CSI 5180. Topics in Artificial Intelligence: Machine Learning for Bioinformatics

Project

Marcel Turcotte

Version: Feb 10, 2025 10:42

🎯 Learning Outcomes

1. **Assess** the methodologies and conclusions of scientific publications in bioinformatics, identifying strengths and potential limitations.
2. **Reproduce** datasets from scientific studies, applying appropriate data cleaning, normalization, and encoding techniques.
3. **Apply** at least two machine learning algorithms to bioinformatics data, including one used in theoriginal study, and assess their performance.
4. **Evaluate** the reproducibility of bioinformatics studies by comparing original results withindependently obtained outcomes, discussing potential discrepancies.
5. **Create** detailed protocols for data analysis and machine learning applications, ensuring clarityand reproducibility for future research.
6. **Work** in teams to plan, execute, and present bioinformatics projects, demonstrating effectivecommunication and project management skills.

This project offers the opportunity to gain a deeper understanding of a specific bioinformatics problem by evaluating the performance of at least two machine learning algorithms. You will conduct this analysis using a real-world dataset sourced from a scientific paper, and critically assessed its methodology.

The task also highlights the critical issue of reproducibility, a growing concern in numerous scientific disciplines. In response, Université Paris-Saclay has introduced a dedicated course titled "Reprohackathon" to address these challenges.

- Cokelaer, Thomas, Sarah Cohen-Boulakia, and Frédéric Lemoine. 2023. "Reprohackathons: Promoting Reproducibility in Bioinformatics through Training." *Bioinformatics* 39 (Supplement_1): i11–20.

In the current academic landscape, an increasing number of publications mandate authors to provide access to their data and code. Unfortunately, reproducibility remains a challenge.

You do not have to exactly replicate the results or to surpass the findings of the original scientific paper. Instead, your evaluation will be based on the development of a robust protocol for conducting your work, as well as the quality of your code and the thoroughness of your analysis.

- Reproducibility for Machine Learning in Life Science.

Submission Requirements & Deadlines

- **February 14, 2025:** Project proposal
- **April 7, 2025:** Report

Select a recent scientific paper in which a machine learning algorithm has been applied to a bioinformatics problem. You can continue to work on the same scientific paper as for your presentation or select a new one. You will find a list of journals and conferences below below.

Reconstruct, either partially or fully, the dataset used in the selected scientific paper. If replicating the dataset proves too challenging due to various constraints, you may simulate the data, but only as a last resort. Refer to our lecture on essential bioinformatics skills to write scripts that automate this process. Clearly outline the steps necessary for data preparation, including tasks such as cleaning, normalization, and encoding.

Subsequently, apply two machine learning algorithms to the dataset and conduct a thorough analysis of the results. One of these algorithms should either be the same or closely related to the one employed in the original publication. Evaluate whether you were able to replicate the results reported in the paper. If not, provide possible explanations for the discrepancies. Compare the performance of the two algorithms: Which approach yields better results, and what factors contribute to its superiority?

Project Proposal

At the project proposal stage, students are required to submit a concise document (approximately 1-2 pages) that includes the following components:

1. **Title:** Clearly state the title of the selected scientific paper.
2. **Background and Motivation:** Provide a brief overview of the paper's context, highlighting its significance in the field of bioinformatics and the rationale for choosing this particular study.
3. **Objectives:** Clearly define the goals of the reproduction study, specifying which results or analyses from the original paper will be targeted.
4. **Methodology:** Outline the approach for reproducing the study, including:
    - **Data Acquisition:** Identify the datasets required and their availability.
    - **Tools and Technologies:** Specify the software, programming languages, and computational resources to be utilized.
    - **Reproduction Plan:** Describe the steps to replicate the experiments, including any anticipated challenges.
5. **Expected Outcomes:** Discuss the anticipated results and any potential discrepancies that might arise compared to the original study.
6. **Timeline:** Provide a tentative schedule outlining key milestones and deliverables throughout the project duration.
7. **References:** List all relevant literature, including the primary paper and any supplementary materials that will inform the reproduction effort.

