**Computational Genomics 02-710/02-510, (Spring 2020)**

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. Projects can be done by  **teams of two or three students**. Your project will be worth 25% of your final class grade, and will have three final deliverables:

* **Project proposal**(2 page maximum including reference list), due **March 23**, worth 10% of the project grade.
* **Poster presentation**, on **April 29**, worth 10% of the project grade.
* **Final report**(8 pages maximum), due **May 9**, worth 80% of the project grade.

**Project Proposal:**

You must turn in a brief project proposal (**2 page maximum including references**) by **March 23**.

You are encouraged to come up 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current research project, but the proposed work must be new and should not be copied from your previous published or unpublished work.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. Alternatively, we list a number of possible projects below as well.

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 pipeline.
* Benchmarking existing methods that address the same or similar problems.

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

You may pick a project from the list of potential project ideas bellow. Of course you can also choose to work on a new problem beyond our list. 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
* Team members (including Andrew IDs)
* 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

**Poster session:**

We will hold a mandatory poster session on the last day of class, April 29, in which each team 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.

**Project Report Expectations and Grading:**

You must turn in a project report (**8 pages maximum)**by end of day, **May 6**.

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
* Team members (including Andrew IDs)
* **[25%] Introduction:**State the motivation, the problem you are addressing, and your approach for solving the problem. Use citations to provide a 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.
* **[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.
* **[30%] Results and Conclusions:**   Provide informative figures and legends, a summary of conclusions, limitations and future directions.
* References

**Regarding printing the posters:**

SCS Computing Facilities has instituted a new procedure for printing posters. The new procedure is intended to make the process of poster printing faster and easier for the SCS community. There will no longer be a need to call Operations in order to print a poster. You can now submit posters via email, to poster@cs.cmu.edu. Simply follow the printing procedures that are documented on the SCS Help pages at: <http://www.cs.cmu.edu/~help/printing/platinum_printing.html> and Operations will print the poster and notify you when it is ready for pickup. Please contact SCS Operations at x8-2608 or send mail to help+@cs.cmu.edu with any questions or concerns. Also, the poster boards we use are 32"x 40" Non SCS students will need to contact their departments about resources for printing posters.

**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, 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 previous year's project are listed as bellow,

 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 Protein-Protein 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

You can also find some project ideas below.

**Project A: Haplotyping blocking and genetic demorgraphical inference**  
  
Genetic polymorphisms such as SNPs and Microsatellite carry important information of human evolution and disease propensity. One of the interesting problems in this area is to infer the haplotype of long sequence of ambiguous genotypes based on haplotypes of small overlapping regions. In this project we want to build a haplotype assembler using a partition-ligation scheme and/or a tiling scheme to stitch together short haplotypes inferred by off-the-shelf haplotype inference algorithm; and then, after determining long haplotypes of a long stretch of markers, find the best block structure using dynamic programming and information theoretic scoring. The resulting blocks will provide essential markers for mapping disease genes and for inferring the evolutionary history of given populations.   
  
Reference:  
  
Niu et al. [Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms](http://www.pubmedcentral.gov/articlerender.fcgi?tool=pubmed&pubmedid=11741196), Am J Hum Genet. 2006 Jan;78(1):174

**Project B: 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 motif 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](http://www.cmb.usc.edu/people/xjzhou/CODENSE.pdf). 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](http://www.cmb.usc.edu/people/xjzhou/nbt1058.pdf).  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](http://bib.oxfordjournals.org/content/13/2/202.full.pdf+html). Briefings in Bioinformatics, doi: 10.1093/bib/bbr033.

**Project C: 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 protein-protein 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](http://www.cs.cmu.edu/~epxing/papers/linkkdd05-12.pdf)**.**Workshop on Link Discovery: Issues, Approaches and Applications**(LinkKDD-2005)**.

**Project D: Dynamic Bayesian networks from time series microbiome datasets.**

Time series micorbiome 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 represent 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.

**Project G: 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](http://www.nature.com/nature/journal/v490/n7418/pdf/nature11412.pdf). Nature 490: 61-70.

**Project H: 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](http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9954/9918). AAAI 2015.   
  
