also supervised

principal component analysis (SPCA) that takes the classes into account to find the greatest degree of variance between classifications [110]. Linear discriminant analysis (LDA) is another linear transformation that reduces the dimensionality of the data that uses supervised information. The dimensionality of the reduced space is limited $C - 1$, where C is the number of classes, whereas supervised PCA does not have this limitation. LDA is also difficult to use on many metagenomic datasets because it suffers from the small sample size problem – thus, making supervised PCA a more appropriate feature extraction technique that takes into account the class separability as well.

Other related techniques include independent component analysis (ICA), nonnegative matrix factorization (NMF), and canonical correlation analysis (CCA) [111]. ICA performs a linear transformation that, unlike PCA, which finds components that maximize the variance and identifies rotations that result in new, transformed features that are mutually statistically independent. In other words, each pair of features in this new feature space will have zero mutual information. NMF approximates a feature matrix X by $X \approx WH$, where each value $x_{i,j}$ is assumed to be Poisson distributed; hence, NMF is appropriate for nonnegative abundance data. CCA is another feature extraction technique. It uses a linear projection on a subset of features, then uses the correlation between the projections [112]. CCA can be applied to both continuous and discrete data, which is beneficial for analyzing not only the metagenomic features from abundance data but also the data associated with the samples. Finally, one of the advantages to CCA, as well as PCA, is that the projections can be computed efficiently using singular value decomposition (SVD).

Many datasets, even those in metagenomics, may not work well for data that lie on a nonlinear lower-dimensional manifold. t-Distributed Stochastic Neighbor Embedding (t-SNE) is a probabilistic nonlinear dimensionality reduction technique. It represents the similarity between any two points x_i and x_j as the conditional probability that x_i and x_j are neighbors, which is Gaussian distributed. It then attempts to learn a lower-dimensional embedding, where the similarities are now heavy tailed – that is, t-distributed. The Kullback-Leibler divergence between the estimated similarities in high- and low-dimensional space is minimized [113]. Visualizing the data with t-SNE can result in compact groups of classes (influenced by adjusted the "perplexity" parameter) in the lower-dimensional embedding. While the nonlinear embedding can be attractive to many metagenomic data analysis problems, there remain some drawbacks to t-SNE. Namely, t-SNE has a poor space complexity that can require a significant amount of memory to find the embedding.

3.2.3 Distance-Based Approaches for Feature Extraction

The remaining approaches are common in the statistical ecology and sequencing domains and are sometimes referred to as "unconstrained ordination" techniques (note that PCA described above can also be described as unconstrained ordination). These include principal coordinates analysis (PCoA) [114], otherwise known as metric multidimensional scaling (MDS), correspondence analysis (CA,

or reciprocal averaging) [115], and nonmetric multidimensional scaling (NMDS) [116, 117]. PCoA is simply an eigenvalue decomposition of a distance matrix. If the chosen distance metric is Euclidean, then PCA and PCoA are equivalent, in which case the components are linear combinations of the original features. With an alternative metric, then the principal *coordinates* are governed by the distance function. CA aims to maximize the correspondence between column and row scores – or, equivalently, sample and feature scores – in a feature matrix with nonnegative elements. This approach is analogous to PCoA with χ^2 distance, a distance metric used to extract relationships between rows and columns. Lastly, NMDS maximizes the rank order between features and hence is less concerned with the underlying pair-wise distances [118].

3.2.4 Neural Network Approaches for Feature Extraction

Neural networks (NNs) are a popular machine learning model and have recently garnered heightened interest in the sequencing domains [119]. Essentially, NNs perform nonlinear adaptive regression. Unsupervised approaches in particular have garnered interest in their ability to extract meaningful features from unlabeled data. One architecture in particular is the denoising autoencoder (DAE), which has recently been shown to perform well when applied to high-dimensional gene expression datasets [120, 121]. Given an input matrix X, the DAE attempts to recover X after X has been corrupted with noise (Fig. 3). The noise enables the DAE to learn robust, potentially generalizable features while preventing it from simply learning the identity function.

Another NN approach involves applying word embeddings, a widely used strategy in the natural language processing domain, directly to sequencing reads. The word2vec model is one of the more popular word embedding models. It gives words continuous vector representation in a lower-dimensional space based on the frequency of pair co-occurrence in a context window of fixed length [122]. We can understand it as mapping each word to a point in a continuous high-dimensional space, such that the points of words with similar semantic meaning are closer to each other in terms of, for example, Euclidean distance. Ng utilized Skip-Gram word2vec to embed short DNA k-mers [123]. He demonstrated that the embedding space extracts useful properties. Specifically, k-mer pairs with high cosine similarity in the embedding space were consistent with high-scoring pairs identified via global sequence alignment.

Word2vec is a shallow, fully connected NN with one hidden layer (Fig. 4). The input and output layer have the same number of nodes which is the number of words in the vocabulary. The number of nodes in the hidden layer is the dimensionality of the embedding space – that is, the size of the reduced feature space. The first step is converting each word into a one-hot vector, thereby giving each word a unique index. Then, training can be performed in one of two varieties: (1) the Skip-Gram model, which uses a word to predict its context (i.e., neighboring words) and (2) the continuous bag-of-words model (CBOW), which uses the context to predict a

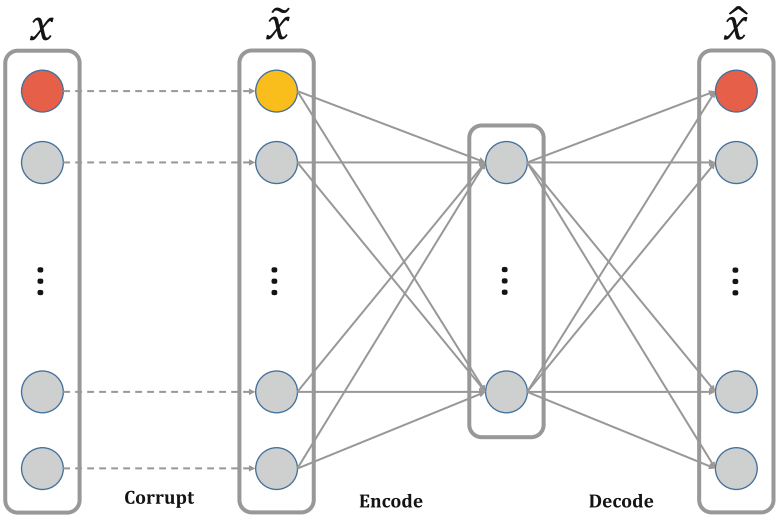

Fig. 3 Denoising autoencoder, where the input data x is corrupted with noise, producing $\tilde{x}$, which is then encoded into a lower-dimensional hidden layer. The hidden layer nodes are then decoded to produce $\hat{x}$, which has the same dimensionality as x. The distance between x and $\hat{x}$ is minimized such that the hidden layer is composed of features (nodes) capable of reconstructing x despite the addition of noise

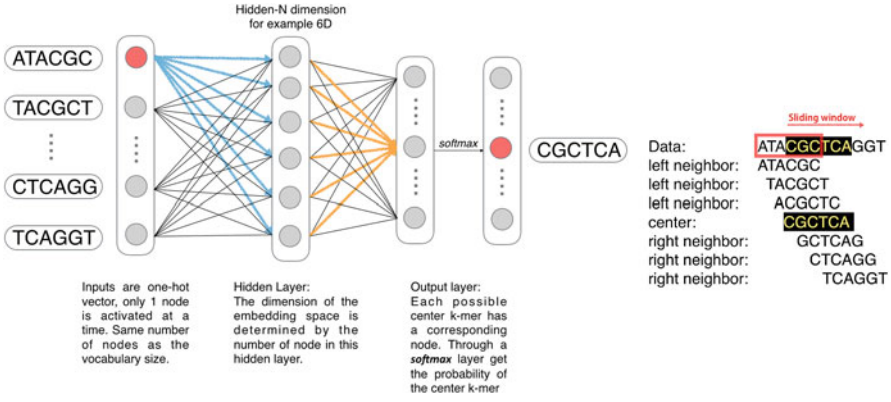

Fig. 4 (left) Neural network architecture for Skip-Graham word2vec. The training process requires the NN to predict the target word given the neighborhood. Words with similar context will activate similar nodes in the hidden layer. For the center k-mer "CGCTCA" and one of its neighbors "ATACGC," the corresponding node in the input and output layer is shown in red. Assuming "ATACGC" is the i-th word in the vocabulary, "CGCTCA" is then the $i + 1$-th word. The weights in blue connect between the input and hidden layers for the input word "ATACGC," i.e., the i-th column in weight matrix V. The weights in yellow connect the hidden and output layers for the output word "CGCTCA," i.e., the $i + 1$-th row of weight matrix U. (right) 6-mer neighborhood for word2vec training

word [123]. Because Skip-Gram does not average context vectors, but updates the weights for each word in the context separately, it can learn better representations for rare words compared to CBOW.

