plyranges: a grammar for manipulating genomics data

Stuart Lee [1] , Michael Lawrence [2] , Di Cook [1]

1 Department of Econometrics and Business Statistics, Clayton, Victoria, Australia
2 Bioinformatics and Computational Biology, Genentech, Inc., South San Francisco, California, United States of America

Abstract

The Bioconductor project has created many powerful abstractions for reasoning about genomics data, such as the *Ranges* data structures for representing genomic intervals. By recognising that these data structures follow 'tidy' data principles we have created a grammar of genomic data manipulation that defines verbs for performing actions on and between genomic interval data. This grammar simplifies performing common genomic data analysis tasks via method chaining, type consistency and results in creating human readable pipelines. We have implemented this grammar as an Bioconductor/R package called plyranges.

Introduction

Genomic data may be naturally represented as sets of pairs of integers corresponding to the start and end points of sequences. Further information such as strandedness and chromosome name may be added to these sets to provide biological context. Because of the flexibility of this representation supplemental information such as measurements obtained from an experimental assay or annotations from a genome database can be joined to their relevant sequences. In the Bioconductor/R packages `IRanges` and `GenomicRanges` these representations have been implemented as a suite of data structures called *Ranges* [1]. These data structures cover many common data types encountered in bioinformatics analyses - a gene can be represented with its coordinates, along with its identifier and the identifiers of its exons; or an RNA-seq experiment may be represented as sets of genes with a matching count column.

The Bioconductor infrastructure for computing with genomic ranges are highly effecient and powerful, however the application programming interface (API) for performing analysis tasks with *Ranges* is complex due to its large number of methods and classes. It also makes resulting scripts written difficult for a non-programmer to read and reason about. An alternative approach would be to implement a domain specific language (DSL) as a fluent interface built on top *Ranges*. The goal of fluent interface is to enable users to write human-readable code via method chaining and consistent function returns. Fluent interfaces fit naturally in the context of Bioinformatics workflows because they enable writing succinct pipelines.

Several other DSLs have been implemented to reason about genomics data. Broadly, these are either implemented as query languages or as command line tools embedded in the unix environment.[1] An example of the former is the Genome Query Language (GQL) and its distributed implementation GenAp which use an SQL-like syntax for fast

[1] other ideas to mention GROK [2]

retrieval of information from genomic databases and BAM files [3]; [4]. Another example is the Genometric Query Language (GMQL) which implements a relational algebra for combining big genomic datasets [5]. The command line application BEDtools develops an extensive algebra for performing arithmetic between two or more sets of genomic regions [6]. It also has a python interface which simplifies construcing scripts for performing analyses based on BEDTools [7].[2]

The abstraction provided by the *Ranges* data structures aligns with the concept of tidy data [8]. The tidy data pattern is useful because it allows us to see how the data relates to the design of an experiment and the variables measured. Consequently, it makes both the modelling and manipulation of data more systematic. The *Ranges* data structure follows this abstraction: it is a rectanglular table corresponding to a single biological context. Each row contains a single observation and each column a variable about that observation.

The tidy data abstraction has motivated the development of `plyranges` a grammar of genomic data manipulation based on the *Ranges* data structures. It implements and extends the grammar defined by the R package `dplyr` [9]. The grammar provides a consistent way of interacting with and analysing genomic data via methods for constructing, grouping, mutating, filtering, and summarising *Ranges* and an algebra for reasoning about actions on *Ranges* and relationships between *Ranges*.

Design and Implementation

The *plyranges* API implements a domain specific language using the existing *IRanges* and *GenomicRanges* packages in Bioconductor as a backend. Consequently, our API still has the speed and effeciency of the aforementioned packages but with a more coherent interface. The API also extends the grammar elements in *dplyr* for performing genomic specific manipulations such as finding overlapping regions or nearest neighbour regions between many *Ranges*. The *plyranges* API is specifically designed to enable fast interactive analysis of *Ranges* objects but can also be used for scripting genomic data workflows.

