# Advances in Machine Learning for Processing and Comparison of Metagenomic Data

Jean-Luc Bouchot[1], William L. Trimble[2], Gregory Ditzler[3], Yemin Lan[4], Steve Essinger[3], and Gail Rosen[3]

[1]Department of Mathematics, Drexel University, 3141 Chestnut street, 19104 Philadelphia, Phone: (215) 895-1849, Fax: (215) 895-1582, jean-luc.bouchot@drexel.edu

[2]Institute for Genomics & Systems Biology, Argonne National Laboratory, University of Chicago, 900 East 57th Street Chicago, IL 60637, Phone: (630) 252-4220, trimble@anl.gov

[3]Department of Electrical & Computer Engineering, Drexel University, 3141 Chestnut street, 19104 Philadelphia, Phone: (215) 895-0400, Fax: (215) 895-1695, gregory.ditzler@gmail.com, sessinger@drexel.edu, gailr@ece.drexel.edu

[4]School of Biomedical Engineering, Science and Health, Drexel University, 3141 Chestnut street, 19104 Philadelphia, Phone: (215) 895-0400, Fax: (215) 895-1695, yeminlan@gmail.com

**Abstract**

Recent advances in next-generation sequencing have enabled high-throughput determination of biological sequences in microbial communities, also known as microbiomes. The large volume of data now presents the challenge of how to extract knowledge–recognize patterns, find similarities, and find relationships–from complex mixtures of nucleic acid sequences currently being examined. In this chapter we review basic concepts as well as state-of-the-art techniques to analyze hundreds of samples which each contain millions of DNA and RNA sequences. We describe the general character of sequence data and some of the processing steps that prepare raw sequence data for inference. We then describe the process of extracting features from the data, in our case assigning taxonomic and gene labels to the sequences. Then we review methods for cross-sample comparisons: 1) using similarity measures and ordination techniques to visualize and measure distances between samples and 2) feature selection and classification to select the most relevant features for discriminating between samples. Finally, in conclusion, we outline some open research problems and challenges left for future research.

## Keywords

Dimensionality reduction; Feature selection; Gene prediction; Taxonomic classification; Gene annotation; Similarity measures; Metagenomic sample comparison;

# 1 Introduction

Metagenomics is the study of nucleic acids extracted from the environment, as opposed to genomics, which studies the nucleic acids derived from single organisms. In a metagenomic study, a sample is collected from the environment, which can be a gram of soil [1,2], milliliter of ocean [3], swab from an object [4], or a sample of the microbes associated with a larger organism, such as humans [4,5], sometimes called the "microbiome". Until now, microbes were usually studied in isolation, whereby researchers literally isolated and cultured the organism to sequence and study its genome and gene functions. However, microbes actually live in communities, cooperating with and competing against each other. While found commonly in soil, water, buildings, etc. in our everyday lives, microbes are also found in unusual places like the extremely cold Antarctica [6], extremely hot springs [7], and the hyper saline Dead Sea [8]. They regulate the global carbon and nitrogen cycles [9, 10] of the Earth and are thought to be responsible for half the oxygen on the Earth [11].

It is now thought that these communities of microbes not only play a large role in the environment but also human health. Microbes are found at almost every interface of the body, including skin, mouth, airways, and even places like the lungs and amniotic fluid once thought to be sterils [12, 13]. It is thought that like environmental ecosystems, the more diverse the human ecosystem, the better we can ward off disease [14, 15]. The concept of microbiome has additionally spurred the hypothesis that human hosts entertain multiple stable ecological community types, termed enterotypes [16], and has been suggested as a forensics tool [17, 18]. Though many groups are pursuing metagenomic sequence data the computational metagenomic methods used to study the communities are underdeveloped, so we discuss recent methods in the paper. However, almost no one has scratched the surface to use the findings of these studies to engineer microbiomes to improve the environment, generate biofuels, and cure disease. This leaves to the imagination – how can personal omics profiling revolutionize medicine [19]?

Currently to profile a metagenomic sample, DNA or RNA is extracted chemically and turned into purified DNA. This is prepared and fed into a machine that determines the sequence of information-containing monomers in DNA fragments. This process is called "sequencing" and has, thanks to technological improvements in the recent decade, become much faster and much cheaper, producing millions of short strings of data, called sequencing **reads**, representing millions of biological molecules.

From this digitization of DNA, biologists can address the questions "Who is there?" and "What are they doing?". [20–23]. The answer to the "who" question, obtained by inferring the name or position in the taxnomy of the organism from which a sequence was likely derived, is called taxonomic classification. The answer to the "what" question, the process of recognizing the biochemical functions of the sampled genes, is called functional annotation. The current algorithms to solve these problems [24, 25] fall short of the speed and accuracy required to process and compare the volumes of data currently being generated. In addition, algorithmic procedures to study how environmental factors affect microbial populations are under active development.

Sequence data comes in two broad kinds–the sequencing of targeted genetic loci, selected by PCR, called **amplicon sequencing**, and the sequencing of random genetic loci, called **shotgun sequencing**. Specific subsets of the 16S ribosomal subunit gene in prokaryotes

and the ribosomal spacer ITS in fungi have been popular targets for amplicon metagenomics. For both sequencing methodologies, sequences are filtered, transformed, and interpreted by explicit or implicit comparison to a database of sequences that are presumed known. This approach is called "closed-reference annotation." The number of possible sequences is extremely large: $4^{100} = 10^{60}$ possible 100-basepair(bp) reads. Annotation provides the first round of dimensionality reduction, mapping from this extremely highly dimensional space to the merely large $(10^6 - 10^8)$ vector space of annotations.

Sequencing techniques naturally divide into targeted and shotgun sequencing. Shotgun data are more complex, require greater sequencing depth, and have thousands more possible annotations. Targeted sequencing, called amplicon sequencing, are generally less expensive and permit larger numbers of samples to be sequenced for similar cost, but are confined to providing examples of sequences of a specific gene, potentially answering only the "Who" question. Shotgun sequences are able to address the "What" question because they consist mostly of protein-coding sequence which has biochemical function, but cost much more to analyze. Though amplicon and shotgun sequence data products are dramatically different, the analytical techniques for comparing samples and drawing inferences post-annotation are very similar.

We describe the process of generating sequence data, a variety of procedures to extract features from the sequence data, the mathematical procedures for sample comparison, visualization, and feature selection, and describe future challenges for the interpretation of sequence data. Since metagenomic data comprise sequences from unknown mixtures of unknown organisms, the field is challenging indeed.

# 2 Preprocessing

Nucleic acid sequences are inferred from highly-multiplexed data acquisition systems which, in batches, capture signals that reveal the sequence of $10^5 - 10^9$ individual molecules, often called templates. Data from sequencing instruments passes through a number of steps before comparison to existing annotated biological sequence data. Collectively called preprocessing, the steps of base calling, sequence filtering, vector removal, sequence compression and assembly, and gene prediction are steps to filter, condition, and compress sequence data before annotation.

## 2.1 Base calling

All modern sequencing techniques use spatial separation to multiplex the large number of templates and have an analog-to-digital conversion that converts the chemical information into digital data. Depending on the particular sequencing technology, the raw signals can be single-color images (pyrosequencing), four-color images (Illumina, PacBio, ABIsolid), or arrays of potential sensors (Iontorrent). The initial stage of recognizing the sequence of bases from the raw instrument output is called **base calling**; this is usually done with vendor-provided software. The basecallers produce the symbols A, C, G, and T, to indicate the four bases, and the symbol N to indicate complete uncertainty in the identity of the underlying base. Sequences containing N (the **ambiguous base symbol**) require special handling and

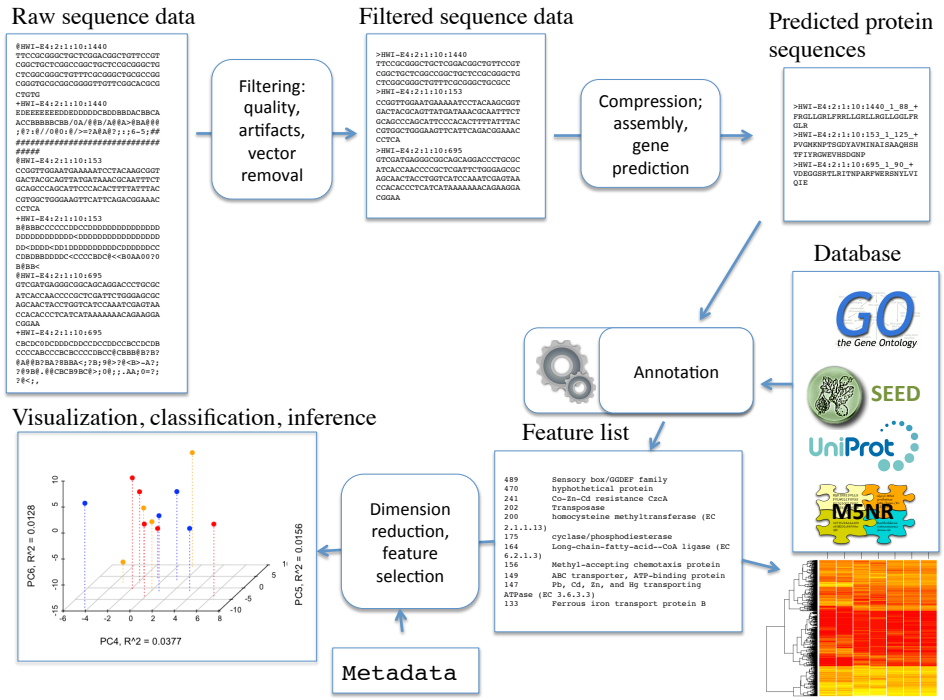


Figure 1: Diagram of generic sequence analysis workflow from raw data to visualization and classification. Each analytical step removes information; this results in a smaller volume and smaller dimensionality of data at later stages of processing.

interpretation. Base calling produces a set of short (25-600 bp) sequences, called reads, and a per-base-pair indication of the base-caller's posterior probability of having reported an accurate base. The per-base quality score is sometimes called a Phred score after an early automated base-calling program that established the encoding standard. [26] The standard format for data from the base caller is FASTQ. FASTQ files have a variety of encoding schemes produced by different vendors; the oldest, "Sanger-style" encoding is preferred for data sharing. [27]

## 2.2 Demultiplexing

It is possible to simultaneously sequence multiple samples by attaching artificial sequences which are recognized as sample identifier labels. Called "multiplex identifiers" or more commonly **barcodes**, these are synthetic sequences that are not part of the biological signal, but permit many samples to be run at lower per-sample depth and cost. The recognition of barcodes and separation of a mixture of samples into sequence data for individual samples is called **demultiplexing** and is an essential part of sequencing protocols which use pooled samples.

## 2.3 Quality filtering

Sequence quality data can be used to filter the raw instrument output to improve the signal-to-noise characteristics of the sequence data. Filtering can be applied to whole reads (accepting or rejecting each read), can selectively discard low-quality basepairs from the ends of reads, or can remove reads with characteristics that suggest sequencing artifacts. A number of recipes for sequence quality filtering exist [28–30], and these generally lower the rate of erroneous sequence recovery without removing large fractions of the sequence data. Some sequence analysis methods analyze sequences in a quality-aware way [31], but most of them depend only on the sequence, not on the qualities. Some workflows discard the quality values at this stage.

## 2.4 Contaminant screening

The template DNA usually requires chemical modifications to correctly interface with the sequencing instrument; kits for these steps are provided by the instrument vendors. These modifications include shearing and size-selecting the template molecules and ligating artificial sequences containing the barcodes, primers for in-instrument PCR, and other artificial nucleic acid constructs that make sequencing possible. These sequences, called **adapters** are not intended to be part of the sequencing output, but occur in a minority of output sequences, ranging from 0.1% to 5%. These unwanted sequences can be removed by similarity searching and filtering; the significance of their effect as signal contaminants is not known.

When metagenomic samples are extracted to study the microbial communities associated with animals and plants, considerable amounts of the eukaryotic host DNA may end up being sequenced. This "host" DNA must be analyzed separately, and it is standard to remove DNA which matches (by similarity) the genome of the host organism before annotation.

## 2.5 Clustering and assembly

After quality filtering, clustering and assembly are two approaches to reduce the size and redundancy of the set of sequences. Clustering and assembly are lossy compression approaches that identify redundancy in the input sequences and use this redundancy to reduce the amount of computational effort on similarity searching. Clustering (using, CD-HIT [32] or UCLUST [33]) is performed for improved speed, replacing a set of similar sequences with one representative. Assembly, on the other hand, both reduces sequence volume and replaces short sequences with inferred longer sequences. The compression afforded by clustering or assembly depends on the sequence-level redundancy of the input data. Some data sets with dominant genomes that comprise a majority of the sequence data have assemblies that reduce to 1% of the input sequence bulk, while others fail to compress significantly. These longer, derived sequences are called **contigs** and represent data from multiple instrument reads. If neither of these approaches is used, the filtered short reads can be annotated.