In the training process, the model will learn two vectors for each word w_i: (1) the i-th column of weight matrix V between the input and hidden layers and (2) the i-th row of weight matrix U between the hidden and output layers. We will refer to them as the input and output word vectors, respectively. The matrix product of V and U^T is the co-occurrence matrix. Levy and Goldberg described the word embedding as a matrix factorization of the co-occurrence matrix [124]. They also showed that a carefully constructed matrix factorization can produce word embeddings similar in quality to word2vec [125]. Also, Landgraf and Bellay showed that Skip-Gram word2vec is equivalent to weighted logistic PCA [126].

Training can be costly in terms of time and memory when the vocabulary is large. To accelerate the training process, Mikolov et al. utilized negative sampling [127]. Instead of updating the entire vocabulary each pass, they randomly sampled a subset of negative samples along with the context words to form a smaller vocabulary. Only the subset's weights are updated during a given pass. Another approach replaces the output weight matrix U with a Huffman tree [128].

4 Diversity Metrics and Constrained Ordination

After taxonomic or functional annotation has been performed, investigators are faced with the difficulty of quantitatively identifying and describing gradients, patterns, and variability within the dataset, particularly between individual samples or sample groups. Such analyses require simultaneous consideration of many, sometimes hundreds or thousands, distinct species or functions for each sample within the dataset. This effort is often further complicated by researchers who wish to include in their analysis information about the samples themselves or the sites from which they were collected, such as nutrient concentrations or availability, sample site location, host species (from which the samples were collected), vegetation composition or coverage, or watershed membership. The high degree of correlation expected between microbial community members and their environment requires the use of multivariate analytical methodologies.

Ordination is one of the most common analytical techniques used to explore the high-dimensional structure of microbial and molecular ecology datasets by using the distance matrix containing the similarity between metagenomic samples. Generally speaking, these methods attempt to identify the major ecological gradients or trends in high-dimensional datasets. Ordination methods can be largely categorized into two classes based upon the nature of the data to be used or the intent of the researcher. Unconstrained ordination methods (described above) employ only community data (i.e., the gene or taxon abundance table) in their calculations. Because unconstrained ordination relies only on species or functional abundances, the results expose or reveal the largest, and potentially most distinctive, gradients within the data. These methods are often used as a form of exploratory data

analysis, where investigators may not possess well-conceptualized hypotheses or are interested in identifying unexpected gradients, patterns, or relationships between taxa, functions, or sample groups.

Constrained ordination methods exploit information about the samples themselves, or their environment, to hopefully explain the source of variation observed within the community dataset. This type of ordination is carried out by constraining the ordination object to be optimally correlated to the values of one or more predictor variables related to an ecologically relevant hypothesis under investigation. Constrained ordination methods can be viewed as analogs of regression models where data describing samples (e.g., sample types, source location, environmental data, etc.) are used directly to describe and subsequently interpret the structure of microbial communities. These methods are commonly used to directly test predetermined hypotheses, such as the effect of nutrient gradients or the impact of ecological disturbances. As with unconstrained ordination, there are many options available for carrying out constrained ordination. Some of the most commonly used are redundancy analysis (RDA) [129], distance-based redundancy analysis (db-RDA) [130], canonical correspondence analysis (CCA, which is distinct from canonical *correlation* analysis described above), and detrended canonical correspondence analysis (DCCA) [131].

With proper caution, constrained ordination methods may also be used during data exploration efforts, especially at the beginning of longer-term or larger-scale studies. Within this context, initial community results can be subjected to constrained ordination with explanatory variables being selected using stepwise variable selection methods such as those suggested by Blanchet et al. [132]. The resulting explanatory variable subset can be compared with results from other data exploration methods such as the BIOENV procedure proposed by Clarke and Ainsworth [133]. The end goal of these efforts is to enable the researcher to determine what explanatory variables may be the most important and will require further study, identify gaps in data collection, and improve or clarify the hypotheses driving the current study.

Ordination has traditionally been applied to manually collected taxa counts or coverage data as well as data describing environmental conditions. The emergence of sequencing technologies has led to the adoption of ordination to carry out similar analyses with both data resulting from both targeted amplicon and metagenomic sequencing. Metagenomic sequencing results are often annotated for both their taxonomic and functional content, providing investigators with two corresponding sources of information and reducing reliance on a single locus for taxonomic annotation and diversity estimates. In some cases side-by-side comparison of results obtained from ordinating the taxonomic and functional annotations have exposed interesting results [134].

Ordination provides a powerful way to probe large complex datasets, but as with any computational or statistical approach, an acute understanding is prerequisite for proper application and interpretation of results. Many decisions must be made regarding the proper choice of ordination method, the distance measures used (if

any), whether or not the data should be transformed prior to calculation, as well as how to handle sample-level data. Further confusion may arise from the periodic development of new, perhaps better, ordination approaches that build upon the methods listed above. Thorough derivations and descriptions of ordination and clustering techniques are available in several well-written books [118, 135, 136] and review articles [137, 138], which may help educate investigators and students about appropriate approaches for answering ecological questions using ordination methods.

5 Statistical Inference

5.1 Multilevel Regression

Researchers may be interested in the relationship between a univariate statistic describing the taxonomic composition of a community (e.g., α-diversity, species richness, or evenness) and sample-level information such as site, temperature, time, or chemical concentration. Elucidating these relationships can be accomplished via linear regression:

$$y \sim \mathrm{N}(X\beta, \sigma^2 I) \tag{11}$$

where y is a vector of length n, X is an $n \times p$ matrix of p sample-level covariates including an intercept term, β is a vector of regression coefficients of length p, and I is an $n \times n$ identity matrix [139]. The coefficients β and variance σ^2 can be estimated via least squares, where $\hat{\beta}$ represents the association between y and X.

Often, however, complex study designs necessitate the use of multilevel regression models, often referred to as mixed-effects models. As an example, suppose metagenomic samples are taken from ten sites, and α-diversity varies depending on which site the sample originated. One approach to model this data may involve coding each site with dummy variables, setting one arbitrary site as a "reference" level. The cost here is nine degrees of freedom, and we are limited in our ability to interpret the regression coefficients, since they can only be interpreted with respect to the reference level [140]. One can imagine that with even more sites, this approach becomes less practical.

An alternative strategy involves letting the intercept vary as a function of site (a random intercept model):

$$y_n = \mu + \alpha_{\mathrm{site}[n]} + \epsilon_n \tag{12}$$

$$\alpha_{\mathrm{site}} \sim \mathrm{N}(\theta, \tau^2) \tag{13}$$

$$\epsilon \sim \mathrm{N}(0, \sigma^2) \tag{14}$$

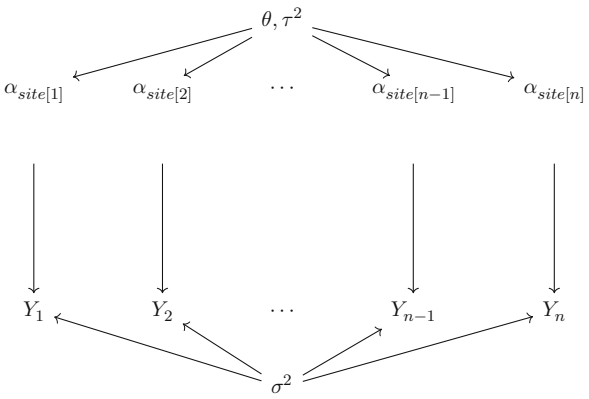

Fig. 5 Multilevel representation of sample-level means [141]. Each sample Y_n is influenced by its own between-site mean $\alpha_{\text{site}[n]}$. Each of these between-site means are generated by the same normal prior distribution with mean θ and variance τ^2. Within-site variability in Y is governed by σ^2. For sites with small sample sizes, $\alpha_{\text{site}[n]}$ will fall closer to θ, whereas sites with larger sample sizes will be represented more by their site-specific average (https://slack-files.com/T1VNV2ABW-F9YTS2K35-a06828b680)

where y_n is the alpha diversity for sample n, μ is the intercept, $\alpha_{\text{site}[n]}$ is the site-specific intercept for sample n, θ is the mean of site-specific intercepts, and τ^2 is their variance. Note that θ and τ^2 do not vary as a function of site. Figure 5 shows how the group-level means, α_{site}, distribute over the N samples. Let's now assume we believed that α-diversity varied as a function of temperature, but the degree of the relationship depended on the site. Here, we can let the slope between temperature and α-diversity vary:

$$y_n = \mu + \alpha_{\text{site}} \times \text{temp} + \epsilon_n \tag{15}$$

As more data become available, and study designs necessitate more complicated regression models, we can combine random effects and build complex multilevel regression models to help describe our community of interest. Such an approach is warranted because it allows an investigator to estimate the degree that specific effects vary by group (such as site) and not only with respect to a reference level. Moreover, because group-level effects share a common prior, a multilevel model can utilize group-level averages to "partially pool" information, thereby dampening the noisy contributions of underpowered group levels [139].

