Biostrings ISC Proposal 2024

Aidan Lakshman

2024-03-11

Signatories

Project team

Aidan Lakshman - PhD Candidate, University of Pittsburgh. See `requirements.Rmd` for more detail.

Contributors

Erik S. Wright - Associate Professor, University of Pittsburgh Hervé Pagès - Bioconductor Core Team, Biostrings Maintainer

Consulted

The Problem

Biostrings is an essential package in the R ecosystem, with over 1 million installations from Bioconductor per year (Fig. 1). A total of 226 Bioconductor packages depend on Biostrings, 254 import Biostrings, and 43 suggest Biostrings. Despite this success, Biostrings has been unsupported financially for over a decade. Package maintenance currently relies upon a small patchwork of volunteer maintainers who struggle to keep pace with basic maintenance. This has resulted in some longstanding bugs and insufficient support to implement planned enhancements. While many issues have been outlined within the Biostrings package, there have been few developers willing to learn the internal code structure of Biostrings to be able to take over maintenance.

As an active Biostrings user and contributor, I have discussed this issue extensively with the package maintainer, Hervé Pagès (see attached Letter of Support). In collaboration, we have drawn a roadmap to sustainable long-term maintenance of the Biostrings package. Here, we present a path forward and request funding support for the labor required to implement it. In this vision, I will take over primary maintenance of the Biostrings package and implement a set of fixes and extensions that put Biostrings on a path toward sustained success.

The proposal

Overview

Biostrings is a core Bioconductor package providing efficient containers for storing, manipulating, and analyzing biological sequences. Biostrings enables access to XString objects that are an essential part of the Bioconductor ecosystem. Presently, Biostrings maintenance is hindered by (i) a list of open issues on GitHub, (ii) a to do list of items developed by package contributors, and (iii) input/output functionality that is becoming outdated with newer technology. In this proposed project, I will address open issues, harden the Biostrings package for long-term sustainability, and add features that will keep Biostrings relevant for modern sequencing technologies. For end-users, this will result in numerous bugfixes, a host of new features

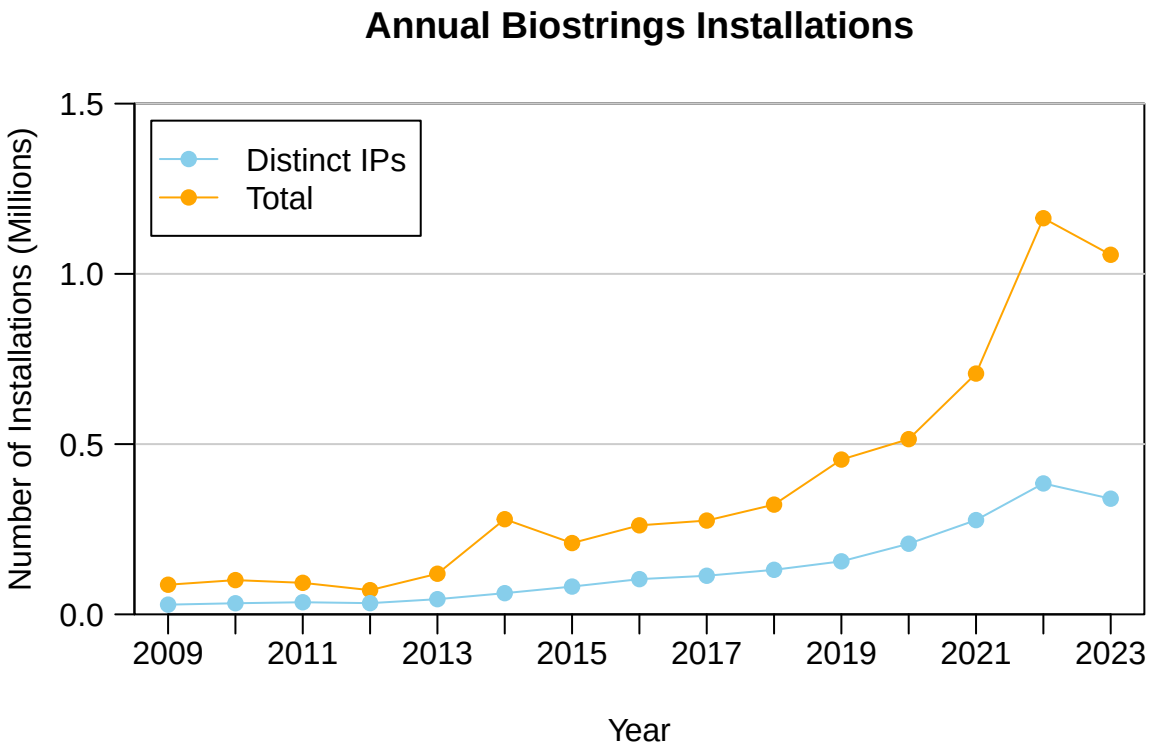

Figure 1: Yearly Biostrings Installations by total and unique IP address.

to support genomic analyses, and a variety of performance improvements to bolster R as one of the top programming languages for bioinformatics.

