Amino Acid k-mers enable assembly- and alignment-free sequence analysis

This manuscript (permalink) was automatically generated from bluegenes/2022-paper-protein-kmers@204f82f on November 22, 2022.

Authors

- **N. Tessa Pierce-Ward**
  0000-0002-2942-5331 · bluegenes · saltyscientist
  Department of Population Health and Reproduction, University of California, Davis · Funded by NSF 1711984; NSF 2018911

- **Olga Borisovna Botvinnik**
  0000-0003-4412-7970 · olgabot · olgabot

- **Taylor E. Reiter**
  0000-0002-7388-421X · taylorreiter · ReiterTaylor
  Department of Population Health and Reproduction, University of California, Davis · Funded by Grant GBMF4551 from the Gordon and Betty Moore Foundation; Grant R03OD030596 from the NIH Common Fund

- **Luiz Irber**
  0000-0003-4371-9659 · luizirber · luizirber
  Graduate Group in Computer Science, UC Davis; Department of Population Health and Reproduction, University of California, Davis · Funded by Grant GBMF4551 from the Gordon and Betty Moore Foundation; Grant R01HG007513 from the NIH NHGRI

- **C. Titus Brown**
  0000-0001-6001-2677 · ctb · ctitusbrown
  Department of Population Health and Reproduction, University of California, Davis · Funded by Grant GBMF4551 from the Gordon and Betty Moore Foundation; Grant R01HG007513 from the NIH NHGRI; Grant 2018911 from the NSF; Grant R03OD030596 from the NIH Common Fund

Abstract

Alignment-free methods for estimating sequence similarity have become critical for scaling sequence comparisons such as taxonomic classification and phylogenetic analysis to large-scale datasets. The majority of these methods rely upon exact matching of DNA k-mers: nucleotide subsequences of length k that can be counted and compared across datasets. However,as k-mer based methods rely on exact sequence matches, DNA k-mer comparisons can suffer from limited sensitivity when comparing highly polymorphic sequences or classifying organisms from groups not well represented in reference databases. Protein-based comparisons have long been the gold-standard approach for taxonomic and functional annotation at larger evolutionary distances. Here, we demonstrate the utility of amino acid k-mers (k_{aa}mers) for sequence comparisons across larger evolutionary distances, including alignment-free estimation of Average Amino Acid Identity (cAAI) and taxonomic classification using sourmash gather. K_{aa}mer comparisons can also be used with sketching methods such as FracMinHash, and will facilitate comparison and classification of a larger fraction of sequenced metagenome data, particularly from understudied and diverse environments.

Background

Advancements in sequencing over the past decades have made it feasible to investigate the vast global diversity of microbial organisms via direct sequencing of bulk DNA from environmental samples (metagenomics). These techniques have expanded and reshaped our understanding of evolutionary relatedness across the tree of life and allowed us to move beyond organismal isolates to investigate the structure and function of microbial communities (CITE).

Metagenomic analyses rely on our ability to make sense of bulk sequencing reads by assigning taxonomic and functional groupings. However, the methods and databases used for characterization impact both the extent and accuracy of classification. As the scale of genomic sequencing continues to grow, fast and low-memory methods for estimating sequence similarity have become critical for conducting tasks ranging from taxonomic classification to phylogenetic analysis on large-scale datasets [1,2]. However, many of these methods struggle with classification specificity, with some methods reporting false positive rates as high as 25% on short read metagenomic datasets prior to thresholding [3]. At the same time, classification techniques often can suffer from limited sensitivity when comparing highly polymorphic sequences or classifying organisms from groups underrepresented in reference databases. For understudied and diverse habitats such as soil, metagenomic classification methods often only categorize a small fraction of metagenomic data, and even well-studied environments such as the human gut can produce significant uncharacterized metagenome content (CITE).

As protein sequence is more conserved than the underlying DNA sequence, protein-based comparisons have long been used for comparisons across larger evolutionary distances [4,**blast?**]. Protein-based metagenomics taxonomic classification approaches typically have increased sensitivity relative to nucleotide methods [5,6,7,8,9,10]. Whole-proteome relatedness indices such as Amino Acid Identity (AAI) can be used to determine whether uncharacterized sequences belong to known taxonomic groups or represent truly novel sequence. AAI has been shown to be a robust measure of overall pairwise relatedness even for highly incomplete datasets, such as those comprised of only ~4% of the genome or 100 genes [11,12]. AAI is most useful when nucleotide comparisons are not longer robust, typically less than ~80% nucleotide identity. As we continue to sequence more of the biosphere, there remains a need for fast and accurate alignment-free sequence comparison tools with protein-level sensitivity.

Alignment-free methods using k-mers, short sequences of length k, can quickly compare and classify metagenomic datasets particularly when used with subsampling methods such as MinHash [1] and FracMinHash [13]. While the majority of k-mer methods utilize nucleotide k-mers, amino acid k-mers (k_{aa}-mers) have shown some promise for functional screening and annotation [10,14,15]. Here, we show that k_{aa}-mer comparisons robustly estimate Average Amino Acid Identity across large evolutionary distances, even while using FracMinHash k-mer subsampling methods. We then use FracMinHash k_{aa}mer sketches to tackle two classification challenges: taxonomic classification of assembled genomes, and compositional analysis of metagenomes. Our results suggest that k_{aa}mer sequence analysis can facilitate large-scale assembly-based and assembly-free metagenomic analyses, even when sequenced organisms are only distantly related to organisms available in reference databases.

Results

K-mer analysis methods enable similarity detection as low as a single shared k-mer between divergent genomes. As a result, exact matching of long nucleotide k-mers has been used for taxonomic classification and similarity detection between closely related genomes (genus-level to strain-level comparisons using k-mer lengths between 21-51) [1,16]. At larger evolutionary distances, accumulated nucleotide divergence limits the utility of exact nucleotide k-mer matching. Amino acid k-mers (kaa-mers) retain the benefits of fast, alignment-free exact k-mer matching, but provide increased tolerance for evolutionary divergence. Here, we evaluate the utility of amino acid k-mers for a wide range of genomic and metagenomic applications, including sequence distance estimation and taxonomic classification.

Amino Acid k_{aa10}mers can be used to discriminate between taxa

or: *Amino Acid k_{aa}mer distribution across taxa informs length selection*

The Genome Taxonomy Database (GTDB) provides a genome-similarity-based taxonomy for bacterial and archaeal genomes [17]. GTDB release `rs202` encompasses 258,407 genomes from 47,895 species. We begin by assessing the prevalance of nucleotide and amino acid k-mers of different k-mer lengths within assemblies included in GTDB.

To make analyses at this scale tractable, we use FracMinHash sketching to randomly subsample all available k-mers, retaining approximately 1/1000 nucleotide k-mers and 1/200 amino acid k-mers [13]. DNA FracMinHash sketches have been shown to representatively subsample genome datasets [13]. For most genomes, both genomic and protein fastas were available for download from NCBI. In remaining cases (n=36,632), genome fastas were translated into protein sequence via Prodigal [18] prior to sketching. We indexed these sketches into `sourmash` databases, which we have made available as part of the `Prepared Databases` section of the `sourmash` documentation, and archived on OSF [https://osf.io/t3fqa/].

For a range of nucleotide and amino acid k-mers lengths, we assessed the fraction of k-mers specific to each taxonomic rank. For nucleotide k-mers, we used lengths of 21, 31, and 51, which are commonly used for analyses at the genus, species, and strain level, respectively. For amino acid k-mers, we focused on k-mer lengths ranging between k=7 and k=11, which roughly correspond to nucleotide k-mer lengths 21-31. K-mers specific to a species were only present in a single species in the database; k-mers specific to a genus were found in at least two genomes of the same genus, etc. K-mers specific to a superkingdom were found in genomes/proteomes from at least two phyla.

