K-mer counts and phylogeny with kmerdb

Matthew Ralston[1,2☯*]

1 University of Delaware, Center for Bioinformatics and Computational Biology, Newark, Delaware, United States of America,
2 Delaware Biotechnology Institute, Newark, Delaware, United States of America,

☯These authors contributed equally to this work.

¤Current Address: Dept/Program/Center, Institution Name, City, State, Country

†Deceased

¶Membership list can be found in the Acknowledgments sections

* Correspondences to ‘mralston.development@gmail.com’

Abstract

Background: With decades since Altschul’s BLAST, it is still important to reflect on the seed region, its algorithmic purpose, and the role of the k-mer in biological sequence relationships and even alignment-free methods. Specifically, k-mers are used to generate alignments, assemblies, and alignment-free sequence inferences across massive amounts of sequencing data being produced at an increasing rate. k-mers are the backbone of sequence assembly by virtue of the De Bruijn graph and again crucial for performance in sequence alignment heuristics like the seed region of BLAST. **Methodology:** kmerdb is a toolkit in C-Python 3.12.4+ that compresses conceivably large k-mer count profiles using a block-access model and the bgzf compression standard distributed with BioPython[1]. This program has additional characteristics such as an indexing functions, count normalization, PCA and t-Stochastic Neighbor Embedding (t-SNE) dimensionality reduction techniques, distance matrices, and k-means and hierarchical clustering features, and a log-odds ratio test for unknown sequences on the available k-mer frequencies. The variety of tools included in kmerdb facilitates any tetra-mer or arbitrary k-mer profile-based analysis, including word frequency analysis, custom seed regions, and assessment of k-mer uniqueness via NumPy compatibility. **Results::** kmerdb is a modern and ergonomic toolkit for k-mers and alignment-free sequence analysis pipelines, which rely on kmer indexing, counting nullomers, unique-kmers, and total-kmer counts on a per-input basis, and creating useful graphics and metrics on inter-species, inter-tissue, or locally disparate k-mer frequencies, and profile similarities/distances. Distributed on the Python Package Index and maintained on GitHub: https://github.com/MatthewRalston/kmerdb Features on the horizon include an aligner similar in strategy to vsearch, assembly and assembly diagnostic information, and a pseudoaligner. Ambitiously, kmerdb seeks to bridge the gap behind Jellyfish (C/C++, 2011 [2]) and kPAL (Python, 2014 [3]) for providing programmatic access to kmer count profiles, count matrices, matrix normalizations and transformations, clustering facilities, and Markov model utilities in the Python programming language.

Author summary

K-mer counts, word-frequencies and alignment-free sequence analysis play crucial roles in modern Bioinformatics toolkits, pipelines, and algorithm design. Access to k-mer profiles and sequence compositional data in Python relies on intricate interfaces to C/C++ functions with Jellyfish [2], or a collection of scripts with kPAL [3]. Here I present a modern Python module and CLI suite for k-mer profiling, alignment-free similarities, clustering of datasets by k-mer frequencies, De Bruijn graphs, and more.

First time with Quarto.

Notes: I

NOTE: [Autogenerated] Quartotemplate for PLOS Comp Biol

This template shows how to use PLOS template from https://plos.org/resources/writing-center/. Each journal has a submission guideline page; please refer to it.

- PLOS Biology
- PLOS Climate
- PLOS Digital Health
- PLOS Computational Biology
- PLOS Genetics
- PLOS Global Public Health
- PLOS Medicine
- PLOS Neglected Tropical Diseases
- PLOS ONE
- PLOS Pathogens
- PLOS Sustainability and Transformation
- PLOS Water

This template file contains some guidelines and recommandation initially given in `plos_latex_template.tex` that can be found in https://github.com/quarto-journals/plos/blob/main/style-guide/plos_latex_template.tex

Metadata

About journal id field

This is an identifier for the target journal. It can be derived from https://plos.org/resources/writing-center/ following submission guidelines link, the identifier is the part of the URL after `https://journals.plos.org/<id>/s/submission-guidelines`