T. Wang et al. (2016) [FastGGM: An Efficient Algorithm for the Inference of Gaussian Graphical Model in Biological Networks](http://www.ploscompbiol.org/article/fetchObject.action?uri=info:doi/10.1371/journal.pcbi.1004755&representation=PDF). PLoS Computational Biology, 12(2): e1004755.

**Project I: eQTL mapping**  
  
Expression quantitative trait locus (eQTL) mapping detects the genetic loci that control gene expressions by examining genetic variant and gene expression data from a large number of individuals. eQTL mapping can be used to study the genetic basis of diseases or the genetic control in different tissue types. This project explores the use of computational methods to determine eQTLs and the biological mechanisms influenced by the eQTLs.   
  
Reference:  
  
J. Becker et al. (2012) [.A systematic eQTL study of cis–trans epistasis in 210 HapMap individuals](http://www.nature.com/ejhg/journal/v20/n1/pdf/ejhg2011156a.pdf). Eur J Hum Genet, 20(1): 97–101.   
  
B. Stranger et al. (2012) [Patterns of Cis Regulatory Variation in Diverse Human Populations](http://www.plosgenetics.org/article/fetchObject.action?uri=info:doi/10.1371/journal.pgen.1002639&representation=PDF). PLoS Genetics, 8(4): e1002639.

**Project J: Population structure inference**  
  
Genome sequences contain information on ancestry, population evolution, and migration. In this project, you will analyze genome sequence data from a population to study the population structure and identify the ancestry information for each individual.   
  
Reference:  
  
B. Engelhardt, M. Stephens (2010) [Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis](http://www.plosgenetics.org/article/fetchObject.action?uri=info:doi/10.1371/journal.pgen.1001117&representation=PDF). PLoS Genetics, 6(9): e1001117.   
  
A. Raj, M. Stephens, and J. Pritchard (2014) [fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets](http://www.genetics.org/content/early/2014/04/14/genetics.114.164350.full.pdf). Genetics 197(2): 573-589.

**Project K: 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.   
  
Resources:  
  
Read-simulation software can be found at <http://deweylab.biostat.wisc.edu/rsem/README.html>   
  
Reference:  
  
Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT & Salzberg SL. [StringTie enables improved reconstruction of a transcriptome from RNA-seq reads](http://www.nature.com/nbt/journal/v33/n3/full/nbt.3122.html), Nature Biotechnology 2015

**Project L: 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](http://genome.cshlp.org/content/20/1/110.full.pdf+html). 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](http://www.nature.com/nature/journal/v443/n7108/abs/nature05113.html). Nature 443   
  
Pertea M1, Pertea GM, Salzberg SL. [Detection of lineage-specific evolutionary changes among primate species](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-274). BMC Bioinformatics. 2011 Jul 4;12:274.

**Project M: 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 sequence​s​. 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​ a​re 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 method​s (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.

**Project N: Comparison and analysis of methods in 3D genome reconstruction from Hi-C 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)

**Project P:** **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 a 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.

**Project Q:** **Protein binding sites prediction**  
  
M​achine 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 RNA-binding proteins by deep learning[J]. Nature biotechnology, 2015, 33(8): 831.

**Project R:** **​Identifying disease relevant SNV​s**  
  
There are a large number of SNVs existing in human genome, some of that leads to disease while others are ​just neural events.​ Many methods including machine learning framework​s​ have been developed to find out the disease driving SNVs. ​Develop a ​computational method to predict whether a SNV​s (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

**Project S: 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>

**Project T: CRISPR analysis**  
  
CRISPR-Cas9 is an exciting new method for editing genomes. It is based on a bacterial defense systems which inserts short segments of viruses into the bacterial genome and uses it to attack invading viruses using complementary base matching. CRISPRs (though not necessarily their goal or how they are used) were initially discovered by a computational analysis of bacteria genomes. In this project you would study the inserts for the same (or very similar) viruses in different bacteria strains. Using methods based on alignment as we discussed in class and more advanced RNA analysis try to determine if these inserts are the same (i.e. from the same virus sequence across all bacteria) or different (i.e. the same virus is represented by different sequences in different bacterial strains). Next you would try to explain what is unique about the specific virus segments that are retained by the bacteria and if they are different, what is common to all the different ones.