Report

The project constitutes 30% of the final grade and will be evaluated based on three components: the project outline (5%), a comprehensive written report (15%), and the accompanying source code (10%). The written report must be detailed enough to enable the replication of the analysis solely from the text.

Despite the requirement for detail, it is crucial to maintain conciseness. For a team of two members, an appropriate report length is 10 pages. The recommended structure for the report is as follows:

- Introduction
  - Background
  - Problem Definition
  - Data Description
    - Source of the Data
    - File Formats Utilized
    - Data Cleaning Procedures, if applicable
    - Data Encoding Strategies and Rationale
- Methods
- Results
- Conclusions
- Comprehensive List of References

Teamwork

Teams should ideally be composed of one or two members. While larger teams are permissible, they are expected to deliver a correspondingly greater volume of work. Collaboration between teams on related yet complementary topics is encouraged, as this can lead to more realistic and comprehensive applications of the project objectives.

Paper Selection

You have the option to select a paper that either aligns with your presentation or differs from it. The decision is entirely yours.

Avoid choosing a scientific review article, as these tend to be overly broad and lack depth. Given the brevity of your presentation, it is unnecessary to cover every detail of the paper. Instead, focus on a specific subset of the content that allows you to deliver a concise and coherent message to your audience.

Prioritize well-established, highly ranked, peer-reviewed journals and conferences for sourcing information. Although excellent research is increasingly available on preprint servers such as bioRxiv and arXiv, these papers have not undergone peer review, making it difficult to assess their scientific validity. Consequently, it is advisable to refrain from relying on preprint papers.

Scientific Journals

This section enumerates the scientific journals that predominantly publish research in the field of bioinformatics. The impact factors of these journals, indicative of their influence and prestige within the academic community, are provided in parentheses. Please note that the impact factors listed here have not been updated since their initial compilation in 2019. It is important to recognize that these values are subject to change over time.

Bioinformatics Research

The following journals are dedicated to bioinformatics research.

- Bioinformatics (5.481)
- Briefings in Bioinformatics (5.134)
- Computational and Structural Biotechnology Journal (4.148)
- BioData Mining (4.0)
- PLOS Computational Biology (3.995)
- Database (3.978)
- BMC Bioinformatics (2.213)
- Bioinformatics Advance (2.4)
- IEEE/ACM Transactions on Computational Biology and Bioinformatics (1.955)
- Bulletin of Mathematical Biology (1.484)
- Computers in Biology and Medicine (2.115)
- Journal of Theoretical Biology (2.049)
- Evolutionary Bioinformatics (1.877)
- Journal of Mathematical Biology (1.846)
- Statistical Applications in Genetics and Molecular Biology (1.77)
- Journal of Proteomics & Bioinformatics (1.57)
- Algorithms for Molecular Biology (1.536)
- Computational Biology and Chemistry (1.331)
- Journal of Data Mining in Genomics & Proteomics (1.16)
- Journal of Computational Biology (1.032)
- Journal of Bioinformatics and Computational Biology (0.931)
- Current Bioinformatics (0.770)

See also: List of bioinformatics journals on Wikipedia, as well as this list of top publications from Google Scholar.

Life Science Journals

The following life science journals are known to publish bioinformatics research on a regular basis.

- Nature Reviews Genetics (40.282)
- Cell (31.398)
- Genome Biology (11.908)
- Nucleic Acids Research (11.561)
- Molecular Biology and Evolution (10.217)
- Molecular Systems Biology (8.447)
- GigaScience (7.463)
- Genome Research (6.2)

High Impact and Interdisciplinary Journals

Major contributions and/or interdisciplinary research are published in journals such as the following.

