Computational Genomics 02-710/02-510, (Spring 2025) Course Project

---

Your class project is an opportunity for you to explore an interesting genomic analysis problem of your choice in the context of a real-world data set. **Each individual is responsible for their own project**, but you may help each other, provided you give appropriate attribution---you are the first author, but if classmates help, you should give them either middle-author credit, or acknowledgment credit.

For students in 02-710, your project accounts for 20% of your final class grade (recall that the problem sets will be 50%, and the midterms will be 30%), and will have four deliverables + peer feedback requirement:

- **Project proposal** (2-page maximum, including references), due **March 12**, worth 10% of the project grade.
- **Project mid-way report** (4-6 pages), due **April 2**, worth 10% of the project grade.
- **Poster presentation**, on **April 23**, in class, worth 10% of the project grade.
- **Final report** (8 pages maximum), due **May 2**, worth 60% of the project grade.
- **Poster peer feedback,** due **April 24**, worth 10% of the project grade. We will randomly assign every student to give feedback on three other posters.

For students in 02-510, a project is not required (so the problem sets will be worth 62.5%, and the midterms 37.5% of your final class grade), but attendance at the poster session and providing peer feedback is required:

- **Poster peer feedback,** due **April 24**, after the **April 23** poster session, worth **1% bonus** of your final class grade. We will randomly assign every student to give feedback on three other posters.

For students in 02-510 who *optionally choose* to do a project, we will compute your grade under both schemes above, and take the higher of the two. Thus, it is never disadvantageous to do the project.

Project Proposal:

You must turn in a brief project proposal (**2-page maximum including references**) by **March 12**. While you will not be bound by anything you write in the proposal, it is meant to help you start thinking about your project.

You are encouraged to come up with a topic directly related to your own current research project, but the proposed work must be new and should not be copied from your previous published or unpublished work. A possible “spin-off” project may include applying a method discussed in class to data you have been analyzing or developing a related method. Another option is to apply methods discussed in class to a new dataset that has not been analyzed using such methods.

Usually projects fall into one of these three basic categories:

- Applying a developed method discussed in class or related to methods we discussed in class to a new biological problem and/or new dataset.

- Developing a new method or data analysis approach.
- Benchmarking existing methods that address the same or similar problems.

Of course, some projects will fall into more than one category.

We've listed some prior year projects below to give you a sense of the types of projects that are in-scope. If you need some additional help picking a project and getting started feel free to discuss it with the course instructors.

Project proposal format:

- Project title
- Project lead (including Andrew ID)
- Project idea. This should be approximately two paragraphs.
- Software you will need to write.
- Papers to read. Include 1-3 relevant papers. You will probably want to read at least one of them before submitting your proposal.

Project Mid-way Report:

So that you can get early feedback on the projects, we are asking you to write a mid-way report, of somewhere between 4-6 pages, due by **April 2**. You should think of this as an early draft of your final report, based on the work you've done thus far, including an **Introduction**, **Methods**, and some **(proposed) Implementation Details** (see Project Report descriptions below for more details).

Although this is only worth 10% of your project grade, it is a good opportunity to get early feedback about the feasibility and scope of your proposed project.

Poster session:

We will hold a mandatory poster session on the last day of class, **April 23**, in which everyone will present the work they have done to the instructors, TAs and students in the class. This serves as a good opportunity for you to obtain feedback on your work and results in order to improve your final project report (see below). Thus, while we will not be grading the posters, the more you can present in terms of work and result the more likely it is that we can provide constructive feedback which will help you further improve your final writeup.

Because of space constraints, the format of the poster session will be to break the class into four groups/mini-sessions A, B, C, and D. Groups A and B will have their posters up in the first half of class, and Groups C and D will have their posters up in the second half of class. Furthermore, in the first quarter of class, Group A will be presenting their posters, in the second quarter of class, Group B will be presenting their posters, in the third quarter, Group C, and in the fourth quarter, Group D. We'll switch over the posters that are up at halftime of course.

This way, every presenter gets the opportunity to see all of the posters, and talk to 3/4th of the presenters. During the poster session, we'll also randomly assign every attendee three posters to give feedback to, due the following day, **April 24**. More details closer to the presentation day.