### 2.6 Gene prediction

Unlike clustering and assembly, which are principally technologically-inspired steps, gene prediction is a computational step which attempts to identify a biological pattern, mimicking the patterns recognized by transcription and translation machinery. Gene predictors take DNA sequences as their input, predict the start and stop sites of genes contained on those sequences, and produce in-silico translations of the genes so identified.

This translation step converts nucleic acid sequences into amino acid sequences, reducing the length of the sequences by a factor of about three. Since individual reads are shorter than typical microbial gene sizes (ca. 900 bp [34]), individual-read gene prediction produces mostly incomplete predicted protein sequences. This factor-of-three reduction in sequence length reflects the fact that gene prediction is a lossy compression step.

Gene prediction tools were developed for the annotation of complete or near-complete genomes, and were later adapted to handle short-read data. GLIMMER uses interpolated Markov models whose parameters are trained on long coding regions and smoothed to give predictions on shorter coding regions. [35] It is well suited for assemblies from single organisms.

For short reads or contigs from mixtures of organisms, one-size-fits-all gene prediction tools are indicated. MetaGeneMark [36, 37] and MetaGeneAnnotator [38] were early applications of Markov models to gene prediction; they deliver good results on error-free data. More recent and more elaborate gene predictors are FragGeneScan(FGS) [39], which uses a hidden-Markov model, and Prodigal [40], which uses dynamic programming, are engineered to perform well on short reads. FGS and Prodigal have more robustness against sequencing error.

Gene prediction accuracy in complete genomes is reported as better than 95%; for short reads (or short contigs) the accuracy is lower. It should be mentioned that gene prediction is not equally sensitive across taxa; some organisms have genes which the gene prediction tools miss 20% of the time. The loss of sensitivity for some organisms is more severe for shorter reads (less than 200 bp). The increased bias and reduced sensitivity of short-reads drives many researchers to perform assembly of short-reads prior to annotation; this exchanges the biases caused by short reads for yet-uncharacterized biases caused by assembly.

## 3 Annotation of genes

At this stage, the biologist's central questions, "Who" and "What" are addressed by classifying the sequences. Annotation is the process of assigning biological meaning to the sequences, usually after gene sequences are identified from bulk reads or from partially-assembled contigs. Annotation consists of comparison, either explicit (similarity searching against a database of sequences) or implicit (searching against models or profiles derived from sets of sequences) against a database of sequences that have already been named. These databases include previously annotated and/or manually curated sequences. Comparing new sequences to existing ones allows each sequence to be associated with the name of the organism, of the protein, of the function, or of the pathways associated with the protein in the database. **Taxonomic classification** refers to annotation that produces in-

Table 1: Summary of the homology-based and composition-based methods for WGS taxonomic classification

| **Features** | **Classifier** | **Published Method** |
|---|---|---|
| Similarity-based | Alignment | BLAST [41], CARMA3 [42], MetaPhyler [43], MetaPhlan [44] |
| | Alignment + Last Common Ancestor | MEGAN [45], MARTA [46], MTR [47], SOrt-ITEMS [48] |
| | Clustering + Alignment | jMOTU [49] |
| Composition-based | Naïve Bayesian | NBC [50, 51] |
| | Support Vector Machines | PhyloPythiaS [52] |
| | Interpolated Markov Models (IMM) | Phymm [53] and Scimm [54] |
| | Miscellaneous | TACOA [55], RAIphy [56], and MetaCluster [57] |
| Mapping | Bloom Filter | FACS [58] |
| | Burrows-Wheeler Transform | Bowtie2 [59], BWA [60], SOAP [61] |
| | Miscellaneous | CLC [62], MAP [63], SMALT [64] |
| Phylogeny | Maximum-Likelihood | EPA [65], pplacer [66], FastTree [67] |
| | Miscellaneous | SAP [68] |
| Hybrid | IMMs+BLAST | PhymmBL [53] |
| | NBC+BLAST | RITA [69] |
| | $k$-mer Clustering + BLAST | SPHINX [70] |
| | Alignment + Phylogeny | PaPaRa [71], AMPHORA [72], MLTreeMap [73], TreePhyler [74], (NAST, Simrank) [75] |

ferences about organism name; **functional classification** concerns itself with identifying biochemical function from protein sequences.

## 3.1 Taxonomic classification

To answer the "Who is there" question (and its quantitative counterpart, "how much of each"), methods are needed that are capable of classifiying newly-observed sequences using information from an existing database of sequences and annotations.

Factors which can impact the classifier's accuracy include read length and sequence novelty. Classifiers are expected to act on short (less than 200 bp) reads and on variable-length, but longer, assembled contigs. Short reads, however, can fail to be unique within the sequence database, thus yielding ambiguous classification. Some of this ambiguity can be overcome by longer reads but some reflects fundamental similarity between annotated sequences. In the opposite direction, sequences can be ambiguous because they lack a good match in the database. The appropriate annotation and interpretaion of these novel sequences remains a serious challenge for taxonomic classification and annotation in general.

There are four primary methods (see Table 1) to perform taxonomic classification of genome fragments: *homology*, *mapping*, *composition*, and *phylogeny* based methods. A fifth category is emerging that combines two or more of these types. However, hybrid methods typically take longer to run [25, 44].

Many current approaches align sequenced fragments to known genomes using sequence similarity. This approach has a rich history from BLAST [41], one of the first optimized alignment tools. While homology-based techniques perform well in identifying protein families and gene homology [76, 77], this computation takes longer and can be more costly than sequencing itself, as noted by [78, 79]. Also, the e-values provided by these tools signify that the sequence does not match by chance, but they do not indicate a particular confidence to the chosen class-label (e.g. if the sequence may belong to two protein families above random chance, which family's assignment is more confident?)

New methods have emerged that perform similarity searches using dramatically fewer computational resources than BLAST and FASTA. Called read mappers, these employ data structures including the Burrows-Wheeler Transform [59–61] and Bloom filters [80]. The Burrows-Wheeler transform(BWT) is a a type of transform that makes the sequence compressible and fast to search. Read mappers based on BWT have quite fast mapping times with high sensitivity and low false positive rates. These techniques promise to be fast but as noted, the false positive rate must be optimized. Bloom filters are probabilistic data structures based on hashing that have efficient lookup times, have no false negatives, and have "manageable" false positive rates.

Composition-based classification approaches use features of length-$k$ motifs, or $k$-mers, and usually build models based on the motif frequencies of occurrence. Intrinsic compositional structure has had many applications in sequence analysis: Markov models [81], in tandem repeat detection [82, 83], inference of evolutionary relationships based on di-, tri-, and tetra-nucleotide compositions [84–91], and the examination of longer oligomers for genomic signatures [92]. Wang et al. [93] use a naive Bayes classifier with 8-mers ($k$-mers of length 8) for 16S recognition. Sandberg et al. pioneered work for whole- genome shotgun sequencing [94]. However, taxonomic classification of WGS sequences has been developed since with a few naive Bayes classifier implementations, support vector machines, interpolated Markov models, and probabilistic analyses of $k$-mers (see Table 1). The advantage of composition-based methods is that they are fast, but they have difficulty assessing the true confidence of their assignments.

Phylogenetic methods attempt to classify a sequence and infer its placement on a phylogenetic tree. These programs aim to address a common question in biology: how is the sequence under study related to the known sequences? These methods infer the position of the new sequence in a tree describing inferred evolutionary relationships between sequences. Note, however, that not all programs compute the branch length. The advantages of phylogenetic methods are to assign taxonomy at upper-level and lower-level ranks, making "novelty-detection" inherent. However, these methods are very computationally intense.

In addition, many "hybrid" techniques are now emerging that attempt to combine usually composition and homology based methods, since they often complement each other. These techniques often combine the fine resolution of composition-based methods with the more general similarity measures of homology. There are some phylogenetic algorithms that do a pre-processing alignment against a precomputed reference alignment of marker genes before phylogenetic placement (see Hybrid Alignment+Phylogeny in Table 1).

## 3.2 Protein similarity searches and databases

The "Who is there" question can be approached using either protein sequences or nucleic acid sequences, but the "What are they doing" question is best answered with proteins. Starting from predicted protein sequences from a measured dataset, protein annotation tries to infer the most likely function of the gene sampled. This has been an extremely prolific area with homology searches being used to identify function from similar protein domains. However, proteins are complicated: similar sequences sometimes have different function and distant sequences sometimes have similar functions. Consequently, much effort has been placed not only in the methods to identify function but also into development of comprehensive

databases. Protein similarity searching is a step that usually requires the most computation– owing to the size of the databases used for comparison.

Researchers can infer the function of unknown metagenomic sequences from the functions of known sequences in the databases that they resemble the most. A variety of functional databases has enabled researchers to view gene composition from various angles. The databases are built upon reference sequences and have different types and different levels of detail of additional information about the sequences. Some organize protein functions into hierarchies at different resolutions, while others emphasize proteins in particular specialties. This section describes some of the databases used to annotate protein-coding genes, which constitute a major part of most metagenomic samples.

Two of the largest reference sequence collections are **the Universal Protein Resource (UniProt) databases** [95], and **NCBI's RefSeq database** [96]. The former are a set of protein databases developed to provide comprehensive knowledge on various protein functions. The UniProtKB/Swiss-Prot database, which is probably the most popular among UniProt databases, has manually curated annotations by expert reviewers and covers a variety of protein functions. The latest release of UniProtKB/Swiss-Prot in November 2012 contains 538,585 sequence entries coming from 12,930 species. There is also a database specifically developed for metagenomic and environmental data called the UniProt Metagenomic and Environmental Sequences (UniMES). It is composed of metagenomic sequences clustered into groups (functional features) using CD-HIT algorithm [97]. RefSeq contains a variety of non-redundant and curated DNA, RNA and protein sequences. While annotations are mainly available for a subset of the database, especially human sequences, RefSeq also profiles conserved domains from NCBI's Conserved Domain Database and protein features from UniProtKB/Swiss-Prot. The database now includes 17,977,767 proteins comprised of 6,003,283,860 amino acids, and includes the complete genomes of 18,512 organisms (release 56 in November 2012). Refseq is continually updated with newly sequenced organisms.

Although tens of millions of genes have now been annotated, the variety of descriptions of these genes can still be hard to summarize or interpret. With this concern in mind, the **Gene Ontology** database [98] strives to standardize the representation of gene and gene product attributes across species and databases. The annotation of unknown sequences falls into a well controlled vocabulary, consequently, it is widely used for interpreting gene functions. While Gene Ontology started out as a database for all eukaryotes, it now contains a prokaryote-specific subset.

The **Clusters of Orthologous Groups of proteins (COGs)** [99] database categorizes the conserved sequences across complete genomes based on orthologous relationships between them. Each COG contains either a single protein or an orthologous group of proteins from multiple genomes, and is classified into one of 23 functional categories. Such high-level classification provides a general view of metagenomic constitutions, and allows for low-resolution comparison between samples. The latest COGs in 2003 contain 4,873 entries, covering proteins from 66 prokaryote and unicellular eukaryote genomes. The eukaryotic version of it, **eukaryotic orthologous groups (KOGs)** [99], includes 7 eukaryotic genomes. To further improve the orthologous groups database by adding more annotated genomes, **eggNOG** [100] inherits the functional categories from COGs/KOGs but has expanded to contain 721,801 orthologous groups of 4,396,591 proteins, covering 1133 species (as of January 2013). This database is constructed through identification of reciprocal best BLAST

matches and triangular linkage clustering. Compared to COGs/KOGs, eggNOG has a finer phylogenetic resolution and a hierarchical classification of protein function, as well as offers more frequent database updates.

Separating orthologous groups, however, is not the only way to explore the diversity of functional roles. Analyzing all biological reactions that can possibly be completed by the gene content is another quest popular among bioinformaticians. The **Kyoto Encyclopedia of Genes and Genomes (KEGG)** [101] provides manually drawn pathways that map genes and other molecules into chains of reactions. It is a good way to examine and compare the functional composition of metagenomes, and to identify genes playing a role in specific biological processes. As of January 2013, the KEGG Pathway has 435 pathway maps in total. The **Metacyc** [102] database is a different pathway database belonging to the **BioCyc** database collection [102]. It contains more than 1,928 experimentally elucidated metabolic pathways from 2,362 organisms (version 16.5 released in November 2012). About 38% of the reactions can be linked to KEGG. However, compared to KEGG pathways that are constructed typically based on reactions from multiple species, Metacyc pathways are often smaller and picture reactions within single organisms. To view the pathways in a different way, **SEED** [103] provides an annotation environment that generalizes pathways into 891 subsystems (as of January 2013). Each subsystem is a set of functions that piece together a specific pathway, biological process or structural complex. As one of the initial attempts to address comprehensive, expert-curated metabolic subsystems for genomic analysis, SEED annotation has been included by almost all of the large-scale analysis pipelines.