Multilevel models can easily be fit via the R package rstanarm. More sophisticated model designs can be implemented in Stan [142], which has interfaces in a variety of programming languages including R, Python, MATLAB, and Julia.

5.2 Multivariate Analysis

With univariate dependent variables, the regression approaches described above are an obvious choice. If, instead, we are interested in measuring the relationship between sample-level covariates and β-diversity, then we can turn to permutational MANOVA, which performs an analysis of variance on a distance matrix using sample-level covariates as predictors [143].

5.3 Differential Abundance Analysis

It is often of interest to detect which genes or taxa best differentiate two or more sample classes. A straightforward approach involves performing hypothesis testing for each variable in an abundance table and then correcting for the false discovery rate via the Benjamini-Hochberg procedure. Such an approach is limited in that it assumes specific assumptions are met prior to performing the analysis, which is typically not the case. These include assumptions regarding normality and mean-variance relationships.

More sophisticated strategies have been developed and applied to sequencing data of similar structure: edgeR and VOOM [144] for gene expression data, DESeq2 [145] for gene expression and 16S amplicon survey data [92], and MetagenomeSeq [146] also for 16S amplicon survey data. MetagenomeSeq, for example, applies either a zero-inflated Gaussian mixture model or a zero-inflated log-normal model to each feature separately. Account for zero inflation is thought to prevent overdispersed fits and also mitigate the detrimental effects of highly sparse sequencing data. DESeq2, on the other hand, first performs a variance-stabilization transformation, followed by fitting a negative binomial generalized linear model (GLM). For metagenomic data, under the right circumstances such that the abundance table consists of counts and hence has yet to be normalized in terms of sample library size imbalance, these approaches may prove viable. Still, however, they were developed with specific distributional assumptions in mind. Considering the plethora of normalization strategies available for metagenomic data, future work is necessary to demonstrate whether readily used differential abundance strategies remain appropriate after a particular normalization is performed and which metagenomic normalization procedures work well in tandem with which differential abundance strategies.

Approaches nevertheless exist that were designed with metagenomic data in mind. One such approach, LEfSe, performs the nonparametric Kruskal-Wallis sum-rank test to identify significant differences in abundances between genes or taxa belonging to a class of interest [147]. The Wilcoxon rank sum then disentangles pair-wise differences between sample subclasses. Linear discriminant analysis is applied last to estimate the effect size of the statistically significant features.

Johnsson et al. applied various differential abundance methods to metagenomic data and evaluated their performance in terms of statistical power, control of the false discovery rate, and uniformity of p-values given the null hypothesis [148]. They found that GLM-based models that combat potential overdispersion perform best. These include DESeq2, edgeR, and an overdispersed Poisson GLM. MetagenomeSeq generally performed well but was found to be inferior to simply performing t-tests to log-transformed features, suggesting that the zero-inflation mixture components have a negligible impact. Also, it was prone to highly biased p-values and consequently type 1 error. Also of note was that performing t-tests on square root transformed features was superior to utilizing non-parametric Wilcoxon rank-sum tests, which, as noted above, are used in LEfSe. The authors speculated this may be due to the latter's susceptibility to ties. It should be stressed that effect of different metagenomic-specific normalization approaches on differential abundance analysis was not explored.

6 Machine Learning and its Application to Metagenomics

6.1 Overview

Machine learning techniques are widely used in different steps in a metagenomic pipeline. For example, Naive Bayes has been applied for taxonomic classification, hidden Markov models (HMM) are often used for functional annotation, and random forest is readily utilized for phenotype prediction. From a research problem perspective, machine learning techniques are helpful in addressing the following questions:

- Who are there (what species are in a sample)?
- What are they doing (what functions are in a sample)?
- What can we infer from the sample (what is the state of the host/environment)?

In the following sections, we will talk about machine learning methods and tools that have been applied to metagenomics.

6.2 A General Machine Learning Review

One of the highly cited definitions of machine learning involves a computer program that is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E [149]. Experience E usually refers to the data collected. Task T usually represents the decision or prediction we want to make. In a metagenomics context, E represents the samples or DNA sequences.

Machine learning methods can be classified into supervised, semi-supervised, and unsupervised learning approaches based on their dependency of labeled data. Supervised learning approaches (classification methods) require a training phase that utilizes the samples and their labels to minimize the cost function of the classifier. Unsupervised learning (clustering) methods, on the other hand, use a distance measure to group samples into clusters. Finally, semi-supervised learning methods act as a compromise between the two. They first use a subset of training data to train the classifier and then use unlabeled examples to improve performance. Table 2 lists the machine learning methods listed throughout this chapter.

In the context of metagenomics, machine learning techniques can be classified based on their application. The following sections will discuss the machine learning techniques that have been implemented in taxonomic classification, DNA binning, functional annotation, and phenotype prediction.

6.3 Taxonomic Classification and DNA Binning

One of the main challenges in metagenomics is the identification of microorganisms in clinical and environmental samples [150]. Taxonomic classification or DNA binning are helpful for researchers to determine the composition of their metagenomic samples. Taxonomic classification is a supervised learning problem, whereas DNA binning has traditionally been unsupervised, but could also be semi-supervised [151].

6.3.1 Naive Bayesian Classifier

A naive Bayes classifier (NBC) is a type of probabilistic classifier that exploits Bayes rule to perform classification. Naive refers to its assumption that features are independent from each other. Here $X = (x_1, x_2, \ldots, x_n)$ is an observation with n features. The probability of X coming from class k is

$$p(C_k|x_1, x_2, \ldots, x_n) = \frac{p(C_k)p(X|C_k)}{p(X)} = \frac{p(C_k)p(x_1|C_k)p(x_2|C_k)\cdots p(x_n|C_k)}{\sum_{k=1}^{|C|} p(C_k)p(x_1|C_k)p(x_2|C_k)\cdots p(x_n|C_k)} \tag{16}$$

So, the estimated class is

$$\hat{c} = \arg \max_{k \in 1,\ldots,K} \; p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \tag{17}$$

NBC is easy to implement and has high accuracy when the features are independent. Rosen et al. proposed a metagenome fragment classification tool using k-mer frequency profiles [152, 153], which has proven to be fast and accurate when trained

Table 2 A list of machine learning algorithms and their implementations

Requires labels	Methods	Abbrev.	Implementation
N	Canonical correlation analysis	CCA	CCA(R), scikit-learn (Python)
N	Denoising autoencoder	DAE	keras
N	Independent component analysis	ICA	fastICA (R), scikit-learn (P)
Y	Linear discriminant analysis	LDA	MASS (R), scikit-learn (P)
N	Nonmetric multidimensional scaling	NMDS	vegan (R)
N	Nonnegative matrix factorization	NMF	nmf (R), scikit-learn (P)
N	Principal component analysis	PCA	prcomp (R), scikit-learn (P)
Y	Supervised principal component analysis	supervised PCA	superpc (R), supervisedPCA-Python (P)
N	Sparse principal component analysis	Sparse PCA	Elasticnet (R), scikit-learn (P)
N	Principal coordinate analysis	PCoA/MDS	vegan (R)
N	t-Distribution stochastic neighbor embedding	t-SNE	Rtsne (R), scikit-learn (P)
N	Word2vec	Word2vec	Gensim (P), keras
Y	Elastic net	EN	Glmnet (R, P)
Y	Hidden Markov model	HMM	HMM (R), Stan, hmmlearn (P)
Y	k-nearest neighbors	k-NN	Caret (R), scikit-learn (P)
Y	Naive Bayes	NB	Caret (R), scikit-learn (P), Stan, NBC[a]
Y	Support vector machine	SVM	Caret (R), scikit-learn (P)
Y	Random forest	RF	Caret (R), scikit-learn (P)
N	k-means/Medoids	k-means/Medoids	kmeans (R), scikit-learn (P), MetaCluster[b]

[a] http://nbc.ece.drexel.edu/
[b] http://i.cs.hku.hk/~alse/MetaCluster/

on large k-mers. Assuming $X = [x_1, x_2, \cdots, x_n]$ is a set of k-mers in a sequence, then, in the training phase, $p(x_i|C_k)$ is estimated by the total number of k-mers x_i that occurs in all the training sequences from class C_k. In the testing phase, the taxa that contains a given query sequence is predicted by the class that maximizes the posterior probability $P(C_k|X)$.