Project Report Expectations and Grading:

You must turn in a project report (**8 pages maximum)** by the end of day, **May 2**.

You could make use of some journal/conference templates to organize your report but make sure your report at least contains the following high-level structure, as this is how a typical scientific paper is structured. Your report will be evaluated on both **the quality of writing** (e.g., does it have a good structure, is it clearly written, was it proofread) as well as **the content** (effort in computational modeling and results). Latex usage is strongly recommended because it saves time in the long run. Your project should include the following:

- Project title
- Project authors (including Andrew IDs); first author should be the person responsible for the paper, but you should give co-authorship credit if someone else in the class helped substantially.
- **[25%] Introduction:** State the motivation, the problem you are addressing, and your approach for solving the problem. Use citations to provide an overview of the recent literature. It may be helpful to read a relevant review article.
- **[30%] Methods:** Explain your computational approach. Describe your model and learning/inference methods in two different subsections. Define all variables and include self-sufficient equations. Articulate new development.
- **[15%] Implementation details:** Give enough detail so the results can be reproduced by someone familiar with the field. Include a description of data processing steps and how you selected constants and/or free parameters (if applicable). Include pseudocode if implementing a new algorithm.
- **[15%] Results:** Provide informative figures and legends, tables, and any other results from your study.
- **[15%] Discussion/Conclusions:** Provide a summary of conclusions, limitations and future directions.
- References

Project suggestions:

Ideally, you will want to pick a problem in a domain of your interest, e.g., DNA sequence analysis, genetic polymorphisms, regulatory networks, epigenomics, single-cell analysis, etc., and formulate your problem in a statistical and/or machine learning framework. For example, you can adapt and tailor standard inference/learning algorithms to your problem, and do a thorough performance analysis. The titles of some projects from previous years are listed below.

- HoriGenT: A novel software to detect Horizontal Gene Transfer
- Investigating Structure in Gene Expression Data: Non-negative Matrix Factorization and other methods
- Comparative genomic analysis of single stranded RNA viruses
- A Workflow for Identifying Transcription Factor Directly from DNase Protected data
- Analysis of clustering methods for lung tissue miRNA

- Uncovering relationships between network topology and co-evolutionary signatures in ProteinProtein Interaction Networks
- Modeling Precision Treatment of Breast Cancer
- Comparison of Sepsis Time Series Gene Expression Data Classification
- Sequence Features of Translation Pause Sites and Slow Translation Regions
- Identifying Inherent Altruistic Biases in Human Genomic Studies
- Prediction of Optimum Sampling Points for Time Series Lung Development data
- Changes in Gene Expression due to Aging and their Relationship with Cancer
- Positive Selection in the Genomes of Humans and Chimpanzees
- Identifying Significantly Linked Proteins in HiC using ChIPSeq
- Improving performance of Random Forest in Clinical Feature Learning
- The Identification of Complementarity Determining Regions of Antibody Sequences

Below are slightly more expanded versions of project descriptions that have been proposed for prior years, so that you have more than just a title to go off of. You should **NOT** propose one of these projects, but you are naturally allowed to incrementally build off of one of the ideas, or to take it in a new direction---that is after all a large part of how science is done!

Discovering network motifs and recurring subgraphs from sequences of biological networks

Network motifs refer to recurring subgraphs and connectivity patterns in a single or multiple networks. They usually represent certain pathway components and bio-regulatory mechanisms, and their occurrence profiles are often unique to different networks and imply intrinsic functionalities of the biological networks. Early research in this area focuses on searching for small motifs in a single network. In this project we want to develop algorithms for searching large and possibly overlapping subgraphs recurring over multiple graphs. We will explore algorithms for constructing multiple networks, and graph theoretical approaches to mine such networks for motifs.

Reference:

Hu H, Yan X, Huang Y, Han J, Zhou XJ (2005) Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics (ISMB 2005), Vol. 21 Suppl. 1 2005, pages 213-221. Supplementary Material/Software

Zhou XJ, Kao MJ, Huang H, Wong A, Nunez-Iglesias J, Primig M, Aparicio OM, Finch CE, Morgan TE, Wong WH (2005) Functional annotation and network reconstruction through cross-platform integration of microarray data. Nature Biotechnology 2005 Feb;23(2):238-43.

