The Journal of Open Source Software

Ziplign: a simple-to-use interactive tool to compare bacterial genomes

Martin Hunt [1,2,3,4¶] **and Zamin Iqbal** [5]

1 European Molecular Biology Laboratory - European Bioinformatics Institute, Hinxton, UK **2** Nuffield Department of Medicine, University of Oxford, Oxford, UK **3** National Institute of Health Research Oxford Biomedical Research Centre, John Radcliffe Hospital, Headley Way, Oxford, UK **4** Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, University of Oxford, Oxford, UK **5** Milner Centre for Evolution, University of Bath, UK ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Abhishek Tiwari

Reviewers:

- @mjsull
- @ammaraziz

Submitted: 01 April 2025
Published: unpublished

Summary

Ziplign is a user-friendly interactive application to visually compare two bacterial genome sequences and their annotation. It requires no command-line use, and is intended to make genome-comparison easily accessible to the biologist. Genome files can be directly drag-and-dropped into Ziplign, or will be automatically downloaded when an accession is provided. All commonly-used file formats and compression are supported. The comparison between genomes is generated using NCBI blast+(Camacho et al., 2009), which is run for the user, and then the two genomes, their annotation, and sequence matches are displayed by Ziplign. A screenshot of Ziplign is shown in Figure 1.

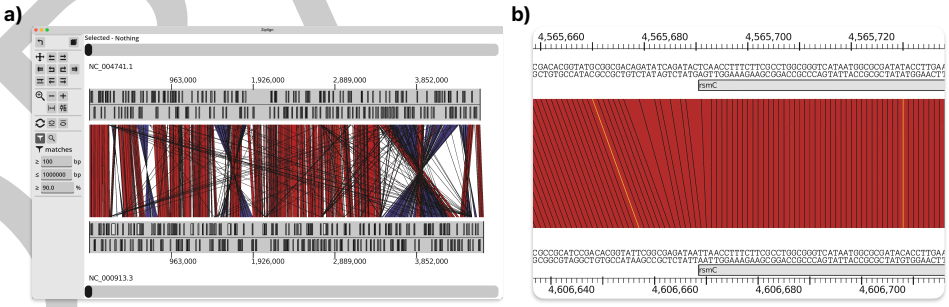

Figure 1: Figure 1 Screenshots of Ziplign comparing the Shigella flexneri 2a genome GCF_000007405.1 (Wei et al., 2003) - shown at the top - with the Escherichia coli K-12 substr. MG1655 GCF_000005845.2 (Riley, 2006) - shown at the bottom. BLAST matches are shown in red when the direction of the match is the same in both genomes, and in blue when they are in opposite directions. To reduce noise in the screenshot, only matches of at least 2000bp and 95% identity are shown (configurable by the user via the panel on the left). Annotation features on the forward/reverse strand are shown in the top/bottom of each contig. a) Default view, showing the complete genomes and their overall structural similarities. b) Zoomed to the base-pair level, matching nucleotides marked with black lines, SNPs are in orange, and non-parallel black lines denote indels in the alignment.

Statement of need

Comparing two bacterial genome sequences is a fundamental task in genomics, used in numerous scenarios: comparing closely related strains to discern differences such as the overall structure

and any rearrangements, the presence or absence of important features such as virulence factors or anti-microbial genes, or to identify horizontal gene transfer. Genome assemblies can be compared to each other or against a reference genome for debugging or determining the most accurate assembly. Whilst many command line tools are available for processing samples at scale and report statistics, it is invaluable to visually and interactively compare two genomes. This is often the simplest way to truly understand the differences between two sequences.

To our knowledge ACT(T. J. Carver et al., 2005) and Mauve(Darling et al., 2004) are the only existing tools for displaying genomes and matches between them in an interactive manner - however both tools are no longer supported. Since ACT is based on Artemis(T. Carver et al., 2012), it incorporates the extensive feature set implemented in Artemis. However, ACT has a number of limitations. It can be difficult for non-technical users to install and use, Java must be installed, the user must provide (most likely via running command line tools) a genome comparison file, multi-sequence genomes are not supported out of the box, and alignment details including SNPs and small insertions/deletions are not shown. Mauve is simple to run but it only shows unique synteny/collinear blocks between genomes, meaning that repeats may not be shown. We tested this using a single contig 1000bp genome compared to a genome comprising two copies of the same contig, where Mauve showed no matches (also tested again using a 10,000bp contig).

