Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Research Article

OXFORD

Genome Analysis

Example article title

Nathan C. Sheffield[1,2,3,4,*]

[1] Center for Public Health Genomics, University of Virginia
[2] Department of Public Health Sciences, University of Virginia
[3] Department of Biomedical Engineering, University of Virginia
[4] Department of Biochemistry and Molecular Genetics, University of Virginia

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Genomic region sets are summaries of different types of functional genomics data, and define locations of interest in the genome such as regulatory regions and transcription factor binding sites. The number of publicly available region sets has increased dramatically, leading to challenges in storage, processing, and retrieval.

Results: We propose a new method to represent genomic region sets as vectors. To create the vectors, we tested three different methods to extract features from region sets: interval unions, term frequency-inverse document frequency, and region set embedding using an adapted word2vec approach. We evaluated each method in two ways: First, by classifying the cell line, antibody, or tissue type of the region set; and second, by assessing whether similarity among embeddings can reflect simulated random perturbations of genomic regions. Our word2vec-based region set embeddings are much smaller than the original genomic region sets, reducing the number of dimensions for our region set representation from more than a hundred thousand to 100 without significant loss in classification performance. The vector representation could identify cell line, antibody, and tissue type with over 90% accuracy. We also found that the vectors could quantitatively summarize simulated random perturbations to region sets. Our evaluations demonstrate that the vectors retain biological information while requiring significantly less disk space and lower run-time to process. We propose that vector representation of region sets is a promising approach for efficient analysis of genomic region data.

Availability: The code is available at: https://github.com/databio/regionset-embedding

Contact: nsheffield@virginia.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Citations

Cite papers using brackets and `bibtex` keys. Example citation: `[@Sheffield2016]` will be rendered like this[1]. Use semicolons to separate multiple citations[1,2].

2 Figures

Refer to a figure using figure labels, so they are numbered automatically, like this: `\ref{abstract}` (See Fig. 1). Wrap a figure using the `pandoc-wrapfig` extension by adding '{0}' to the end of the caption (Fig. 2).

Duis in tempor mauris, a lobortis nisl. Integer arcu lorem, vehicula sed ante commodo, maximus eleifend nisi. Aenean efficitur molestie lorem, ac pharetra felis euismod nec. Duis vitae ligula facilisis, dignissim justo eget, elementum est. Nulla quis mi a justo porta pellentesque eget sit amet purus. Ut ac vestibulum ante, in efficitur massa. Cras feugiat in urna facilisis ultrices. Nullam vestibulum, lacus eget pretium pharetra, augue ligula consectetur diam, eget condimentum ipsum magna sed augue.

Vivamus eu rhoncus neque. Quisque egestas venenatis odio a mattis. Ut ligula turpis, facilisis a cursus eget, semper quis dolor. Integer varius est ipsum, porttitor ornare eros placerat eget. Nulla aliquet nisi arcu, sed vestibulum urna faucibus pretium. Maecenas laoreet diam non urna

Fig. 1: Example full-width figure

tincidunt iaculis a ut ex. Aenean sem enim, laoreet id accumsan sed, faucibus vitae diam. Aenean facilisis tincidunt risus. Mauris sit amet hendrerit est, sit amet maximus augue.

Duis in tempor mauris, a lobortis nisl. Integer arcu lorem, vehicula sed ante commodo, maximus eleifend nisi. Aenean efficitur molestie lorem, ac pharetra felis euismod nec. Duis vitae ligula facilisis, dignissim justo eget, elementum est. Nulla quis mi a justo porta pellentesque eget sit amet purus. Ut ac vestibulum ante, in efficitur massa. Cras feugiat in urna facilisis ultrices. Nullam vestibulum, lacus eget pretium pharetra, augue ligula consectetur diam, eget condimentum ipsum magna sed augue.

Fig. 2: Example wrapped figure

Duis in tempor mauris, a lobortis nisl. Integer arcu lorem, vehicula sed ante commodo, maximus eleifend nisi. Aenean efficitur molestie lorem, ac pharetra felis euismod nec. Duis vitae ligula facilisis, dignissim justo eget, elementum est. Nulla quis mi a justo porta pellentesque eget sit amet purus. Ut ac vestibulum ante, in efficitur massa. Cras feugiat in urna facilisis ultrices. Nullam vestibulum, lacus eget pretium pharetra, augue ligula consectetur diam, eget condimentum ipsum magna sed augue.