## 3.3 Protein domain annotation

Protein domain annotation is an alternative approach to sequence similarity searches. Rather than relying on pairwise sequence alignments, domain annotation uses models that represent conserved protein regions. These conserved regions are inferred from multiple sequence alignments (of database proteins) and have higher theoretical sensitivity than sequence comparisons lacking weights informed by biological conservation. Conserved protein regions, or domains/families, are represented by Hidden Markov Models (HMMs) built from multiple sequence alignment. Instead of sequence similarity, which searches for local alignments, protein domain annotation scans for small functional regions the genes may carry.

**The Pfam database** [104] is a large collection of protein domains/families containing 13,672 entries (Release 26.0 in November 2011). It is mainly comprised of sequences from UniProt Knowledge Database [95], NCBI GenPept Database [105] and Protein Data Bank [106]. Because of its wide coverage of proteins and intuitive naming strategies, Pfam has become a preferable annotation resource for protein domains/families. **TIGRFAMs** [107] are a similar set of models designed to be complementary to Pfam. The development of this database aims to provide functionally accurate names for genes being annotated, so that a well-informed annotator would assign the same protein name across different species with good confidence [108]. HMMER3 [109] is a software package that applies accurate probabilistic models for searching the HMM profile, and can be used to annotate sequences with Pfam or TIGRfam-derived clusters.

**FIGfam** [110] is another collection of protein families. Unlike Pfam's sensitive identification of all possible domains on the given sequences, FIGfam strictly requires full-length

sequence similarity and common domain structure within a protein family, sacrificing sensitivity for more accurate sequence binning. It currently contains more than 100,000 entries, coming from over 950,000 manually annotated proteins of many hundred bacteria and archaea. Two other multiple sequence alignment bases, TIGRFAM [107] and PIRSF [111], have a similar requirement for binning proteins into families. However, these two databases cannot compete with FIGfam in the number of manually curated sequences, coverage of families, and computation time [110].

It should be mentioned that there are many other protein family databases similar to the ones above but that focus on maintaining annotations for particular classes of proteins, such as **PROSITE** [112] that focuses on the biological functions of domains, **BLOCKS+** [113] that focuses on the family's characteristic and distinctive sequence features and **SMART** [114] that focuses on domains found in signaling proteins, extracellular and nuclear domains. Additionally, some databases use protein secondary structure in characterizing protein functions, such as in **SCOP** [115] and **CATH** [116] databases, aiming to provide a comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. There are also databases that take into consideration both sequence alignment and secondary protein structural information, such as **SUPFAM** [117] and **ProDom** [118].

## 3.4 ID mapping and computational complexity

As functional databases become more diverse, considerable effort has been made to connect similar entries of various databases, in order to gain comprehensive understanding of numerous gene annotations and avoid non-necessary repeat of annotation work.

Some databases incorporate annotations from other databases, such as **InterPro** [119] that integrates protein signatures from a variety of sources including Pfam, PROSITE, ProDom, PIRSF, CATH, TIGRFAM *etc.*, and the **M5NR database** [120] that integrates Gene Ontology, KEGG, SEED, UniProt, eggNOG *etc.* About 80% of the proteins in UniProtKB/Swiss-Prot have Pfam matches. Also, the entry mapping between many functional databases is made available on the UniProt website, including the mapping between most entries in UniProt, PIRSF, RefSeq, KEGG, eggNOG, BioCyc *etc.* In addition, the UniProt Gene Ontology Annotation (UniProt-GOA) database [121] provides high-quality Gene Ontology annotations to proteins in the UniProt Knowledgebase.

While gene annotation provides direct insight into the functional composition of metagenomes, the proportion of sequences that can be annotated are relatively limited. A gene annotation example for human gut metagenomes published in a comparative metagenomics study shows that most functional annotations are only able to assign less than 50% of the samples (Table 2) [122], revealing a need to improve current annotation methods and characterize genes with yet unknown functions.

Another challenge for similarity searching is the expensive computational cost compared with taxonomy classification, using searching tools either against reference sequences or HMMs. Using the dataset 6-19-DNA-flx(MG-RAST accession 4440276.3) as an example, a marine metagenomic sample collected from coastal waters in Norway. This dataset represents 68Mbases of metagenomic sequence data which reduces to 255,702 predicted protein sequences. It takes 34.0 CPU hours on a local machine (single 800MHz AMD Phenom(tm) II X6 1045T processor) to BLAST it against UniProtKB/Swiss-Prot database, resulting in

Table 2: Percentage of metagenome samples covered by gene annotation resources.

| Metagenomic samples | Sequences annotated | Gene annotation resource |
|---|---|---|
| MetaHit dataset (124 samples) | 77.1% | Taxonomic annotation |
| | 50.2% | Pfam |
| | 40.7% | KEGG Ontology |
| | 18.7% | KEGG Pathway |
| A multi-source dataset (52 samples) | 52.0% | Pfam |
| | 47.5% | Gene Ontology |

28.8% of the sequences annotated (e-value 10e-5), while it takes 59.9 CPU hours to search Pfams, annotating 30.4% of the sequences.

Shotgun metagenomic datasets at present comprise between $10^7$ and $10^{11}$ base pairs per sample, sometimes amounting to billions of reads. The databases to be searched have typically $10^9 - 10^{10}$ amino acids. Searching ever-larger sets of sequences for matches against growing databases of "known" sequences makes annotation the most computationally expensive step in sequence analysis [123]. With more metagenomic data made available by the use of high-throughput sequencing, faster annotation procedures are urgently needed.

### 3.5 Existing pipelines for metagenomic annotation

Several large-scale pipelines provide gene annotation for metagenomic data across various functional databases. MG-RAST [124], for example, is a fully automated analysis pipeline designed specifically for metagenomic data. It provides functional annotation on KEGG, eggNOG, COG and SEED subsystems on multiple levels of resolution. IMG/M [125] is another automated annotation pipeline. It can be used for annotating COGs, Pfams and KEGG pathways, with a preference for larger assembled contigs compared to MG-RAST. RAMMCAP (Rapid Analysis of Multiple Metagenomes with a Clustering and Annotation Pipeline) [126] provides implementation of HMMER for Pfam and TIGRFAM annotation, as well as BLAST for COG annotation.

## 4 Cross sample analysis

This section focuses on the last building blocks (from Figure 1) of metagenomic sample analysis. Here we are concerned with inferring information based on the community data matrix, as detailed in Figure 2.

We will first review some ways to compare data and project them for visualization purposes and then introduce methods to select a subset of features that should be the most useful for further processing.

### 4.1 Initial manipulation of annotation

In cross sample analysis, one is interested in comparing samples and finding the features that make different samples and samples with different mutated distinct. Samples represent

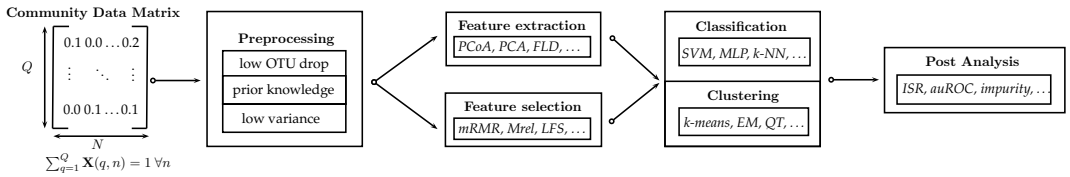


Figure 2: As a last step in the analysis of metagenomic data, feature selection methods are important in understanding microbial communities. This corresponds to a zoom in on the two last boxes of Fig 1

mixtures of microorganisms; these mixtures might contain slightly different lineages of a particular species of organisms, might contain the same organisms in different proportions, or might contain completely different organisms over varying environments. We will first examine the $k$-mer representation, an annotation-independent approach to creating features from sequence data. We then examine distance metrics, the mathematical procedures to compare datasets, and probability measures on phylogenetic trees, biologically-relevant distance measures. Finally we introduce some projection methods for dimensionality reduction for the visualization of high dimensional data.

### 4.1.1 Alignment-free sequence comparison: the $k$-mer representation

While sequence alignment is the traditional approach to analyzing DNA sequences, there are alternative approaches, called "alignment-free" sequence analysis methods, reviewed in [127]. One such is the $k$-mer representation of sequences, but the general principles apply to any vector space of features. We use the $k$-mer representation here as an example.

$k$-mers (alternatively called $n$-mers, $\ell$-tuples, $n$-grams in natural language processing, or "words") are subsets of a biological sequence of fixed length. The integer $k$ denotes the length of these subsets. Given an alphabet $\Omega = \{A, C, G, T\}$ a DNA sequence corresponds to a word $x$ of length $n$ on that alphabet, i.e. $x \in \Omega^n$. However, $n$ being very large in the case of genome analysis (in the order of magnitude of $10^6$-$10^7$), we prefer to decompose it in sub-words or sub-sequences of a given length $k$. There are two perspectives to represent the data: a set theory point of view and a frequentist or probabilistic point of view. Table 5 gives examples of the two representations in terms of dimers (kmers with $k = 2$) for two short DNA sequences.

Assume an input sequence $x$ of size $n$ and let us denote by $V_k(x) = \{\omega \in \Omega^k : \exists l \in \{1, \cdots, n\}, x_{l \cdots l+k-1} = \omega\}$. This is the set of all $k$-mers found anywhere in the read $x$.

$k$ is a parameter which can be chosen by the researcher. As $k$ increases the set of possible $k$-mers (numbering $4^k$) increases exponentially. Since biological sequences are finite, for large enough $k$ (in the range 8-14) the set of $k$-mers not encountered always becomes larger than the set of $k$-mers actually observed. Table 3 illustrates this behavior. It shows a calculation of the number of distinct $k$-mers contained in the 12817 bp woolly mammoth mitochondrial sequence (GenBank accession DQ188829.) Table 4 shows a similar calculation on the much larger (5.5 Mbase) genome of the bacterium *E.coli* O157:H7 (Genbank accession AE005174). As we can see from these two tables, the larger the size of the $k$-mers, the less likely a random $k$-mer is to occur. The probability distribution of these subsequences, and of numbers in the

$k$-mer representation, is an area of current study useful for interpreting comparisons between these representations.

Table 3: Evolution of the set of found $k$-mers with $k$ getting bigger.

| Mammoth (12817 nucleotides) | $\|V_k(x)\|$ | Max | Ratio |
|---|---|---|---|
| $k = 4$ | 256 | 256 | 100% |
| $k = 6$ | 3460 | 4096 | 84.47% |
| $k = 9$ | 15276 | 262 164 | 5.83% |

Table 4: Evolution of the set of found $k$-mers with $k$ getting bigger.

| E.coli (5528445 nucleotides) | $\|V_k(x)\|$ | Max | Ratio |
|---|---|---|---|
| $k = 4$ | 256 | 256 | 100% |
| $k = 6$ | 4096 | 4096 | 100% |
| $k = 8$ | 65450 | 65536 | 99.87% |
| $k = 10$ | 925235 | 1048576 | 88.24% |

#### 4.1.2 Distances and divergences

Once features (whether from taxonomic classification, protein annotation, or $k$-mer analysis) have been extracted from a set of sequences, the next most fundamental operation is the pairwise comparison of samples.

A straightforward tool for comparing two set-based representations $x$ and $y$ is known as the Jaccard index [128] and defined as:

$$J(x, y) = \frac{|V_k(x) \cap V_k(y)|}{|V_k(x) \cup V_k(y)|} \tag{1}$$

This represents the fraction of common elements over the set of all elements in these two sequences. We clearly have that $J(x, y) = 0$ if no $k$-mer from $x$ is found in $y$ and $J(x, y) = 1$ if all the $k$-mers of both sequences are found in the other. Note that this index only indicates if $x$ and $y$ contain similar sets of $k$-mers but does not imply similar ordering of the two sequences. Another well-known similarity based on set representations is the Sørensen similarity index. It is defined as:

$$S(x, y) = \frac{2 \cdot |V_k(x) \cap V_k(y)|}{|V_k(x)| + |V_k(y)|} \tag{2}$$

It can be seen as ratio of the the number of $k$-mers in common to the (potentially doubled) number of different $k$-mers found. Once again a 1 corresponds to the case where the sets of $k$-mers are exactly the same and 0 happens when there are no common elements. This similarity is also used in the context of frequency representation and known as the Bray-Curtis index in this case (see Eq. 3 for the details).

Another remark about the Jaccard (or the Sørensen) index is that it lacks information regarding the frequency of each $k$-mer which implies that we have to work on largest subsequences and hence makes it harder to process. Clearly a probabilistic representation of such

$k$-mers overcome that problem. A sequence of length $n$ has up to $m = n-k+1$ $k$-mers taken from the $4^k$ different options. Denote by $m_j(x),\ j \in \{1, \cdots, 4^k\}$ the number of occurrences of $k$-mer $j$ then the frequency representation of $x$ is given by the $4^k$ dimensional vector $\vec{x} = \{m_j(x)/m\}_{j=1}^{4^k}$. This representation can also be viewed as a probability distribution over the set of words of size $k$. It tells us *that the probability of picking $k$-mer $j$ by randomly picking a sub-sequence of size $k$ in the DNA sequence $x$ is $\vec{x}_j$.*

Another similarity coefficient often used for sequence comparison (and later for sample comparison) is known as the Bray-Curtis dissimilarity, sometimes inaccurately called a "distance". The Bray-Curtis dissimilarity is not a metric; it does not satisfy the triangle inequality, but is symmetrical. It is defined as

$$BC(x,y) = \frac{2 \cdot \sum_{i=1}^{4^k} \min(m_i(x), m_i(y))}{m(x) + m(y)} \tag{3}$$

It can be seen as the proportion of common $k$-mers given the sum of $k$-mers contained in either of the reads.