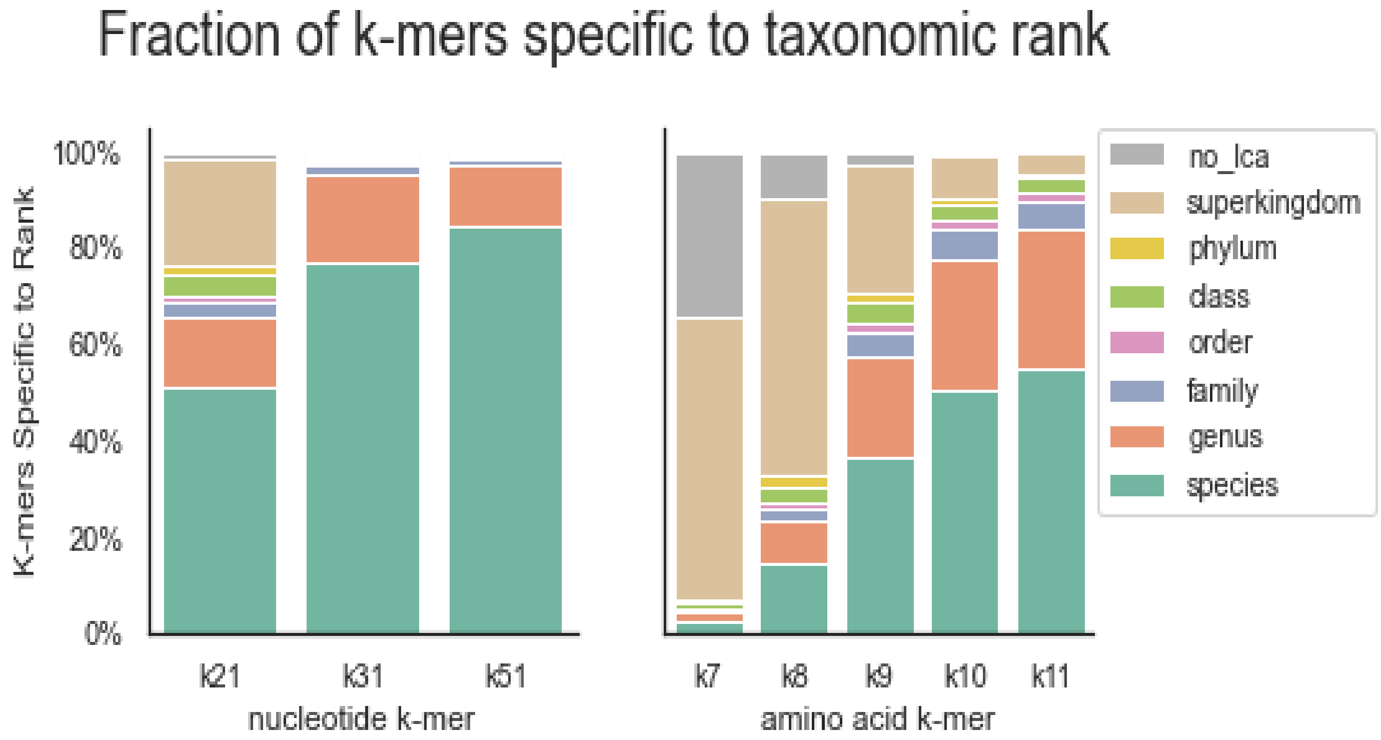

Figure 1: Fraction of k-mers specific to taxonomic rank

For the GTDB-RS202 database, the majority of nucleotide k-mers are specific to (unique at) a specific genome, species, or genus. Few k-mers are shared across superkingdoms, though these do exist at a k-mer length of 21. In contrast, all protein k-mer sizes contain a portion of k-mers that are shared across genera and above. At a protein k-mer size of 7, over 80% of k-mers are present in genomes found in more than one phylum, while at a protein k-size of 10, the number of genome-specific k-mers is closer to that observed for nucleotide k-mers. The differences observed between nucleotide and amino acid k-mers, as well as across different k-mer lengths suggests that these different k-mer sizes may provide resolution (CTB: do we want to say specificity here?) at different taxonomic ranks. We choose amino acid k-mer lengths 7 and 10 for our primary analyses, and have set a default kaa-mer length of 10 within `sourmash`.

Evolutionary Paths Dataset

To rigorously assess the utility of protein k-mers for comparisons at an array of evolutionary distances, we selected a subset of GTDB genomes that would allow standardized comparisons across taxonomic ranks and overcome the database-inclusion limitations mentioned above.

For each genus with at least two species clusters in GTDB, one representative genome was randomly selected as an “anchor” genome. Then, one additional genome was selected from the GTDB representative genomes matching the anchor’s taxonomy at each higher taxonomic rank. This “evolutionary path” consists of seven genomes: an anchor genome, a genome matching anchor taxonomy down to the genus level, one matching anchor taxonomy to the family level, one matching to the order level, and so on. This creates a gradient of similarity from genus to superkingdom.

Path selection using the representative genomes in GTDB rs202 resulted in 4095 paths comprised of 9213 unique genomes (8790 Bacteria, 333 Archaea). These paths include genome comparisons across 40 phyla (36 Bacteria, 4 Archaea), covering roughly a quarter of the 169 phyla (149 Bacteria, 20 Archaea) in GTDB release rs202. While paths are limited to taxonomies with at least two GTDB representative genomes for each taxonomic rank, these paths provide a rich resource for comparisons at increasing evolutionary distances. (CTB: are these available for download?)

For DNA comparisons, each genome was sketched from the available genome downloaded from genbank. For protein comparisons, we conducted both protein comparisons and translated comparisons. In both workflows, all anchor genomes were sketched from available proteomes (either downloaded or generated via Prodigal, as above). For the direct protein assessment, comparison proteomes were also sketched from the available proteome. For these sketches, k-mer containment results are equally valid in both directions, so we report the mean containment for the comparison alongside the Jaccard Index. For the second workflow, comparison genomes were 6-frame translated to build protein kaa-mers. As 6-frame translation introduces erroneous k_{aa}-mers, we report only the containment estimate relative to the anchor proteome (CTB: perhaps note that the intuition here is that for long k, only correct k-mers will match). We term this "anchor containment", where the trusted genome is the "anchor" upon which we base the comparison. We conduct k-mer comparisons using `sourmash` FracMinHash sketches default fractional scaling: 1/1000 k-mers from DNA sketches and 1/200 k-mers for protein sketches (including 6-frame translated sketches). For amino acid k-mers, we focus on k-mer lengths k=7 and k=10, which are closest to nucleotide k-mer lengths 21 and 31. To verify results and estimate the impact of FracMinHash scaling, we also conducted all comparisons using all available k-mers (no subsampling).

Protein k-mers facilitate alignment-free comparisons at increased evolutionary distances

We begin by assessing k-mer containment across the 6 comparisons (each genome compared with the anchor genome) within each of 4095 evolutionary paths. When plotted by the rank of the lowest common ancestor, the dynamic range of containment values is much larger for kaa-mer comparisons. While DNA k-mers can provide resolution at the genus level, log-transformed containment values for protein k-mers continue to decrease, providing resolution for comparisons even between genomes of different phyla. Average containment estimated from proteome sequence is very similar to anchor containment estimated from 6-frame translation of genome sequence, suggesting that either value can be used for this type of comparison. We obtained similar results when comparing all available k-mers, suggesting that these results are not affected by FracMinHash scaling (*Supplemental Figure XX*).