6.3.2 k-Nearest Neighbors

k-nearest neighbors (k-NN) classifies query samples $x_0 \in R^m$ based on distance. Given the labeled data, $T = \{x_i, y_i \mid i \in [1, N]\}$, $x_i \in R^m$, a new observation will be assigned to a class where the majority of the first K nearest labeled samples originate. The following is the classification workflow for 1-NN and K-NN:

1. 1-NN: (1) find the nearest instance in the training set $\min_{x_i}(\|x_o - x_i\|)$; (2) test data label is the same as the nearest instance via $y_0 = y_i$.
2. K-NN: (1) find the k nearest instances in the training set $\min_{x_1,\cdots,x_k}(\|x_o - x_i\|)$; (2) let the k nearest instances vote via $y_0 = \texttt{Mode}(y_1, \cdots, y_k)$.

Borozan et al. used K-NN to perform classification in their taxonomic lineage prediction tool, and they regarded K-NN as one of the simplest and most intuitive classification algorithms [154].

6.3.3 Clustering

The k-means clustering algorithm is used to partition N observations into k clusters. The observations' affiliations are determined by some distance measure, such as Euclidean distance. Hence, the observations that are close to each other will be grouped together, and the observations that are distant from each other will be assigned to different clusters. To converge to an optimum quickly, this clustering process utilizes an expectation-maximization (EM) procedure. This is an iterative refinement approach that assigns observations into k clusters by comparing the distance between the observations and k centroids (usually initialized randomly) and then updates the centroids with the new cluster assignment until convergence. The objective function is given by

$$\texttt{arg}\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{18}$$

Many tools perform DNA binning by clustering sequences based on a predefined distance metric. Wang et al. used k-means to cluster sequences [155].

An alternative approach, CD-HIT, uses a greedy search algorithm to cluster the sequences [156]. First, it sorts sequences based on their length. The longest sequence will be representative of the first cluster formed. Then, the second

longest sequence is compared to this cluster's representatives. It will be assigned to this cluster if the distance between it and the representative is within a user-selected distance threshold; otherwise, a new cluster will be created and the sequence becomes its representative. This process will be repeated for all remaining sequences until all sequences are assigned to either an existing cluster or a newly created one. In addition to CD-HIT, there are some other sequence clustering tools such as DNACLUST [157] and UCLUST [158]. These tools can be faster than CD-HIT under some circumstance. For example, UCLUST, by default, operates in an inexact mode that reduces the search space by only comparing the sequence to the representatives from a subset of all clusters.

6.4 Functional Annotation and Prediction

6.4.1 Hidden Markov Model

Hidden Markov models (HMMs) describe a sequential observation and their underlying latent states. The observed sequence of length L can be described as $\mathbf{x} = x_1, x_2, \cdots, x_L$ [159]. Its latent state sequence is $y = y_1, y_2, \cdots, y_L$. Each symbol x_n takes on a finite number of possible values from the set of observations $\mathbf{O} = O_1, O_2, \cdots, O_N$, and each state y_n takes one of the values from the set of states $\mathbf{S} = 1, 2, \cdots, M$, where N and M denote the number of distinct observations and states in the model. This model can be described by two matrices, the transition matrix and the emission matrix. One entry in the transition matrix is the probability of entering state j in the next time given in state i:

$$t(i, j) = P(y_{n+1} = j | y_n = i) \tag{19}$$

One entry in the emission matrix is the probability of observing x given state i:

$$e(x|i) = P(x_n = x | y_n = i) \tag{20}$$

The probability that an HMM will generate an observation $\mathbf{x}$ with underlying state sequence $\mathbf{y}$ is [159]:

$$p(x, y|t, e) = p(x|y, t, e)p(y|t, e) \tag{21}$$

where,

$$p(x|y, t, e) = p(x_1|y_1)p(x_2|y_2)\cdots p(x_L|y_L) \tag{22}$$

$$p(y|t, e) = p(y_0)p(y_1, y_2)p(y_2|y_3)\cdots p(y_{L-1}|y_L) \tag{23}$$

Rho et al. proposed to use HMMs to model nucleotide sequences to predict a given gene [160].

6.4.2 Logistic Regression

Logistic regression usually takes real value input and outputs a value between 0 and 1. This is accomplished by the logistic function:

$$\hat{y}(\mathbf{x}) = \frac{1}{1 + e^{-\sum_i^n \beta_i \cdot x_i}} \tag{24}$$

where $\mathbf{x} = [x_1, x_2, \cdots, x_n]$ is an input vector and $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_n]$ are the weights estimated by the model. Since the output of this function is between 0 and 1, one can consider the output to be the probability of being classified into class C, i.e., $p(y = C|\mathbf{x}; \boldsymbol{\beta}) = 1/1+e^{-\sum_i^n \beta_i \cdot x_i}$. Hence, we can determine the best parameter β using a maximum likelihood approach:

$$\max_{\beta} L(Y|\mathbf{X}; \boldsymbol{\beta}) \tag{25}$$

where $L(Y|\mathbf{X}; \boldsymbol{\beta})$ is the product of the probabilities that all labeled samples get classified into the correct class, i.e.,

$$L(Y|\mathbf{X}; \boldsymbol{\beta}) = \prod_{i=1}^{n} \hat{y}(\mathbf{x_i})^{y(\mathbf{x_i})}(1 - \hat{y}(\mathbf{x_i}))^{1-y(\mathbf{x_i})} \tag{26}$$

where $y(\mathbf{x_i})$ is the true label for sample $\mathbf{x_i}$. The optimal parameter to maximize the likelihood can be found using a gradient descent algorithm, which iteratively updates the parameters by the estimated derivative of the function given the current parameter such that the likelihood tends to increase after each update. In the testing phase, observations $\mathbf{x_i}$ will be classified into class C if the output $\hat{y}(\mathbf{x_i})$ is greater than 0.5 or a predefined threshold; otherwise, it will be classified as the alternative class $\neg C$.

Noguchi et al. used logistic regression to analyze the GC content of a given sequence and estimate the mono-codon and di-codon frequencies [161].

6.5 Phenotype Prediction

6.5.1 Random Forest

A decision tree is a supervised learning technique that looks at each feature individually to make a binary decision, thereby splitting samples into branches. The information gain is maximized during this process to help the classifier make an accurate decision. It is widely used because the decision process is interpretable, and the performance is often promising. Random forest (RF) is an ensemble learning extension of a decision tree where the decision is made by majority vote of many

decision trees. Each tree is fit to only a subset of features, forcing the classifier to learn robust, potentially generalizable subsets of features, particularly when compared to simpler decision tree approaches. The following is the construction workflow of random forest [162]:

1. Draw $ntree$ bootstrap samples from the original data.
2. Grow a tree for each bootstrap dataset. At each node of the tree, randomly select $mtry$ variables for splitting. Grow the tree so that each terminal node has no fewer than $nodesize$ cases.
3. Aggregate information from the $ntree$ trees for new data prediction such as majority voting for classification.
4. Compute an out-of-bag (OOB) error rate by using the data not in the bootstrap sample.

RF has been applied in many metagenomic pipelines to predict phenotype given a high-dimensional abundance table. It naturally finds useful features and is robust to overfitting. Additional applications of RF can be found in Chen and Ishwaran [162].