Table 5: Set and probabilistic description of two DNA subsequences when considering only 2-mers (representation for bigger $k$-mers is impossible). The last column counts the occurrences of each possible 2-mers (which implies a certain number of 0s)

| Sequence $x$ | $k$-mer set $V_2(x)$ | probability vector $\vec{x}$ |
|---|---|---|
| GTACGTACACACA | $\{GT, TA, AC, CG, CA\}$ | $\frac{1}{12}(0,4,0,0,3,0,1,0,0,0,0,2,2,0,0,0)$ |
| ATAGACATAGATA | $\{AT, TA, AG, GA, AC, CA\}$ | $\frac{1}{12}(0,1,2,3,1,0,0,0,2,0,0,0,3,0,0,0)$ |

Two DNA sequences (or two entire datasets) can be compared by evaluating the divergence of two probability measures. $f$-divergences, introduced independently in [129,130], are particularly appropriate to this task. Divergences are reviewed in [131]. Basseville [132](in French) offers a quite exhaustive survey of such measures as well.

**Definition 4.1 ($f$-divergence measures)** *Given two probability distributions $P$ and $Q$ absolutely continuous with respect to a reference measure $\mu$ over the set $\Omega$ and denote by $p$ and $q$ their probability density; moreover, let $f$ be convex and such that $f(1) = 0$, then the $f$ divergence of $P$ given $Q$ is defined as*

$$D_f(P||Q) = \int_\Omega f\left(\frac{p(x)}{q(x)}\right) q(x)\, d\mu(x) \tag{4}$$

For discrete probability distributions (as are the case in bioinformatics) the previous formulation is equivalent to the following:

$$D_f(P||Q) = \sum_i f\left(\frac{p(i)}{q(i)}\right) q(i) \tag{5}$$

This framework includes many of the metrics often used in DNA comparison such as the Hellinger distance and Kullback-Leibler divergence as special cases. Table 6 gives some examples of such $f$-divergences. Such divergence measures usually lack symmetry.

Table 6: Some examples of $f$-divergence measures

| Name | Formula | $f$ | Reference |
|---|---|---|---|
| Kullback-Leibler | $D_{KL}(P\|\|Q) = \sum_i q(i) \ln\left(\frac{q(i)}{p(i)}\right)$ | $t \mapsto -\ln(t)$ | [133] |
| Hellinger | $D_H(P\|\|Q) = \sum_i \left(\sqrt{p(i)} - \sqrt{q(i)}\right)^2$ | $t \mapsto \left(\sqrt{t}-1\right)^2$ | [134] |
| Bhattacharyya | $D_{BC}(P\|\|Q) = -\sum_i \sqrt{p(i)q(i)}$ | $t \mapsto -\sqrt{t}$ | [135] |

### 4.1.3 Unifrac distances

The distances and divergences so far have been mathematical (Sørensen, Hellinger) and information-theoretic (Kullback-Leibler) in origin, but not biological. When the features represent organisms with known taxonomic relationships (that is, the samples have been projected onto a common phylogenetic tree), distances between samples can be constructed with awareness of the phylogenetic relationships. (The procedure for building a phylogenetic tree is not developed in this section, as it would bring us far beyond the scope of this chapter.) The first widely-adopted metric using the phylogenetic tree is the **UniFrac** measure [136]. This metric can be formulated as follows.

Consider a phylogenetic tree $T$, that can either be taken as a universal reference or built based from the two samples. It can be seen as a set of nodes $\{n_i\}_i$ and a hierarchical relationship $\{t_j\}_j$. Now, let $A$ and $B$ denote respectively the first and second samples. Both can be represented as a set of branches $\{a_j\}_j$ (respectively $\{b_j\}_j$) assumed to be in $T$. We will abusively write $A$ and $B$ for both the samples and the branches representing them in the phylogenetic tree but the meaning should be clear in the context. The uniform fraction (a.k.a. UniFrac) of similarity is defined as the fraction of branches belonging to only a single sample over the overall length based on both of them. Figure 3 illustrates this fact. Assume you are given two samples, one containing species 1 to 5 (represented as the leaves on the tree) and another one containing species 3 to 9. The tree shows a potential phylogenetic relation between all the species where the dashed blue connections correspond to phylogenetic relation that are present on both of the samples. The UniFrac similarity distance between these two samples can be seen as the proportion of solid lines given the whole tree (or in other words, the size of the tree without the dashed part over the size of the whole tree).

Formally, we have:

$$UniFrac(A,B) := \frac{|A| + |B| - 2 \times |A \cap B|}{|A| + |B| - |A \cap B|} \tag{6}$$

Unifrac can be generalized by considering that each $t_j$ has a given length $l_j$ (depending on the inferred time between two branches or any phylogenetic related metrics) and by weighting the branches by their frequencies or abundances. The yields the weighed UniFrac measure [137]:

$$wUF(A,B) := \sum l_i \left| \frac{A_i}{m} - \frac{B_i}{n} \right| \tag{7}$$

where $A_i$ (respectively $B_i$) is the number of descendant of branch $i$ in $A$ (respectively $B$), i.e. $A_i = |\{t_j : t_i \in Parents_A(t_j)\}|$. $m$ and $n$ denote the respective number of reads in

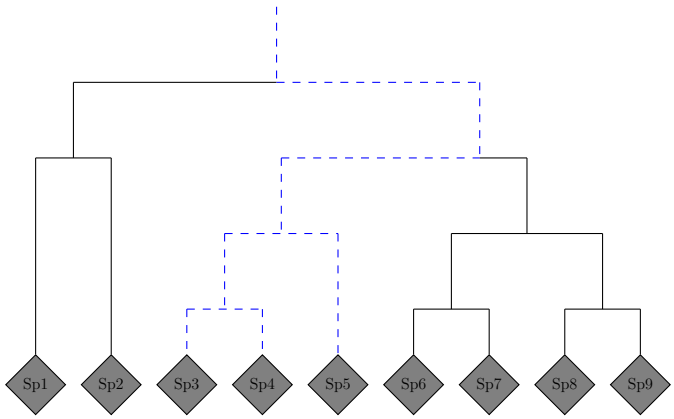


Figure 3: Example of a basic phylogenetic tree and how to compute the UniFrac distance (See text for details)

sample $A$ and $B$. This formulation corresponds to weighting a certain branch by its relative importance in the total length of the tree. The original $UniFrac$ metric can be recovered by assuming that $\frac{A_i}{m}$ and $\frac{B_i}{n}$ take values in $\{0,1\}$ whether the current branch is in the analyzed sample or not.

With this new approach, we can generalize even further by noticing that a sample corresponds to a probability distribution on the phylogenetic tree [138]. We can understand this distribution as *What is the probability that picking a random read from a sample corresponds to this particular taxa?* Now keeping this probabilistic view in mind opens new perspectives towards phylogenetic tree-based sample comparison as we can now use any similarity or divergence measures between probability distributions and adapt them to work on a tree structure.

In their novel work, Evans and Matsen [138] showed that applying the Kantorovitch-Rubinstein metric (also known as the Earth Mover's Distance in computer science or the Wasserstein distance in mathematics) yields a generalization of the previous weighted and unweighted UniFrac. We refer to [138] for the mathematical derivations and only give its formulation:

$$Z(P,Q) := \int_T |P(\tau(y)) - Q(\tau(y))| \, \lambda(dy) \tag{8}$$

In this expression, $P$ and $Q$ represent both probability measures of $A$ and $B$ respectively on the phylogenetic tree $T$. The notation $\tau(y)$ denotes the subgraph starting at node $y$ (part of the tree that is below $y$). $\lambda$ denotes the equivalent to the Lebesgue measure on the tree $T$ (which contains a distance metric derived from the length of the branches).

This generalization can even go one step further by integrating any pseudo-metric $f$ instead of the absolute value:

$$\widehat{Z}_f(P,Q) := \int_T f\left(P(\tau(y)) - Q(\tau(y))\right) \lambda(dy) \tag{9}$$

It is clear that this definition yields the classical UniFrac metric if $f$ takes the value one when exactly one of the probabilities $P$ and $Q$ is greater than 0.

### 4.1.4 Dimensionality reduction for visualization purposes

Ecologists employ ordination methods when a visual relation is desired based on similarity of a set of multivariate objects [139,140]. Typically, each object represents a sample site and is a table representing the abundances of organisms within the community. This composition generally varies among the sites and may be structured by environmental variables termed gradients. An ideal ordination technique would display all sample sites in the same order, as they exist along the environmental gradient, with inter-sample distances proportional to their separation along the gradient [141]. Various ordination methods are available for both direct (constrained) and indirect (unconstrained) analyses. The difference between the classes of methods depends on whether environmental gradient measurements are included (direct) or omitted (indirect). Popular methods methods discussed below include PCA, PCoA, NMDS, CA, CCA, RDA and DCA [139–144].

Metagenomic annotations are a class of high-dimensional data that can be explored and examined using dimension reduction. In such cases we are given a set of $n$ samples (e.g. samples that are collected from different regions of the body). Each sample, which we call $x$, is described by a certain number $K$ of features (e.g. the frequencies or abundances of annotated microbial species) and all of them are gathered in what is called a **data matrix**: $X = \left[x^{(1)}, x^{(2)}, \cdots, x^{(n)}\right] \in \mathbb{R}^{K\times n}$. The number of features can be between 10 and $10^7$, making visualization impossible (since displaying more than three dimensions is challenging at best). Picking just three variables to display randomly is not a good choice–it throws away most of the data without attempting to identify and preserve aspects of the data that may prove interesting. For concreteness, consider the case where the features are the abundances of each of $K = 150$ types of known bacteria. If we decide to choose a subset of these dimensions of size $m = 3$ for visualization or further investigation, there are $\binom{K}{m} = \frac{K!}{m!(K-m)!} = 551300$ possible candidates to choose from. We will present two examples of data reduction techniques that try to choose "interesting" subsets of the feature space that are commonly used for visualization purposes: **Principal Component Analysis (PCA)** and **Principal Coordinate Analysis (PCoA)**. Both rely on projections onto orthogonal axes representing the most variance of the data as possible.

**PCA** is a method that projects the data onto axes which contain most of the variance. This is done by calculating the eigenvalue decomposition of the covariance matrix and keeping only the projections of the data onto the few most important eigenvectors. This procedure is called **spectral decomposition** and the eigenvectors corresponding to the largest eigenvalues are called the **principal components**. They point in the directions with the highest variability of the data. This method reduces the data to a sum of typical mixtures of the different bacteria in a sample, but it will not give the most relevant bacteria given an outside paramter. The whole PCA decomposition can be seen as a combination of an average microbiome mixture over all samples plus some small variations according to some calculated mixtures.

The large dimensionality of feature vectors makes spectral decomposition of the covariance matrix intractable in most cases. Consequently, PCA as a dimensionality reduction technique is applied to the analysis of $(n \times n)$ resemblance or distance matrices.

**PCoA** is yet another method for dimensionality reduction useful for visualization [141]. First introduced [145] as a method that would preserve the distances between objects even when representing them using fewer components, it performs PCA on a modified version of the distance matrix. A main advantage of using PCoA over PCA is that it allows the user to tune the similarity of distance metric used for comparison; this fact yields a more flexible tool regarding the dynamics or range of the different variables. A resemblance matrix is built by comparing each element of the data matrix to one another using the chosen comparison metric $d$:

$$\forall i, j \in \{1, \cdots, n\},\ D_{ij} = d(x_i, x_j)$$

This matrix is then squared and centered as follows:

$$\begin{aligned} A_{ij} &= D_{ij}^2 \\ B_{ij} &= A_{ij} - \overline{A_{i\cdot}} - \overline{A_{\cdot j}} + \overline{A} \end{aligned}$$

where $\overline{A_{i\cdot}}$, $\overline{A_{\cdot j}}$ and $\overline{A}$ denote the means taken of each rows, columns and the whole matrix, respectively. This matrix transformation does not affect the distance relationships between the samples and hence keeps the structure of the data. But on the other hand it centers the data on the centroid of the samples.

Finally, a spectral decomposition is applied to the $B$ matrix and its eigenvectors are scaled by the square root of the eigenvalues. Usually only a few eigenvectors cover most the variability of the data. Note that when using the Euclidean distance as a metric for the resemblance matrix, PCoA yields equivalent result to the PCA approach based on the spectral decomposition of the covariance matrix.