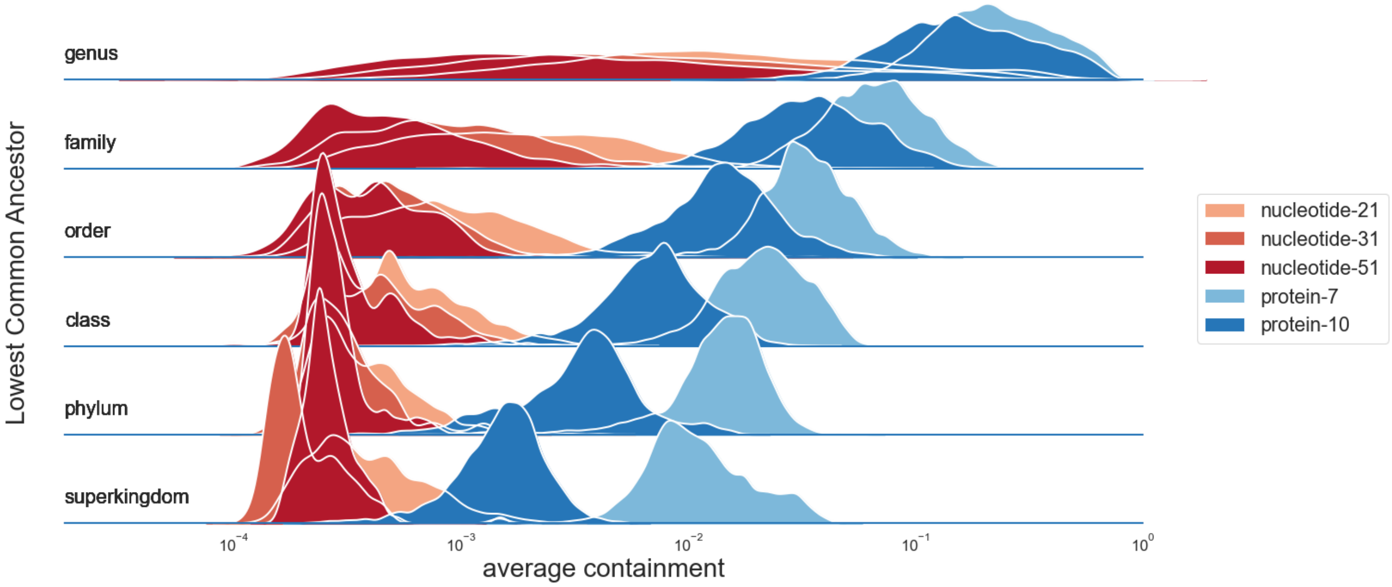

Figure 2: Protein k-mers are shared at higher taxonomic ranks Default scaled values 1000, 200. CTB: provide log scale on x axis.

FracMinHash k$_{aa}$-mer containment estimates average proteome identity and coverage

The Jaccard estimate from MinHash k-mer comparisons can be transformed into the “Mash Distance,” which estimates the rate of sequence mutation under a simple evolutionary model assuming equal and random nucleotide substitution at any position across a genomic sequence [1,19]. Despite potential issues with assuming such a simple mutational model, Mash Distance estimates have been shown to be reliable for high quality genomes with high similarity (>90% ANI), and have permitted sequence distance estimation at much larger scales than is tractable for traditional alignment-based sequence identity estimation [1,10,20]..

Compared with Jaccard, the Containment Index permits more accurate estimation of genomic distance, particularly for genomes of very different lengths [10,21,22]. As we recently described ([23]), we can use the same simple mutational model to obtain a point estimate of sequence identity between two genomes. This estimate is highly correlated with mapping-based Average Nucleotide Identity (ANI) estimates even when using only a small fraction of k-mers (e.g. fractional scaling = 1/1000) [23].

When FracMinhash sketches are instead generated with k$_{aa}$-mers, we can use a similar approach to estimate average amino acid identity (AAI) between two proteomes. Traditional AAI represents the average amino acid identity of all genes shared between two proteomes, and has shown lasting utility for phylogenomic comparisons and taxonomic classification [11,12]. However, alignment-based AAI analyses are not tractable for large-scale comparisons.

Under a simple mutational model assuming equal and random *amino acid* substitution at any position across a proteome, we can use the Fractional Containment Index $C_{\text{frac}}(A, B)$ (estimated at k$_{aa}$-mer length $k\ aa$) to obtain a containment-based estimate of Amino Acid Identity ($cAAI$). As with nucleotide k-mer comparisons, we also derive confidence intervals around this point estimate to account for the variance of FracMinHash subsampling (see Methods and [23] for details).

$$cAAI = C_{\text{frac}}(A, B)^{1/k\ aa}$$

To assess the utility of $cAAI$ for phylogenetic comparisons, we tranform the k$_{aa}$mer containment values of the “Evolutionary Paths” dataset (above) into $cAAI$ values. For each pairwise comparison, we then also estimate AAI with programs leveraging three different alignment algorithms: EzAAIb (BLASTp), EzAAIm (MMSeqs2), and CompareM (DIAMOND). As BLAST-based alignment remains the gold-standard method, we compare each AAI and $cAAI$ value to the AAI values generated with EzAAIb’s BLASTp approach.

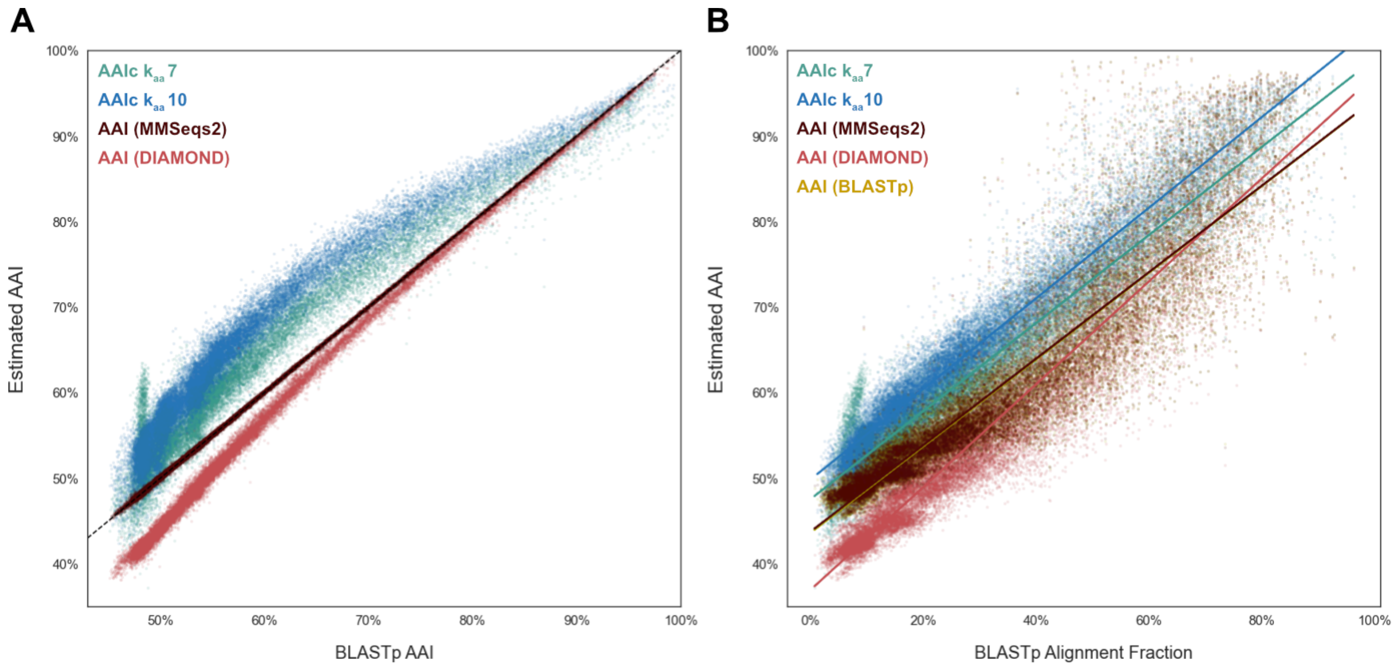

Figure 3: FracMinHash cAAI vs BLASTp alignment-based AAI

When compared with alignment-based AAI (Figure 3A), $cAAI$ is more sensitive, finding increased similarity across the entire range of comparisons. Across much of the AAI range, k$_{aa}$7 $cAAI$ is closer to AAI than k$_{aa}$10, but k$_{aa}$7 $cAAI$ values become unreliable at the low end of AAI range, with a spike in $cAAI$ values observed below 50% AAI. This suggests that k$_{aa}$7 containment does not provide sufficient resolution to distinguish protomes in the 40-60% AAI range. In contrast, despite their difference from AAI, k$_{aa}$10 $cAAI values provide resolution across the full range of pairwise comparisons. The three alignment-based methods are highly correlated, though DIAMOND-based AAI estimates consistently underestimate proteome similarity, particularly at the lower end of the range.