6.5.2 Support Vector Machine

The support vector machine (SVM) finds a hyperplane or a set of hyperplanes that best separates labeled data in some geometric space. In the testing phase, samples are assigned to classes based on their location in this space. Normally, the linear SVM separates the space linearly, but when data are not linearly separable, the "kernel trick" enables the data to be projected into a higher dimensional space, thereby potentially rendering the data linearly separable. Hence, the core of the SVM model is a linear SVM algorithm. The following is an overview of the application of this model in a binary classification problem. Given some training data $\mathcal{D}$, a set of n points have the form

$$\mathcal{D} = \left\{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^m,\ y_i \in \{-1, 1\}\right\}_{i=1}^{n} \tag{27}$$

where y_i is either 1 or -1, indicating the class to which the point $\mathbf{x}_i$ belongs. Each $\mathbf{x}_i$ is an m-dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having $y_i = 1$ from those having $y_i = -1$. If the training data are linearly separable, we can select two hyperplanes that completely separate the data and then try to maximize the distance between the data and hyperplanes. The region bounded by them is called "the margin." These hyperplanes can be described by the equations

$$\begin{cases} \boldsymbol{\omega} \cdot \mathbf{x} - b = 1 \\ \boldsymbol{\omega} \cdot \mathbf{x} - b = -1 \end{cases} \tag{28}$$

It is an optimization problem to find these two hyperplane. Maximizing the distance between the data and hyperplane is equivalent to minimizing $||\boldsymbol{\omega}||$. While ensuring all positive and negative samples are separated, we have a constraint: $\forall(\mathbf{x}_i, y_i) \in \mathcal{D},\ y_i(\boldsymbol{\omega} \cdot \mathbf{x}_i - b) \geq 1$. Again, in the testing phase, samples are assigned to classes by the subspaces segmented by the hyperplanes. Capriotti et al. developed a method based on kerneled SVMs to predict whether a new phenotype derived from a single nucleotide polymorphism can be related to a genetic disease in humans [163].

6.5.3 Elastic Net

A sparsity-promoting approach related to Lasso, but more robust for highly correlated features, is the elastic net (EN) [164]. It has been shown to perform well in both regression and classification, particularly with high-dimensional data where the number of features greatly outnumbers the number of samples [165]. Whereas the Lasso regularization penalty involves the L1-norm, EN compromises between the L1- and L2-norm (readers may recognize L2-norm penalized regression as ridge regression). EN is formally given by

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta^{\mathsf{T}} \mathbf{x}_i)^2 + \lambda \left(\frac{1}{2}(1-\alpha)|\theta|_2^2 + \alpha|\theta|_1 \right) \tag{29}$$

where $\alpha \in [0, 1]$ controls the relative contributions of the L1- and L2-norms. Notice that when $\alpha = 1$, the EN reduces to the Lasso, whereas when $\alpha = 0$, it reduces to ridge regression; thus, EN can be considered a generalization of the two regularization approaches.

7 Discussion and Conclusion

Techniques can generally be broken down into two main categories: (1) techniques that directly work with DNA/RNA sequences to classify attributes about them (taxonomy and function) and (2) techniques that facilitate comparative analyses. Some fundamental preprocessing steps – such as normalization, feature selection, and feature extraction – can be applied to single samples; however, most of these preprocessing steps are designed for multiple samples, as most studies use many samples and are usually limited by cost.

For the sequence identification problems, the longest-standing methods are those that extend read sequence length and identify its taxonomic origin and functional annotation. Assembly, with the most successful methods involving de Bruijn graphs, was one of the first algorithms to be developed because it was

key in the Human Genome Project (even from the longer Sanger sequencing reads) [166]. In metagenomics, the problem is more complex since any read can come from any one of thousands of organisms in a sample, culminating in a demultiplexing (read binning) step before assembly. Taxonomic/functional binning and classification, with the rich history of k-mer-based clustering, alignment, and profile HMMs, have also been extensively studied but are still being investigated for their extensions to metagenomics (not just annotating whole genomes). The problem with metagenomic data is mainly that sequences come from a collection of possible species' origins and many sequences are from unknown (or not yet sequenced) species.

Comparative analysis is the most active area of development, notably the rich areas of statistical inference and machine learning, which are utilized to make cross-sample and even cross-study comparisons. Early attempts leveraged ordination, but the substantial growth of the machine learning field has provided researchers with an immense resource of potential tools, particularly classification algorithms, allowing one to apply discriminative and generative functions to discern groups of samples. Moreover, deep neural networks show much promise for learning complex relationships and hence are areas of active research.

We closed this chapter with a discussion of general machine learning approaches, since these techniques can be applied to not only sequence identification but also to comparative analysis and phenotype prediction. Machine learning and statistical inference can help researchers disentangle the complexity that make other models with strict assumptions fail. We show examples of where these algorithms have been applied. Still, when using learning algorithms, one must think about how much training data is available and whether supervised versus unsupervised learning is suitable. Also, sometimes there are many confounding factors, where feature selection or normalization may simplify and denoise the data. If there is inherent structure in data, hierarchical models which capture this structure should be used. If prediction is the goal, supervised approaches should be considered. Finally, no matter what method is used, researchers should be aware of class imbalance and model overfitting and try to mitigate these effects through carefully designing training/validation/testing regimes. There are many considerations that researchers should consider when analyzing complex metagenomic data, and these should be identified early and examined throughout analyses.

In this book chapter, important techniques in metagenomic analyses are reviewed. However, good benchmarking data and infrastructure is not available to ensure that future methods improve upon the state of the art. Therefore, there is not only much work to be done to improve metagenomic software, but there is need to standardize the way we assess these methods.

References

1. Handelsman J, Rondon M, Brady S, Clardy J, Goodman R. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol. 1998;5(10):R245–9.
2. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev. 2004;68(4):669+.
3. Pace N, Stahl D, Lane D, Olsen G. The analysis of natural microbial-populations by ribosomal-RNA sequences. Adv Microb Ecol. 1986;9:1–55.
4. Simon C, Daniel R. Metagenomic analyses: past and future trends. Appl Environ Microbiol. 2011;77(4):1153–61.
5. Streit W, Schmitz R. Metagenomics – the key to the uncultured microbes. Curr Opin Microbiol. 2004;7(5):492–8.
6. Tringe SG, Hugenholtz P. A renaissance for the pioneering 16S rRNA gene. Curr Opin Microbiol. 2008;11(5):442–6.
7. Ward N. New directions and interactions in metagenomics research. FEMS Microbiol Ecol. 2006;55(3):331–8.
8. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005;71(12):8228–35.
9. Solden L, Lloyd K, Wrighton K. The bright side of microbial dark matter: lessons learned from the uncultivated majority. Curr Opin Microbiol. 2016;31:217–26.
10. Vieites JM, Guazzaroni ME, Beloqui A, Golyshin PN, Ferrer M. Metagenomics approaches in systems microbiology. FEMS Microbiol Rev. 2009;33(1):236–55.
11. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Nat Acad Sci. 1977;74(11):5088–90. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.74.11.5088.
12. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, et al. Diversity of the human intestinal microbial flora. Science. 2005;308(5728):1635–8.
13. Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, et al. Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics. 2006;7:1–13.
14. Ley RE, Peterson Da, Gordon JI. Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell. 2006;124(4):837–48. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16497592.
15. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Sci Trans Med. 2009;1(6):6ra14. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2894525&tool=pmcentrez&rendertype=abstract.
16. Venter J, Remington K, Heidelberg J, Halpern A, Rusch D, Eisen J, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004;304(5667):66–74.
17. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage-host relationships. FEMS Microbiol Rev. 2016;40(2):258–72.
18. Forbes JD, Knox NC, Ronholm J, Pagotto F, Reimer A. Metagenomics: the next culture-independent game changer. Front Microbiol. 2017;8:1069. Available from: http://dx.doi.org/10.3389/fmicb.2017.01069.
19. Hurwitz BL, U'Ren JM, Youens-Clark K. Computational prospecting the great viral unknown. FEMS Microbiol Lett. 2016;363(10):1–12.
20. Kimura N. Metagenomic approaches to understanding phylogenetic diversity in quorum sensing. Virulence. 2014;5(3):433–42.
21. Mathieu A, Vogel TM, Simonet P. The future of skin metagenomics. Res Microbiol. 2014;165(2):69–76.
22. Sangwan N, Xia F, Gilbert JA. Recovering complete and draft population genomes from metagenome datasets. Microbiome. 2016;4:2–11.

