RSeqAn: Headers and wrappers for the SeqAn library in R

August Guang

7 December 2018

Summary

`RSeqAn` provides R with access to SeqAn (Döring et al. 2008, @Reinert2017) header files. 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 2013) and Tophat (Trapnell, Pachter, and Salzberg 2009). 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 (Eddelbuettel and Balamuta 2018). 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. However, to date these packages have not utilized SeqAn.

An R package that does aim to use SeqAn runs into 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 data types in order to make the library more efficient. Any data types in SeqAn must thus be extended with Rcpp through the templated functions `Rcpp::as` and `Rcpp::wrap`.

In `RSeqAn` v1.0 we provide access to SeqAn header files and add support for mingw so that it can be used in R for Windows as well. The development version also includes some wrappers to convert between R objects and SeqAn data types with a goal of eventually providing all necessary wrappers for easy interface with SeqAn. `RSeqAn` thus provides an easy way to write R bioinformatics packages that are much more efficient without need for the user to install SeqAn themselves. `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.

Package	replications	elapsed	relative	user.self	sys.self	user.child	sys.child
system.file(package="ShortRead", "extdata", "E-MTAB-1147")							
ShortRead	100	13.268	1	11.803	1.197	0	0
qckitfastq	100	4.870	1	4.661	0.083	0	0
system.file(package="qckitfastq", "extdata", "test.fq.gz")							
ShortRead	100	6.372	1	5.656	0.481	0	0
qckitfastq	100	0.186	1	0.099	0.019	0	0

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, W., V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, et al. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor." *Nature Methods* 12 (2):115–21. http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html.

Langmead. 2013. "Bowtie2." *Nature Methods*. https://doi.org/10.1038/nmeth.1923.Fast.

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.

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