Unlike traditional AAI methods, $cAAI$ need not identify the best reciprocal matches between genes found in each proteome. $cAAI$ is based entirely on containment, the fraction of k-mers found in each proteome that is shared with the comparison proteome. When we instead compare AAI and $cAAI$ values to the BLASTp-reported alignment fraction, we see that $cAAI$ k$_{aa}$10 is best correlated with alignment fraction (R^2=0.94). Again, k$_{aa}$7 loses resolution for highly divergent proteomes. For alignment-based methods, we observe a slightly lower correlation between alignment fraction and AAI (R^2: 0.92 BLASTp, 0.86 MMSeqs2, 0.89 DIAMOND). This can be both a strength and shortcoming of AAI as a method, as traditional AAI values only consider the alignable fractions of the two proteomes.

Given these propeties, containment-based $cAAI$ is closest to an alignment-adjusted version amino acid identity. While it does not exactly mimic alignment-based AAI estimates, it is able to represent both alignment fraction and identity information in a single value.

AAI from 6-frame translated sequences

As above, we utilize anchor containment for comparisons involving 6-frame translated sketches.

Protein k-mer containment and AAI can be used for taxonomic classification

Given that protein k-mers facilitate similarity estimation across these larger evolutionary distances, we next assess the utility of protein k-mers for taxonomic assignment, both for metagenome breakdown/classification and for assembled genomes.

Metagenome breakdown using protein k-mers

As developed in Irber et al., 2022 [13], minimum set cover of nucleotide k-mers can be used to find the set of genomes that cover all known k-mers in a metagenome. This approach, implemented in `sourmash gather`, works by using k-mer containment relative to reference genomes ("anchor containment", as above) and "assigning" metagenome k-mers iteratively to the reference genome with highest containment. Anchor containment is then re-estimated using the remaining unassigned query k-mers until all known k-mers have been assigned. This step provides us with an ordered list of reference genomes, each of which represent a non-overlapping portion of the metagenome. The taxonomy of these matched reference genomes thus represents the closest match for each of these non-overlapping portions of the metagenome. In addition to reporting these exact matches, we can aggregate these taxonomic assignments of these matches to obtain taxonomic summarization at each rank.

Here, we assess the utility of protein k-mers for this application using the same metagenome samples described in Irber et al., 2022 [13]. As metagenome samples are unassembled, we use the 6-frame translation approach described above to obtain protein k-mers for comparison. No modification to the min-set-cov approach is required, as it already relies upon anchor containment to the reference genomes.

add figure: genome-grist mg breakdown, nucl k-mers, prot k-mers, nucl mapping

do we need an additional metagenome w/more divergent genomes, to show advantage of protein methods?

Robust taxonomic classification from protein k-mers

We use a similar approach for taxonomic classification of assembled genomes from protein k-mer containment. We apply the same minimum set cover approach to find the set of reference genomes that cover all known k-mers in our sample (in this case, a genome itself rather than a metagenome). If the most contained reference genome is sufficiently similar (passes default or user-defined threshold) to our query, we can annotate the query with taxonomic information from this reference genome. If not, we can use the genome-based lowest common ancestor approach to classify the query genome to the taxonomic rank where it contains sufficient similarity to matched reference genome sequence.

We select two sets of genomes: first, a set of 1000 genomes from the MGNify project ("MGNify-1000"), which are assembled from human gut and likely to be well-represented in existing databases. We next choose a set of 885 microbial ("Delmont-885"; 820 *Bacteria*, 65 *Archaea*) metagenome-assembled genomes (MAGs) assembled from TARA Oceans metagenomes [24]. As the marine environment is understudied relative to human gut, these genomes are more challenging for classifiers as they are less likely to have close relatives available in reference databases.

To assess the utility of protein k-mers for genome classification, we conduct this classification using three k-mer approaches: direct nucleotide k-mers, 6-frame translated protein k-mers, and direct protein k-mers from prodigal-translated proteomes. Where reference taxnonomic lineages were available (MGNify-1000), we compared our results directly to these annotations. With experimental genomes where no reference taxonomic lineage is available, we assessed our annotation relative to `gtdb-tk` classification [25].

Dataset	Exact Match	Higher Rank	Unclassified (sourmash)	Unclassified (GTDB-Tk)
MGNify-1000	95.7%	4.3%	N/A	N/A

Dataset	Exact Match	Higher Rank	Unclassified (sourmash)	Unclassified (GTDB-Tk)
Delmont-885	73.5%	26.5%	1 (0.1%)	15 (1.7%)

discuss/utilize/assess AAI threshold for tax classif?

Discussion

Protein sequences are more conserved than their underlying DNA sequence, allowing amino acid k-mer comparisons to match across larger evolutionary distances. Protein sequence matching is also less impacted by sequencing errors due to codon degeneracy. Our results show that amino acid k-mers can be used for global proteome comparisons, including estimation of sequence similarity (AAI) and taxonomic classification from either unassembled read datasets or assembled proteomes.

k_{aa}mer Containment comparisons for alignment-free and assembly-free protein analyses

As the majority of genome and proteome data is sequenced at the nucleotide level, comparisons of amino acid sequence are typically limited to assembly-based workflows, where a genome assembly has been translated into predicted Open Reading Frames (ORFs). As amino acid k_{aa}mers do not utilize any additional assembly information, we can also conduct comparisons directly from read datasets, bypassing assembly altogether. In many cases, we are interested in comparing a novel sequencing dataset to one or more reference proteomes. In this case, 6-frame translation of read sequences is sufficient for accurate comparisons to a k_{aa}mer database generated from reference (assembled) proteome data [10]. This method relies on the biological properties of k_{aa}mers: not all potential k_{aa}mers are represented in biological databases, particularly at longer k_{aa}mer lengths such as k_{aa7} and k_{aa10} (info theory paper). Containment comparisons allow us to focus only on the translated read k_{aa}mers that match to the reference database, ignoring the 5/6ths of k_{aa}mers that will originate from incorrect reading frames. When using assembled genomes as a test case, our results show that this 6-frame translation method performs equally well when compared with generating k_{aa}mers from the corresponding Prodigal-translated or RefSeq-downloaded proteomes. By using only the k-mer containment estimate relative to reference proteomes, we can obtain accurate Amino Acid Identity estimates directly from DNA sequence. In this way, we can use the more permissive nature of protein analyes for assembly-free genome and metagenome assignment and comparisons. Of course, when both samples are proteomes that are equally trusted (e.g. neither set of protein k-mers is being 6-frame translated from genome sequence), then the average containment considers the entire set of protein sequence from both proteomes.

k_{aa10} enables whole-proteome comparisons out to domain-level

Selection of k_{aa}mer size is critical for the utility of these comparisons. Longer k_{aa}mer lengths (k_{aa}=10) provide sufficient resolution for comparisons across taxonomic groups, including out to comparisons between genomes in different domains. In contrast, shorter k_{aa}mer lengths (k_{aa}=7) are more likely to be shared across taxonomic groups, with the majority of unique k_{aa7}mers shared across phyla within GTDB. These k-mers lose resolution at wider evolutionary distances (<62% AAI), where insufficient k_{aa}mers are unique to each taxonomic group to allow for taxonomic classification and robust similarity estimation via k_{aa}mer containment. However, this property can be advantageous for functional similarity estimation, where k_{aa7}mers may provide the ability to match gene sequence even across wider evolutionary distances. Indeed, this functionality has already begun to be explored, with k~aa7-mer functional annotation server [14].

