Using Google Colab for Easily-Implemented User-Customizable Bioinformatics Web Tools: A Case Study Through OKtool for Overlapping K-mer Retrieval

Andrew Gao (andrewgao22@gmail.com)
Canyon Crest Academy https://orcid.org/0000-0002-3695-0046
Sarah Gao
Canyon Crest Academy

Method Article

Keywords: bioinformatics, colab, server, tools, web development, computational biology, kmer, sequence analysis, epitopes

Posted Date: June 6th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1724216/v1

Abstract

Bioinformatics web tools and servers are a powerful way for researchers to share valuable computational tools. They are ideal because users can run analyses in their browsers without having to download, set up, and run code on their local system. However, bioinformatics tool development requires web development proficiency and server costs which can be a barrier to entry. Google Colaboratory (Colab) is an online interactive code environment available to anyone with a free account. Colab is easy for both the developer and user. It allows researchers to make code available without needing web development skills. Users can customize code easily through Colab, which is beneficial since current bioinformatics web tools do not allow the user to edit code and perform customized operations. Also, Colab allows for easy collaboration and editing of code without concerns about platform compatibility between computers. Due to its user-friendly and customizable properties, we propose using Colab for developing bioinformatics tools. In this case study, we demonstrate implementation of the OKtool on Colab, a Python-based tool for calculating overlapping k-mers of desired length from a list of protein or DNA sequences. Applications include motif analysis and antigenic epitope analysis. It is a highly-customizable tool fully implemented in Colab that provides a FASTA file of k-mers derived from any given list of sequences. Overall, our work demonstrates the potential for bioinformatics tool development on Colab. OKtool is freely accessible at oktool.cloud.

Availability

Short link: oktool.cloud

Full link: https://colab.research.google.com/drive/15Z1ZETuJcXW59D4XKDWi9OKwtaEZC1nz?usp=sharing

Github link: https://github.com/andrewgcodes/OKtool

Implementation

Simply open the Google Colab link to run the code. Step by step instructions are provided. No downloads are necessary. If unable to access Colab, download the Github IPYNB file and open in Jupyter on your system. The only needed library is Pandas, which can be installed via pip.

Introduction

Web tools and servers have become a staple of the bioinformatics field, with hundreds of Internet-based tools available to researchers [1]. For instance, popular web tools and servers include the Expasy suite of tools, BLAST, Clustal Multiple Sequence Alignment, PEP-FOLD, and the Gene Expression Omnibus GEO2R [2–6]. Web-based software are valuable to researchers because they streamline many aspects of the research process. While many programs such as BLAST can be run locally, this can be difficult due to the wide variety of operating systems and associated challenges in keeping software updated to be

compatible with these operating systems. In comparison, browser-based web services can function on almost any device. Furthermore, browser-based services do not necessarily need to leverage the user's local computing resources, instead drawing from cloud computing. Services such as Galaxy can help researchers with limited computing resources, expanding access to methods such as protein structure prediction [7].

Despite the benefits of bioinformatics web services, there is a barrier to entry. In general, web development skills and proficiency with frameworks like R Shiny, Django Flask, or HTML/CSS/Javascript are necessary [8–9]. Many researchers lack the experience with web development to properly implement web tools. Also, hosting web services can be difficult to maintain for many years and costly to operate. In fact, a 2021 article reports that only about 50% of bioinformatics web services published in 2010 were still accessibl [10].

Google Colaboratory (Colab) is a free interactive browser code-editing environment focused on Python, a popular programming language in science. Colab is free and allows users to work with Python in a IPYNB format, like Jupyter Notebooks which organize code in blocks. Colab notebooks are easily shared and collaborators can be added. Importantly, non-collaborators can execute and modify code without changing the code for others. This enables customization of web tools to a user's specific needs. Colab reduces the need for researchers to learn web development. Recently, Mirdita et al. successfully implemented AlphaFold2 on Colab, demonstrating the practical utility of Colab as a platform for bioinformatics tools and its potential for increasing accessibility [11].

In the present study, we implement an overlapping k-mer retrieval algorithm in Colab and create a user-friendly interface. K-mers are widely used in bioinformatics for DNA and epitope analysis [12–13]. For instance, we previously used similar code to obtain overlapping 5-mers (pentapeptides) for the monkeypox virus cell surface-binding protein E8L and identify potential epitope similarity with the human proteome [14]. Currently, to our knowledge there is no web tool that enables users to easily submit sequences and obtain k-mers of a specified length in FASTA format.

Implementation

A Google Colab Python notebook was created. We used an algorithm to iterate through a Pandas Dataframe column of sequences and obtain all k-mers of a given length using slicing and list comprehensions and add the k-mers to a FASTA file. We used the Form Field functionality within Colab to create a user-friendly panel to specify parameters such as k-mer length, output file name, and the name of the file column where input sequences are found.