We have desgined the API to be fluent. Every function call corresponds to an action on a *Ranges* object (they are named verbs) and where possible functions have few arguments. Each verb is constructed to enable a tab completion based workflow. Both of these aspects reduce the cognitive load on a new user since most manipulations can be performed with a vocabulary of several verbs, rather than having to memorise functions with many arguments that are nouns (as is required in the existing Bioconductor packages). This is also has the advantage of allowing users to write human readable code because verbs describe what the code is doing rather than how its doing it.

Workflows can be composed by chaining verbs together via the pipe operator,`%>%` (exported from the R package *magrittr*). This is possible because every function call is endomorphic: when the input is *Ranges* object the output will also be a *Ranges* object. One advantage of this static typing is that it does not require any additional learning of classes beyond *Ranges* and the *DataFrame* classes. This is similar to the bedtools API, where the output is usually a BED file. However, it strongly deviates from the design of the *Ranges* Bioconductor packages, where many methods return a new class upon return. The Bioconductor design enables effecient computing as users are exposed to low-level features of its API which *plyranges* trys to abstract away. Method chaining via the pipe operator can also be difficult to debug, as there multiple points of failure.

[2] probably should also mention something about their cons, and more detaiil

Working with Ranges

Construction and Import/Output

The *plyranges* API provides the methods `as_granges()` and `as_iranges()` for constructing Ranges from tabular data structures, such as data.frames in base R. These methods use non-standard evaluation so columns in a data.frame can be modified before a Ranges object is created.

The package also provides a consistent framework for reading and writing files from and to common genomic data formats, using the `rtracklayer` package as a backend. The methods are implemented in the `read_/write_` family of functions, currently *plyranges* can read and write BAM, BED, BEDPE, narrowPeaks, GFF/GTF, WIG and BigWig files.

Manipulation of Ranges

The *plyranges* API exports the six core verbs from the *dplyr* package and modifys them for use with simple *Ranges* and *Genomic Ranges*. The verb `mutate()` takes a Ranges object and a set of name-value pairs and generates a new Ranges object that with modified or new metadata columns or modified core components (start, end, width, seqnames, strand). The use of `mutate()` means that a user no longer needs knowledge of the accessors of the Ranges object, as they can modify them in place. The `filter()` function takes a Ranges object and logical expressions and restricts Ranges object to where the logical expression evaluates to true. The `summarise()` function takes a Ranges object and a set of name-value pairs and aggregates the Ranges according to functions evaluated in the name-value pairs. As `summarise()` is an aggregation it may break the structure of the of a Ranges object, hence it returns a DataFrame object. The `select()` function determines which metadata columns are returned and the order they are returned in. The `arrange()` function sorts a Ranges object by named variables. The `group_by()` function creates an implicit grouping of Ranges object according to variables in the Ranges object. This modifies the actions of `mutate()`, `summarise()` and `filter()` (and also the set operations, `reduce_ranges()` and `disjoin_ranges()`), so they are performed on each partition created by the grouping.

Arithimetic on Ranges

The API has an expressive algebra for performing arithmetic on Ranges via the verbs `set_width()` and `stretch()`. As the names suggest `set_width()` modifies the width of a Ranges object, while `stretch` extends the start and end of a Ranges object. These can be chained with the anchoring functions `anchor_start()`, `anchor_end()`, `anchor_center()`, `anchor_3p()` or `anchor_5p()`, which fix the coordinates of a Ranges object in place. Moreover, the `shift_` and `flank_` family of functions can be used to shift all coordinates in a Ranges object or generate flanking regions from a Ranges object to the left, right, upstream or downstream of the input. Unlike, the Bioconductor API, *plyranges* makes it expclit via function calls whether to take into account the strand information of a *Ranges* object.