Journal	id
PLOS Biology	plosbiology
PLOS Climate	climate
PLOS Digital Health	digitalhealth
PLOS Computational Biology	ploscompbiol
PLOS Genetics	plosgenetics

Journal	id
PLOS Global Public Health	globalpublichealth
PLOS Medicine	plosmedicine
PLOS Neglected Tropical Diseases	plosntds
PLOS ONE	plosone
PLOS Pathogens	plospathogens
PLOS Sustainability and Transformation	sustainabilitytransformation
PLOS Water	water

Example :

```yaml
format:
  plos-pdf:
    journal:
      id: water
```

Once your paper is accepted for publication,

Do not include track change in LaTeX file and leave only the final text of your manuscript. PLOS recommends the use of latexdiff to track changes during review, as this will help to maintain a clean tex file. Visit https://www.ctan.org/pkg/latexdiff?lang=en for info or contact us at latex@plos.org.

This should not be a problem using Quarto but still a recommandation from the journal

There are no restrictions on package use within the LaTeX files except that no packages listed in the template may be deleted.

Please do not include colors or graphics in the text. Color can be used to apply background shading to table cells only.

The manuscript LaTeX source should be contained within a single file (do not use`\input`, `\externaldocument`, or similar commands).

Please contact latex@plos.org with any questions submission guidelines. For anything Quarto related, please open an issue in https://github.com/quarto-journals/plos. If this is related to the LaTeX template, this could also be a good idea to contact PLOS directly.

Figures and Tables

Please include tables/figure captions directly after the paragraph where they are first cited in the text.

Figures

However, do not include graphics in your manuscript

- Figures should be uploaded separately from your manuscript file.
- Figures generated using LaTeX should be extracted and removed from the PDF before submission.
- Figures containing multiple panels/subfigures must be combined into one image file before submission.

This means that, depending on how you create your figure, a manual post processing will be required.

For figure citations, please use “Fig” instead of “Figure”. This has been made the default in this Quarto format:

```yaml
crossref:
  fig-title: Fig
```

Also, place figure captions after the first paragraph in which they are cited.

See PLOS figure guidelines at https://journals.plos.org/plosone/s/figures and in your specific journal guideline.

Tables

Tables should be cell-based and may not contain:

- spacing/line breaks within cells to alter layout or alignment
- do not nest tabular environments (no tabular environments within tabular environments)
- no graphics or colored text (cell background color/shading OK)

See PLOS table guidelines at http://journals.plos.org/plosone/s/tables and in your specific journal guideline.

For tables that exceed the width of the text column, use the adjustwidth environment as illustrated in the example table in text below. If you are in this case, you’ll either need to manually post process the `.tex` file and recreate the PDF, or you need to include LaTeX tables directly.

Also, place tables after the first paragraph in which they are cited.

Equations, math symbols, subscripts, and superscripts

Below are a few tips to help format your equations and other special characters according to our specifications. For more tips to help reduce the possibility of formatting errors during conversion, please see our LaTeX guidelines at http://journals.plos.org/plosone/s/latex

- For inline equations, please be sure to include all portions of an equation in the math environment. For example, `x$^2$` is incorrect; this should be formatted as x^2 (or x^2 if the romanized font is desired).

- Do not include text that is not math in the math environment. For example, `CO2` should be written as `CO\textsubscript{2}` giving CO_2 instead of `CO$_2$`.

- Please add line breaks to long display equations when possible in order to fit size of the column.

- For inline equations, please do not include punctuation (commas, etc) within the math environment unless this is part of the equation.

- When adding superscript or subscripts outside of brackets/braces, please group using `{}`. For example, change `"[U(D,E,\gamma)]^2"` to `"{[U(D,E,\gamma)]}^2"`.

- Do not use `\cal` for caligraphic font. Instead, use `\mathcal{}`

Title and headings

Please use “sentence case” for title and headings (capitalize only the first word in a title (or heading), the first word in a subtitle (or subheading), and any proper nouns).

PLOS does not support heading levels beyond the 3rd, meaning no 4th level headings. Header 4 levels `####` is used for the *Supporting information* section

Abstract and author summary

Abstract must be kept below 300 words.