Some care should be taken with this method. It works well when used with a metric distance measures (i.e. a binary positive definite form fulfilling the triangle inequality). Since some beloved discrepancies do not fulfill the triangle inequality, they should not be used in this framework if one wishes to transform the data into another space. This should not be a major concern if the only purpose of the transformation is the visualization of the data. Figure 4 shows some PCoA plots for the Human microbiome study, comparing visualizations based on Hellinger distance (left hand side) and Euclidean distance (to the right). Note that PCA would give similar results to the PCoA with the Euclidean distance.

### 4.1.5 Testing for differences

Once we can visually see that the sample groupings are separable, we may wish to test whether the visual differences are statistically significant or not. The **ANalysis Of SIMilarity (ANOSIM)** tool was first introduced in [146, 147] to solve this problem, as a hypothesis test based the distance matrix used in PCoA.

Assume you are given labels associated to the different samples (for instance the site where they were taken from). We order the set of all different distances in an increasing order and replace the distance matrix $D$ by the rank matrix $R_D$ where the components of $R_{Dij}$ correspond to the rank of $D_{ij}$. Then compute the **within site similarity** $R_w$ as the average of all ranks between two samples coming from a same site on the one hand and the **between class similarity** as the average of all ranks between samples with different labels.

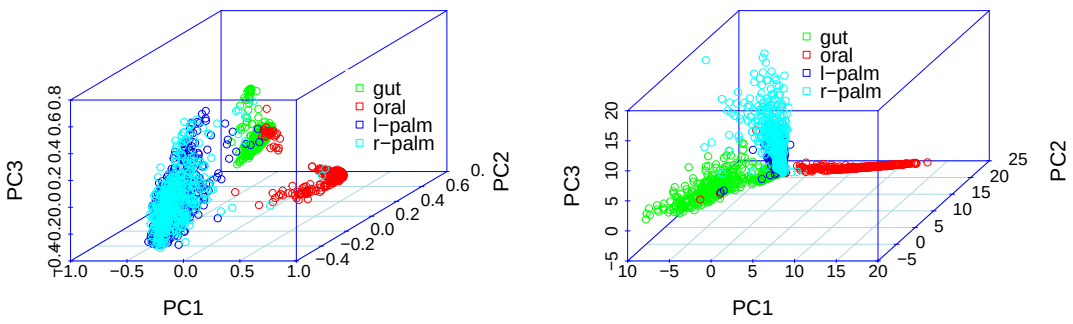


Figure 4: PCoA plots of the 15 month study of the human microbiome [4] (**left**) Hellinger and (**right**) Euclidean.

The ANOSIM test calculates

$$R = 4\frac{R_b - R_w}{n(n-1)} \tag{10}$$

where $n$ denotes the number of samples. The higher the value of this number the better it is. Moreover note that the value has to be in the range $[-1, 1]$ where values closer to 1 mean that the data are highly separable.

To compare the significance of the data to a null hypothesis, we can construct the null hypothesis by permuting randomly the labels of the samples and then comparing the significance of the original $R$ against the null hypothesis $R$ to see of the original $R$ has greater significance.

## 4.2 Feature selection

Section 4 discussed several dimensionality reduction techniques, all of which are based on the projection of observations onto a lower dimensional space. For example, PCA projected the observations onto a lower dimensional space that maximized the variation in the data. In this section, we describe **feature selection** tools that identify features that can best discriminate between multiple classes in a data set. That is, we seek to identify a subset of features $\mathcal{F}_\theta$ of the original $\mathcal{F}$ that provide the best discrimination between multiple classes in the data set. This parallel way of processing the data is illustrated on Fig. 2. As we can see, projection based methods introduced in the previous section and feature selection algorithms developed in this section can be done parallel to each other and may be combined for further classification of data.

As an example, we may want to determine differences between healthy patients' and unhealthy patients' gut microbiomes. The goal here is to determine which organisms carry information that can differentiate between the healthy and unhealthy populations. From a machine learning perspective, this is a feature selection problem; however, from a biological perspective, the selection of organisms allows the biologist the opportunity to identify a set of species that is responsible for differentiating healthy and unhealthy patients. It is important

to note that there may be additional factors that influence the results, but may not be in the feature set.

Selecting highly informative features is the primary objective of any feature selection method; however, other objectives are typically used in feature selection as well. Clearly, selecting features that are relevant is of top priority, but many feature selection methods have redundancy tools integrated into their selection objective as well. That is, they select informative features that are not redundant with one another. Many biologists may not be worried about redundancy if the sole purpose of the their data analysis is to find the most informative features, because they simply do care that they may be redundant. Classification scenarios generally need to have some form of redundancy built into the feature selection algorithm to boost the prediction accuracy.

In section 4.2 and its subsections, we have solely focused on filter-based feature selection methods, rather than wrapper-based feature selection. The interested reader is encouraged to pursue recent literature for the differences between various feature selection methods [148–151].

### 4.2.1 A Forward Selection Algorithm

In feature selection, we have an objective function $\mathcal{J}$ that we seek to maximize, and this function is dependent upon a subset of features $\mathcal{F}_\theta$. The goal of the forward selection algorithm is to find $k$ features in $\mathcal{F}$ that maximize the objective function. A simple forward selection algorithm to achieve such a task is shown in Figure 5.

The algorithm begins by initializing $\mathcal{F}_\theta$ to an empty set, and the method takes in a number of features to select ($k$) and the original feature set ($\mathcal{F}$). Let $X_j$ be a random variable for feature $j$ and $Y$ be the variable that determines the class label (e.g., healthy vs. unhealthy). The first step is to find the feature $X_j$ which maximizes the objective function $\mathcal{J}$ that takes in the arguments $X_j$, $Y$, and $\mathcal{F}_\theta$. The feature, $X_j$, that maximizes equation (11) is added to $\mathcal{F}_\theta$ and removed from $\mathcal{F}$. This process is repeated until the cardinality of $\mathcal{F}_\theta$ is $k$.

The feature selection algorithm in Figure 5 is a simple method to select the $k$ features that maximize the objective function; however, the algorithm makes a key assumption that is often not true – feature independence. This algorithm assumes that all features are independent of each other, which is generally not the case with many real-world data sources. Nevertheless, this approach has been shown to be quite robust to a number of problems even when the assumptions are violated in practice [150, 152–154].

### 4.2.2 Information-Theoretic Feature Selection

The feature selection algorithm presented in Section 4.2.1 relied on an objective function; however, such a function has not yet been discussed in detail.

So far the objective function depends on $X_j$, $Y$, and $\mathcal{F}_\theta$, but what is the form of the function? Intuitively, it should promote features that are capable of describing the $Y$. That is, find the features that carry the most information about $Y$. In this section we focus solely on *information-theoretic* objective functions.

**Input**: Feature set $\mathcal{F}$, an objective function $\mathcal{J}$, $k$ features to select, and initialize an empty set $\mathcal{F}_\theta$

1. Maximize the objective function
$$X_j = \arg\max_{X_j \in \mathcal{F}} \mathcal{J}(X_j, Y, \mathcal{F}_\theta) \tag{11}$$
2. Update relevant feature set such that $\mathcal{F}_\theta \leftarrow \mathcal{F}_\theta \cup X_j$
3. Remove relevant feature from the original set $\mathcal{F} \leftarrow \mathcal{F} \backslash X_j$
4. Repeat until $|\mathcal{F}_\theta| = k$

Figure 5: Generic forward feature selection algorithm for a filter-based method.

The fundamental unit of information is entropy[1], which measures the level of uncertainty in a random variable $X$. Mathematically, this is given by,

$$H(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x) \tag{12}$$

where $p_X$ is the marginal probability distribution on the random variable $X$ with possible outcomes in the set $\mathcal{X}$. The antilog of the entropy, an information metric, can be interpreted as the number of equiprobable outcomes in a distribution with the same information content. Different outcomes in the set $\mathcal{X}$ contribute different amounts to the overall entropy. Datasets that are evenly distributed have higher entropy than datasets that are skewed toward a handful of values[2]. Maximum entropy is achieved with a uniform distribution on $p_X(x)$. Figure 6 shows that this is the case for a Bernoulli random variable. Similar to standard probabilities, entropy can be conditioned on a second random variable $Y$, which gives us conditional entropy.

$$H(X|Y) = -\sum_{y \in \mathcal{Y}} p_Y(y) \sum_{x \in \mathcal{X}} p_{X|Y}(x|y) \log p_{X|Y}(x|y) \tag{13}$$

**Conditional entropy** can be interpreted as the amount of information that is left in $X$ after the outcome of the random variable $Y$ is observed. Using the definitions of entropy and conditional entropy gives rise to mutual information, which is given by,

$$\begin{aligned} I(X;Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)} \\ &= H(X) - H(X|Y) \end{aligned} \tag{14}$$

[1]Entropy is measured in bits, nats, or bans depending on whether $\log_2$, $\log_e$, or $\log_{10}$ is used in the calculation, respectively.

[2]Imagine flipping a coin such that $\mathbb{P}(X = \texttt{heads}) = 0.99$ and $\mathbb{P}(X = \texttt{tails}) = 0.01$. Such a random variable has low entropy because there is little uncertainty in the outcome of $X$.

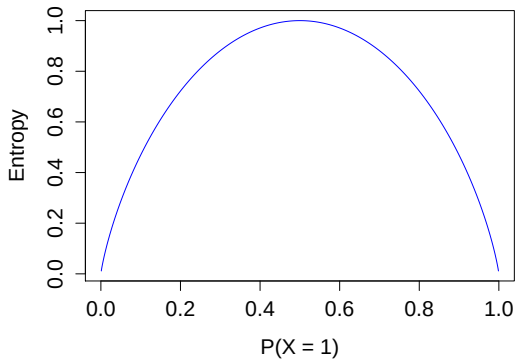


Figure 6: Entropy of a Bernoulli random variable. Maximum entropy, measured in bits, is achieved when the distribution on $X$ is uniform.

where $I(X;Y)$ is the remaining uncertainty in $X$ after the uncertainty about $X$ given what we known about $Y$ is removed. Note that $I(X;Y) = 0$ if the random variables $X$ and $Y$ are independent of each other. This should be quite intuitive since independent variables would be expected to share any information about each other. Lastly, define the **conditional mutual information**,

$$\begin{aligned} I(X;Y|Z) &= \sum_{z\in\mathcal{Z}} p_Z(z) \sum_{y\in\mathcal{Y}} p_{X,Y|Z}(x,y|z) \log \frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)} \\ &= H(X|Z) - H(X|Y,Z) \end{aligned} \tag{15}$$

which is the amount of information left between $X$ and $Y$ after $Z$ is observed. We now have discussed the appropriate tools in information theory that can allow us to design an objective function for equation (11) that accounts for feature relevancy and redundancy (both conditional and unconditional). As an example, equation (16) presents the objective function for the **Joint Mutual Information** feature selection method.

$$\mathcal{J}_{\mathsf{JMI}}(X_t, Y, \mathcal{F}_\theta) = I(X_t;Y) - \frac{1}{|\mathcal{F}_\theta|} \sum_{X_j\in\mathcal{F}_\theta} [I(X_t;X_j) - I(X_t;X_j|Y)] \tag{16}$$

There are three terms in $\mathcal{J}_{\mathsf{JMI}}(X_t, Y, \mathcal{F}_\theta)$ that controls the features that are selected. The first term $I(X_t;Y)$ is simply the amount of information shared between $X_t$ (i.e., the feature under test), and the class label $Y$. The second term $I(X_t;X_j)$ is a measure of redundancy and since $I(X_t;X_j) \geq 0$, the quantity decreases $\mathcal{J}_{\mathsf{JMI}}(X_t, Y, \mathcal{F}_\theta)$. Hence, a measure of removing redundant features. The third term shows that the $\mathcal{J}_{\mathsf{JMI}}(X_t, Y, \mathcal{F}_\theta)$ can be increased by having some level of conditional redundancy between features.

### 4.2.3 Measuring Feature Consistency

The reliability or consistency of a feature selection method remains an important question. Would the feature selection algorithms always return the same relevant features if we were

to use the forward selection method on cross-validation or bootstrap data sets? To answer this question, a simple consistency index may be applied to measure the similarity between multiple sets.

Typically some form of cross-validation or bootstrapping is applied to a data set and feature selection algorithm to determine the relevant features; however, what if the features selected vary slightly over each of the validation/bootstrap trials? In such situations, we need a way to quantify the **consistency** of the relevant feature set. Kuncheva developed a consistency index that meets three primary criteria for an index: ($a$) the consistency index is a monotonically increasing function of increasing elements in common with two sets union, ($b$) the index is bounded, and ($c$) the index should have a constant value for independently drawn subsets of features of the same cardinality [155]. Using these criteria, Kuncheva derived the following definition of consistency.

