Quantitative Biology 2023, 11(2): 105–108
https://doi.org/10.15302/J-QB-023-0327

PERSPECTIVE

Empowering beginners in bioinformatics with ChatGPT

Evelyn Shue[1], Li Liu[2,3], Bingxin Li[4], Zifeng Feng[5], Xin Li[6], Gangqing Hu[1,*]

[1] Department of Microbiology, Immunology & Cell Biology, West Virginia University, Morgantown, WV 26506, USA
[2] College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA
[3] Biodesign Institute, Arizona State University, Tempe, AZ 85281, USA
[4] Finance Department, John Chambers College of Business and Economics, West Virginia University, Morgantown, WV 26506, USA
[5] Department of Economics and Finance, The University of Texas at El Paso, El Paso, TX 79902, USA
[6] Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA
* Correspondence: michael.hu@hsc.wvu.edu

Received February 28, 2023; Revised March 20, 2023; Accepted March 20, 2023

The impressive conversational and programming abilities of ChatGPT make it an attractive tool for facilitating the education of bioinformatics data analysis for beginners. In this study, we proposed an iterative model to fine-tune instructions for guiding a chatbot in generating code for bioinformatics data analysis tasks. We demonstrated the feasibility of the model by applying it to various bioinformatics topics. Additionally, we discussed practical considerations and limitations regarding the use of the model in chatbot-aided bioinformatics education.

Keywords: bioinformatics; education; scientific data analysis; ChatGPT

The recent emergence of large language modes (LLMs), such as ChatGPT, has sparked great interests in their potential use in facilitating programming-heavy data analysis, such as bioinformatics. With its remarkable conversational and programming abilities [1] (see rtutor.ai website for examples), ChatGPT holds great promise for helping students overcome the programming hurdle. However, as an advanced artificial intelligence (AI) system, the behavior of a chatbot heavily relies on prompts provided by human operators. To fully harness this potential to assist scientific data analysis, prompts used to instruct a chatbot must be carefully crafted so that responses from the chatbot are valid and results are robust.

Inspired by adaptive learning in educational literature [2], we proposed the OPTIMAL model to facilitate chatbot-aided scientific data analysis: Optimization of Prompts Through Iterative Mentoring and Assessment with an LLM chatbot (Supplementary Fig. S1). The model involves a series of iterative steps to improve communication with a chatbot for scientific data analysis and enhance students' learning outcomes. Students will first review the scientific question, analysis task, computational methods, and expected outputs. They will receive guidance on creating a set of draft prompts that describe the data analysis task at various levels of details. Students then converse with the chatbot by inputting the prompts to generate code and execute the chatbot-produced code. In the event error messages are issued after running the code, students must evaluate the error messages and determine the best way to proceed, such as instructing the chatbot to revise the code given the error messages or debugging the code manually. This process iterates until the code no longer issues errors and outputs a result for critical assessment. For an unexpected result, the prompts will be reevaluated and refined, repeating until the expected result is obtained. At the end of the session, students should reflect on the entire communication process and review the code to identify any missing details to be

added to the initial prompts. This step may require students to consult relevant manuals and summarize the analytic methods to ensure accuracy and reproducibility. In the end, the iteration and final review expect to yield clear, focused, and concise prompts, as well as a reference code for the desired data analysis. As a proof-of-concept, we applied the model to three case studies from different topics of bioinformatics and summarized the findings as follows.

Short sequencing reads alignment and visual inspection in next generation sequencing analysis: Alignment, the process of determining genomic positions of short sequencing reads, is a fundamental step in deep sequencing data analyses. In this case study, we visually inspected the quality of one chromatin immunoprecipitation followed by sequencing dataset generated by Encyclopedia of DNA Elements [3] (see Supplementary Table S1, Fig. S2A). To this end, we instructed ChatGPT to generate code to align the short reads to the human reference genome and summarize the alignments into count numbers across the genome. The results were visually assessed by loading the summarized alignments to Integrative Genomics Viewer [4]. The initial prompts included key details of the analyses and bioinformatics tools. The interaction involved two iterations where we instructed the chatbot to handle error messages generated from running the code. Analyzing the final code and reflecting on the entire interactions identified additional details missing from the initial prompts.