Here we introduce Ziplign, which fills the need for an easy-to-install and simple-to-use genome comparison tool. It is heavily inspired by ACT, with a very similar user interface, but is significantly easier to install and use.

Usage and availability

Ziplign is intended for microbiologists with no command line experience. As such, no use of a terminal is required. First, two genomes must be provided, either with an NCBI accession or by dragging-and-dropping local files. FASTA, FASTQ, GFF3, EMBL and GenBank file formats are supported, optionally with gzip, bzip2 or xz compression. Genome sequences and annotation are automatically downloaded when an accession is used. Ziplign runs blastn from the NCBI blast+ suite to generate the matches between the two genomes. The blastn options are configurable by the user.

Genomes are displayed at the top and bottom of the window, with BLAST matches shown between them (Figure 1). The view can be zoomed and panned using mouse, trackpad or keyboard controls, or with buttons in the control panel on the left. Features include searching by nucleotide sequence or annotation, contig reordering, and reverse complementing contigs. Ziplign can save and load an entire "project" - the genomes, annotations, and BLAST matches - using a single binary file, removing the need to store the original files.

Ziplign is available for Windows 11, macOS, and Linux operating systems from GitHub https://github.com/martinghunt/ziplign, under the MIT license. Comprehensive documentation is hosted on ReadTheDocs https://ziplign.readthedocs.io/en/.

Implementation

Ziplign is primarily written in GDScript, the scripting language of the free, open source, MIT licensed, game engine Godot (https://godotengine.org, https://github.com/godotengine/godot). This handles the graphical user interface (GUI), and displaying and interacting with all the genome and comparison data. Bioinformatics tasks such as parsing sequence/BLAST files and downloading genomes are processed using a separate command line program called zlhelper https://github.com/martinghunt/zlhelper, also with the MIT license, written in the Go programming language. All command line programs are hidden from the user, so that the only interaction is simply with the GUI.

Acknowledgements

The authors thank Daniel Anderson, Jane Hawkey and Leah Roberts for testing Ziplign, and Thomas Hunt for making the Ziplign icon. Martin Hunt was supported by the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with the UK Health Security Agency (UKHSA) (NIHR200915), and supported by the NIHR Biomedical Research Centre, Oxford. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health or the UKHSA.

References

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M.-A., Barrell, B. G., & Parkhill, J. (2005). ACT: The Artemis comparison tool. *Bioinformatics*, *21*(16), 3422–3423. https://doi.org/10.1093/bioinformatics/bti553

Carver, T., Harris, S. R., Berriman, M., Parkhill, J., & McQuillan, J. A. (2012). Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, *28*(4), 464–469. https://doi.org/10.1093/bioinformatics/btr703

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research*, *14*(7), 1394–1403. https://doi.org/10.1101/gr.2289704

Riley, M. (2006). Escherichia coli K-12: A cooperatively developed annotation snapshot–2005. *Nucleic Acids Research*, *34*(1), 1–9. https://doi.org/10.1093/nar/gkj405

Wei, J., Goldberg, M. B., Burland, V., Venkatesan, M. M., Deng, W., Fournier, G., Mayhew, G. F., Plunkett, G., Rose, D. J., Darling, A., Mau, B., Perna, N. T., Payne, S. M., Runyen-Janecky, L. J., Zhou, S., Schwartz, D. C., & Blattner, F. R. (2003). Complete Genome Sequence and Comparative Genomics of *Shigella flexneri* Serotype 2a Strain 2457T. *Infection and Immunity*, *71*(5), 2775–2786. https://doi.org/10.1128/IAI.71.5.2775-2786.2003