E. Wong, B. Baur, S. Quader and C. Huang (2011) Biological network motif detection: principles and practice. Briefings in Bioinformatics, doi: 10.1093/bib/bbr033.

Protein function prediction from interaction network using graph theoretic and statistical latent-space modeling approaches

Local and global connectivities of an element in a network are often indicative of its functions; and such predictions often going beyond the traditional approaches that are based on physical and sequence properties biological element, but seeks a combination of such properties with its interaction contexts in biological processes, as reflected in the network, and such predictions can often be context-specific. In this project explore algorithms to infer biological functions of proteins from proteinprotein interaction networks and other protein attributes.

Reference:

E. Airoldi, D. Blei, E.P. Xing and S. Fienberg, A Latent Mixed Membership Model for Relational Data.Workshop on Link Discovery: Issues, Approaches and Applications **(LinkKDD-2005)**.

Dynamic Bayesian networks from time series microbiome datasets.
Time series microbiome data measures the levels of taxa at different time points in specific conditions. Using time series data we would like to learn a graphical model that represents the set of interactions and maybe causality events that take place between these microbial communities. In this project you will explore ways to use time series datasets for determining the structure and parameters of the network underlying the observed changes over time.

Reference:

J. Lugo-Martinez, D. Ruiz-Perez, G. Narasimhan and Ziv Bar-Joseph. Dynamic interaction network inference from longitudinal microbiome data. Microbiome, 7(1):54, 2019.
McGeachie, Michael J., et al. "Longitudinal prediction of the infant gut microbiome with dynamic Bayesian networks." Scientific reports 6 (2016): 20359.

Cancer pathway subtype analysis

Personalized medicine is already becoming a reality in cancer treatment. Signatures for cancer subtypes have been found in gene expression, epigenetic, and genome sequence data. In this project, you will explore the use of computational tools to identify cancer subtypes from various types of genomic data and to classify tumor data into subtypes.

Reference:

The Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. Nature 490: 61-70.

Gene network analysis

In this project, you will construct gene regulatory networks from genomic data. In particular, Gaussian graphical models have been extremely popular as a computational tool for constructing a gene network from gene expression data. You will explore different variants of Gaussian graphical models

to construct a gene network, to identify gene modules, and to interpret the learned network and modules.

Reference:

Grechkin et al. (2015) Pathway graphical lasso. AAAI 2015.

T. Wang et al. (2016) FastGGM: An Efficient Algorithm for the Inference of Gaussian Graphical Model in Biological Networks. PLoS Computational Biology, 12(2): e1004755.

Calculating the relative abundance of different transcript isoforms

Simulate RNAseq data from a variety of complex splicing scenarios and investigate performance of different quantification methods. You can either compare existing implementations or implement some strategies from scratch.

Reference:

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT & Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, Nature Biotechnology 2015

Identifying genomic regions under accelerated evolution

Use Rphast models and genomic multiple alignment to identify regions that show accelerated evolution in different species or species groups.

Resources:

Rphast is an R package found at: http://compgen.cshl.edu/rphast/

Reference:

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20: 110–121.

Pollard KS, Salama SR, Lambert N, Lambot M-A, Coppens S, Pedersen JS,Katzman S, King B, Onodera C, Siepel A, et al. 2006b. An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443

Pertea M1, Pertea GM, Salzberg SL. Detection of lineage-specific evolutionary changes among primate species. BMC Bioinformatics. 2011 Jul 4;12:274.

Predicting enhancers using genomics features

In recent studies, various computational methods have been developed to predict enhancers based on functional genomic features and/or DNA sequences. However, different features and models have been used in different studies. There is still limited knowledge of the general rule in accurate prediction of enhancer elements. In this project you are encouraged to use available datasets and existing models from the published papers to explore which features or combination of features are important for enhancer identification in a single cell type or across different cell types. You can also develop your own computational methods (e.g., DNN).

Reference:

Kleftogiannis, Dimitrios, Panos Kalnis, and Vladimir B. Bajic. "DEEP: a general computational framework for predicting enhancers." Nucleic acids research 43.1 (2014): e6-e6.