Author Summary must be kept between 150 and 200 words and first person must be used.

For PLOS ONE, author summary won’t be included as it is not valid for submission.

Supplementary information syntax

Use this markdown syntax to create the supplementary information block with a custom block of class `.supp`

```
::: {.supp}
## SI TYPE {#id}

First paragraph is a title sentence that will be bold. (required)

Optionnaly, add descriptive text after the title of the
item. No third paragraph is allowed
:::
```

They need to be referenced in text using `nameref` by using this syntax `[id](.nameref)` where `ìd` will be the id used on the header.

References

Within Quarto, `natbib` will be used with `plos2015.bst`, which expect numeric style citation. Use brackets for references, e.g `[@ref]`.

Quarto features limitation

Some features are not working with this format in PDF:

- Callouts
- Code highlighting customization (border left, background color)

Following content of this document is from the LaTeX template content to demo journal style.

Introduction

Lorem ipsum dolor sit [?] amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam Eq 1 sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. [**?**] Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

$$\mathrm{P_Y} = \underbrace{H(Y_n) - H(Y_n|\mathbf{V}_n^{Y})}_{S_Y} + \underbrace{H(Y_n|\mathbf{V}_n^{Y}) - H(Y_n|\mathbf{V}_n^{X,Y})}_{T_{X\rightarrow Y}} \tag{1}$$

Materials and methods

Etiam eget sapien nibh

Nulla mi mi, Fig 1 venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Figure 1. Bold the figure title. Figure caption text here, please use this space for the figure panel descriptions instead of using subfigure commands. A: Lorem ipsum dolor sit amet. B: Consectetur adipiscing elit.

Results

Results and Discussion can be combined.

Nulla mi mi, venenatis sed ipsum varius, Table **??** volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, S1 Video vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

% Place tables after the first paragraph in which they are cited.

Table 2. Table caption Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam.

Heading1				**Heading2**			
$cell1row1$	cell2 row 1	cell3 row 1	cell4 row 1	cell5 row 1	cell6 row 1	cell7 row 1	cell8 row 1
$cell1row2$	cell2 row 2	cell3 row 2	cell4 row 2	cell5 row 2	cell6 row 2	cell7 row 2	cell8 row 2
$cell1row3$	cell2 row 3	cell3 row 3	cell4 row 3	cell5 row 3	cell6 row 3	cell7 row 3	cell8 row 3

Table notes Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl. Ut ornare mauris tellus, vel dapibus arcu suscipit sed.

““

LOREM and IPSUM nunc blandit a tortor

3rd level heading

Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit. Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat.

1. react
2. diffuse free particles
3. increment time by dt and go to 1

Sed ac quam id nisi malesuada congue

Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

- First bulleted item.
- Second bulleted item.
- Third bulleted item.

Discussion

Nulla mi mi, venenatis sed ipsum varius,see Table **??** volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero [**?**].

Conclusion

CO_2 Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit.

Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat. Ut neque ipsum, luctus id lacus ut, laoreet scelerisque urna. Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl. Ut ornare mauris tellus, vel dapibus arcu suscipit sed. Nam condimentum sem eget mollis euismod. Nullam dui urna, gravida venenatis dui et, tincidunt sodales ex. Nunc est dui, sodales sed mauris nec, auctor sagittis leo. Aliquam tincidunt, ex in facilisis elementum, libero lectus luctus est, non vulputate nisl augue at dolor. For more information, see S1 Appendix.

Supporting information

S1 Fig. Bold the title sentence. Add descriptive text after the title of the item (optional).

S2 Fig. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 File. Lorem ipsum.

S1 Video. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Appendix. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Table. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

Acknowledgments

Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae.

References

1. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–1423.

2. Marcais G, Kingsford C. Jellyfish: A fast k-mer counter. Tutorialis e Manuais. 2012;1(1-8):1038.

3. Anvar SY, Khachatryan L, Vermaat M, van Galen M, Pulyakhina I, Ariyurek Y, et al. Determining the quality and complexity of next-generation sequencing data without a reference genome. Genome biology. 2014;15(12):555.