Users should upload a CSV file where one column contains the input sequences, with one sequence in each row. The CSV file can have other columns. Users upload CSV files through the upload panel on the left-hand side of Colab. We link a Youtube tutorial by Chanin Nantasenamat that explains the upload process [15]. Also, we provide an example input file containing 15 epitopes for Epstein Barr virus antigen 6 obtained from the IEDB [16].

Next, the user specifies five input parameters, the input file name, the desiredKmerLength, the column where sequences are found, the output file name, and the fasta type (detailedFASTA or simpleFASTA). The user clicks the Run circular button to run the code.

Finally, the user clicks the Run button on the second code block to begin the k-mer retrieval. We include a statement in the k-mer code to report any length errors. That is, if the desired k-mer length is longer than a sequence, it will be skipped and Colab will output a message such as “GPPAA is too short for the desired k-mer length”. The user can download the resulting FASTA file using the left-hand panel of Colab, where the input file was originally uploaded.

In order to make OKtool easily shareable, we purchased the short domain name “oktool.cloud” and set it to redirect to the Colab notebook URL. We also documented nearly every line of code to enhance readability and ease of customization for the user. Finally, recognizing that not every person can obtain a Google account because of geographical restrictions, we also post our code on a public Github repository.

Conclusion

In the current study, we demonstrate using Google Colaboratory for easily implementing a bioinformatics web service without knowledge of web development. It is user-friendly, easy to maintain and update, and free. A limitation is that a free Google account is required, which may not be available in certain regions. Accordingly, we provide a Github link to the OKtool source code file. OKtool is accessible at oktool.cloud and is well-documented with step-by-step instructions. We encourage researchers interested in making bioinformatics code interactively available to others to look into Colab.

Declarations

Conflicts of Interest

The authors have no relevant conflicts of interest to report.

References

1. Fox, Joanne A., et al. "The Bioinformatics Links Directory: a compilation of molecular biology web servers." Nucleic acids research 33.suppl_2 (2005): W3-W24.
2. Artimo, Panu, et al. "ExPASy: SIB bioinformatics resource portal." Nucleic acids research 40.W1 (2012): W597-W603.
3. Johnson, Mark, et al. "NCBI BLAST: a better web interface." Nucleic acids research 36.suppl_2 (2008): W5-W9.
4. Chenna, Ramu, et al. "Multiple sequence alignment with the Clustal series of programs." Nucleic acids research 31.13 (2003): 3497-3500.

5. Maupetit, Julien, Philippe Derreumaux, and Pierre Tuffery. "PEP-FOLD: an online resource for de novo peptide structure prediction." Nucleic acids research 37.suppl_2 (2009): W498-W503.
6. Barrett, Tanya, et al. "NCBI GEO: archive for functional genomics data sets—update." Nucleic acids research 41.D1 (2012): D991-D995.
7. Afgan, Enis, et al. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update." Nucleic acids research 44.W1 (2016): W3-W10.
8. Fehlmann, Tobias, et al. "Aviator: a web service for monitoring the availability of web services." Nucleic Acids Research 49.W1 (2021): W46-W51.
9. Jia, Lihua, et al. "Development of interactive biological web applications with R/Shiny." Briefings in Bioinformatics 23.1 (2022): bbab415.
10. Kern, Fabian, Tobias Fehlmann, and Andreas Keller. "On the lifetime of bioinformatics web services." Nucleic acids research 48.22 (2020): 12523-12533.
11. Mirdita, M., Schütze, K., Moriwaki, Y. et al. ColabFold: making protein folding accessible to all. Nat Methods (2022). https://doi.org/10.1038/s41592-022-01488-1
12. Ounit, Rachid, et al. "CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers." BMC genomics 16.1 (2015): 1-13.
13. Richer, Josh, Stephen Albert Johnston, and Phillip Stafford. "Epitope identification from fixed-complexity random-sequence peptide microarrays." Molecular & cellular proteomics 14.1 (2015): 136-147.
14. Gao, Andrew, and Sarah Gao. "In Silico Identification of Non-cross-reactive Epitopes for Monkeypox Cell Surface-Binding Protein." (2022).
15. Nantasenamat, Chanin. “How to Upload Files to Google Colab.” Www.youtube.com, 24 Apr. 2020, www.youtube.com/watch?v=6HFlwqK3oeo&t=174s. Accessed 31 May 2022.
16. Vita, Randi, et al. "The immune epitope database (IEDB): 2018 update." Nucleic acids research 47.D1 (2019): D339-D343.