**Definition 4.2 (Consistency [155])** *The consistency index for two subsets $\mathcal{A} \subset \mathcal{F}$ and $\mathcal{B} \subset \mathcal{F}$, such that $r = |\mathcal{A} \cap \mathcal{B}|$ and $|\mathcal{A}| = |\mathcal{B}| = k$, where $1 \leq k \leq |\mathcal{F}| = K$, is*

$$\mathcal{I}_{\mathcal{F}}(\mathcal{A}, \mathcal{B}) = \frac{rK - k^2}{k(K-k)} \tag{17}$$

**What Does All This Mean?** This section has described the information theoretic tools to select / design an objective function for a feature selection method. If the end goal is to strictly find highly informative features than maximizing $I(X_j; Y)$ is sufficient. However, for many classification problems incorporating redundancy and conditional redundancy is quite beneficial over methods that do not use redundancy, though using redundancy terms does not guarantee improved performance.

# 5 Understanding microbial communities

In the section above, we have described the methods used comparing metagenomic samples and identifying relevant features. These machine learning techniques, when applied to real metagenomic problems, provide us with more opportunities tounderstand the lives of microbes in their environments.

For example, by assessing the biodiversity across metagenomes, we can learn about the functional capabilities of microbes in a community and evaluate hypotheses about survival strategies under environmental shift. Systems investigated already with these tools include analyzing the redundancy of microbes in infected human lungs of cystic fibrosis patients [156], the role of microbes in human breast milk in colonization of the infant gastrointestinal tract and maintenanceof mammary health [157], and the communities of microbes on human skin to examine how antibiotic exposure and lifestyle changes alter the skin microbiomeselectively [158].

Moreover, the use of metagenomic methods allows a window into the interaction between microbes. One study, [159] compared human metagenomes across different body sites and identified 3,005 significant co-occurrence and co-exclusion relationships between bacterial branches. This is an informative way for us to learn about potential microbial interactions. There are also a slew of on-going studies that link microbes by similar genes or pathways

they share, in search of microorganism cooperation and competition. Another tool has been developed for analyzing the topology of metabolic networks and calculating the metabolic overlap between species, which provides a way of estimating the competitive potential between bacterial species [160]. Although this is not readily a tool for metagenomes, it shows us that extracting metabolic information has potential to answer more biological questions than we currently do.

In addition to the comparison between microbes, many people are interested in the symbiosis between microbes and their human host. By comparing the functional capability of microbiomes and their host, we can learn about microbes strategy to maintain the symbiosis, such as providing nutrients, degradation of toxins and immune enhancement [161].

We are glad to observe the increase of not only metagenomic studies, but also the increase of metatranscriptomic, metaproteomic and metabolomic data. The incorporation of microbial genomes, transcripts, proteins and metabolites into the machine learning techniques introduced provides more information and can possibly lead to personalized medicine [19].

# 6 Open problems and challenges

There are a plethora of open problems in metageome analysis. We will highlight several that we believe are important to fully exploit the information in the sequence of a sample and its relationship to its environment.

Metagenomic annotation can typically provide only an approximate estimate of the taxonomic [25, 93, 162] and functional content [78, 124, 163] of an environmental sample (e.g., 16S rRNA surveys and "light-sequencing" whole-genome shotgun (WGS) studies). The high coverage of deep WGS sequencing offers the promise of providing the identity and relative abundance even low-abundance organisms, rather than that of just the most-abundant organisms. Currently, low abundance taxa cannot be studied rigorously, due to lack of effective confidence estimation procedures in taxonomic / gene identification: reads originating from known organisms or gene families can be falsely labeled due to sequencing errors causing techniques to miss their presence all-together; conversely, reads from novel low-abundance organisms or genes can be mistaken as errors. *Therefore, it is important to be able to assign a confidence to the probability of detecting an organism in a sample.*

High coverage also enables the sampling of genetically novel organisms as well as informing how these organisms interact with their environment. However, while 18000 genomes have been sequenced and their annotation nearly completed, many more–in fact vast majority of – species have not been sequenced. This makes deciphering whether a metagenomic read originates from "novel" organism a formidable task. In fact, due to the pangenome and the flexible definition of a species, strain classification is practically impossible. *A very important open problem is to identify strains by comparing the assigned gene and taxonomic labels to previously-known and annotated data about an environment, and offer suggestions about possible horizontal-gene transfer, genetic modification/evolution, and level of error that might have affected a read.*

Metagenomes are described using thousands of features; features are explanatory variables that represent species, metabolic pathways, orthologous protein groups, protein families or other functional categories. Methods are needed to exclude features whose abun-

dance/expression remain stable over certain physiologies, or those that have little or very indirect impact on physiological changes of interest. *An interesting open problem is to develop a computationally-identifiable set of metagenome features that are predictive of physiologies.* We hypothesize that different metagenomic datasets will require different feature selection methods which makes assumptions about the underlying biology.

When comparing multiple samples, dimensionality reduction techniques are used to visualize the data. However, due to loss of information from dimension reduction, distortion occurs. In particular, ordination methods suffer from the "arch effect"; a mathematical artifact that has no real relationship to community structure [144]. It arises because the second ordination axis is constrained to be uncorrelated with the first axis, but is not constrained to be independent of it. To circumvent the arch effect it is necessary to ensure that subsequent axes do not have a systematic relation to the first axis. Detrended correspondence Analysis (DCA) has been developed to address the arch effect, but it turn destroys information in higher axes that could be related to additional environmental gradients [164]. While this technique may be useful for single gradient ordination, the arch effect prohibits the ordination of multiple gradients since axes must be independent in order to be interpreted separately. Given that many communities are structured by multiple gradients the development of corresponding methods is fertile ground for research. Adjacently, the selection of proper ordination techniques given the sampling depth and magnitude of the gradient (effect size) is also an open area of research. It has been shown in simulation that an improper choice of method can lead to erroneous results [143]. Therefore, analysis of methods for discerning samples that are clustered tightly versus those that are differentiated by a continuum would be most welcome to the field.

# Acknowledgements

This work was supported in part by the National Science Foundation (NSF) CAREER award number #0845827, NSF award number #1120622, and Department of Energy (DOE) Office of Science (BER) award #DE-SC0004335.

# References

[1] J. Rousk, E. Bååth, P. C. Brookes, C. L. Lauber, C. Lozupone, J. G. Caporaso, R. Knight, and N. Fierer, "Soil bacterial and fungal communities across a pH gradient in an arable soil," *ISME Journal*, vol. 4, pp. 1340–1351, 2010.

[2] R. M. Bowers, S. McLetchie, R. Knight, and N. Fierer, "Spatial variability in airborne bacterial communities across land-use types and their relationship to the bacterial communities of potential source environments," *ISME Journal*, vol. 5, pp. 601–612, 2011.

[3] S. Williamson, D. Rusch, S. Yooseph, A. Halpern, K. Heidelberg, J. Glass, C. Andrews-Pfannkoch, D. Fadrosh, C. Miller, G. Sutton, M. Frazier, and J. C. Venter, "The

Sorcerer II global ocean sampling expedition: Metagenomic characterization of viruses within aquatic microbial samples," *PLoS Biology*, no. 1, 2008.

[4] J. G. Caporaso, C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, J. I. Gordon, and R. Knight, "Moving pictures of the human microbiome," *Genome Biology*, vol. 12, no. 5, 2011.

[5] E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, and R. Knight, "Bacterial community variation in human body habitats across space and time," *Science*, vol. 326, pp. 1694–1697, 2009.

[6] T. Varin, C. Lovejoy, A. D. Jungblut, W. F. Vincent, and J. Corbeil, "Metagenomic analysis of stress genes in microbial mat communities from antarctica and the high arctic," *Applied and Environmental Microbiology*, vol. 78, no. 2, pp. 549–559, 2012.

[7] D. J. Jiménez, F. D. Andreote, D. Chaves, J. S. Montaña, C. Osorio-Forero, H. Junca, M. M. Zambrano, and S. Baena, "Structural and functional insights from the metagenome of an acidic hot spring microbial planktonic community in the colombian andes," *PloS one*, vol. 7, no. 12, p. e52069, 2012.

[8] I. Bodaker, I. Sharon, M. T. Suzuki, R. Feingersch, M. Shmoish, E. Andreishcheva, M. L. Sogin, M. Rosenberg, M. E. Maguire, S. Belkin, *et al.*, "Comparative community genomics in the dead sea: an increasingly extreme environment," *The ISME journal*, vol. 4, no. 3, pp. 399–407, 2009.

[9] E. M. Hunter, H. J. Mills, and J. E. Kostka, "Microbial community diversity associated with carbon and nitrogen cycling in permeable shelf sediments," *Applied and Environmental Microbiology*, vol. 72, no. 9, pp. 5689–5701, 2006.

[10] J. McCarren, J. W. Becker, D. J. Repeta, Y. Shi, C. R. Young, R. R. Malmstrom, S. W. Chisholm, and E. F. DeLong, "Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea," *Proceedings of the National Academy of Sciences of the USA*, vol. 107, no. 38, pp. 16420–16427, 2010.

[11] G. Rocap, F. W. Larimer, J. Lamerdin, S. Malfatti, P. Chain, N. A. Ahlgren, A. Arellano, M. Coleman, L. Hauser, W. R. Hess, *et al.*, "Genome divergence in two prochlorococcus ecotypes reflects oceanic niche differentiation," *Nature*, vol. 424, no. 6952, pp. 1042–1047, 2003.

[12] O. Koren, D. Knights, A. Gonzalez, L. Waldron, N. Segata, R. Knight, C. Huttenhower, and R. E. Ley, "A guide to enterotypes across the human body: Meta-analysis of microbial community structures in human microbiome datasets," *PLoS computational biology*, vol. 9, no. 1, p. e1002863, 2013.

[13] A. A. Pragman, H. B. Kim, C. S. Reilly, C. Wendt, and R. E. Isaacson, "The lung microbiome in moderate and severe chronic obstructive pulmonary disease," *PloS one*, vol. 7, no. 10, p. e47305, 2012.

[14] J. D. van Elsas, M. Chiurazzi, C. A. Mallon, D. Elhottovā, V. Krištůfek, and J. F. Salles, "Microbial diversity determines the invasion of soil by a bacterial pathogen," *Proceedings of the National Academy of Sciences*, vol. 109, no. 4, pp. 1159–1164, 2012.

[15] K. Findley, J. Oh, J. Yang, S. Conlan, C. Deming, J. A. Meyer, D. Schoenfeld, E. Nomicos, M. Park, N. I. S. C. C. Sequencing, *et al.*, "Topographic diversity of fungal and bacterial communities in human skin," *Nature*, vol. advance online publication, p. 2013/05/22/online, 2013.

[16] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, *et al.*, "Enterotypes of the human gut microbiome," *Nature*, vol. 473, pp. 174–180, 2011.

[17] N. Fierer, C. L. Lauber, N. Zhou, D. McDonald, E. K. Costello, and R. Knight, "Forensic identification using skin bacterial communities," *Proceedings of the National Academy of Sciences*, vol. 107, no. 14, pp. 6477–6481, 2010.

[18] G. Ditzler, R. Polikar, and G. Rosen, "Forensic identification with environmental samples," in *IEEE Intl. Conference on Acoustics, Speech, and Signal Processing*, 2012.

[19] R. Chen, G. I. Mias, J. Li-Pook-Than, L. Jiang, H. Y. Lam, R. Chen, E. Miriami, K. J. Karczewski, M. Hariharan, F. E. Dewey, *et al.*, "Personal omics profiling reveals dynamic molecular and medical phenotypes," *Cell*, vol. 148, no. 6, pp. 1293–1307, 2012.

[20] J. Handelsman, *Committee on Metagenomics: Challenges and Functional Applications*. The National Academies Press, 2007.

[21] J. Raes, K. U. Foerstner, and P. Bork, "Get the most out of your metagenome: computational analysis of environmental sequence data," *Current Opinion in Microbiology*, vol. 10, pp. 1–9, 2007.

[22] J. A. Eisen, "Environmental shotgun sequencing: Its potential and challenges for studying the hidden world of microbes," *PLoS Biology*, vol. 5, no. 3, 2007.

[23] W. Valdivia-Granda, "The next meta-challenge for bioinformatics," *Bioinformation*, vol. 2, no. 8, pp. 358–362, 2008.

[24] X. Xiao, E. R. Dow, R. Eberhart, Z. B. Miled, and R. J. Oppelt, "Gene clustering using self-organizing maps and particle swarm optimization," in *International Parallel and Distributed Processing Symposium*, 2003.

[25] A. Bazinet and M. Cummings, "A comparative evaluation of sequence classification programs," *BMC Bioinformatics*, 2012.

[26] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, "Base-calling of automated sequencer traces using phred. i. accuracy assessment.," *Genome Research*, vol. 8, pp. 175–185, Mar. 1998.

[27] P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.," *Nucleic acids research*, vol. 38, pp. 1767–1771, Apr. 2010.