Overlapping Ranges

A common operation to perform between two *Ranges* objects is to find overlaps or nearest neighbours. The *plyranges* API recasts these operations as 'joins' or 'pairing' operations. For overlaps, there are three join operations: `join_overlap_intersect()`, `join_overlap_inner()` and `join_overlap_left()` which are shown in figure (). These operations consider any overlap between two input ranges and return any corresponding

metadata from both Ranges objects as metadata. The intersect join takes the intersect of the start and end coordinates of overlapping intervals of the query and subject Ranges (for GenomicRanges it also accounts for sequence name), when there is a overlap the metadata corresponding to the query and subject Ranges are returned. Similarly, inner join takes the start and end coordinates of the query Ranges that overlap the subject Ranges and returns metadata of the overlapping query and subject Ranges. Finally, the left join performs a left outer join between the query and subject Ranges, it returns all genomic intervals from the query ranges, and returns missing values in metadata columns when there is no overlap. A user may also restrict or group by overlaps with the `filter_by_overlaps()`, `filter_by_non_overlaps()` and `group_by_overlaps()`. All overlap methods can be modified with the `within` suffix (which changes the type of overlap from 'any' to 'within') or the `directed` suffix (which takes into account the strand of a GenomicRanges object.).

For nearest neighbours, the `plyranges` API provides `join_nearest()`, `join_precede()`, and `join_follow()` functions. These functions are similar to the overlapping functions, in that they return the query ranges that are nearest (or precede or follow) the subject ranges and add metadata from the subject ranges when the query is a nearest neighbour of the subject. Like the overlap joins, these functions can modified with suffixes to find nearest neighbours that are left, right, upstream or downstream of the subject.

The pairing operations, `pair_overlap()`, `pair_nearest()`, `pair_follow()`, and `pair_precede()` are similar to the join operation but instead of returning a Ranges, they pair up the subject and query Ranges objects into a DataFrame, alongside their metadata columns. This data structure is similar to the *Pairs* data structure in the *S4Vectors* package or the BED-PE file format.

Non-standard evaluation

Results

Availablilty and Future Work

The *plyranges* package is avaiable on the Bioconductor project website `https://bioconductor.org` or can be accessed via Github `https://github.com/sa-lee/plyranges`. We aim to continue developing the `plyranges` package and extend it for use with more complex data structures such as the *SummarizedExperiment* class, which can be used for analysing transcriptomic and variant data. As the `plyranges` interface encourages tidy data practices it integrates well with the principles of the grammar of graphics, we aim to use it for the visualisation of genomic ranges.

References

1. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9. doi:10.1371/journal.pcbi.1003118

2. Ovaska K, Lyly L, Sahu B, Jänne OA, Hautaniemi S. Genomic region operation kit for flexible processing of deep sequencing data. IEEE/ACM Trans Comput Biol Bioinform. 2013;10: 200–206. doi:10.1109/TCBB.2012.170

3. Kozanitis C, Heiberg A, Varghese G, Bafna V. Using genome query language to uncover genetic variation. Bioinformatics. 2014;30: 1–8. doi:10.1093/bioinformatics/btt250

4. Kozanitis C, Patterson DA. GenAp: A distributed SQL interface for genomic data. BMC Bioinformatics. 2016;17: 63. doi:10.1186/s12859-016-0904-1

5. Kaitoua A, Pinoli P, Bertoni M, Ceri S. Framework for supporting genomic operations. IEEE Trans Comput. 2017;66: 443–457. doi:10.1109/TC.2016.2603980

6. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

7. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: A flexible python library for manipulating genomic datasets and annotations. Bioinformatics. 2011;27: 3423–3424. doi:10.1093/bioinformatics/btr539

8. Wickham H. Tidy data. Journal of Statistical Software, Articles. 2014;59: 1–23. doi:10.18637/jss.v059.i10

9. Wickham H, Francois R, Henry L, Müller K. Dplyr: A grammar of data manipulation [Internet]. 2017. Available: `https://CRAN.R-project.org/package=dplyr`