FracMinHash k_{aa}-mer sketches support whole-proteome analysis at scale

All of the k_{aa}mer methods described here function with all dataset k_{aa}mers. However, our results show that leveraging sketching methods such as FracMinHash and that retain as few as 5% of microbial proteome k-mers ($scaled = 200$) maintains accuracy of k_{aa}mer containment comparisons while reducing runtime, memory, and storage requirements. Smaller proteomes such as viral proteomes and read-level or contig-level analysis may require different fractional scaling or different approach to ensure sufficient k_{aa}mers for accurate comparisons. While we have focused on FracMinHash sketching, these k_{aa}mer comparisons can be used with any sketch that enables containment estimation. Comparisons between sets of similar sizes or without 6-frame translation of protein k-mers can also use Jaccard to estimate $cAAI$, though Containment comparisons will always be as or more accurate than Jaccard comparisons [21,23].

$cAAI$ estimates Amino Acid Identity and Alignment Fraction

Amino Acid Identity (AAI) is a measure of average protein similarity between the homologous regions of two proteomes. Traditional AAI methods use BLAST or BLAST-like alternatives to identify homologous fragments for comparison. K-mer Jaccard and containment have been used for estimating average nucleotide identity between genomes [1,19,23]. While several studies have proposed utilization of more complex evolutionary models [26], an m-dependent model taking into account the non-independence of mutated k-mers is able to closely approximate alignment-based estimates of Average Nucleotide Identity (ANI) [23,27]. Applying this same simple mutational model to amino acid k_{aa}mer containment yields $cAAI$ values that strongly correlate with alignment-based AAI values.

For many alignment-based AAI approaches, it is important to report both the percent identity of matched regions and the fraction of the genomes that were mapped. This prevents believing artificially high similarity values when only small fractions of the genomes overlap. While alignment AAI measure sequence similarity, k_{aa}$mer containment, the fraction of shared dataset k_{aa}$mers, is closer to a proxy for the fraction of shared proteome. Indeed, $cAAI$ is strongly correlated with the alignment fraction reported by BLASTp AAI (as executed by EzAAI [28]), suggesting that $cAAI$ represents an aggregated similarity measure that encompasses both amino acid similarity and the fraction of shared protein sequence. K_{aa}$mer length selection remains critical, as discussed for direct k_{aa}mer Containment comparisons. while $k\ aa7\ cAAI$ is more closely related to BLASTp values, k_{aa}mer similarity saturates at distances greater than ~65% AAI. While $cAAI\ kaa10$ are slightly offset from traditional AAI values, k_{aa10} comparisons enable consistent whole-proteome similarity estimation even out to the widest evolutionary distances. As with Average Nucleotide Identity comparisons, different AAI approaches vary slightly in the AAI reported for a given pair of proteomes, suggesting that comparisons are best made between values produced by the same method [29].

The majority of AAI estimation software has focused been alignment-based comparisons, which has limited comparisons at scale. Larger-scale comparisons have leveraged similarity estimation across single-copy universally conserved genes [30]. A recent extension of this concept introduced $\widehat{AAI}$ which uses tetramer frequency of universal single-copy proteins to estimate AAI, enabling AAI estimation for hundreds of thousands of comparisons [31]. While FracMinHash k_{aa}mer subsampling allows $cAAI$ to operate at a similar scale, $cAAI$ encompasses information from across the whole proteome, which averages similarity across fast and slow-evolving genes, does not require prior selection of appropriate universal protein sets, and allows AAI estimation directly from read datasets. However, speed of $cAAI$ comparisons depends on the FracMinHash k_{aa}mer subsampling rate and can be impacted by the quality of the reference proteomes used for comparison. *For GTDB-wide* *comparisons, FracMinHash subsampling of XX% of k_{aa}mers per proteome allowed XXX,XXX pairwise comparisons in XX time.*

k_{aa}mer Taxonomic Assignment is database-dependent

(but protein helps with sensitivity + min-set-cov helps with specificity)

K-mer based taxonomic assignment relies upon matching k-mers found in previously sequenced reference proteomes. While this approach will always be database-dependent and improved by presence of closely-related proteomes in the database, protein-based matching allows for classification at larger evolutionary distances. While protein matching increases the sensitivity by matching across synonymous substitutions in the DNA sequence, classification LCA approaches often suffer from sensitivity/specificity trade-offs. Here, the use of `sourmash gather` minimum set cover approach assigns each protein k-mer to its most likely/parsimonious match based on presence of other proteome k-mers present in the query genome/metagenome.

We expect classification at the amino acid level to be most useful for organisms not well represented in reference databases. In these cases, the increased sensitivity of k_{aa}mers can find any available similarity in the database. While this similarity may not be sufficient to provide a species or even genus-level annotation, even higher-level taxonomic information can be useful when attempting to understand and classify novel metagenomic samples.

Limitations

While k_{aa}mer containment allows protein analysis directly from DNA reads, it cannot be used for comparisons between two 6-frame translated datasets. In this case, there is no way to distinguish between true biological k_{aa}mers and noisy k_{aa}mers introduced by translation into all potential open reading frames. AAI is often most useful when comparing to known or assembled genomes, and does not have a direct application for comparisons between two unassembled metagenomic datasets. If desired, it is possible to use a k_{aa}mer Containment strategy to select the most likely translation frame for each read, which can be used for downstream analyses. If instead, k_{aa}mers are generated for each translation frame of each read separately, Containment comparisons can again be used to find the fraction of these k_{aa}mers that can be matched to the k_{aa}mers in the reference database. The translation frame with the highest percent of matched k_{aa}mers is most likely to be the coding frame for that read. These strategy can also be used to predict non-coding reads, where few, if any, translated k_{aa}mers match to reference database k_{aa}mers. This method works best when there are closely related organisms present in the reference database [15,32].

Future directions and utility

Here we have focused on k_{aa}mer containment comparisons for whole-proteome comparisons, both for Amino Acid Identity estimation and Taxonomic Classification. We anticipate k_{aa}mer comparison methods will be especially useful for poorly sequenced environments where many organisms are not well represented in reference databases, where protein-level similarity to a common ancestor can result in classification to higher taxonomic ranks as needed. Protein comparisons are also critical for sequence comparisons across polymorphic sequences, including viral proteome comparisons or metapangenome analysis [32]. Both standard k_{aa}mers and k_{aa}mers generated from reduced amino acid alphabets can be used for a myriad of other applications as well, including functional annotation and clustering [10,14,33], gene expression analyses, metapangenomics [32], and single-cell eukaryotic transcriptomics [15]. FracMinHash k_{aa}mer subsampling may be useful for a subset of these applications in order to allow analysis at scale without loss of accuracy.

Conclusions

Protein search has long been used for comparisons conducted at increased evolutionary distances. Using genomes from the Genome Taxonomy Database (GTDB) we showcase the utility of amino acid k_{aa}-mer comparisons for alignment-free and assembly-free proteome similarity estimation and taxonomic classification. Subsampling k_{aa}mers using FracMinHash sketching makes global protein similarity assessment tractable for the current scale of sequencing.