- Nature (40.137)
- Science (37.205)
- Nature Machine Intelligence (18.8)

- Nature Communications (12.353)
- Proceedings of the National Academy of Sciences of the United States of America (PNAS) (9.661)
- PLOS One (2.766)

Conferences

Bioinformatics Conferences

- Intelligent Systems for Molecular Biology (ISMB), 2024
- European Conference on Computational Biology (ECCB) 2024
- Research in Computational Molecular Biology (RECOMB), 2024
- ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB)
- IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM), 2024
- Pacific Symposium on Biocomputing (PSB), 2025
- International Conference on Bioscience, Biochemistry and Bioinformatics (ICBBB), 2024
- Machine Learning in Computational Biology, 2024 (formerly a workshop of NeurIPS)
- Bioinformatics and Biomedical Engineering 2024 Part 1, 2024 Part 2
- Advanced Intelligent Computing in Bioinformatics 2024
- International Conference on Bioinformatics Research and Applications, 2024, past event
- International Conference on Bioinformatics and Computational Biology (ICBCB), 2023, past events
- International Conference on Biomedical and Bioinformatics Engineering (ICBBE), 2023. past events
- Several conference proceedings are featured as special issues in BMC Bioinformatics. The provided link grants access to an extensive collection of these proceedings.

For conferences that lack a permanent website, I have included only the most recent event for which the proceedings are accessible. In certain instances, there are considerable delays in the publication of conference proceedings. For additional venues, you may also consult WikiCFP.

Machine Learning Conferences

High-quality bioinformatics research is frequently presented at top-tier machine learning conferences. Nevertheless, the broad scope of these conferences can make it difficult to locate papers specifically focused on bioinformatics.

- Annual Conference on Neural Information Processing Systems (NeurIPS)
- International Conference on Machine Learning (ICML)
- International Conference on Learning Representations (ICLR)
- ACM SIGKDD Conference on Knowledge Discovery and Data Mining
- AAAI Conference on Artificial Intelligence

Lists

- Awesome-LLMs-meet-genomes, a resource suggested by Kaixi Xu.

Selection Challenge

If you encounter difficulties in selecting a suitable scientific publication, consider examining machine learning repositories for datasets that are associated with relatively recent publications.

Ensure that the dataset you choose includes data types discussed in the initial three lectures. If you have any uncertainties, please seek clarification.

- UC Irvine Machine Learning Repository
  - Codon Usage: data and paper.
  - Mice Protein Expression: data and paper.
  - 9mers from cullpdb: data and paper.
  - Molecular Similarity Perception Based on Machine-Learning Models: data and paper.
  - Glioma Grading Clinical and Mutation Features: data and paper
  - Gene expression cancer RNA-Seq: data and paper (there might be more papers)
- OpenML
- ELVIRA Biomedical Data Set Repository
- PPT-DB! PPT-DB is a database of protein property databases

Suggestions

- Predicting synthetic mRNA stability using massively parallel kinetic measurements, biophysical modeling, and machine learning
- TSignal: a transformer model for signal peptide prediction, data
- A comprehensive benchmarking for evaluating TCR embeddings in modeling TCR-epitope interactions, data
- EnrichRBP: an automated and interpretable computational platform for predicting and analysing RNA-binding protein events, data
- Integrating representation learning, permutation, and optimization to detect lineage-related gene expression patterns
- Predicting synthetic mRNA stability using massively parallel kinetic measurements, biophysical modeling, and machine learning

Resources

If you do use AI assistance, thoroughly document your interactions. Include the tools and their versions in your report, along with a transcript of all interactions. Most AI assistants keep a record of your conversations. The recommended practice is to create a new conversation specifically for the presentation and consistently reuse this conversation throughout your work on the presentation. Ensure that this conversation is solely dedicated to the presentation. Submit this conversation transcript in the reference section of your summary.

Questions

- You may ask your questions in the Assignment topic of the discussion forum on Brightspace.
- Alternatively, you can email the teaching assistant. However, using the forum is strongly preferred, as it allows your fellow students to benefit from the questions and the corresponding answers provided by the teaching assistants.

References

Cokelaer, Thomas, Sarah Cohen-Boulakia, and Frédéric Lemoine. 2023. "Reprohackathons: promoting reproducibility in bioinformatics through training." *Bioinformatics* 39 (Supplement_1): i11–20. https://doi.org/10.1093/bioinformatics/btad227.

Samuel, Sheeba, and Daniel Mietchen. 2024. "Computational reproducibility of Jupyter notebooks from biomedical publications." *GigaScience* 13: giad113. https://doi.org/10.1093/gigascience/giad113.