Annals of Biomedical Engineering
https://doi.org/10.1007/s10439-023-03324-9

BMES BIOMEDICAL ENGINEERING SOCIETY

LETTER TO THE EDITOR

Code Interpreter for Bioinformatics: Are We There Yet?

Lei Wang[1] · Xijin Ge[2] · Li Liu[3,4] · Gangqing Hu[1]

Received: 12 July 2023 / Accepted: 14 July 2023

Abstract

The Code Interpreter feature in ChatGPT has the potential to democratize data analysis for non-specialists. As bioinformaticians, we are impressed by its performance in data manipulation and visualization. However, bioinformatics tasks often require execution of third-party packages, access to annotation knowledgebase, and handling large datasets. Code Interpreter's exclusive support for Python, no installation option for additional packages, inability to utilize external resources, and limited storage capacity could pose obstacles to its wide adoption in bioinformatics applications. To address these limitations, we advocated for the necessity of locally deployable, API-based systems for chatbot-aided bioinformatics applications.

Keywords Bioinformatics · ChatGPT · Code interpreter

Since its release in December 2022, ChatGPT has attracted substantial interest from the bioinformatics and computational biology communities due to its extensive knowledge in biology [1–3] and impressive capabilities in programming [4–7]. In a typical workflow, the human operator provides the chatbot with a natural language description of the data and desired analysis, initiating a request for code generation through an interactive web-based chat session [6]. The code is then copied and pasted into a local programming environment for execution. Any error messages from running the code are then fed back to the chatbot for correction. While ChatGPT has been used to address many bioinformatics data analysis tasks [4, 6], managing a local environment for code execution represents a significant barrier, especially for people without computational expertise.

Released on July 6, 2023, Code Interpreter beta is a new feature accessible to ChatGPT Plus users. This plugin integrates code generation and execution into one environment. Coupled with new features such as data upload and download, Code Interpreter represents a significant step toward personalized data analysis using natural languages. Following its release, numerous posts on social media have showcased complex coding tasks accomplished by this plugin. However, the extent to which Code Interpreter can assist with complex data analysis tasks that require interdisciplinary knowledge, such as in bioinformatics, remains elusive.

As bioinformaticians experienced with ChatGPT-assisted coding [5, 6], we first evaluated the data visualization capabilities of Code Interpreter in the context of gene expression data analysis. Our analysis began with a gene expression matrix from a previous work [8], with rows and columns representing genes and samples, respectively. Once the file was uploaded, we conducted three commonly utilized data visualizations: principal component analysis (Fig. 1a), an MA plot to highlight differentially expressed (DE) genes (Fig. 1b), and a heatmap visualization of gene expression augmented with hierarchical clustering analysis (Fig. 1c). As anticipated, Code Interpreter facilitated a seamless experience, enabling code generation, execution, and refinement all within one interface (Supplementary File 1). We acquired expected results after several iterations for each task and downloaded the results. These examples underscored the impressive data visualization capabilities of Code Interpreter. In addition, Code Interpreter can autonomously

Associate Editor Stefan M. Duma oversaw the review of this article.

✉ Gangqing Hu
michael.hu@hsc.wvu.edu

[1] Department of Microbiology, Immunology & Cell Biology, West Virginia University, Morgantown, WV 26506, USA

[2] Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007, USA

[3] College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA

[4] Biodesign Institute, Arizona State University, Tempe, AZ 85281, USA

Published online: 23 July 2023

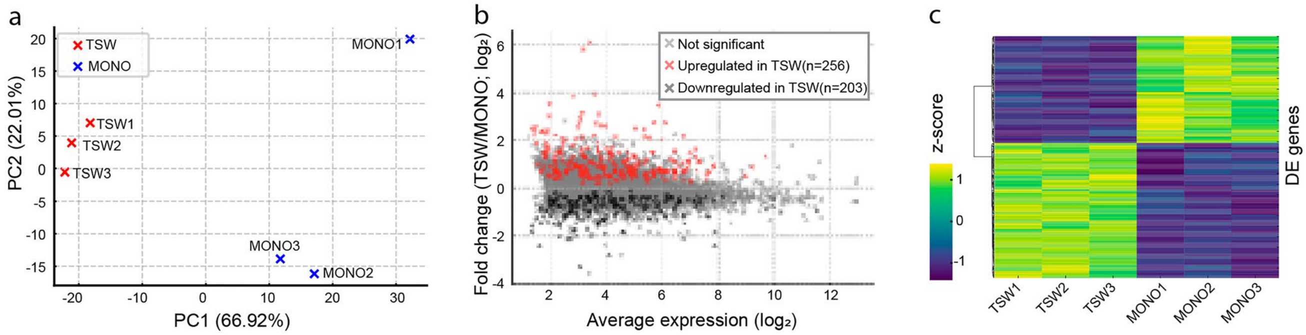

Fig. 1 Three examples of gene expression data visualization with Code Interpreter. **a** PCA analysis showing two groups of samples. **b** A MA plot with DE genes highlighted ($p < 0.001$; t-test). **c** Heatmap visualization of gene expression values (z-score normalized) across samples for DE genes identified in panel **b**.

regenerate code based on error messages and self-inspection of outputs (Supplementary File 2).