In summary, this proposal details a new era for Biostrings. The project will transition maintenance to a new developer, and in the process, ensure the package is robust and maintainable for years to come.

Detail

There are three categories of changes needed to sustain the utility of Biostrings in the long-term. First, outstanding user-contributed GitHub issues must be addressed. Most of these are straightforward but require effort to complete. I anticipate up to 4 months for this first Aim. Second, current and past Biostrings developers have kept a list of to do items for the package. Prior to this proposal submission, I worked with Hervé Pagès to narrow this list to the items that are the most important items for Biostrings' long-term success. I anticipate up to 6 months will be required to complete all items on this list. Third, modern DNA sequencing technologies have advanced markedly since Biostrings was first introduced. Changes to import and export of sequences are needed to sustain the project long-term. I expect these changes will take 2 months. These three categories of changes are described in the three Aims below.

Aim 1: Package cleanup and hardening

The goal of this aim is the prepare the Biostrings codebase for future improvements and facilitate ongoing maintenance. Biostrings has had minimal maintenance for the past 5 years, and as a result lacks many processes that ease continued development. The goals of the Aim are the following:

1. Clean up outdated documentation and bug reports. Many internal documentation files and scripts have `TODO` tags or warnings that are out of date. Additionally, many bug reports on GitHub have been resolved but remain open. This makes ongoing maintenance challenging, as it requires additional developer effort to determine if a bug still exists before addressing it. This task will clean up the codebase, update outdated documentation, and clean up resolved bug reports.
2. Implement unit testing. Much of this project depends on improving and adding additional functionality to Biostrings. However, as mentioned in the Problem Definition section, Biostrings is a critical package with over a million downloads per year. As such, robust unit testing is essential to ensure that other proposed changes do not break existing functionality for end users. Robust testing suites would also make it easier for maintainers to review community-contributed code, thus making it easier to involve the community in Biostrings.
3. Resolving outstanding bug reports. While many bug reports on GitHub are resolved, there are many that remain unaddressed. This task will resolve any remaining bug reports. Part of this task will involve reaching out to bug reporters for additional detail, as some bug reports are years old without any updates (see here or here examples of very old bug reports).

Successful completion of this Aim will result in a cleaner Biostrings GitHub repository, resolution of outstanding user-submitted bugs, and a robust testing pipeline for future submissions. I expect this Aim to take three months; most of the GitHub issues are relatively quick to fix, so the majority of the time will be dedicated to building a robust testing infrastructure.

Aim 2: Completing critical items on the Biostrings developers' TODO list

This will focus on items related to XStringSets in the TODO file. More detail to come.

Aim 3: Enhancing input/output for modern sequencing technologies

Biostrings was initially developed during a time when sequencing produced megabases (~1M nucleotides) of data per run. Modern sequencing technologies easily produce gigabases (> 1B nucleotides) per run. Hence, Biostrings' input/output needs improvement to scale alongside next generation sequencing technologies. At present, Biostrings can only read and write sequences from gzip compressed files. There is an `open_input_files` function that allows reading sequences in batches, but it does not use the standard R connections interface and is cumbersomely slow on large compressed files.

To enhance Biostrings, I will add functionality for reading from standard gzip, bzip2, and xz connections in R. This will involve overhauling the `readXStringSet` functions within Biostrings. Furthermore, I will enable writing to alternative output file compression types (bzip2 and xz), while allowing for different compression levels. At present there is a `compression_level` argument in `writeXStringSet`, but it is unused by the function. I will focus on improving the speed of reading and writing from files so that large file sizes are no longer problematically slow. Collectively, these enhancements will propel Biostrings (and by extension, the R programming language itself) into the future of big biological data.

Project plan

I plan to complete the proposed work in a one-year time period. As funding commences June 1, this means the duration of the proposal is June 1, 2024 - June 1, 2025. A summary of the timeline of aims is included below, and a detailed description follows.

Since the package is hosted on Bioconductor, large milestones are already clearly defined. Bioconductor releases new versions in October and April, which will each act as milestones for delivery of the first two Aims. This grant will conclude shortly before useR! and the annual Bioconductor conference (typically held in July), allowing me an opportunity to highlight the work done and the ISC program at the end of my award.