Methods

Large-scale k-mer comparisons with FracMinHash sketches

FracMinHash sketching, as implemented in sourmash [13,34,35], is a MinHash variant that uses a scaling factor to subsample the unique k-mers in the dataset to the chosen fraction (1/ `scaled`). As k-mers are randomized prior to systematic subsampling, FracMinHash sketches are representative subsets that can be used for comparisons across datasets sketched with consistent k-mer lengths and scaling factors.

While FracMinHash sketches can be used to estimate both the Jaccard Index [1] and Containment Index [21], containment has been shown to permit more accurate estimation of genomic distance when genomes or datasets differ in size [21,22,36,37]. We focus here on the utility of containment comparisons for similarity estimation. Containment comparisons are directional: the containment of genome A in sample B is the interection of k-mers in A and B divided by the k-mers in genome A (and vice versa). Thus, two containment values can be estimated for a given pairwise comparison. The choice of which containment value to use (or whether to average the two values) depends on the particular comparison. FracMinHash containment has been shown to be an unbiased estimator of the true containment index, as long as the sketches contain sufficient k-mers for comparison or utilize a high-quality estimation of the true cardinality of the dataset [13,23].

Sourmash supports sketching from either nucleotide or protein input sequence, to generate either nucleotide or protein FracMinHash sketches. We generated nucleotide and protein sketches directly from genome and protome files, respectively. All genome sequences were sketched with sourmash v4.2.1 using the `sourmash sketch dna` command, k-mer sizes of 21,31,51, a scaling factor of 1000. All proteome sequences were sketched with sourmash v4.2.1 using the `sourmash sketch protein` command at protein k-sizes (*kaa-mer sizes?*) of 7-12 and a scaling factor of 200. Sourmash also supports 6-frame translation of nucleotide sequence to amino acid sketches. To assess the utility of these translated sketches, genome sequences were also sketched with the `sourmash sketch translate` command at protein k-sizes (*kaa-mer sizes?*) of 7-12 and a scaling factor of 200.

In select cases, we also conducted comparisons using all available k-mers, rather than using FracMinHash sketch subsampling. While `sourmash` sketching is not optimized for this use case, we can generate these complete k-mer sketches using the same `sourmash` commands with a scaling factor of 1 (`scaled` =1).

Anchor Containment analysis for protein comparisons directly from DNA sequence

For protein k-mer comparisons to be useful, any DNA queries must be translated into protein sequence. This typically limits amino acid comparisons to assembly-based workflows, as assemblies can be reliably translated into predicted Open Reading Frames (ORFs). With k-mer methods, we can utilize direct 6-frame translation, which is assembly-free but does not attempt to find the correct open reading frame. Assuming a single open reading frame, only ~1/6th of the k-mers generated by 6-frame translation will belong to true ORFs. The remaining erroneous k-mers greatly impact the Jaccard Index (set similarity) when comparing samples, but only impact the containment index in one direction (relative to the 6-frame translated set). The containment estimate relative to reference proteomes will be an accurate comparison directly from DNA sequence. We term this “anchor” containment, where the trusted genome is the “anchor” upon which we base the comparison. Since 6-frame translation should always yield excess k-mers relative to genomes of similar size, this desired containment will generally be the larger of the two containment values (maximum containment).

To facilitate these comparisons within `sourmash`, we have implemented “maximum containment,” a shorthand method to select the greater of the two containment values. The maximum containment method may also provide advantages for genomes with potential contamination, as containment will always be relative to the smaller, and presumably less contaminated, genome. However, highly incomplete genomes may overestimate similarity with this method, so we suggest using containment relative to the more trusted sample if known, or considering both containment values when comparing two genomes of approximately equal quality. Note that comparing two 6-frame translated datasets is not recommended, as there is no mechanism to exclude erroneous k-mers introduced during translation.

Estimating Average Amino Acid Identity

MinHash Sketch Jaccard has been shown to correlate well with ANI at high sequence identities (>=90% sequence identity) [1]. Recently, Blanca et al, 2021 [38] presented a method to increase the accuracy of sequence similarity estimation from MinHash Jaccard by recognizing that k-mers generated from mutated sequence are not independent. Hera et al, 2022 [23] extended this approach to estimate sequence identity from FracMinHash Containment estimates. Each of these methods assumes a simple mutational model, with equal substitution probability for each nucleotide, and then estimates sequence identity based on k-mer comparisons. Here, we note that there is nothing unique to nucleotide sequence included in these equations. By applying the same equations to comparisons between amino acid k-mer sketches, we can estimate average Amino Acid Identity (AAI) between proteomes. For this application, we maintain the assumption of a simple mutational model of equal substitution probability at each position, but recognize that it now applies to any amino acid, rather than any nucleotide.

The debiased Fractional Containment Index is an unbiased estimator of the true containment index $C(A,B)$.

$$C_{\text{frac}}(A,B) = \frac{|\mathbf{FRAC}_s(A) \cap \mathbf{FRAC}_s(B)|}{|\mathbf{FRAC}_s(A)|\left(1-(1-s)^{|A|}\right)} \backslash\text{mathbbm}1_{|\mathbf{FRAC}_s(A)|>0}$$

The equation for sequence similarity estimation (ANI or AAI) from FracMinHash Containment is reproduced here for completeness. See [23] for these and other analytical details.

Immplementation of ANI and AAI estimation

We provide an implementation of Fractional Containment to average sequence identity (ANI/AAI) in the software package `sourmash`, which is implemented in Python and Rust and developed under the BSD license [34,35]. ANI and AAI values can be reported from sequence comparisons The distance estimation equations can be found in the `distance_utils.py` file and ANI/AAI values can be reported from a variety of `sourmash` comparison and search commands as of version `4.4`.

sourmash is available at github.com/sourmash-bio/sourmash. The results in this paper were generated with sourmash v4.4.1.

FracMinHash Amino Acid Identity Correlates with Alignment-based Methods

To assess whether k-mer methods can be used to approximate AAI, we ran generated alignment AAI values for each pairwise comparison using methods that leverage different mapping algorithms: EzAAIb (BLASTp), EzAAIm (MMSeqs2), and CompareM (DIAMOND). As BLAST-based alignment remains the gold-standard method, we compare all AAI values the BLAST AAI values.

EzAAI v1.12 [28] was used to run BLAST-based and MMSeqs-based Amino Acid Identity. The EzAAI workflow begins with PRODIGAL-based translation of genome sequence [4], followed by reciprocal BLAST [**blast?**] or MMSeqs2 [39] alignment. For both, we utilized EzAAI default parameters: 40% coverage threshold, 40% sequence identity threshold. CompareM v0.1.2 ([40]; run with `--sensitive` parameter for DIAMOND mapping) was used to obtain Average Amino Acid Identity between the anchor proteome and each additional proteome in its evolutionary path. CompareM reports the mean and standard deviation of AAI, as well as the fraction of orthologous genes upon which this estimate is based. Briefly, CompareM calls genes for each genome or proteome using PRODIGAL [4] and conducts reciprocal best-hit mapping via DIAMOND [18]. By default, CompareM requires at least 30% percent sequence identity and 70% percent alignment length to identify orthologous genes. As DIAMOND alignment-based homology identification may correlate less well with BLAST-based homology under 60% sequence identity [41/], **we also ran compareM with a percent sequence identity threshold of 60% to obtain a set of high-confidence orthologous genes for AAI estimation. We report correlation between FracMinHash AAI estimation and each of these compareM parameter sets in XX *(TBD)*.** *CompareM was also used to obtain AAI values directly from each genome, using PRODIGAL to translate sequences prior to gene calling. These results [were not significantly different from proteome-based AAI estimation??] (Supplemental XX).*