Phylogeny inference by DNA sequences in molecular evolution: Phylogenetic inference is an essential yet challenging subject in molecular biology curricula. To demonstrate how ChatGPT can assist students in phylogenetic analyses, we asked the chatbot to generate R code to build a phylogenetic tree for nine species (see Supplementary Table S2, Fig. S2B). This case study started with multiple alignments of protein-coding sequences of the TP53 tumor suppressor gene. The initial prompts included a description of major steps to build an unrooted tree. With two rounds of iterations and human feedback on error messages from running the code, the chatbot wrote workable code to generate a reasonable unrooted phylogenetic tree. We then instructed the chatbot to use a designated species as an outgroup to root the tree. For this complicated task, the chatbot failed to find a valid solution and began to make up functions that did not exist. In this situation, human intervention was required to correct the code after multiple failed iterations.

Robust circles fitting in computer vision: Biomedical imaging captures and examines images in biotechnology and medicine. Human vision systems can recognize many different objects in various challenging environments, such as cluttered backgrounds and extreme poses. Circles are arguably the simplest geometric objects for training a computer to recognize. However, instructing a computer to fit circles is a nontrivial task that requires advanced mathematical preparation and proficient computer programming skills. More importantly, students often face the challenge of decomposing a complex problem into several more manageable sub-problems (*i.e.*, a divide-and-conquer approach). In contrast to the previous case studies, this one illustrated a scenario in which describing all analyses in one prompt failed to generate workable code. Instead, we demonstrated how ChatGPT could serve as a virtual teaching assistant to teach the divide-and-conquer approach to a student (see Supplementary Table S3, Fig. S2C). Using a chain-of-thought (CoT) prompt, the student can gradually learn how to solve more and more challenging circle fitting problems, from single to multiple, from clean to noisy observation, and from analytical to numerical, as well as how to incorporate Bayesian priors into the solution algorithms. The results of this experiment was a sophisticated circle-fitting algorithm that cannot be obtained through iterations but could be achieved through CoT prompt design [5].

Our firsthand experience with ChatGPT has identified several practical considerations for implementing the OPTIMAL model in education settings. To streamline the iteration process, it is crucial to clearly define how the chatbot should respond to prompts, such as acting as an expert in bioinformatics and being proficient in a designated language, outputting code with a minimal number of lines, and resetting the thread upon request. To effectively use the model, it is essential to possess a good understanding of the key concepts and steps involved in a specific data analysis task for crafting the initial prompts. Other prerequisite knowledge and skills include the ability to install software, execute commands, and interpret code with the aid of user manuals and other resources.

Merely using the chatbot as a code-generating tool may limit creative thinking. Therefore, reviewing the code at the end of each session is just as important as optimizing the prompts. At this stage, the focus is on being familiar with the code and identifying missing details in the initial prompts. A well-crafted prompt should be robust across different chat sessions by yielding consistent results. Beginners may start communicating with the chatbot using natural language. When transitioning into intermediate or advanced levels, they may include code in the prompts to keep the chatbot on track.

In addition to overcoming the programming hurdle, another significant impact of the model is to enhance students' abilities in critical thinking and evaluation of the chatbot's response. We have observed instances

where ChatGPT produced erroneous functions, misused certain options, and faked author name of a package. While many of these errors can be detected by running the code, it is crucial to cross-reference with the manual to ensure a precise understanding of the functions, options, and chatbot's comments on code, as well as an accurate description of the methods.

A successful application of our model to a specific bioinformatics data analysis task is expected to generate a set of prompts and their associated code, which we refer to as a reference code. While the results from running the reference code should be deterministic if not involving any nondeterministic algorithm, ensuring the robustness of the prompts requires a systematic validation approach. One possible method is to have multiple users run the prompts in new chat sessions and compare the results to the reference one. In research projects where a reference is often absent, the results should be validated through external methods such as literature and existing or additional experiments the same as conventional data analysis. To ensure reproducibility, the prompts, code, and input files used for the project should be made publicly accessible upon publication.

Relevant to the robustness, the same prompts may not generate the same code in a new session. The uncertainty may result from the existence of multiple solutions to the same question or ambiguities/missing details in the prompts, giving the chatbot flexibility to make choices. Educators should be mindful of these uncertainties to control for their disruptions during lectures. On the other side, the uncertainties offer great opportunities for training critical and creative thinking. For example, by comparing new code to the reference, students may learn alternative solutions to improve their bioinformatics skills. Moreover, uncertainties arising from ambiguities in the prompts provide an excellent chance for further refinement through iterations. Nevertheless, novice students must be informed of these uncertainties and potential solutions such as adjusting the temperature setting that controls the chatbot's response randomness to reduce their anxiety, making chatbot-assisted learning an enjoyable experience.