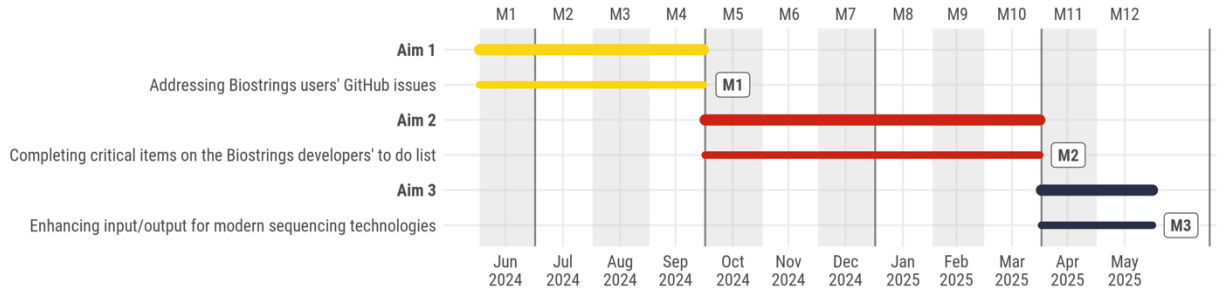

Figure 2: Figure 2. Timeline of aims and milestones(M) throughout the project.

Start-up phase

This project has a short start-up phase. I have already coordinated with Hervé Pagès to identify critical tasks for each of the Aims (see "Proposal"). The codebase and dissemination method have already been created (GitHub and Bioconductor, respectively), and I have acquired write access to the codebase from Hervé Pagès. Additionally, I have experience contributing to Biostrings in the past, so I am relatively familiar with the codebase and contribution pipeline. As a result, I would be ready to begin work on Aim 1 from the moment the grant is awarded.

Technical delivery

More detail to come.

Aim 1: Finished by October 2024, in time for Bioconductor release 3.20.

Aim 2: Finished by April 2025, in time for Bioconductor release 3.21.

Aim 3: Finished by June 2025 and merged into Bioconductor version 3.22 (development). Present on improvements at conferences, finalize changes in time for Bioconductor release 3.22 in October 2025.

Other aspects

This project will be highlighted through a variety of methods.

Changes made will be released via Bioconductor's semiannual version releases. I will accompany these updates with social media outreach and blog posts to my website (pending approval to host at R-bloggers.com).

As mentioned previously, this proposal will conclude in June 2025, shortly before the annual useR! and Bioconductor conferences. I will apply to present the work done in this proposal at both of these conferences. This will also provide a good opporunity to highlight the ISC's contribution to this project.

Finally, I plan to submit a paper on the updated Biostrings to the R Journal. As Biostrings is a critical package in the R ecosystem, an updated paper would be of interested to the readership of the R Journal. I plan to write this paper in conjunction with Erik Wright and Hervé Pagès, and submit in mid-2025.

Requirements

The primary requirement for making this project happen is a developer with the time and willingness to maintain Biostrings, and the technical ability to make that happen. A secondary requirement is funding to support the developer during their work on this package.

People

I (Aidan Lakshman) will be the primary developer on this project. I am a PhD student in the Department of Biomedical Informatics at the University of Pittsburgh. In my work, I am a developer of the SynExtend package that is dependent upon Biostrings. I have already made several major contributions to R (e.g., dendrapply, wilcox) and Biostrings (e.g., AAStrings), and I developed a strong working relationship with the current Biostrings maintainer, Hervé Pagès, in the process. If funded, I will conduct this project during the last year of my PhD, and then the package will be in a state where I can continue to support it longer-term in conjunction with others in the Bioconductor community.

Auxiliary supporters of this proposal are my PhD advisor, Erik Wright, and current Biostrings maintainer Hervé Pagès. Erik Wright is a past contributor to Biostrings, and is supportive of me committing 20% of my work hours to this project. Both Erik Wright and Hervé Pagès will provide advising throughout the project to ensure contributed code is high quality and to prepare me to become a long-term maintainer of Biostrings. Their primary contribution will take the form of code reviews and suggestions for additional improvements.

I have already met with both Erik Wright and Hervé Pagès to develop the structure of this proposal–letters of support from both are included.

Processes

Tools & Tech

No specific tools or technology are needed to deliver this project. The Biostrings codebase is hosted on GitHub and is released regularly on Bioconductor. All software development can be done on my personal computer.

Funding

I respectfully request $8,000 to cover my labor cost associated with the 20% effort for 1 year needed to complete this project. This amount is based on my current $40,000 per year stipend as a graduate student at the University of Pittsburgh.

Summary

Success

Definition of done

Measuring success

Future work

Key risks