ECEN 766

Dwight Look College of ENGINEERING
TEXAS A&M UNIVERSITY

Algorithms in Structural Bioinformatics

Guidelines to Final Projects

Yang Shen
Department of Electrical & Computer Engineering
Texas A&M University

Timeline

- **By Feb. 19: Read** this guideline and attached doc of some project ideas; **Teaming**
- **Feb. 21-28**: **Meet** with instructor (15 minutes for each individual or each team*)
  Please make appointments through Canvas -> Calendar -> Find Appointment
  - Your ideas and my feedback; More meetings or emails follow as you need
- **Mar. 10**: 2-page **proposal** due (submit to Canvas) **10/40pts**
  - Problem definition, significance, current methods/research, remaining gaps to fill
  - Proposed methods/research, rationale, data, assessment, timelines and milestones
  - As many meetings / as much help as you need by then
- **Apr. 4**: 2-page half-way results **update** (submit to Canvas) **10/40pts**
  - Progress, barriers and new ideas, current results, expected outcomes and updated timelines
- **Apr. 23,25,30**: 15-min final **presentation** for each team* (including 2-min Q&A)
  - Introduction: define problems and identify challenges
  - Methods: describe your approach and the rationales
  - Actual / Expected results: assess/compare/benchmark your approach; Insights; Conclusions
- **May 3**: Final project **report** due (6-8 page paper format; see Canvas) **20/40pts**
- **May 8**: Grades due by 6pm and viewable from 10pm (Commencement next days)
- Afterwards: possibilities to extend to a conference / journal paper
  - *NeurIPS main & ICML Workshops* in May / *NeurIPS Workshops* in Sep.
  - ICML ML4MHD, IMLH; NeurIPS AI4Science, MLSB, AI4D3, DGM4H …
- Suggested team size: no more than three. Expectations increase as team sizes increase.

* Members with the best presentation(s) each gets **1 bonus point / 40pts**.

Thought Process

- ~~What final project does the instructor want?~~
- What final project would be **useful** to *you*?
  - To connect to present or future research
  - To gain research experiences
  - To strengthen a skill or expand your skill sets
  - To build your resume
  - …
- And interesting enough? (See Page 6)

Now choosing/pitching a topic

- **Application** (very broad)
  - Sequence analysis; Predicting protein/RNA secondary structure/fold/tertiary structures/properties, protein/DNA/RNA/drug interaction, protein function, mutational effects; protein/drug design; omics-data integration; Biomed imaging or sensoring; Network/System analysis … and Your own interests.
- **Algorithm** (very broad) / **Data Modality (text/image/graph/geometry)**
  - Alignment or database search for sequence data; optimization (LP, DP, combinatorial optimization, derivative-free optimization); machine learning [classification, regression, clustering, un/semi/self-supervised (incl. contrastive learning), discriminative vs generative, shallow vs **deep learning**, **NLP, CV, CNN/RNN/GNN/Language Models/Multimodal Representation Learning or Alignment/Generative AI**], UQ, Bayesian active learning; information theory; network analysis; systems steady-state/dynamics …
- Think from both perspectives and find a topic where they interact. If you have more thoughts in one perspective, bring it to me and I might be able to help with the other.
- There will be a working document of sample projects as well

2020 Journal Club Presentations

Who	Which	What Content	When*
ZO	1	1D Sequence -> Categorical Protein Folds (1D CNN)	Mar. 26
RD	2	2D Face Images -> Classifying Syndromes/Phenotypes (2D CNN)	Mar. 31
MG	3	2D Histopathology Images -> Diagnosis and Predicting Gene Mutational Status (2D CNN (Inception v3))	Mar. 31
NT	4	3D Optical Coherence Tomography (OCT) Images -> Diagnosis (3D CNN (3D U-Net))	Apr. 2
YY	5	3D Protein Structures -> Energy Functions (Energy-based model with 3D CNN)	Apr. 2
XZ	6	1D Protein Sequence -> 1D Backbone Torsional Angles (RNN)	Apr. 7
RA	7	EHR -> Sequence -> Classification (RNN(LSTM))	Apr. 7
FZ	8	1D Protein Sequences -> Unsupervised Representation -> Supervised Predictions for Structure and Function Properties (Transformer)	Apr. 9
RR	10	Desired transcript. Profiles -> Drugs in 1D seq. (conditional GAN)	Apr. 14
SZ	11	Desired property -> Drugs in graphs (RL + GCN for policy)	Apr. 16

What directions to pursue?

- Challenges in applications (lecture, shared papers, other references …)
- Emerging new algorithms (good at …)
- Let applications and algorithms inform each other!

What types of projects are interesting (and useful)

- Given an application problem, **benchmark** a few established algorithms (at least one new to the application) performances, analyze, derive **new insights** into the problem or solving the problem, and propose future directions.
- Given a challenge in an application problem, use **new insights or hypotheses** to propose a **new algorithm**, compare the results to old ones, analyze, and comment more on exploring and exploiting the new insights or hypotheses.