We vision that a repository of well-defined prompts for typical bioinformatics tasks, along with sample inputs, reference chatbot code, and expected results, would be immensely valuable to beginners. Familiarizing themselves with sample prompts can serve as a steppingstone for students to improve their ability to customize prompts with greater specificity to fit their evolving requirements. The repository also serves as a platform for the community to further validate the robustness of the prompts. However, the challenge remains that prompt engineering tailored for chatbot-aided bioinformatics data analysis in biomedical research, or the broader health science is just an emerging field of research.

The OPTIMAL model, like any other model, has its limitations that need to be addressed. The prompt-optimization iteration may not effectively converge to produce a valid solution without in-depth human intervention, especially for advanced data analysis with customized code. Meanwhile, the iterative model may not apply to problems that can be broken down into smaller subproblems that resemble the original problem (known as recursion). An alternative solution is to use the CoT prompt design.

In addition to bioinformatics, case studies in economics and finance (Supplementary Tables S4 and S5, Fig S2D and E) supported a potential extension of the model to scientific data analysis in other disciplines. However, all these case studies were mainly performed by experienced researchers in scientific data analysis with teaching experience. To examine the impact on students, controlled experiments conducted in a classroom setting are needed. Further research is necessary to evaluate the effectiveness of the OPTIMAL model in improving students' learning outcomes in data analysis relative to traditional lecturing. Lastly, an extension of the model to innovative bioinformatics research remains to be explored. As an ongoing effort, we are working on applying the model to new algorithm developments in single-cell gene expression data analysis.

Disadvantaged students beginning to learn bioinformatics face additional barriers like limited access to tutoring services, lack of interactions with instructors, difficulty forming study groups with academically advanced peers, and so on. With a positive outlook, we argue that ChatGPT has the potential to address this knowledge-dissemination disparity. In this human-AI integrative learning process, students may serve as mentors to guide the chatbot in bioinformatics data analysis and, at the same time, learn coding skills from the chatbot. Students may interact with ChatGPT as if studying with a peer that responds instantaneously. This is extremely valuable for academically challenged students who often struggle to find capable peers.

In conclusion, the OPTIMAL model represents a promising step forward in chatbot-aided education in bioinformatics data analysis for beginners. While the concept of ChatGPT-aided education is relatively new [6], our case studies from different disciplines demonstrated ChatGPT's potential to enhance students' coding skills and critical creative thinking. Such benefits of practicing bioinformatics with a chatbot are likely to extend from the classroom to a lifelong learning experience, especially for beginners.

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at https://doi.org/10.15302/J-QB-023-0327.

Fig. S1: The OPTIMAL model for LLM chatbot-assisted scientific data analysis; Fig. S2: Summary of case studies applying the OPTIMAL model to chatbot-assisted data analysis in five distinctive fields; Table S1: Case study for short sequencing reads alignment and visual inspection; Table S2: Case study for phylogeny inference by DNA sequences; Table S3: Case study for robust circles fitting; Table S4: Case study for household income vs. high school graduation rates; Table S5: Case study for time series analysis of trading data

ACKNOWLEDGEMENTS

NIH-NIGMS grants P20 GM103434, U54 GM-104942, and 1P20 GM121322 to GH; NIH-NLM grant R01LM013438 to LL. We thank Dr. Jackie J.D. Han from Peking University, Dr. Heather Henderson from West Virginia University, and Dr. Dong Xu from University of Missouri for insightful discussions. The writing was polished by ChatGPT.

COMPLIANCE WITH ETHICS GUIDELINES

Evelyn Shue, Li Liu, Bingxin Li, Zifeng Feng, Xin Li and Gangqing Hu declare that they have no conflict of interest.

This article is a perspective article and does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Chatterjee, J. and Dethlefs, N. (2023) This new conversational AI model can be your friend, philosopher, and guide... and even your worst enemy. Patterns, 4, 100676
2. Durlach, P. J. and Lesgold, A. M. (2012) Adaptive Technologies for Training and Education. Cambridge: Cambridge University Press
3. The ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature, 489, 57−74
4. Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform., 14, 178−192
5. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ihter, B., Xia, F., Chi, E. H., Le, Q. V. and Zhou, D. (2023) Chain of thought prompting elicits reasoning in large language models. In: 36th Conference on Neural Information Processing Systems (NeurIPS 2022)
6. Pavlik, J. V. (2023) Collaborating with chatgpt: Considering the implications of generative artificial intelligence for journalism and media education. Jour. Mass Comm. Educ., 78, 84−93