Taxonomic Classification with Sourmash `Gather` and `Taxonomy`

To take advantage of the increased evolutionary distance comparisons offered by protein k-mers, we apply compositional analysis with sourmash gather [13] to protein sequences (amino acid input and 6-frame translation from nucleotides). Sourmash gather is conducted in two parts: first (preselection), gather searches the query against all reference genomes, building all genomes with matches into a smaller, in-memory database for use in step 2. Second (decomposition), gather does iterative best-containment decomposition, where query k-mers are iteratively assigned to the reference genome with best containment match. In this way, gather reports the minimal list of reference genomes that contain all of the k-mers that matched any reference in the database. For queries with high sequence identity (ANI) to reference matches, we classify the query sequence as a member of the reference taxonomic group, as in [13]. However, when ANI between the query and the top reference match exceeds the taxonomic rank threshold (e.g. species default 95%), we use a least/lowest common ancestor (LCA) approach to report likely taxonomy at a higher taxonomic rank. Briefly, as gather reports non-overlapping genome matches, we can sum the k-mer matches for all genomes with shared taxonomies at the next higher taxonomic rank to report the best query containment at that rank. As this gather-LCA approach first uniquely assigns k-mers to their best reference genome, it bypasses the impact of increasing database size on taxonomic assignment observed for other LCA-based k-mer classification approaches [42].

These taxonomic utilities are implemented in the `sourmash taxonomy` module, and classifications were run on gather output via `sourmash tax genome`.

Workflows and Computing Resources

Reproducible workflows associated with this paper are available at https://github.com/bluegenes/2022-protein-kmers-workflow (ADD DOI for release), with datasets available at OSF (XX). All workflows were executed using snakemake >= 5.26 [43)] on the FARM cluster at UC Davis, using practices outlined in [44].

References

1. **Mash: fast genome and metagenome distance estimation using MinHash**
Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, Adam M Phillippy
Genome Biology (2016-12) https://doi.org/gfx74q
DOI: 10.1186/s13059-016-0997-x · PMID: 27323842 · PMCID: PMC4915045

2. **Improved metagenomic analysis with Kraken 2**
Derrick E Wood, Jennifer Lu, Ben Langmead
Genome Biology (2019-11-28) https://doi.org/ggfk55
DOI: 10.1186/s13059-019-1891-0 · PMID: 31779668 · PMCID: PMC6883579

3. **Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets**
Daniel M Portik, CTitus Brown, NTessa Pierce-Ward
Bioinformatics (2022-02-02) https://doi.org/hhqs
DOI: 10.1101/2022.01.31.478527

4. **Fast and sensitive protein alignment using DIAMOND**
Benjamin Buchfink, Chao Xie, Daniel H Huson
Nature Methods (2015-01) https://doi.org/gftzcs
DOI: 10.1038/nmeth.3176 · PMID: 25402007

5. **A review of methods and databases for metagenomic classification and assembly**
Florian P Breitwieser, Jennifer Lu, Steven L Salzberg
Briefings in Bioinformatics (2019-07-19) https://doi.org/gdq95k
DOI: 10.1093/bib/bbx120 · PMID: 29028872 · PMCID: PMC6781581

6. **Fast and sensitive taxonomic assignment to metagenomic contigs**
M Mirdita, M Steinegger, F Breitwieser, J Söding, E Levy Karin
Bioinformatics (2021-09-29) https://doi.org/gnnprm
DOI: 10.1093/bioinformatics/btab184 · PMID: 33734313 · PMCID: PMC8479651

7. **Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT**
FABastiaan von Meijenfeldt, Ksenia Arkhipova, Diego D Cambuy, Felipe H Coutinho, Bas E Dutilh
Genome Biology (2019-12) https://doi.org/ggfm6r
DOI: 10.1186/s13059-019-1817-x · PMID: 31640809 · PMCID: PMC6805573

8. **MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs**
Daniel H Huson, Benjamin Albrecht, Caner Bağcı, Irina Bessarab, Anna Górska, Dino Jolic, Rohan BH Williams
Biology Direct (2018-01) https://doi.org/gnnprp
DOI: 10.1186/s13062-018-0208-7 · PMID: 29678199 · PMCID: PMC5910613

9. **Fast and sensitive taxonomic classification for metagenomics with Kaiju**
Peter Menzel, Kim Lee Ng, Anders Krogh
Nature Communications (2016-09) https://doi.org/f8h4b6
DOI: 10.1038/ncomms11257 · PMID: 27071849 · PMCID: PMC4833860

10. **Mash Screen: high-throughput sequence containment estimation for genome discovery**

Brian D Ondov, Gabriel J Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B Buck, Adam M Phillippy
Genome Biology (2019-12) https://doi.org/ghtqmb
DOI: 10.1186/s13059-019-1841-x · PMID: 31690338 · PMCID: PMC6833257

11. **Toward a More Robust Assessment of IntraspeciesDiversity, Using Fewer GeneticMarkers**
Konstantinos T Konstantinidis, Alban Ramette, James M Tiedje
Applied and Environmental Microbiology (2006-11) https://doi.org/dcmw9q
DOI: 10.1128/aem.01398-06 · PMID: 16980418 · PMCID: PMC1636164

12. **Uncultivated microbes in need of their own taxonomy**
Konstantinos T Konstantinidis, Ramon Rosselló-Móra, Rudolf Amann
The ISME Journal (2017-11) https://doi.org/gbprgw
DOI: 10.1038/ismej.2017.113 · PMID: 28731467 · PMCID: PMC5649169

13. **Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers**
Luiz Irber, Phillip T Brooks, Taylor Reiter, NTessa Pierce-Ward, Mahmudur Rahman Hera, David Koslicki, CTitus Brown
Bioinformatics (2022-01-12) https://doi.org/gn34zt
DOI: 10.1101/2022.01.11.475838

14. **Flexible protein database based on amino acid k-mers**
Maxime Déraspe, Sébastien Boisvert, François Laviolette, Paul H Roy, Jacques Corbeil
Scientific Reports (2022-12) https://doi.org/gqhr3c
DOI: 10.1038/s41598-022-12843-9 · PMID: 35650262 · PMCID: PMC9160020

15. **Single-cell transcriptomics for the 99.9% of species without reference genomes**
Olga Borisovna Botvinnik, Venkata Naga Pranathi Vemuri, NTessa Pierce, Phoenix Aja Logan, Saba Nafees, Lekha Karanam, Kyle Joseph Travaglini, Camille Sophie Ezran, Lili Ren, Yanyi Juang, … CTitus Brown
Bioinformatics (2021-07-10) https://doi.org/gns4sg
DOI: 10.1101/2021.07.09.450799

16. **MetaPalette: a <i>k</i> -mer Painting Approach for Metagenomic Taxonomic Profiling and Quantification of Novel Strain Variation**
David Koslicki, Daniel Falush
mSystems (2016-06-28) https://doi.org/gg3gbd
DOI: 10.1128/msystems.00020-16 · PMID: 27822531 · PMCID: PMC5069763

17. **A complete domain-to-species taxonomy for Bacteria and Archaea**
Donovan H Parks, Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J Mussig, Philip Hugenholtz
Nature Biotechnology (2020-09-01) https://doi.org/ggtbk2
DOI: 10.1038/s41587-020-0501-8 · PMID: 32341564

18. **Prodigal: prokaryotic gene recognition and translation initiation site identification**
Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, Loren J Hauser
BMC Bioinformatics (2010-12) https://doi.org/cktxnm
DOI: 10.1186/1471-2105-11-119 · PMID: 20211023 · PMCID: PMC2848648

19. **An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data**
Huan Fan, Anthony R Ives, Yann Surget-Groba, Charles H Cannon
BMC Genomics (2015-12) https://doi.org/f7s6tp

