CSE 566 Spring 2023: Course Project

The course project gives you a chance to get a feel for bioinformatics research and get more in-depth knowledge in a certain area. It is intended to give you the opportunity to perform a small research project related to ideas in the class. It is a project where you define your own goals and extend some of the results we learned about in class. You do not have the same harsh constraints as one would have in order to publish a research paper: your results do not have to show improvement over state of the art in order to get a good grade. This is a research project where finding out that your idea did not work is not a failure. However, you must come up with a research question and some ideas and pursue them. Your research question should be solid and interesting and your idea should be novel with respect to what we covered in class and with respect to the papers which you have read.

Choice of topic:

You must propose your own project. A good way to come up with a project is to extend and/or combine the papers presented by yourself and your classmates during class. You might also try to combine what you learned in class with other research problems. You are encouraged to pursue any connections you see to your own research. Your choice of project can be practical, i.e., involving real or simulated data, or it may be theoretical, i.e., proving theoretical results.

Asking a research question:

You should come up with a clear research question. A “research question” is something that is unsolved, and if solved, will sharpen state-of-the-art technologies or expand human’s knowledge base. Proposing such a research question is an important part of the course project. There are three types of research questions you may consider.

The first category is “can we make it better?”. There are many existing, well-defined, and/or well-studied research tasks that will be introduced in class and in the paper presentations. Probably methods have been developed but not perfect. Therefore, a natural question to ask is, can we improve them or design better methods. By “better”, we mean either improved accuracy (can be improved theoretical guarantee or evaluated with benchmarking datasets) or improved running time or memory footprint (again, can be either theoretical or experimentally evaluated).

The second category is “can we apply an existing algorithm/data structure to a new problem?”. You learned some bioinformatics problems and solutions for them. Probably from your other classes/experiences you know problems and solutions in other domains (for example, databases, networking, security, etc). Could you apply a technique from

this class to solve a problem in databases? Or, could you apply a technique from databases to solve a problem you learned in this class? Bioinformatics makes broad connections with other fields. If you are currently pursuing another research project as part of your Ph.D. or masters, could you make the connection?

The third category is more broad but can be summarized as "testing a hypothesis", that is, you try to reveal some new knowledge. Below are some examples:

- Papers often do not do a thorough evaluation of a method's performance. You can evaluate its performance in detail, testing how it behaves with respect to parameters. The hypothesis would be: the algorithm is sensitive to certain parameters.
- Some papers provide algorithms without implementations. You could implement the algorithm and evaluate its performance. The hypothesis would be: the algorithm performs well on certain datasets.
- When papers compare their tools to others, they often end up doing better than the competition. However, an independent and unbiased evaluation can offer surprising results. The hypothesis would be: the methods perform differently under a fair evaluation.
- We had several ideas of how to come up with a problem formulation that is well-formulated and useful. Many papers, however, skip this type of analysis and jump straight to the solution. You can take a problem we looked at in class or that was studied by one of the presented papers and define and/or analyze the problem formulation. The hypothesis would be: different formulations result in different results.
- When you evaluate the performance of an algorithm, you may consider to test: What is the observed running time and memory usage on simulated and/or real data? What is the accuracy of the results compared to the ground truth on real data? What is the robustness of the algorithms to its parameters and to violations of its assumptions on the input data? What are the boundaries of the parameter space and input data where the algorithm stops performing well?

Coming up with a research plan:

Once you have asked a research question, you then need to come up with some ideas to answer it. This is a creative endeavor and part of the learning experience.

- If your question falls in above category 1, then think about which part of the existing algorithm is the bottleneck and whether it can be replaced by a more efficient module. Also think about whether certain signal/information is not considered by the existing methods and you then extend the algorithm to take it into account.

- If your question falls in above category 2, then think about how the new application can be formulated in order to apply the existing algorithm, what assumptions need to be made, etc.
- If your question falls in above category 3, then design experiments that you think will lead to a convincing conclusion for your hypothesis.

These things are meant as ideas and guidelines. If you have an idea that falls outside the range of the above, that's great and I encourage you to pursue it. If you are not sure if it is appropriate, just talk to me after class. When you submit your project plan, I will give you feedback about the scope of your project so you can adjust it as necessary.

Project Plan:

Once you have chosen a project, you should prepare a 200-400 word project plan describing both your research question and your research plan/idea. Your project plan must set a clear deliverable for the end of the course.

Presentation:

You will make a short presentation near the end of class. Your target audience is your classmates and me. Though you will make your presentation before the project is due, you are expected to have results that show near-completion of the project.

Report:

The report should describe your project in detail, including the research question, ideas, results, and conclusions. It must be no more than 10 pages long, at least 1 inch margins all around and at least 11pt font.

Grading:

Your grade will be based on the content of your presentation (40%) and your report (60%). Some projects are more risky and challenging than others; these factors will be taken into account while grading (for example, category 1 requires more intellectual efforts than the other two). The grade of your presentation (40%) will again be a combination of classmates' evaluation (20%) and instructor's evaluation (20%). The evaluation of your report will be solely done by the instructor.

The evaluation of your project presentation will be primarily based on:

- whether your research question is interesting and impactful
- whether your research ideas are novel
- whether your research plan is convincing
- whether your results are solid and can support your conclusion
- clarity, coherence, and organization of your presentation
- stay close to the time limit

- quality of your answers to audience's questions
- clarity, coherence, and organization of your slides

The classmates' evaluation score will be collected immediately in class (via physical paper sheets). By the end of the day (11:59pm), you should submit your comments/feedback, including your critical judgements about the presentation (to justify your score), and your constructive suggestions, via google form (link: https://forms.gle/tgn5Xc4GvDkFPWXM8). Late feedback will NOT be accepted. Your evaluations and quality of your feedback is one of the major considerations for the class participation. The instructor will collect and send the anonymized comments/feedback to the presenter shortly.

Key Dates (tentative):

- **before 3/27:** discuss with me in-person
- **3/27**: project plan due
- **4/17**: project presentations
- **4/30**: report due

Useful resources

- UCSC genome browser: http://genome.ucsc.edu/cgi-bin/hgGateway
  - What do we know about a specific genome location? Look it up here.
- BLAT: sequence alignment: http://genome.ucsc.edu/cgi-bin/hgBlat
  - Does this sequence match anything previously known to us?
- Sequence and annotation downloads: http://hgdownload.cse.ucsc.edu/downloads.html
- Bioinformatics discussion forums
  - www.biostars.org
  - www.seqanswers.com
- More comprehensive list of resources: http://anil.cchmc.org/BioInfoRes.html
- Major conferences in the field are RECOMB and ISMB. You can view proceedings at:
  - http://www.informatik.uni-trier.de/~ley/db/conf/recomb/index.html
  - http://www.iscb.org/ismb-proceedings
- Popular journals are Genome Research , Bioinformatics
- "Hot" datasets
  - 1000 genomes: http://www.1000genomes.org/
    - § DNA sequencing data from over 1000 healthy individuals
  - Cancer Genomics Browser and Downloads: https://genome-cancer.ucsc.edu/