[28] A. Minoche, J. Dohm, and H. Himmelbauer, "Evaluation of genomic high-throughput sequencing data generated on illumina HiSeq and genome analyzer systems," *Genome Biology*, vol. 12, pp. R112+, Nov. 2011.

[29] S. M. Huse, J. A. Huber, H. G. Morrison, M. L. Sogin, and D. M. M. Welch, "Accuracy and quality of massively parallel DNA pyrosequencing.," *Genome biology*, vol. 8, pp. R143+, July 2007.

[30] M. Cox, D. Peterson, and P. Biggs, "SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data," *BMC Bioinformatics*, vol. 11, pp. 485+, Sept. 2010.

[31] D. R. Kelley, M. C. Schatz, and S. L. Salzberg, "Quake: quality-aware detection and correction of sequencing errors.," *Genome Biology*, vol. 11, no. 11, pp. R116+, 2010.

[32] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.," *Bioinformatics (Oxford, England)*, vol. 22, pp. 1658–1659, July 2006.

[33] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST.," *Bioinformatics (Oxford, England)*, vol. 26, pp. 2460–2461, Oct. 2010.

[34] K. T. Konstantinidis and J. M. Tiedje, "Trends between gene content and genome size in prokaryotic species with larger genomes," *Proceedings of the National Academy of Sciences of the USA*, vol. 101, pp. 3160–3165, Mar. 2004.

[35] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White, "Microbial gene identification using interpolated Markov models," *Nucleic Acids Research*, vol. 26, pp. 544–548, Jan. 1998.

[36] J. Besemer and M. Borodovsky, "Heuristic approach to deriving models for gene finding.," *Nucleic Acids Research*, vol. 27, pp. 3911–3920, Oct. 1999.

[37] W. Zhu, A. Lomsadze, and M. Borodovsky, "Ab initio gene identification in metagenomic sequences," *Nucleic Acids Research*, vol. 38, p. e132, July 2010.

[38] H. Noguchi, T. Taniguchi, and T. Itoh, "MetaGeneAnnotator: Detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes," *DNA Research*, vol. 15, pp. 387–396, Dec. 2008.

[39] M. Rho, H. Tang, and Y. Ye, "FragGeneScan: predicting genes in short and error-prone reads," *Nucleic Acids Research*, vol. 38, p. e191, Nov. 2010.

[40] D. Hyatt, G.-L. L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification.," *BMC Bioinformatics*, vol. 11, no. 1, pp. 119+, 2010.

[41] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, pp. 403–410, 1990.

[42] W. Gerlach and J. Stoye, "Taxonomic classification of metagenomic shotgun sequences with CARMA3," *Nucleic Acids Research*, vol. 39, no. 14, p. e91, 2011.

[43] B. Liu, T. Gibbons, M. Ghodsi, and M. Pop, "Metaphyler: Taxonomic profiling for metagenomic sequences," in *IEEE BIBM*, 2010.

[44] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower, "Metagenomic microbial community profiling using unique clade-specific marker genes," *Nature Methods*, 2012.

[45] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "MEGAN analysis of metagenomic data," *Genome Research*, vol. 17, no. 3, pp. 377—386, 2007.

[46] M. Horton, N. Bodenhausen, and J. Bergelson, "MARTA: a suite of java-based tools for assigning taxonomic status to DNA sequences," *Bioinformatics*, vol. 26, no. 4, pp. 568–9, 2010.

[47] F. Gori, G. Folino, M. Jetten, and E. Marchiori, "MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks," *Bioinformatics*, vol. 27, no. 2, pp. 196–203, 2011.

[48] H. M. Monzoorul, T. S. Ghosh, D. Komanduri, and S. S. Mande, "Sort-items: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences.," *Bioinformatics*, vol. 25, no. 14, pp. 1722–30, 2009.

[49] M. Jones, A. Ghoorah, and M. Baxter, "jMOTU and Taxonerator: turning DNA barcode sequences into annotated operational taxonomic units," *PLoS ONE*, vol. 6, no. 4, p. e19259, 2011.

[50] G. L. Rosen, E. R. Reichenberger, and A. M. Rosenfeld, "NBC: the naïve bayes classification tool webserver for taxonomic classification of metagenomic reads," *Bioinfomatics*, vol. 27, no. 1, pp. 127–129, 2011.

[51] G. Rosen and T. Y. Lim, "NBC update: The addition of viral and fungal databases to the naïve bayes classification tool," *BMC Research Notes*, 2012.

[52] K. R. Patil, P. Haider, P. B. Pope, P. J. Turnbaugh, M. Morrison, T. Scheffer, and A. C. McHardy, "Taxonomic metagenome sequence assignment with structured output models," *Nature Methods*, vol. 8, pp. 191–192, 2011.

[53] A. Brady and S. Salzberg, "Phymm and Phymmbl: metagenomic phylogenetic classification with interpolated Markov models," *Nature Methods*, vol. 6, no. 9, pp. 673–676, 2009.

[54] D. R. Kelley and S. L. Salzberg, "Clustering metagenomic sequences with interpolated Markov models," *BMC Bioinformatics*, vol. 11, no. 544, 2010.

[55] N. N. Diaz, L. Krause, A. Goesmann, K. Niehaus, and T. W. Nattkemper, "TACOA – taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach," *BMC Bioinformatics*, 2009.

[56] O. U. Nalbantoglu, S. F. Way, S. H. Hinrichs, and K. Sayood, "RAIphy: Phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles," *BMC Bioinformatics*, vol. 12, no. 1, 2011.

[57] H. C. M. Leung, S. Yiu, B. Yang, Y. Pend, Y. Wang, Z. Liu, J. Chen, J. Qin, R. Li, and F. Chin, "A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio," *Bioinformatics*, vol. 27, no. 11, pp. 1489–1495, 2011.

[58] H. Stranneheim, M. Kaller, T. Allander, B. Andersson, L. Arvestad, and J. Lundeberg, "Classification of DNA sequences using bloom filters," *Bioinformatics*, vol. 26, no. 13, pp. 1595–1600, 2010.

[59] B. Langmead and S. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, pp. 357–359, 2012.

[60] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler Transform," *Bioinformatics*, vol. 25, no. 1754-60, 2009.

[61] R. Li, C. Yu, Y. Li, T. Lam, S. Yiu, K. Kristiansen, and J. Wang, "SOAP2: an improved ultrafast tool for short read alignment," *Bioinfomatics*, vol. 25, no. 15, pp. 1966–1967, 2009.

[62] CLCBio, "http://www.clcbio.com/desktop-applications/features/."

[63] J. Reumers, P. D. Rijk, H. Zhao, A. Liekens, and D. Smeets, "Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing," *Nature Biotechnology*, vol. 30, pp. 61–68, 2012.

[64] SMALT, "http://www.sanger.ac.uk/resources/software/smalt/."

[65] S. A. Berger, D. Krompass, and A. Stamatakis, "Performance, accuracy, and web server for evolutionary place- ment of short sequence reads under maximum likelihood," *Syst Biology*, vol. 60, no. 3, pp. 291–302, 2011.

[66] F. A. Matsen, R. B. Kodner, and E. V. Armbrust, "pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree," *BMC Bioinformatics*, vol. 11, no. 538, 2010.

[67] M. N. Price, P. S. Dehal, and A. Arkin, "Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix," *Mol Biol Evol*, vol. 26, no. 7, pp. 1641–50, 2009.

[68] K. Munch, W. Boomsma, J. P. Huelsenbeck, E. Willerslev, and R. Nielsen, "Statistical assignment of DNA sequences using Bayesian phylogenetics," *Syst Biology*, vol. 57, no. 5, pp. 750–7, 2008.

[69] N. J. Macdonald, D. H. Parks, and R. G. Beiko, "RITA: Rapid identification of high-confidence taxonomic assignments for metagenomic data," *Nucleic Acids Research*, vol. doi:10.1093/nar/gks335, 2012.

[70] M. H. Mohammed, T. S. Ghosh, N. K. Singh, and S. S. Mande, "SPHINX - an algorithm for taxonomic binning of metagenomic sequences.," *Bioinformatics*, vol. 27, pp. 22–30, oct 2010.

[71] S. A. Berger and A. Stamatakis, "Aligning short reads to reference alignments and trees," *Bioinformatics*, vol. 27, no. 15, pp. 2068–2075, 2011.

[72] M. Wu and J. A. Eisen, "A simple, fast, and accurate method of phylogenomic inference," *Genome Biology*, vol. 9, no. 10, p. R151, 2008.

[73] M. Stark, S. A. Berger, A. Stamatakis, and C. von Mering, "MLTreeMap – accurate maximum likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies," *BMC Genomics*, vol. 11, no. 461, 2010.

[74] F. Schreiber, P. Gumrich, R. Daniel, and P. Meinicke, "Treephyler: fast taxonomic profiling of metagenomes.," *Bioinformatics (Oxford, England)*, vol. 26, pp. 960–1, Apr. 2010.

[75] T. DeSantis, P. Hugenholtz, K. Keller, E. Brodie, N. Larsen, Y. Piceno, R. Phan, and G. Andersen, "NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes," *Nucleic Acids Research*, vol. 34, pp. W394–W399, 2006.

[76] A. Bateman, L. Coin, R. Durbin, R. Finn, V. Hollich, S. Griffiths-Johns, A. Khanna, M. Marshall, S. Moxon, E. Sonnhammer, D. Studholme, C. Yeats, and S. R. Eddy, "The Pfam protein families database," *Nucleic Acids Research*, vol. 36, pp. 281–288, 2008.

[77] L. Koski and G. B. Golding, "The closest BLAST hit is often not the nearest neighbor," *Journal of Molecular Evolution*, vol. 52, no. 6, pp. 540–2, 2001.

[78] J. H. Huson, S. Mitra, H. J. Ruscheweyh, N. Weber, and S. C. Schuster, "Integrative analysis of environmental sequences using megan 4," *Genome Research*, vol. 21, no. 9, pp. 1552–1560, 2011.

[79] J. Gilbert, F. Meyer, and M. Bailey, "The future of microbial metagenomics (or is ignorance bliss?)," *ISME J*, vol. 5, pp. 777–779, 2011.

[80] J. Martin, S. Sykes, S. Young, K. Kota, R. Sanka, N. Sheth, J. Orvis, E. Sodergren, Z. Wang, G. M. Weinstock, and M. Mitreva, "Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities," *PLoS ONE*, vol. 7, no. 6, p. e36427, 2012.

[81] A. V. Lukashin and M. Borodovsky, "Genemark.hmm: new solutions for gene finding," *Nucleic Acids Research*, vol. 26, no. 4, pp. 1107–1115, 1997.

[82] G. L. Rosen, "Examining coding structure and redundancy in DNA," *IEEE Engineering in Medicine and Biology Magazine*, vol. Special Issue on Communication Theory, Coding Theory, and Molecular Biology, pp. 62–68, 2006.

[83] M. Akhtar, J. Epps, and E. Ambikairajah, "Signal processing in sequence analysis: Advances in eukaryotic gene prediction," *IEEE Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 310–321, 2008.

[84] S. Karlin and C. Burge, "Dinucleotide relative abundance extremes: a genomic signature," *Trends in Genetics*, vol. 11, pp. 283–290, 1995.

[85] S. K. J. Mrázek and A. M. Campbell, "Compositional biases of bacterial genomes and evolutionary implications," *Journal of Bacteriology*, vol. 179, pp. 3899–3913, 1997.

[86] H. Nakashima and et al., "Genes from nine genomes are separated into their organisms in the dinucleotide composition space," *DNA Research*, vol. 5, pp. 251–259, 1998.

[87] D. Pride, R. J. Meinersmann, T. M. Wassenaar, and M. J. Blaser, "Evolutionary implications of microbial genome tetranucleotide frequency biases," *Genome Research*, vol. 13, pp. 145–158, 2003.

[88] T. Abe and et al., "Informatics for unveiling hidden genome signatures," *Genome Research*, vol. 13, pp. 693–702, 2003.

[89] T. Abe and et al., "Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples," *DNA Research*, vol. 12, pp. 281–290, 2005.

[90] B. Fertil, M. Massin, S. Lespinats, C. Devic, P. Dumee, and A. Giron, "GENSTYLE: exploration and analysis of DNA sequences with genomic signature," *Nucleic Acids Research*, vol. 33, 2005.

[91] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glockner, "Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences," *BMC Bioinformatics*, vol. 5, no. 163, 2004.

[92] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil, "Genomic signature: characterization and classification of species assessed by chaos game representation of sequences," *Molecular Biology and Evolution*, vol. 16, pp. 1391–1399, 1999.

[93] Q. Wang, G. Garrity, J. M. Tiedje, and J. Cole, "Naïve Bayes classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," *Applied Environmental Microbiology*, pp. 5261–5267, 2007.

[94] R. Sandberg, G. Winberg, C.-I. Bränden, A. Kaske, I. Ernberg, and J. Cöster, "Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier," *Genome Research*, vol. 11, no. 8, pp. 1404–1409, 2001.