23. Schmieder R, Edwards R. Insights into antibiotic resistance through metagenomic approaches. Future Microbiol. 2012;7(1):73–89.
24. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
25. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6.
26. Giardine B, Riemer C, Hardison R, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 2005;15(10):1451–5.
27. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17(3):377–86.
28. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9.
29. Rosen GL, Reichenberger ER, Rosenfeld AM. NBC: the Naive Bayes classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics. 2011;27(1):127–9.
30. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–41.
31. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261–67.
32. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004;428(6978):37–43.
33. Sedlar K, Kupkova K, Provaznik I. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. Computational and Structural Biotechnology Journal 2017;15:48–55. Available from: http://doi.org/10.1016/j.csbj.2016.11.005.
34. Mende DR, Waller AS, Sunagawa S, Järvelin AI, Chan MM, Arumugam M, et al. Assessment of metagenomic assembly using simulated next generation sequencing data. PLoS One. 2012;7(2):1–11.
35. Vázquez-Castellanos JF, García-López R, Pérez-Brocal V, Pignatelli M, Moya A. Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. BMC Genomics. 2014;15(1):37. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-37.
36. Mande SS, Mohammed MH, Ghosh TS. Classification of metagenomic sequences: methods and challenges. Brief Bioinform. 2012;13(6):669–81.
37. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ. 2014;2:e603. Available from: https://peerj.com/articles/603.
38. Ribeca P, Valiente G. Computational challenges of sequence classification in microbiomic data. Brief Bioinform. 2011;12(6):614–25.
39. Mohammed M, Ghosh TS, Singh NK, Mande SS. SPHINX – an algorithm for taxonomic binning of metagenomic sequences. Bioinformatics. 2010;27(1):22–30.
40. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence. Nature. 1998, p. 537–544. Available from: http://dx.doi.org/10.1038/31159.
41. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechno. 2013;31(6):533–8.
42. Alneberg J, Bjarnason BS, De Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11(11):1144–6.
43. Miller IJ, Chevrette MG, Kwan JC. Interpreting microbial biosynthesis in the genomic age: biological and practical considerations. Marine Drugs. 2017, 1–24. Available from: http://dx.doi.org/10.3390/md15060165.

44. Lykidis A, Chen CL, Tringe SG, McHardy AC, Copeland A, Kyrpides NC, et al. Multiple syntrophic interactions in a terephthalate-degrading methanogenic consortium. ISME J. 2011;5(1):122–30.
45. Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, Romero H, Simón-Soro A, Pignatelli M, et al. The oral metagenome in health and disease. ISME J. 2012;6(1):46–56.
46. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464(7285):59–65.
47. Sangwan N, Xia F, Gilbert JA. Recovering complete and draft population genomes from metagenome datasets. Microbiome. 2016;4(1):8. Available from: http://www.microbiomejournal.com/content/4/1/8.
48. Mohammed MH, Ghosh TS, Reddy RM, Reddy CV, Singh NK, Mande SS. INDUS – a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences. BMC Genomics. 2011;12(Suppl 3). Available from: http://www.hubmed.org/display.cgi?uids=22369237.
49. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. 2011;6(3):1–11.
50. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9(8):811–4. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3443552&tool=pmcentrez&rendertype=abstract.
51. Liu B, Gibbons T, Ghodsi M, Pop M. MetaPhyler: taxonomic profiling for metagenomic sequences. In: Proceedings – 2010 IEEE international conference on bioinformatics and biomedicine, BIBM 2010; 2010, p. 95–100.
52. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger Sa, Kultima JR, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 2013;10(12):1196–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24141494.
53. Nayfach S, Pollard KS. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome Biol. 2015;16(1):51. Available from: http://genomebiology.com/2015/16/1/51.
54. Freitas TAK, Li PE, Scholz MB, Chain PSG. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. Nucleic Acids Res. 2015;43(10):e69(1–14).
55. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17(3):377–86. Available from: http://www.hubmed.org/display.cgi?uids=17255551.
56. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16(1):236. Available from: http://www.biomedcentral.com/1471-2164/16/236.
57. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053813&tool=pmcentrez&rendertype=abstract.
58. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics (Oxford, England). 2013;29(18):2253–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23828782%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3753567.
59. Sobih A, Tomescu AI, Mäkinen V. Metaflow: metagenomic profiling based on whole-genome coverage analysis with min-cost flows. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9649; 2016. p. 111–121.
60. Rosen G, Garbarine E, Caseiro D, Polikar R, Sokhansanj B. Metagenome fragment classification using N-mer frequency profiles. Adv Bioinform. 2008;2008:205969. Available from: http://www.hubmed.org/display.cgi?uids=19956701.
61. Darling AE, Jospin G, Lowe E, Matsen FA, Bik HM, Eisen JA. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ. 2014;2:e243. Available from: https://peerj.com/articles/243.

62. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol. 2017;18(1):182. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1299-7.
63. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep. 2016;6:1–14. Available from: http://dx.doi.org/10.1038/srep19233.
64. Prakash T, Taylor TD. Functional assignment of metagenomic data: challenges and applications. Brief Bioinform. 2012;13(6):711–27. Prakash, Tulika Taylor, Todd D eng Research Support, Non-U.S. Gov't Review England 2012/07/10 06:00 Brief Bioinform. 2012;13(6):711–27. https://doi.org/10.1093/bib/bbs033.Epub2012Jul6.
65. Carr R, Borenstein E. Comparative analysis of functional metagenomic annotation and the mappability of short reads. PLoS One. 2014;9(8):e105776. Carr, Rogan Borenstein, Elhanan eng DP2 AT007802/AT/NCCIH NIH HHS/ P30 DK089507/DK/NIDDK NIH HHS/ DP2 AT007802-01/AT/NCCIH NIH HHS/Comparative Study Research Support, N.I.H., Extramural 2014/08/26 06:00 PLoS One. 2014;9(8):e105776. https://doi.org/10.1371/journal.pone.0105776. eCollection 2014.
66. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45. O'Leary, Nuala A Wright, Mathew W Brister, J Rodney Ciufo, Stacy Haddad, Diana McVeigh, Rich Rajput, Bhanu Robbertse, Barbara Smith-White, Brian Ako-Adjei, Danso Astashyn, Alexander Badretdin, Azat Bao, Yiming Blinkova, Olga Brover, Vyacheslav Chetvernin, Vyacheslav Choi, Jinna Cox, Eric Ermolaeva, Olga Farrell, Catherine M Goldfarb, Tamara Gupta, Tripti Haft, Daniel Hatcher, Eneida Hlavina, Wratko Joardar, Vinita S Kodali, Vamsi K Li, Wenjun Maglott, Donna Masterson, Patrick McGarvey, Kelly M Murphy, Michael R O'Neill, Kathleen Pujar, Shashikant Rangwala, Sanjida H Rausch, Daniel Riddick, Lillian D Schoch, Conrad Shkeda, Andrei Storz, Susan S Sun, Hanzhen Thibaud-Nissen, Francoise Tolstoy, Igor Tully, Raymond E Vatsan, Anjana R Wallin, Craig Webb, David Wu, Wendy Landrum, Melissa J Kimchi, Avi Tatusova, Tatiana DiCuccio, Michael Kitts, Paul Murphy, Terence D Pruitt, Kim D eng Intramural NIH HHS/ Research Support, N.I.H., Intramural England 2015/11/11 06:00 Nucleic Acids Res. 2016;44(D1):D733–45. https://doi.org/10.1093/nar/gkv1189. Epub 8 Nov 2015.
67. UniProt Consortium. Reorganizing the protein space at the universal protein resource (UniProt). Nucleic Acids Res. 2012;40:D71–5.
68. Gasteiger E, Jung E, Bairoch A. SWISS-PROT: connecting biomolecular knowledge via a protein database. Curr Issues Mol Biol. 2001;3(3):47–55. Gasteiger, E Jung, E Bairoch, A Eng Review England 2001/08/08 10:00 Curr Issues Mol Biol. 2001;3(3):47–55.
69. Alberti A, Poulain J, Engelen S, Labadie K, Romac S, Ferrera I, et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. Sci Data. 2017;4:170093. Alberti, Adriana Poulain, Julie Engelen, Stefan Labadie, Karine Romac, Sarah Ferrera, Isabel Albini, Guillaume Aury, Jean-Marc Belser, Caroline Bertrand, Alexis Cruaud, Corinne Da Silva, Corinne Dossat, Carole Gavory, Frederick Gas, Shahinaz Guy, Julie Haquelle, Maud Jacoby, E'krame Jaillon, Olivier Lemainque, Arnaud Pelletier, Eric Samson, Gaelle Wessner, Mark Acinas, Silvia G Royo-Llonch, Marta Cornejo-Castillo, Francisco M Logares, Ramiro Fernandez-Gomez, Beatriz Bowler, Chris Cochrane, Guy Amid, Clara Hoopen, Petra Ten De Vargas, Colomban Grimsley, Nigel Desgranges, Elodie Kandels-Lewis, Stefanie Ogata, Hiroyuki Poulton, Nicole Sieracki, Michael E Stepanauskas, Ramunas Sullivan, Matthew B Brum, Jennifer R Duhaime, Melissa B Poulos, Bonnie T Hurwitz, Bonnie L Pesant, Stephane Karsenti, Eric Wincker, Patrick eng Research Support, Non-U.S. Gov't England 2017/08/02 06:00 Sci Data. 2017;4:170093. https://doi.org/10.1038/sdata.2017.93.
70. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.

71. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genetics. 2000;25(1):25–9.
72. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. BMC Bioinform. 2003;4:41–7.
73. Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, et al. The Genome portal of the department of energy joint Genome Institute. Nucleic Acids Res. 2012;40: D26–32.
74. Kanehisa M, Goto S, Kawashima YSM, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014;42: D199–205.
75. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
76. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2010;38:D473–9.
77. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005;33:5691–702.
78. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
79. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, et al. IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res. 2008;36:D534–8.
80. Markowitz V, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, et al. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res. 2012;40:D115–22.
81. Aziz RK, et al. The RAST server: rapid annotations using subsystems technology. BMC Genomics. 2008;9(75):1–15.
82. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17(3):377–86.
83. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol. 2012;8(6):1–17.
84. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44(D1):D279–85. Finn, Robert D Coggill, Penelope Eberhardt, Ruth Y Eddy, Sean R Mistry, Jaina Mitchell, Alex L Potter, Simon C Punta, Marco Qureshi, Matloob Sangrador-Vegas, Amaia Salazar, Gustavo A Tate, John Bateman, Alex eng 108433/Z/15/Z]/Wellcome Trust/United Kingdom BB/L024136/1/Biotechnology and Biological Sciences Research Council/United Kingdom Howard Hughes Medical Institute/ Research Support, Non-U.S. Gov't England 2015/12/18 06:00 Nucleic Acids Res. 2016;44(D1):D279–85. https://doi.org/10.1093/nar/gkv1344. Epub 15 Dec 2015.
85. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. Nucleic Acids Res. 2013;41(D1):E344–7. 062BE Times Cited:260 Cited References Count:14.
86. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 2013;41(Database issue):D377–86. Mi, Huaiyu Muruganujan, Anushya Thomas, Paul D eng GM081084/GM/NIGMS NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England 2012/11/30 06:00 Nucleic Acids Res. 2013;41(Database issue):D377–86. https://doi.org/10.1093/nar/gks1118. Epub 27 Nov 2012.
87. Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, de Castro E, et al. HAMAP in 2013, new developments in the protein family classification and annotation system. Nucleic Acids Res. 2013;41(Database issue):D584–9. Pedruzzi, Ivo Rivoire, Catherine Auchincloss, Andrea H Coudert, Elisabeth Keller, Guillaume de Castro, Edouard Baratin, Delphine Cuche,

Beatrice A Bougueleret, Lydie Poux, Sylvain Redaschi, Nicole Xenarios, Ioannis Bridge, Alan eng 5R01GM080646-07/GM/NIGMS NIH HHS/ 8P20GM103446-12/GM/NIGMS NIH HHS/ 5G08LM010720-03/LM/NLM NIH HHS/ 2P41 HG02273/HG/NHGRI NIH HHS/ 3R01GM080646-07S1/GM/NIGMS NIH HHS/ SP/07/007/23671/British Heart Foundation/United Kingdom 1 U41 HG006104-03/HG/NHGRI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England 2012/11/30 06:00 Nucleic Acids Res. 2013 Jan;41(Database issue):D584–9. https://doi.org/10.1093/nar/gks1157. Epub 27 Nov 2012.

88. Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D. The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res. 2005;33(Database issue):D212–5. Bru, Catherine Courcelle, Emmanuel Carrere, Sebastien Beausse, Yoann Dalmar, Sandrine Kahn, Daniel eng Research Support, Non-U.S. Gov't England 2004/12/21 09:00 Nucleic Acids Res. 2005;33(Database issue):D212–5. https://doi.org/10.1093/nar/gki034.
89. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009;37(Database issue):D211–5. Hunter, Sarah Apweiler, Rolf Attwood, Teresa K Bairoch, Amos Bateman, Alex Binns, David Bork, Peer Das, Ujjwal Daugherty, Louise Duquenne, Lauranne Finn, Robert D Gough, Julian Haft, Daniel Hulo, Nicolas Kahn, Daniel Kelly, Elizabeth Laugraud, Aurelie Letunic, Ivica Lonsdale, David Lopez, Rodrigo Madera, Martin Maslen, John McAnulla, Craig McDowall, Jennifer Mistry, Jaina Mitchell, Alex Mulder, Nicola Natale, Darren Orengo, Christine Quinn, Antony F Selengut, Jeremy D Sigrist, Christian J A Thimma, Manjula Thomas, Paul D Valentin, Franck Wilson, Derek Wu, Cathy H Yeats, Corin eng BB/F010508/1/Biotechnology and Biological Sciences Research Council/United Kingdom 087656/Wellcome Trust/United Kingdom GM081084/GM/NIGMS NIH HHS/ Wellcome Trust/United Kingdom BB/F010435/1/Biotechnology and Biological Sciences Research Council/United Kingdom Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England 2008/10/23 09:00 Nucleic Acids Res. 2009;37(Database issue):D211–5. https://doi.org/10.1093/nar/gkn785. Epub 21 Oct 2008.
90. Nayfach S, Pollard KS. Toward accurate and quantitative comparative metagenomics. Cell. 2016;166(5):1103–16. Available from: http://dx.doi.org/10.1016/j.cell.2016.08.007.
91. Manor O, Borenstein E. MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. Genome Biol. 2015;16(1):53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25885687%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4391136.
92. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014;10(4):1–11.
93. Silverman JD, Washburne AD, Mukherjee S, David LA. A phylogenetic transform enhances analysis of compositional microbiota data. eLife. 2017;6:1–20.
94. Li H. Microbiome, metagenomics, and high-dimensional compositional data analysis. Ann Rev Stat Appl. 2015;2(1):73–94. Available from: http://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-010814-020351?journalCode=statistics.
95. Kurtz ZD, Mueller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol. 2015;11(5):1–25.
96. Gloor GB, Reid G. Compositional analysis: a valid approach to analyze microbiome high throughput sequencing data. Can J Microbiol. 2016;703(April):2015–0821. Available from: http://www.nrcresearchpress.com/doi/abs/10.1139/cjm-2015-0821#.VxVj4pMrJIX.
97. Kumar MS, Slud EV, Okrah K, Hicks SC, Hannenhalli S, Corrada Bravo H. Analysis and correction of compositional bias in sparse sequencing count data. bioRxiv. 2017;1–34. Available from: http://www.biorxiv.org/content/early/2017/05/27/142851?%3Fcollection=.
98. Tsilimigras MCB, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann Epidemiol. 2016;26(5):330–5. Available from: http://dx.doi.org/10.1016/j.annepidem.2016.03.002.

99. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. App Environ Microbiol. 2005;71(12):8228–8235.
100. Lozupone C, Hamady M, Kelley S, Knight R. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol. 2007;73(5):1576–1585.
101. Zvelebil M, Baum J. Understanding bioinformatics. New York: Garland Science; 2008.
102. Cover TM, Thomas JA. Elements of information theory. New York: Wiley-Interscience; 2006.
103. Kira K, Rendell L. A practical approach to feature selection. In: National conference on artificial intelligence; 1992.
104. Hall MA. Correlation-based feature selection for discrete and numeric class machine learning. In: Proceedings of the seventeenth international conference on machine learning; 2000, p. 359–366. Available from: http://www.ime.unicamp.br/~wanderson/Artigos/correlation_based_feature_selection.pdf.
105. Brown G, Pocock A, Zhao MJ, Luján M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J Mach Learn Res. 2012;13:27–66.
106. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc. 1996;58(1):267–88.
107. Bates S, Tibshirani R. Log-ratio Lasso: scalable, sparse estimation for log-ratio models. 2017;1–24. Available from: http://arxiv.org/abs/1709.01139.
108. Ditzler G, Morrison JC, Lan Y, Rosen G. Fizzy: feature selection for metagenomics. BMC Bioinform. 2015;16(358):1–8.
109. Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. J Comput Graph Stat. 2006;15(2):262–86.
110. Blair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. J Am Stat Assoc. 2006;101(473):119–37.
111. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Elements. 2009;1:337–87. Available from: http://www.springerlink.com/index/10.1007/b94608.
112. Hotelling H. Relations between two sets of variates. Biometrika. 1936;28(3):321–77.
113. van der Maaten L, Hinton GE. Visualizing high-dimensional data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
114. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966;53(3/4):325. Available from: http://www.jstor.org/stable/2333639?origin=crossref.
115. Hirschfeld HO. A connection between correlation and contingency. Math Proc Camb Philos Soc. 1935;31(4):520–24. Available from: http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=1737020%5Cnhttp://journals.cambridge.org/action/displayFulltext?type=1&fid=2109508&jid=&volumeId=&issueId=04&aid=1737020&bodyId=&membershipNumber=&societyETOCSession=.
116. Kenkel NC, Orloci L. Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. Ecology. 1986;67(4):919–928.
117. Kruskal JB. Nonmetric multidimensional scaling: a numerical method. Psychometrika. 1964;29(2):115–29.
118. Legendre P, Legendre L. Numerical ecology. Amsterdam: Elsevier Science; 2008.
119. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. bioRxiv. 2017;142760. Available from: https://www.biorxiv.org/content/early/2017/05/28/142760.full.pdf+html.
120. Tan J, Doing G, Lewis KA, Price CE, Chen KM, Kyle C, et al. System-wide automatic extraction of functional signatures in Pseudomonas aeruginosa with eADAGE. bioRxiv. 2016, p. 1–25.
121. Xie R, Wen J, Quitadamo A, Cheng J, Shi X. A deep auto-encoder model for gene expression prediction. BMC Genomics. 2017;18(S9):845. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4226-0.
122. Mikolov T, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. CrossRef Listing of Deleted DOIs. 2000;1:1–9. Available from: http://www.crossref.org/deleted_DOI.html.

123. Ng P. dna2vec: consistent vector representations of variable-length k-mers. 2017;1–10. Available from: http://arxiv.org/abs/1701.06279.
124. Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization. Adv Neural Inf Process Syst (NIPS). 2014;2177–85. Available from: http://papers.nips.cc/paper/5477-neural-word-embedding-as-implicit-matrix-factorization.
125. Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings. Trans Assoc Comput Linguist. 2015;3:211–25. Available from: https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/570.
126. Landgraf AJ, Bellay J. word2vec skip-gram with negative sampling is a weighted logistic PCA. 2017;1–5. Available from: http://arxiv.org/abs/1705.09755.
127. Mikolov T, tau Yih W, Zweig G. Linguistic regularities in continuous space word representations. In: North American Chapter of the Association for Computational Linguistics. 2015.
128. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. CoRR. 2013;abs/1301.3781. Available from: http://arxiv.org/abs/1301.3781.
129. Rao C. The use and interpretation of principal component analysis in applied research; 1964. Available from: http://www.jstor.org/stable/25049339.
130. Legendre P, Andersson MJ. Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. Ecol Monogr. 1999;69(1):1–24.
131. ter Braak CJ. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology. 1986;67(5):1167–79.
132. Blanchet G, Legendre P, Borcard D. Forward selection of spatial explanatory variables. Ecology. 2008;89(9):2623–32.
133. Clarke KR, Ainsworth M. A method of linking multivariate community structure to environmental variables. Marine ecology progress series. 1993;92:205–219.
134. MacKelprang R, Waldrop MP, Deangelis KM, David MM, Chavarria KL, Blazewicz SJ, et al. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature. 2011;480(7377):368–71.
135. Borcard D, Gillet F, Legendre, Legendre P. Numerical ecology with R. Springer. 2011.
136. McCune B, Grace JB. Analysis of ecological communities. Gleneden Beach: MjM Software Design; 2002.
137. Ramette A. Multivariate analyses in microbial ecology. Fems Microbiology Ecology 2007;62(2):142–160. Available from: http://doi.org/10.1111/j.1574-6941.2007.00375.x.
138. Ter Braak CJF. Canonical community ordination. Part I: basic theory and linear methods. Ecoscience. 1994;1:127–40.
139. Gelman A, Stern H. The difference between significant and not significant is not itself statistically significant. Am Stat. 2006;60(4):328–31.
140. Zuur AF, Ieno EN, Elphick CS. A protocol for data exploration to avoid common statistical problems. Methods Ecol Evol. 2010;1(1):3–14. Available from: http://doi.wiley.com/10.1111/j.2041-210X.2009.00001.x.
141. Hoff PD. A first course in Bayesian statistical methods, vol. 64; 2009. Available from: http://books.google.com/books?id=9tv0taI8l6YC%5Cnhttp://www.amazon.com/Bayesian-Statistical-Methods-Springer-Statistics/dp/0387922997.
142. Team SD. Stan modeling language. User's guide and reference manual. 2017; p. 1–488. Available from: http://mc-stan.org/manual.html%5Cnpapers2://publication/uuid/C0937B19-1CC1-423C-B569-3FDB66090102.
143. Paliy O, Shankar V. Application of multivariate statistical techniques in microbial ecology. Mol Ecol. 2016;25(5):1032–57.
144. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053721&tool=pmcentrez&rendertype=abstract.
145. Love MI, Anders S, Huber W. Differential analysis of count data – the DESeq2 package, vol. 15; 2014. Available from: http://biorxiv.org/lookup/doi/10.1101/002832%5Cnhttp://dx.doi.org/10.1186/s13059-014-0550-8.

146. Paulson J. MetagenomeSeq: statistical analysis for sparse high-throughput sequencing. BioconductorJp. 2014;1–20. Available from: http://bioconductor.jp/packages/2.14/bioc/vignettes/metagenomeSeq/inst/doc/metagenomeSeq.pdf.
147. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12:R60(1–18).
148. Jonsson V, Österlund T, Nerman O, Kristiansson E. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC Genomics. 2016;17(1):78. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4727335&tool=pmcentrez&rendertype=abstract.
149. Mitchell TM. Machine learning. 1st ed. New York: McGraw-Hill, Inc.; 1997.
150. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol. 2017;18(1):182. Available from: https://doi.org/10.1186/s13059-017-1299-7.
151. Chatterji S, Yamazaki I, Bai Z, Eisen J. CompostBin: a DNA composition-based algorithm for binning environmental shotgun reads. ArXiv e-prints. 2007 Aug.
152. Rosen G, Garbarine E, Caseiro D, Polikar R, Sokhansanj B. Metagenome fragment classification using N-mer frequency profiles. Adv Bioinform. 2008;2008(205969):1–12164: e79(1–11).
153. Rosen GL, Reichenberger ER, Rosenfeld AM. NBC: the Naive Bayes classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics. 2011;27(1):127–9. Available from: +http://dx.doi.org/10.1093/bioinformatics/btq619.
154. Borozan I, Watt S, Ferretti V. Integrating alignment-based and alignment-free sequence similarity measures for biological sequence classification. Bioinformatics. 2015;31(9):1396–404.
155. Wang Y, Leung H, Yiu S, FY C. MetaCluster 4.0: a novel binning algorithm for NGS reads and huge number of species. J Comput Biol. 2012;19(2):241–9.
156. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9. Available from: +http://dx.doi.org/10.1093/bioinformatics/btl158.
157. Ghodsi M, Liu B, Pop M. DNACLUST: accurate and efficient clustering of phylogenetic marker genes. BMC Bioinformatics. 2011;12(1):271. Available from: https://doi.org/10.1186/1471-2105-12-271.
158. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–1. Available from: +http://dx.doi.org/10.1093/bioinformatics/btq461.
159. Yoon BJ. Hidden Markov models and their applications in biological sequence analysis. Curr Genomics. 2009;10(6):402–15.
160. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 2010;38(20):e191. Available from: +http://dx.doi.org/10.1093/nar/gkq747.
161. Noguchi H, Park J, Takagi T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res. 2006;34(19):5623–30. Available from: +http://dx.doi.org/10.1093/nar/gkl723.
162. Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics. 2012;99(6):323–9.
163. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006;22(22):2729–34. Available from: +http://dx.doi.org/10.1093/bioinformatics/btl423.
164. Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the Lasso and generalizations. Boca Raton: CRC; 2015; p. 362.
165. Hughey JJ, Butte AJ. Robust meta-analysis of gene expression using the elastic net. Nucleic Acids Res. 2015;43(12):e79(1–11). Available from: http://doi.org/10.1093/nar/gkv229.
166. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science (N Y). 2001;291(5507):1304–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11181995.