DOI: 10.1186/s12864-015-1647-5 · PMID: 26169061 · PMCID: PMC4501066

20. **High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries**
Chirag Jain, Luis M Rodriguez-R, Adam M Phillippy, Konstantinos T Konstantinidis, Srinivas Aluru
Nature Communications (2018-12) https://doi.org/gfknmg
DOI: 10.1038/s41467-018-07641-9 · PMID: 30504855 · PMCID: PMC6269478

21. **Improving MinHash via the containment index with applications to metagenomic analysis**
David Koslicki, Hooman Zabeti
Applied Mathematics and Computation (2019-08) https://doi.org/ghtqrv
DOI: 10.1016/j.amc.2019.02.018

22. **Dashing: fast and accurate genomic distances with HyperLogLog**
Daniel N Baker, Ben Langmead
Genome Biology (2019-12) https://doi.org/ggkmjc
DOI: 10.1186/s13059-019-1875-0 · PMID: 31801633 · PMCID: PMC6892282

23. **Debiasing FracMinHash and deriving confidence intervals for mutation rates across a wide range of evolutionary distances**
Mahmudur Rahman Hera, NTessa Pierce-Ward, David Koslicki
Bioinformatics (2022-01-12) https://doi.org/gn342h
DOI: 10.1101/2022.01.11.475870

24. **Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes**
Tom O Delmont, Christopher Quince, Alon Shaiber, Özcan C Esen, Sonny TM Lee, Michael S Rappé, Sandra L McLellan, Sebastian Lücker, AMurat Eren
Nature Microbiology (2018-07) https://doi.org/gdvhp5
DOI: 10.1038/s41564-018-0176-9 · PMID: 29891866 · PMCID: PMC6792437

25. **GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database**
Pierre-Alain Chaumeil, Aaron J Mussig, Philip Hugenholtz, Donovan H Parks
Bioinformatics (2019-11-15) https://doi.org/ggc9dd
DOI: 10.1093/bioinformatics/btz848 · PMID: 31730192 · PMCID: PMC7703759

26. **On the transformation of MinHash-based uncorrected distances into proper evolutionary distances for phylogenetic inference**
Alexis Criscuolo
F1000Research (2020-11-10) https://doi.org/gjn4jw
DOI: 10.12688/f1000research.26930.1 · PMID: 33335719 · PMCID: PMC7713896

27. **The Statistics of <i>k</i> -mers from a Sequence Undergoing a Simple Mutation Process Without Spurious Matches**
Antonio Blanca, Robert S Harris, David Koslicki, Paul Medvedev
Journal of Computational Biology (2022-02-01) https://doi.org/gp7hqc
DOI: 10.1089/cmb.2021.0431 · PMID: 35108101

28. **Introducing EzAAI: a pipeline for high throughput calculations of prokaryotic average amino acid identity**
Dongwook Kim, Sein Park, Jongsik Chun
Journal of Microbiology (2021-05) https://doi.org/gqhr2f
DOI: 10.1007/s12275-021-1154-0 · PMID: 33907973

29. **All ANIs are not created equal: implications for prokaryotic species boundaries and integration of ANIs into polyphasic taxonomy**
Marike Palmer, Emma T Steenkamp, Jochen Blom, Brian P Hedlund, Stephanus N Venter
International Journal of Systematic and Evolutionary Microbiology (2020-04-01)
https://doi.org/gqhr2g
DOI: 10.1099/ijsem.0.004124 · PMID: 32242793

30. **The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level**
Luis M Rodriguez-R, Santosh Gunturu, William T Harvey, Ramon Rosselló-Mora, James M Tiedje, James R Cole, Konstantinos T Konstantinidis
Nucleic Acids Research (2018-07-02) https://doi.org/ghjn7m
DOI: 10.1093/nar/gky467 · PMID: 29905870 · PMCID: PMC6031002

31. **FastAAI: Efficient Estimation of Genome Average Amino Acid Identity and Phylum-level relationships using Tetramers of Universal Proteins**
Konstantinos Konstantinidis, Carlos Ruiz-Perez, Kenji Gerhardt, Luis Rodriguez-R, Chirag Jain, James Tiedje, James Cole
In Review (2022-03-23) https://doi.org/gqhr2h
DOI: 10.21203/rs.3.rs-1459378/v1

32. **Protein k-mers enable assembly-free microbial metapangenomics**
Taylor E Reiter, NTessa Pierce-Ward, Luiz Irber, Olga Borisovna Botvinnik, CTitus Brown
Bioinformatics (2022-06-27) https://doi.org/gqfjbh
DOI: 10.1101/2022.06.27.497795

33. **Snekmer: A scalable pipeline for protein sequence fingerprinting using amino acid recoding (AAR)**
Computational Biology @Pacific Northwest National Laboratory
(2022-10-17) https://github.com/PNNL-CompBio/Snekmer

34. **Large-scale sequence comparisons with sourmash**
NTessa Pierce, Luiz Irber, Taylor Reiter, Phillip Brooks, CTitus Brown
F1000Research (2019-07-04) https://doi.org/gf9v84
DOI: 10.12688/f1000research.19675.1 · PMID: 31508216 · PMCID: PMC6720031

35. **sourmash: a library for MinHash sketching of DNA**
C Titus Brown, Luiz Irber
The Journal of Open Source Software (2016-09-14) https://doi.org/ghdrk5
DOI: 10.21105/joss.00027

36. **Beware the Jaccard: the choice of <b>similarity measure</b> is important and non-trivial in genomic colocalisation analysis**
Stefania Salvatore, Knut Dagestad Rand, Ivar Grytten, Egil Ferkingstad, Diana Domanska, Lars Holden, Marius Gheorghe, Anthony Mathelier, Ingrid Glad, Geir Kjetil Sandve
Briefings in Bioinformatics (2020-09-25) https://doi.org/gjnvx4
DOI: 10.1093/bib/bbz083 · PMID: 31624847

37. **The minimizer Jaccard estimator is biased and inconsistent***
Mahdi Belbasi, Antonio Blanca, Robert S Harris, David Koslicki, Paul Medvedev
Bioinformatics (2022-01-17) https://doi.org/gpm78w
DOI: 10.1101/2022.01.14.476226

38. **The statistics of <i>k</i> -mers from a sequence undergoing a simple mutation process without spurious matches**

Antonio Blanca, Robert S Harris, David Koslicki, Paul Medvedev
Bioinformatics (2021-01-17) https://doi.org/fq3g
DOI: 10.1101/2021.01.15.426881

39. **MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets**
Martin Steinegger, Johannes Söding
Nature Biotechnology (2017-11) https://doi.org/ggctnw
DOI: 10.1038/nbt.3988 · PMID: 29035372

40. **CompareM**
Donovan Parks
(2022-07-26) https://github.com/dparks1134/CompareM

41. **AAI: BLAST vs Diamond**
LM Rodriguez-R
https://rodriguez-r.com/blog/aai-blast-vs-diamond/

42. **RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification**
Daniel J Nasko, Sergey Koren, Adam M Phillippy, Todd J Treangen
Genome Biology (2018-12) https://doi.org/ggc9db
DOI: 10.1186/s13059-018-1554-6 · PMID: 30373669 · PMCID: PMC6206640

43. **Sustainable data analysis with Snakemake**
Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O Twardziok, Alexander Kanitz, … Johannes Köster
F1000Research (2021-01-18) https://doi.org/gjjkwv
DOI: 10.12688/f1000research.29032.1 · PMID: 34035898 · PMCID: PMC8114187

44. **Streamlining data-intensive biology with workflow systems**
Taylor Reiter, Phillip T Brooks†, Luiz Irber†, Shannon EK Joslin†, Charles M Reid†, Camille Scott†, CTitus Brown, NTessa Pierce-Ward
GigaScience (2021-01-13) https://doi.org/gjfk22
DOI: 10.1093/gigascience/giaa140 · PMID: 33438730 · PMCID: PMC8631065