He, Yupeng, et al. "Improved regulatory element prediction based on tissue-specific local epigenomic signatures." Proceedings of the National Academy of Sciences 114.9 (2017): E1633-E1640.

Comparison and analysis of 3D genome inference from chromatin interaction data

Hi-C technology has been widely used in studying genome organization in the cell nucleus. Computational methods have been developed to identify chromatin interactions and topologically associating domains (TADs) from Hi-C data. However, the methods vary in their performance in different applications. In this project you can explore the advantage/disadvantage and differences of existing methods. You can compare the methods with analysis of their performance. You can also develop improvement or modification of an existing method or methods.

Reference:

Forcato, Mattia, et al. "Comparison of computational methods for Hi-C data analysis." Nature methods 14.7 (2017): 679.

Dali R, Blanchette M. "A critical assessment of topologically associating domain prediction tools." Nucleic Acids Res. (2017)

Predicting DNA/RNA structure

Much like protein, DNA and RNA has secondary and tertiary structure as well. A lot of methods have been developed to predict their structure based on sequence-level feature. In this project, you can make a comparison between the existing methods on their accuracy and efficiency. Also, you are encouraged to develop new computational methods.

Reference:

Steffen P, Voß B, Rehmsmeier M, et al. RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics, 2005, 22(4): 500-503.

Roll J, Zirbel C L, Sweeney B, et al. JAR3D Webserver: Scoring and aligning RNA loop sequences to known 3D motifs. Nucleic acids research, 2016, 44(W1): W320-W327.

Protein binding sites prediction

Machine learning algorithms have been developed to predict TF binding from genomic sequences. In this project, you are encouraged to develop a new framework to predict protein binding sites based on existing data. Evaluate your algorithm by comparing to ChIP-seq data and other annotations/resources.

Reference:

Sai Zhang, Jingtian Zhou, Hailin Hu, Haipeng Gong, Ligong Chen, Chao Cheng, Jianyang Zeng; A deep learning framework for modeling structural features of RNA-binding protein targets, Nucleic Acids Research, Volume 44, Issue 4, 29 February 2016, Pages e32,

Zeng H, Edwards M D, Liu G, et al. Convolutional neural network architectures for predicting DNA–protein binding. Bioinformatics, 2016, 32(12): i121-i127.

Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of DNA-and RNAbinding proteins by deep learning[J]. Nature biotechnology, 2015, 33(8): 831.

Identifying disease relevant SNVs

There are a large number of SNVs existing in the human genome, some of that leads to disease while others are just neural events. Many methods including machine learning frameworks have been developed to find out the disease driving SNVs. Develop acomputational method to predict whether a SNVs (coding or non-coding, or both) are relevant to disease.

Reference:

Jiaxin Wu, Yanda Li, Rui Jiang, Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies, PLoS Genetics, 10(3): e1004237, 2014

Dimensionality reduction of high throughput RNA-seq data

An RNA-seq experiment typically measures gene expression for thousands of genes at a time (mice and humans have around 20,000 genes). In addition to this, single-cell RNA-seq technologies have

made it possible to run tens of thousands of cells at the same time. Storage and analysis of this large amount of data is a challenge. Develop computational methods to reduce the dimensions of RNA-seq data and compare it with methods such as PCA or recursive feature elimination for tasks such as clustering, pseudo-time analysis, or cell-type inference.

Reference:

Chieh Lin, Siddhartha Jain, Hannah Kim, Ziv Bar-Joseph; Using neural networks for reducing the dimensions of single-cell RNA-Seq data, Nucleic Acids Research, Volume 45, Issue 17, 29 September 2017, Pages e156, https://doi.org/10.1093/nar/gkx681

CRISPR perturbation followed by single-cell RNA-seq

To study the genetic effects on gene expression regulation, CRISPR-Cas9 has been used to introduce genetic perturbations into cells, followed by single-cell RNA-sequencing on these cells. You can apply or develop computational tools to pre-process the scRNA-seq data and to identify the genes whose expression levels are affected by the variants introduced by CRISPR.

References:

Atray Dixit, et al., Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167.7 (2016): 1853-1866.