(You can use an established algorithm or software as reference, and just change one or few components based on your rationale developing a new algorithm)

(Find a paper with a dataset or maybe even source codes to work with)

Examples of a not-so-interesting project design and very interesting ones

- “I tried algorithm A for application B and it works” 😑
- “I found algorithm A works for application B because of the algorithmic component X addresses well the challenge Y from the application. And here is my justification.” 😊😊😊😊
- “To address the challenge Y from the application B, I introduced the algorithmic component X into algorithm A because I hypothesize that … The results improved which justifies my hypothesis / The results did not improve which suggests X did not address Challenge Y well or other confounding factors. This motivated me to try another component X’ or another method A’ because … and here’s the new conclusion. 😊😊😊😊😊

Past Projects: Summary
(including ECEN689 2015-17)

- 69 participants including 7 undergraduates
- 6 departments in the Colleges of Engineering, Science, and Medicine
- 55 course projects
  - 2015 AstraZeneca Sanger Drug Combination DREAM Challenge (4 students): Consortium paper published in Nature Communication
  - 2017 Kaggle Data Science Bowl / Image-based lung cancer detection (1 student): 108th/1972 as of April 24, 2017
  - Topics are evolving and I expect even more deep learning this year …
  - Opportunities this year to work with and publish with senior Ph.D. students (*NeurIPS & ICML Workshops* in May / *NeurIPS Workshops* in Sep.)

SUPPLEMENT: PAST PROJECTS

Past Projects (I)

- *“A survey in global optimization with multi-start”*
- *“Multithread performance comparisons of genetic, metropolis, and particle swarm methods*
- *“Improve the performance of HIV-1 protease cleavage prediction using prior knowledge”*
- *“Predicting HIV-1 protease cleavage site with different features of octamers”*
- *“Drug-target interaction prediction from genomic and pharmacological data”*
- *“AstraZeneca-Sanger drug combination prediction DREAM challenge”*

Past Projects (II)

- *“Classification of Glioblastoma Multiforme via machine learning”*
- *“Predicting molecular binding to Thrombin using machine learning methods”*
- *“RNA structural alignment through interaction networks”*
- *“Functional module identification through novel network querying”*
- *“Direct Coupling Analysis (DCA) in coevolution and protein contact prediction”*

Past Projects (III)

- *"Feature selection for high-dimensional data"*
- *"Prediction of acceptor sites in DNA Splicing"*
- *"Machine learning with mice protein data"*
- *"Robust Optimization for Classification"*
- *"A Survey on Missing Data Analysis"*
- *"Investigating feature encoding schemes for HIV-1 protease cleavage site prediction"*

Past Projects (IV)

- *"Appreciating the role of the membrane in the structural prediction of intermembrane proteins"*
- *"Bayesian top scoring pairs algorithm for feature selection in the high-dimensional biological data"*
- *"Identification of therapeutic targets in cancer through PBNs and Bayesian networks"*
- *"Data Science Bowl 2017"* (Lung cancer detection from images)
- *"Sequence alignment with solvent accessibility"*

Past Projects (V)

- *“Classification of highly correlated biological data using feature selection and feature extraction”*
- *“Prediction of hospital readmission with feature selection”*
- *“Conservation analysis of biological system using rank revealing methods”*
- *“A new Bayesian approach for protein docking”*
- *“Predicting protein function by associating functional inter-relationship”*

Past Project (VI)

- *“Developing improved mortality predictive models using EMR (Electronic Medical Records)”*
- *“Additional structural information for protein contact prediction”*
- *“Classification of microarray data”*
- *“Design of a chimeric peptide to inhibit uPA•uPAR”*
- *“A case study with continuous glucose measurements to minimize adverse glycemic events using targeted intervention delivery”*

Past Project (VII)

- *"Addressing Data Uncertainty in DeepSEA"*
- *"Comparison of Feature Selection Methods"*
- *"Automatic Diagnosis and Classification of Patient Status"*
- *"Prediction of Drug-Kinase binding affinity"* (2019 IDG-DREAM Drug-Kinase Binding Prediction Challenge)
- *"Drug Target Interaction prediction using network information"*
- *"Prediction of Molecular Consequences Induced by Missense Mutations Using Deep Generative Models"*
- *"Predicting Important Genes in the Type 1 Diabetes Disease Pathway"*

Past Project (VIII)

- *“Improved Multi-Layer Multi-Leaf Collimator Optimization”*
- *“Cross Modality Learning on Proteins”*
- *“Pre-Ictal Seizure Forecasting with Scalp EEG features using Majority Vote Classifiers”*
- *“Diagnose Pneumonia Based On Chest X-Ray Images Using Deep Learning”*
- *“The T-F-M Battle”*
- *“Network-based Gene Classification: Observations and Discussions”*

Past Project (IX)

- *“Adaptive Feature Selection for Human Activity Recognition”*
- *“Detecting Breast Cancer using Convolutional Neural Networks”*
- *“Identifying Mutated Nuclear Receptor Ligands through Sequence-Based Deep Learning”*
- *“Graph TALE” (GNN-based protein function prediction)*
- *“Human-Virus Protein Protein Interaction Prediction using Self-Supervised Graph Learning Methods”*
- *“A Comparison between Joint Sequence(1D)-Fold(3D) Embedding-based Generative Models for Protein Design”*
- *“Droplet Microfluidics Detection, Classification, and Segmentation using Different Deep Neural Networks”*

(Update) Spring 2024 Final Projects

- *Attacks and Defenses in Privacy-Preserving Machine Learning*
- *Predicting Drug Response in Breast Cancer Using Deep Learning Neural Networks*
- *Interprets and Predicts Single-Cell Responses to Perturbations*
- *Utilizing DNA Language Models to Optimize CROTON (CRISPR Gene-Editing Outcome Prediction)*
- *Latent Diffusion for Protein-Ligand Docking*
- *Drug Docking Augmentation Using Energy Based Models*
- *Protein Residue Annotation with Language Model*
- *mRNA Vaccine Design for Infectious Diseases*
- *Enhancing Protein Function Prediction Using Sequence and Structure Information*
- *Variant Effect Predictions Leveraging Structure Information*
- *Classifying Clinically Relevant Variants using a Protein Language Model and Secondary Structural Information*