Comprehensive evaluation of genomic redundancy and taxonomic coherence in large biological databases

This manuscript (permalink) was automatically generated from dib-lab/2022-paper-genomic-tax-redundancy@eede4db on November 19, 2022.

Authors

- **John Doe**
  XXXX-XXXX-XXXX-XXXX · johndoe · johndoe
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Jane Roe** ✉
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

✉ — Correspondence possible via GitHub Issues or email to Jane Roe <jane.roe@whatever.edu>.

Abstract

A central challenge in bacterial genomics and taxonomy is that genomic databases are increasingly large and contain redundant content. This impacts the accuracy of taxonomic profilers and metagenome analysis tools that rely on these databases. Here, we probe the practical and theoretical limits of genomic identification and taxonomic classification by exploring *unicity distance* and *Shannon entropy*. Unicity distance provides an estimate of how many k-mers are required to precisely identify a single reference genome from within a database, while Shannon entropy of k-mers describes the informativeness of a k-mer for taxonomic classification. We show that approximately 30% of genomes in GTDB rs207 have infinite unicity and that 99% of k-mers can resolve taxonomy at the species, genus, or family level. We conclude that unicity distance and Shannon entropy provide simple metrics for evaluating genomic redundancy and taxonomic coherence of large genomic reference databases.

Introduction

Introduction goes here.

Results

Hackmd for tables here: https://hackmd.io/GvngZ4gHQE-9ERB4Gd71HQ

Many k-mers are genome specific

(also see unicity distance at bottom.)

Shannon entropy of k-mers can be used to measure taxonomic informativeness

We measured the species distribution in GTDB rs207 for 21.2 million hashes, representing 21.2 billion 31-mers, and calculated the Shannon entropy for each hash (equationXX) at the species, genus, and family levels. Per Table 1, 99.1% of hashes uniquely identify a specific family within the GTDB taxonomy.

Table 1: Entropy measurements for GTDB taxonomy using 318k genomes from rs207 genomes.

Taxonomic level (GTDB)	# perfectly informative hashes	cumulative % total
species	21,150,287	92.8%
genus	1,262,281	98.3%
family	170,249	99.1%

Make point that nucleotide k-mers are not necessarily specific beyond family (ref protein paper).

Add fourth row to tables.

Shannon entropy can summarize the taxonomic cohesion of taxonomies based on genomic relationships.

We can also calculate the Shannon entropy with respect to different taxonomies. Table 2 uses the NCBI taxonomy with the same genomes used above. Here we see that approximately 4.5% of hashes cannot be used to distinguish between different families - a full 5 times as many as with the GTDB taxonomy. These 1.0 million hashes represent approximately 10 billion k-mers, or approximately 2,000 bacterial genomes worth of sequence.

Table 2: Entropy measurements for same 318k GTDB rs207 genomes as in Table 1, but using NCBI taxonomic labels.

Taxonomic level (NCBI)	# perfectly informative hashes	cumulative % total
species	20,744,791	91.0%
genus	779,234	94.4%
family	245,718	95.5%

Many k-mers with non-zero entropy come from a few specific genomes

Explore taxonomic incoherence and database contamination.

Find challenging genomes. Explore low H values.

Unicity distance can be used to estimate genomic redundancy

We next ask, how many genomes can be distinguished from each other using a combinatorial collection of k-mers? To do this, we estimate the *unicity distance* of each genome in the database, where the unicity distance is defined as the smallest set of hashes capable of uniquely identifying an individual genome. (k=31, scaled=1000)

Table 3 shows that approximately 29.2% of the genomes in GTDB rs207 cannot be distinguished uniquely by *any* combination of 31-mers at a scaled of 1000, while some substantial amount (estimated as 15.3%) can be precisely identified using a single hash.

(Do higher resolution k-mer analysis of some of these; talk about error rates, etc.)

(Compare also with k-mer informativeness; can we tie entropy computation at top back to number of genomes with unicity of 1, and cross validate?)

Table 3: Estimated unicity distances with hashes for 318k GTDB rs207 genomes using FracMinHash as implemented in sourmash (k=31, scaled=1000).

Unicity distance	Number of genomes	Percent of genomes
1	48,630	15.3%
infinite	92,564	29.2%

Discussion

Species-level classification should be straightforward with k-mers

Taxonomic classification to the species level is straightforward, largely because GTDB taxonomy is closely tied to genomic content and most of the genomic redundancy lies within species and genus level. Thus GTDB taxonomy largely encapsulates this redundancy. The entropy measurements demonstrate that it is possible to choose an informative subset of k-mers that would robustly classify at the species level, and that doing so would not compromise sensitivity. LCA-style approaches such as those used by Kraken should work even if we use genus and family level k-mers, while eliminating those above.

Shared genomic content at higher levels confounds taxonomic classification methods. While surely some shared genomic content is real, our analysis suggests that significant portions of it are contamination.

Classification below the strain level

Detecting genomes from sequences is easy with k-mers, but significant redundancy prevents straightforward classification to the genome/strain level. Here leveraging combinatorial application of k-mers provides significant leverage; this is how sourmash achieves high precision. Nonetheless sourmash cannot distinguish a full 30% of the genomes in the database from each other.

Some implications are that it should be possible to use information from both short and long reads to classify robustly to the species level, but it is unlikely to work below that (at the strain level). This is because many reads will map to shared content within a species, and some of that shared content may not distinguish a particular genome from others in the pangenome at the resolution of the available reads.

A simple thought experiment also suggests that reduced-representation /slimmed-down databases will not support strain-level classification. Suppose that a technique exists that can classify reads to a strain level. First, choose a read that belongs to two or more different strains; there is no way to identify which strain this belongs to. Second, choose a read that belongs to a strain that is not represented in the database; while it clearly belongs to a known species, there is no way to identify which. Classifiers should be using all available information and it is clearly possible to do so, viz sourmash.

(Probably need to spend some time here talking about core vs accessory genomes.)

Approaches such as sourmash can try to operate “above” individual k-mers, but will also be stymied by infinite unicity. Here combinatorial uses of k-mers (via e.g. containment) may be able to resolve strains, but will need to do so at higher resolution than sourmash’s current parameters. Here approaches such as Agamemnon may be useful.

A method to evaluate, compare, and study taxonomic lables

The GTDB and NCBI taxonomies are not entirely consonant and our studies using entropy suggest that a substantial portion of the NCBI taxonomy is confused. Comparing Table 1 and Table 2, we see that 5x as many k-mers belong to genomes that do not share the same family-level labels. This difference is due solely to the taxonomic labels. It is not necessarily surprising that NCBI is so different

since GTDB is directly constructed using content-based phylogeny, but it does suggest that there are many places where the NCBI taxonomy should be examined closely.

(drill down; contamination, etc.)

Shannon entropy and unicity on k-mers are robust ways to study, evaluate, and summarize large databases

Exploration of these results suggest that k-mer size and scaled do not dramatically affect our conclusions. (Confirm me, please :).

Conclusion

This is easy mode: this kind of taxonomic classification is “just” a database lookup. What messes up taxonomic classification with k-mers is (1) biology (redundancy and laterally transferred genetic elements) and (2) humans (taxonomy). (1) is resolvable to a significant extent with combinatorics. (2) can be tackled with better metrics and systematic improvement. Here we provide measures that assist with both.

Despite this, biological questions remain that are out of scope of this paper: correctness and completeness of reference databases matters. And we really also say nothing about generalizability, where we know that we have problems.

Methods

Methods go here.

References