RSeqAn: Headers and wrappers for the SeqAn library in R

August Guang

5 February 2019

Summary

R is used heavily in the bioinformatics community for processing, analyzing, and visualizing biological sequence data due to the strong support for data exploration and analysis. One common complaint about R is its lack of speed relative to other languages, which have to do with properties of the R kernel (Sridharan 2015). A solution to this is to rewrite key functions in C++ instead, then use Rcpp (Eddelbuettel and Balamuta 2018) to interface with R. Many packages in R are sped up with C++ code: as of November 2018, out of 13525 packages on CRAN, 1493 of those use Rcpp. On Bioconductor (Huber et al. 2015), a repository specifically for bioinformatics packages, there is a similar proportion of packages using Rcpp: 150 out of 1649.

Writing certain functions in C++ also allows the researcher to take advantage of high-performance C++ libraries. One such library that has proven powerful and popular for bioinformatics applications is SeqAn (Döring et al. 2008; Reinert et al. 2017). SeqAn is an open source C++ library of efficient algorithms and data structures for the analysis of sequences with a focus on biological data. It has been used for many popular bioinformatics tools, including Bowtie2 (Langmead and Salzberg 2012) and Tophat (Trapnell, Pachter, and Salzberg 2009). Its capabilities include efficient storage and I/O of sequence data, fast algorithms for pattern matching, and much more.

To date, no R package has taken advantage of SeqAn. This is due to two issues. First, R for Windows is built with the mingw compiler, which the SeqAn development team does not offer support for. Second, SeqAn implements several custom data types in order to make the library more efficient. This means that any data type in SeqAn needs to have have R to C++ conversion and C++ to R conversion functions written for it before it can be used in R.

`RSeqAn` solves both of these problems, thus allowing any R researcher to use the capabilities of SeqAn in their own work without need for the user to install SeqAn themselves or deal with interfacing between R and C++. `RSeqAn` can

be installed from Bioconductor, and documentation for it can be found in the package as well as online.

Benchmarked Example

As a proof of concept for its utility, we have benchmarked a function (searching for adapter contamination) from `qckitfastq`, a package that uses RSeqAn for quality control on bioinformatics data compared to `ShortRead` (Morgan et al. 2009), another package on Bioconductor serving the same function that is purely written in R (Table 1). As can be seen, computing adapter content through `qckitfastq` is much faster than `ShortRead`.

Table 1: Benchmark results of qckitfastq, which uses RSeqAn, and ShortRead, which does not against a test file from each package.

File	Package	replications	elapsed	user.self	sys.self
E-MTAB-1147	ShortRead	100	13.3	11.58	1.37
	qckitfastq	100	5.159	4.902	0.091
test.fq.gz	ShortRead	100	6.612	5.739	0.567
	qckitfastq	100	0.195	0.105	0.019

Acknowledgments

Thanks to Ashok Ragavendran for the concept of using SeqAn in R packages to speed up functions. Thanks to Fernando Gelin for help with the documentation and style. This publication was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM109035.

References

Döring, Andreas, David Weese, Tobias Rausch, and Knut Reinert. 2008. "SeqAn an efficient, generic C++ library for sequence analysis." *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-9-11.

Eddelbuettel, Dirk, and James Joseph Balamuta. 2018. "Extending R with C++: A Brief Introduction to Rcpp." *American Statistician*. https://doi.org/10.1080/00031305.2017.1375990.

Huber, Wolfgang, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S. Carvalho, Hector Corrada Bravo, et al. 2015. "Orchestrating high-throughput genomic analysis with Bioconductor." *Nature Methods*. https://doi.org/10.1038/nmeth.3252.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast gapped-read alignment with Bowtie 2." *Nature Methods*. https://doi.org/10.1038/nmeth.1923.

Morgan, Martin, Simon Anders, Michael Lawrence, Patrick Aboyoun, Hervé Pagès, and Robert Gentleman. 2009. "ShortRead: A Bioconductor Package for Input, Quality Assessment and Exploration of High-Throughput Sequence Data." *Bioinformatics* 25:2607–8. https://doi.org/10.1093/bioinformatics/btp450.

Reinert, Knut, Temesgen Hailemariam Dadi, Marcel Ehrhardt, Hannes Hauswedell, Svenja Mehringer, René Rahn, Jongkyu Kim, et al. 2017. "The SeqAn C++ template library for efficient sequence analysis: A resource for programmers." *Journal of Biotechnology*. https://doi.org/10.1016/j.jbiotec.2017.07.017.

Sridharan, Shriram. 2015. "Profiling R on a Contemporary Processor." *Proceedings of the VLDB Endowment*. https://doi.org/10.14778/2735471.2735478.

Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. 2009. "TopHat: Discovering splice junctions with RNA-Seq." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp120.