[95] T. U. Consortium, "Reorganizing the protein space at the universal protein resource (uniprot)," *Nucleic Acids Research*, vol. 40, no. D1, pp. D71–D75, 2012.

[96] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins," vol. 35, pp. D61–5–, 2007.

[97] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," vol. 22, pp. 1658–9–, 2006.

[98] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. the gene ontology consortium," vol. 25, pp. 25–9–, 2000.

[99] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale, "The cog database: an updated version includes eukaryotes," vol. 4, pp. 41–, 2003.

[100] S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, L. J. Jensen, C. von Mering, and P. Bork, "eggnog v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges," vol. 40, pp. D284–9–, 2012.

[101] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "Kegg: Kyoto encyclopedia of genes and genomes," vol. 27, pp. 29–34–, 1999.

[102] R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, A. Pujar, A. G. Shearer, M. Travers, D. Weerasinghe, P. Zhang, and P. D. Karp, "The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases," vol. 40, pp. D742–53–, 2012.

[103] R. Overbeek, T. Begley, R. M. Butler, J. V. Choudhuri, H. Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Ruckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein, "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes," vol. 33, pp. 5691–702–, 2005.

[104] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn, "The pfam protein families database," vol. 40, pp. D290–301–, 2012.

[105] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, S. Krasnov, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Karsch-Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye, "Database resources of the national center for biotechnology information," vol. 40, pp. D13–25–, 2012.

[106] P. W. Rose, C. Bi, W. F. Bluhm, C. H. Christie, D. Dimitropoulos, S. Dutta, R. K. Green, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, A. G. Ramos, J. D. Westbrook, J. Young, C. Zardecki, H. M. Berman, and P. E. Bourne, "The rcsb protein data bank: new resources for research and education," vol. 41, pp. D475–D482–, 2013.

[107] D. H. Haft, J. D. Selengut, R. A. Richter, D. Harkins, M. K. Basu, and E. Beck, "Tigrfams and genome properties in 2013," vol. 41, pp. D387–95–, 2013.

[108] A. Bateman and D. H. Haft, "Hmm-based databases in interpro," vol. 3, pp. 236–45–, 2002.

[109] S. R. Eddy, "A new generation of homology search tools based on probabilistic inference," vol. 23, pp. 205–11–, 2009.

[110] F. Meyer, R. Overbeek, and A. Rodriguez, "Figfams: yet another set of protein families," vol. 37, pp. 6643–54–, 2009.

[111] A. N. Nikolskaya, C. N. Arighi, H. Huang, W. C. Barker, and C. H. Wu, "Pirsf family classification system for protein functional and evolutionary analysis," vol. 2, pp. 197–209–, 2006.

[112] C. J. Sigrist, E. de Castro, L. Cerutti, B. A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, and I. Xenarios, "New and continuing developments at prosite," vol. 41, pp. D344–7–, 2013.

[113] S. Henikoff, J. G. Henikoff, and S. Pietrokovski, "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations," vol. 15, pp. 471–9–, 1999.

[114] I. Letunic, T. Doerks, and P. Bork, "Smart 7: recent updates to the protein domain annotation resource," *Nucleic Acids Research*, vol. 40, no. D1, pp. D302–D305, 2012.

[115] A. Andreeva, D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin, "Data growth and its impact on the scop database: new developments," vol. 36, pp. D419–25–, 2008.

[116] M. Knudsen and C. Wiuf, "The cath database," vol. 4, pp. 207–12–, 2010.

[117] S. B. Pandit, R. Bhadra, V. S. Gowri, S. Balaji, B. Anand, and N. Srinivasan, "SUPFAM: a database of sequence superfamilies of protein domains," vol. 5, pp. 28–, 2004.

[118] C. Bru, E. Courcelle, S. Carrere, Y. Beausse, S. Dalmar, and D. Kahn, "The prodom database of protein domain families: more emphasis on 3d," vol. 33, pp. D212–5–, 2005.

[119] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, V. Buillard, L. Cerutti, R. Copley, E. Courcelle, U. Das, L. Daugherty, M. Dibley, R. Finn, W. Fleischmann, J. Gough, D. Haft, N. Hulo, S. Hunter, D. Kahn, A. Kanapin, A. Kejariwal, A. Labarga, P. S. Langendijk-Genevaux, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, A. N. Nikolskaya, S. Orchard, C. Orengo, R. Petryszak, J. D. Selengut, C. J. Sigrist, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats, "New developments in the interpro database," vol. 35, pp. D224–8–, 2007.

[120] A. Wilke, T. Harrison, J. Wilkening, D. Field, E. M. Glass, N. Kyrpides, K. Mavrommatis, and F. Meyer, "The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools," *BMC Bioinformatics*, vol. 13, no. 1, p. 141, 2012.

[121] E. C. Dimmer, R. P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M. J. Martin, B. Bely, P. Browne, W. Mun Chan, R. Eberhardt, M. Gardner, K. Laiho, D. Legge, M. Magrane, K. Pichler, D. Poggioli, H. Sehra, A. Auchincloss, K. Axelsen, M. C. Blatter, E. Boutet, S. Braconi-Quintaje, L. Breuza, A. Bridge, E. Coudert, A. Estreicher, L. Famiglietti, S. Ferro-Rojas, M. Feuermann, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, J. James, S. Jimenez, F. Jungo, G. Keller, P. Lemercier, D. Lieberherr, P. Masson, M. Moinat, I. Pedruzzi, S. Poux, C. Rivoire, B. Roechert, M. Schneider, A. Stutz, S. Sundaram, M. Tognolli, L. Bougueleret, G. Argoud-Puy, I. Cusin, P. Duek-Roggli, I. Xenarios, and R. Apweiler, "The uniprot-go annotation database in 2011," vol. 40, pp. D565–70–, 2012.

[122] Y. Lan, A. Kriete, and G. Rosen, "Selecting age-related functional characteristics in the human gut microbiome," *BMC Microbiome*, vol. 1, no. 1, pp. 1–12, 2013.

[123] J. Wilkening, A. Wilke, N. Desai, and F. Meyer, "Using clouds for metagenomics: a case study," in *IEEE International Conference on Cluster Computing (Cluster 2009), AUG 31-SEP 04, 2009 New Orleans, LA*, pp. 1–6, IEEE, 2009.

[124] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards, "The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes," *BMC Bioinformatics*, vol. 9, no. 386, 2008.

[125] V. M. Markowitz, N. N. Ivanova, E. Szeto, K. Palaniappan, K. Chu, D. Dalevi, I. M. Chen, Y. Grechkin, I. Dubchak, I. Anderson, A. Lykidis, K. Mavromatis, P. Hugenholtz, and N. C. Kyrpides, "Img/m: a data management and analysis system for metagenomes," vol. 36, pp. D534–8–, 2008.

[126] W. Z. Li, "Analysis and comparison of very large metagenomes with fast clustering and functional annotation," vol. 10, pp. –, 2009.

[127] S. Vinga and J. Almeida, "Alignment-free sequence comparisona review," *Bioinformatics*, vol. 19, pp. 513–523, Mar. 2003.

[128] P. Jaccard, *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901.

[129] I. Csiszár, "Eine informationstheoretische Ungleichung und ihre anwendung auf den Beweis der ergodizität von Markoffschen Ketten," *Publ. Math. Inst. Hungar. Acad.*, vol. 8, pp. 95–108, 1963.

[130] S. Ali and S. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 131–142, 1966.

[131] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *Information Theory, IEEE Transactions on*, vol. 52, pp. 4394–4412, Oct. 2006.

[132] M. Basseville, "Information : entropies, divergences et moyennes," Rapport de recherche PI-1020, May 1996.

[133] S. Kullback, *Information theory and statistics*. Dover Pubns, 1997.

[134] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, pp. 145–151, Jan. 1991.

[135] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

[136] C. Lozupone and R. Knight, "UniFrac: a new phylogenetic method for comparing microbial communities," *Applied Environmental Microbiology*, vol. 71, no. 12, 2005.

[137] C. Lozupone, M. Hamady, S. Kelley, and R. Knight, "Quantitative and qualitative $\beta$ diversity measures lead to different insights into factors that structure microbial communities," *Applied Environmental Microbiology*, vol. 73, no. 5, 2007.

[138] S. N. Evans and F. A. Matsen, "The phylogenetic kantorovich-rubinstein metric for environmental sequence samples," *Journal of the Royal Statistical Society*, 2010.

[139] R. H. Jongman, C. J. Ter Braak, and O. F. van Tongeren, *Data Analysis in Community and Landscape Ecology*. Pudoc, Wageningen, 1987.

[140] C. J. Ter Braak and I. C. Prentice, "A theory of gradient analysis," *Advances in Ecological Research*, vol. 34, pp. 235–282, 2004.

[141] P. Legendre and L. Legendre, *Numerical ecology*, vol. 20. Elsevier, 2012.

[142] M. Fasham, "A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines, and coenoplanes," *Ecology*, vol. 58, no. 3, pp. 551–561, 1977.

[143] J. Kuczynski, Z. Liu, C. Lozupone, D. McDonald, N. Fierer, and R. Knight, "Microbial community resemblance methods differ in their ability to detect biologically relevant patterns," *Nature Methods*, vol. 7, no. 10, pp. 813–821, 2010.

[144] H. Gauch, R. Whittaker, and T. Wentworth, "A comparative study of reciprocal averaging and other ordination techniques," *The Journal of Ecology*, vol. 65, no. 1, pp. 157–174, 1977.

[145] J. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis," *Biometrika*, vol. 53, no. 3-4, pp. 325–338, 1966.

[146] K. CLARKE, "Non-parametric multivariate analyses of changes in community structure," *Australian journal of ecology*, vol. 18, no. 1, pp. 117–143, 1993.

[147] K. Clarke and R. Warwick, *Change in marine communities: an approach to statistical analysis and interpretation*. Plymouth marine laboratory, Natural environment research council, 1994.

[148] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications*. Springer, 2006.

[149] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[150] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.

[151] Y. Saeys, I. Inza, and P. Larra naga, "A review of feature selection techniques in bioinformatics," *Oxford Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[152] W. Duch, K. Grqbczewski, T. Winiarski, J. Biesiada, and A. Kachel, "Feature selection based on information theory, consistancy, and separability indices," in *International Conference on Neural Information Processing*, pp. 1951–1955, 2002.

[153] G. Ditzler, R. Polikar, and G. Rosen, "Information theoretic feature selection for high dimensional metagenomic data," in *IEEE Gen. Sig. Proc. and Stat. Workshop (GENSIPS)*, 2012.

[154] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[155] L. I. Kuncheva, "A stability index for feature selection," in *International Conference on Artifical Intelligence and Application*, pp. 390–395, 2007.

[156] F. A. Stressmann, G. B. Rogers, C. J. van der Gast, P. Marsh, L. S. Vermeer, M. P. Carroll, L. Hoffman, T. W. Daniels, N. Patel, B. Forbes, *et al.*, "Long-term cultivation-independent microbial diversity analysis demonstrates that bacterial communities infecting the adult cystic fibrosis lung show stability and resilience," *Thorax*, vol. 67, no. 10, pp. 867–873, 2012.

[157] K. M. Hunt, J. A. Foster, L. J. Forney, U. M. Schütte, D. L. Beck, Z. Abdo, L. K. Fox, J. E. Williams, M. K. McGuire, and M. A. McGuire, "Characterization of the diversity and temporal stability of bacterial communities in human milk," *PLoS ONE*, vol. 6, no. 6, p. e21313, 2011.

[158] E. A. Grice, H. H. Kong, S. Conlan, C. B. Deming, J. Davis, A. C. Young, G. G. Bouffard, R. W. Blakesley, P. R. Murray, E. D. Green, *et al.*, "Topographical and temporal diversity of the human skin microbiome," *Science*, vol. 324, no. 5931, pp. 1190–1192, 2009.

[159] K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower, "Microbial co-occurrence relationships in the human microbiome," *PLoS computational biology*, vol. 8, no. 7, p. e1002606, 2012.

[160] A. Kreimer, A. Doron-Faigenboim, E. Borenstein, and S. Freilich, "NetCmpt: a network-based tool for calculating the metabolic competition between bacterial species," *Bioinformatics*, vol. 28, no. 16, pp. 2195–2197, 2012.

[161] L. Dethlefsen, M. McFall-Ngai, and D. A. Relman, "An ecological and evolutionary perspective on human–microbe mutualism and disease," *Nature*, vol. 449, no. 7164, pp. 811–818, 2007.

[162] D. Koslicki, S. Foucart, and G. Rosen, "Quikr: A method for rapid reconstruction of bacterial communities via compressive sensing," *Bioinfomatics*, 2013. accepted for publication.

[163] "Fizzy feature selection tool." https://github.com/gditzler/qiime/tree/fizzy.

[164] M. Hill and H. Gauch, "Detrended correspondence analysis: An improved ordination technique," *Plant Ecology*, vol. 42, no. 1, pp. 47–58, 1980.