The subsequent tasks unveiled two major limitations that might impede a broad application of Code Interpreter in the field of bioinformatics. We requested ChatGPT conduct a Gene Ontology (GO) enrichment analysis, which requires downloading GO annotations and installing additional packages. ChatGPT responded stating that the environment does not support internet access and new package installation (Supplementary File 1). We encountered similar issues with other common bioinformatics tasks such as multiple sequence alignment (Supplementary File 2), gene ID conversion (Supplementary File 3), alignment of short sequencing reads to a reference genome (Supplementary File 4), prediction of DE genes based on a count matrix (Supplementary File 5), and phylogeny inference (Supplementary File 6). Although Code Interpreter currently has 343 pre-installed packages (Supplementary File 7), very few are specifically dedicated to bioinformatics data analysis. Furthermore, as the field of bioinformatics constantly evolves with the release of new tools, it is not feasible to pre-load all desired packages in Code Interpreter.

Our experiences also highlighted other limitations of Code Interpreter in bioinformatics data analysis. Firstly, the current version only supports Python programming. Some research fields may have preferences in programming languages such as R for evolutionary analysis, MATLAB for computer vision, and C/C++ for high-throughput sequence processing. Secondly, although the chat history can be archived, files and links generated by Code Interpreter are not preserved. This would necessitate re-uploading files and repeating the analysis to continue with new tasks. Thirdly, Code Interpreter allows for the upload of file with size up to 100 MB. While this size is sufficient for most applications, bioinformatics data analysis, especially those based on deep sequencing, often require a much larger storage. For instance, a single fastq file for an RNA-Seq library typically exceeds 1 GB in size. Fourthly, lack of support for parallel processing hampers the analysis of large datasets, resulting in slow performance. Lastly, users of the Code Interpreter must be mindful of the risk of data leakage. While renaming sample labels may partially mitigate this issue, certain genomic sequence data such as those containing single-nucleotide polymorphisms pose a risk of re-identification [9]. At a minimum, data uploading to the Code Interpreter should comply with existing ethical regulations, such as the Health Insurance Portability and Accountability Act in the United States and the General Data Protection Regulation in Europe.

In conclusion, the Code Interpreter feature of ChatGPT enables users, including those without coding experiences, to analyze their data through natural language. But in its current version, its utility in bioinformatics data analysis is hindered by many limitations, some of which could be addressed in future versions. These include internet access for downloading genome annotations, preinstallation of bioinformatics-specific packages, expansion of storage capacity, and support for additional programming languages. However, regarding privacy and security, it will be beneficial to develop applications like the Code Interpreter but can be deployed locally. In this scenario, code generation would be accomplished through API communication with remote language models without data uploading. The code would then be tested in a local and secure execution environment with all necessary packages installed. We foresee a future where there is a coexistence of locally deployable, API-based systems, such as RTutor [10], and online web interface-based plugins such as Code Interpreter that meets different needs.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10439-023-03324-9.

Acknowledgements NIH-NIGMS grants P20 GM103434, U54 GM-104942, and 1P20 GM121322 (GH). NIH-NIGMS Grant R01HG010805 and P20GM135008 to XG. NIH-NLM Grant No.

R01LM013438 to LL. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The writing was polished by ChatGPT.

Author Contributions GH contributed to conceptualization, formal analysis, and writing—original draft; LW contributed to formal analysis; and XG and LL contributed to formal analysis, writing—review and editing.

Data availability Prompts and scripts to support the conclusions are in Supplementary Files of the manuscript.

Declarations

Competing interests The authors declared no competing interests.

References

1. Hou, W., and Z. Ji. GeneTuring tests GPT models in genomics. *BioRxiv*. 2023. https://doi.org/10.1101/2023.03.11.532238.
2. Duong, D., and B. D. Solomon. Analysis of large-language model versus human performance for genetics questions. *Eur. J. Hum. Genet.* 2023. https://doi.org/10.1038/s41431-023-01396-8.
3. Kung, T. H., et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit. Health*.2(2):e0000198, 2023.
4. Merow, C., et al. AI chatbots can boost scientific coding. *Nat. Ecol. Evol.* 7:960, 2023.
5. Perkel, J. M. Six tips for better coding with ChatGPT. *Nature*. 618(7964):422–423, 2023.
6. Shue, E., et al. Empowering beginners in bioinformatics with ChatGPT. *Quant. Biol.* 11(2):105–108, 2023.
7. Xu, D. ChatGPT opens a new door for bioinformatics. *Quant. Biol.* 11(2):204–206, 2023.
8. Dziadowicz, S., et al. Bone marrow stroma-induced transcriptome and regulome signatures of multiple myeloma. *Cancers*. 14(4):927, 2022.
9. Bernier, A., H. Liu, and B. M. Knoppers. Computational tools for genomic data de-identification: facilitating data protection law compliance. *Nat. Commun.* 12(1):6949, 2021.
10. Ge, S.X. *RTutor, Chat with your data via AI*. [cited 2023 07/11/2023]. https://RTutor.ai, 2023.