3 Tables

3.1 One-column table

Flag	Indication
1	CONTENT-ALL-A-IN-B
2	CONTENT-ALL-B-IN-A
4	LENGTHS-ALL-A-IN-B
8	LENGTHS-ALL-B-IN-A
16	NAMES-ALL-A-IN-B
32	NAMES-ALL-B-IN-A
64	CONTENT-A-ORDER
128	CONTENT-B-ORDER

. Table 1: Compatibility flags Parameter combinations used in the analysis and their results.

3.2 A two-column table

You can do a two-column table using the `\begin{table*}` environment. See Table 2.

3.3 Markdown tables

You can use markdown tables, too…sort of. Pandoc renders markdown tables with the `longtable` package. But longtable is not compatible with a two-column template. So, there are a few hacks and workarounds, but nothing works really well. The best thing I have found works *sometimes* – but then occasionally it just gobbles up text and figures silently. So, I suggest using latex templates until this issue is solved:

https://github.com/jgm/pandoc/issues/1023

Another issue is that Captions are preceded by the *Table* keyword. Unfortunately, I can't figure out how to put the caption below the table (it's above it by default).

3.4 Lorem ipsum

In hac habitasse platea dictumst. Mauris ut aliquet nunc, id mattis velit. Nunc commodo enim sed orci ultrices sodales. Fusce nec sem est. Nam euismod erat at neque facilisis iaculis. Cras rutrum elementum erat eu egestas. In sit amet est vitae ligula semper vestibulum sit amet quis justo. Sed porta dolor ac scelerisque congue.

Vivamus convallis arcu et lacus egestas tempus. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Integer eleifend efficitur risus sit amet suscipit. Curabitur consectetur sapien eget nulla maximus, quis lobortis nisl porta. Sed hendrerit semper placerat. Nulla sagittis orci arcu, a tincidunt lorem lacinia sit amet. Nullam nec fringilla odio. In mollis vitae nibh ut sollicitudin. Aliquam finibus tellus quis sollicitudin cursus. Cras lobortis, tortor ac sodales tempus, lacus ipsum aliquam mauris, sed placerat neque ante in arcu. Aliquam erat volutpat. Donec eu sodales odio, eu cursus libero.

Sed in porttitor leo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Sed tristique malesuada ligula, vulputate commodo mi lacinia ut. Ut consectetur, mauris vitae hendrerit mattis, nisi urna condimentum nibh, vitae rhoncus nulla tellus sit amet dui. Vivamus tempus magna eget quam posuere interdum. Nulla turpis augue, consequat quis euismod a, tincidunt sed ligula. Integer ac iaculis lacus, nec posuere augue. Fusce vitae dictum felis. In hac habitasse platea dictumst. Etiam suscipit magna turpis, eget volutpat lectus placerat quis. Mauris sed cursus erat. Sed a pellentesque felis. Ut in blandit dolor, vitae lobortis justo. Aenean turpis felis, pulvinar fringilla vulputate et, venenatis in lorem. Donec vulputate, nunc non imperdiet ullamcorper, justo nunc placerat elit, ut pretium justo metus a mi.

4 Embedded LaTeX

You can insert `latex` in-line in the markdown document: $rList[I_E] \leq q.start$

Or you can create separate environments like this:

4.1 Algorithm examples

These examples use the `algorithmic` environment (from the `algorithmcx` package:)

Require: $n \geq 0$
Ensure: $y = x^n$
  $y \Leftarrow 1$
  $X \Leftarrow x$
  $N \Leftarrow n$
  while $N \neq 0$ **do**
    if N is even **then**
      $X \Leftarrow X \times X$
      $N \Leftarrow \frac{N}{2}$ ▷ This is a comment
    else if N is odd **then**
      $y \Leftarrow y \times X$

Fig. 3: Example double-column figure

parameter set	add	drop	shift	Jaccard mean	Coverage mean	Euclidean mean	Cosine mean
add1	0.1	0.0	0.0	0.909	0.981	0.939	0.988
add2	0.2	0.0	0.0	0.833	0.964	0.914	0.977
add3	0.3	0.0	0.0	0.769	0.951	0.895	0.966
drop1	0.0	0.1	0.0	0.900	0.950	0.883	0.954
drop2	0.0	0.2	0.0	0.800	0.900	0.834	0.905
drop3	0.0	0.3	0.0	0.700	0.850	0.796	0.852
shift1	0.0	0.0	0.2	0.941	0.902	0.979	0.998
shift2	0.0	0.0	0.5	0.860	0.756	0.966	0.996
shift3	0.0	0.0	0.8	0.785	0.610	0.957	0.994
add_drop1	0.1	0.1	0.0	0.942	0.933	0.874	0.946
add_drop2	0.1	0.2	0.0	0.840	0.886	0.831	0.901
add_drop3	0.1	0.3	0.0	0.737	0.838	0.795	0.852
add_drop4	0.2	0.1	0.0	0.783	0.920	0.865	0.939
add_drop5	0.2	0.2	0.0	0.878	0.886	0.827	0.898
add_drop6	0.2	0.3	0.0	0.772	0.828	0.795	0.852
add_drop7	0.3	0.1	0.0	0.736	0.910	0.857	0.932
add_drop8	0.3	0.2	0.0	0.693	0.867	0.824	0.894
add_drop9	0.3	0.3	0.0	0.807	0.828	0.795	0.851
shift_drop1	0.0	0.1	0.2	0.850	0.857	0.882	0.953
shift_drop2	0.0	0.1	0.5	0.779	0.718	0.879	0.950
shift_drop3	0.0	0.1	0.8	0.714	0.579	0.877	0.949
shift_drop4	0.0	0.2	0.2	0.758	0.812	0.833	0.904
shift_drop5	0.0	0.2	0.5	0.765	0.767	0.832	0.902
shift_drop6	0.0	0.2	0.8	0.642	0.548	0.830	0.900
shift_drop7	0.0	0.3	0.2	0.665	0.767	0.795	0.851
shift_drop8	0.0	0.3	0.5	0.615	0.643	0.794	0.849
shift_drop9	0.0	0.3	0.8	0.568	0.518	0.793	0.847

. Table 2: Parameter combinations used in the analysis and their results.

$N \Leftarrow N - 1$
end if
end while

1: **repeat** ▷ forever
2: this
3: **until** you die.

This example uses the `algorithm` environment:

Algorithm 1 Euclid's algorithm

1: **procedure** Euclid(a, b) ▷ The g.c.d. of a and b
2: $r \leftarrow a \bmod b$
3: **while** $r \neq 0$ **do** ▷ We have the answer if r is 0
4: $a \leftarrow b$
5: $b \leftarrow r$
6: $r \leftarrow a \bmod b$
7: **end while**
8: **return** b ▷ The gcd is b
9: **end procedure**

Maecenas vitae sodales est, venenatis ullamcorper magna. Integer id orci ut arcu venenatis mattis. Pellentesque eget risus non lectus interdum efficitur. In pharetra odio in tellus eleifend commodo. Morbi facilisis mauris ac eros gravida pretium. Nam sit amet nisi massa. Morbi at turpis in leo dictum suscipit. Ut interdum, orci sed laoreet venenatis, odio dui consectetur tortor, sit amet vulputate ipsum neque ac ante. Vivamus vitae mi interdum, dignissim leo lobortis, ultricies leo. Aenean facilisis sagittis urna in blandit. Sed sit amet consectetur purus. Mauris bibendum efficitur magna, vitae egestas lacus pretium dignissim. Nullam eu magna est. Suspendisse vel lobortis metus.

Suspendisse potenti. Donec gravida ut mauris vel scelerisque. Nullam gravida maximus porttitor. Duis dictum nisl sed neque tristique sodales. Maecenas lacinia dolor eget ligula volutpat maximus. Etiam placerat lobortis enim ut iaculis. Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Fusce porta venenatis metus, vehicula sagittis ligula faucibus vel. Nunc nibh ipsum, vulputate sed bibendum non, euismod at sem. Praesent mi nulla, ornare vitae est a, euismod facilisis mauris.

Mauris a orci vehicula, aliquam orci in, cursus eros. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras semper vel enim eu dapibus. Maecenas placerat arcu nec metus tincidunt pharetra. Aenean rhoncus lacinia elit et cursus. Ut et metus vel augue sagittis volutpat quis nec nisi. Ut massa nisi, maximus vitae faucibus ut, eleifend ut odio.

Nam aliquam ex non accumsan efficitur. Nullam vehicula lorem vitae porttitor pellentesque. Fusce a tristique mi, sed congue velit. Nullam at ornare quam. Proin hendrerit accumsan ipsum, sed viverra velit vehicula sit amet. Donec non lectus diam. Sed condimentum non velit vel suscipit. Sed odio ex, vestibulum ullamcorper odio sit amet, lobortis accumsan risus. Nulla facilisi. Mauris eleifend viverra metus, ac varius lacus scelerisque non.

1. Sheffield, N. C. & Bock, C. LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32,** 587–589 (2016).

2. Sheffield, N. C., Nagraj, V. & Reuter, V. simpleCache: R caching for reproducible, distributed, large-scale projects. *The Journal of Open Source Software* **3,** 463 (2018).

5 References