# Form A: Initial Class Project Proposal

[1]**Due: Oct 05.** Please submit as a PDF file to Gradescope.
See "Initial Project Proposal" section in the class project guide for further info.

Graduate students (registered under BIOCB 6840):

- Must use Form A.
- Must do original research.
- One-on-one meeting with the instructor about the class project is not required but can be requested if desired.
- The project must be relevant to the course.

Undergraduate students (registered under BIOCB 4840/CS 4775):

- If you have a rough project idea, use Form A.
- If you're still unsure about the project topic, use Form B.
- No expectation of original research.
- If you're in a research group, one-on-one meeting with the instructor about the class project is not required but can be requested if desired.
- See the guideline for the potential project types.

1. Name(s) and NetID(s):

Esther Yu (yy465)
Hanqing Li (hl698)
Jingyu Xu (jx62)
Rainney Wan (rw476)
Sicheng Ma (sm2287)

2. (Tentative) Project title:

Comparison of DNA motif clustering with HMM, Bayesian, Tree-based Algorithms, and Expectation maximization on TF-binding DNA Dataset

3. Base paper(s):

1, Hammock: a hidden Markov Model based peptide clustering algorithm[1]
2, A novel Bayesian DNA motif comparison method for clustering and retrieval [2]
3, TreeCluster: Clustering biological sequences using phylogenetic tree[3]
4, Bailey, T.L., Elkan, C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization[4]

4. Project keywords (minimum 5):

Motif clustering, phylogenetic tree, Bayesian statistics, Expectation Maximization, Unsupervised Learning, HMM, DNA motif, TF

5. Brief project description (minimum 100 words):

Over the decades, next-generation sequencing techniques have been used to produce large DNA datasets containing millions of DNA tags. To understand the biological context of

the sequencing results, bioinformaticians have developed numerous algorithms for motif discovery and clustering. In this project, we will compare between three different approaches regarding their clustering results on a TF-binding ChIP dataset. Specifically, we will implement motif clustering algorithms based on HMM, Bayesian statistics, and phylogenetics tree.

The HMM algorithm, Hammock[1], was originally developed for peptide clustering. Here, we will apply this towards DNA sequence data. Hammock's main framework was to initialize small clusters, extend each cluster based their profile HMM, merge clusters. The second and third step are conducted repeatedly.

While the Hammock paper applied HMM-HMM similarity measurement in merging clusters, we might also compute the similarity between motif clusters using Bayesian approaches[2]. Continuously computing pair-wise similarity score and merging highly similar motifs would finally generalize all the motifs into a phylogenetic tree, which is suited for clustering.

The expectation maximization (EM) algorithm is included in this course. Applying unsupervised learning to EM can extends EM for discovering new motifs in a set of biopolymer sequences where little or nothing is known in about any motifs( even unaligned). It is interesting to compare the results from using unsupervised learning on unaligned subsequences and using EM on known sequences.

We have also learned other tree construction algorithms in our class, and each tree could be further processed into clusters of motifs[3]. It would also be interesting to compare different clustering results derived from different tree construction algorithms.

(1) Krejci, A.; Hupp, T. R.; Lexa, M.; Vojtesek, B.; Muller, P. Hammock: A Hidden Markov Model-Based Peptide Clustering Algorithm to Identify Protein-Interaction Consensus Motifs in Large Datasets. *Bioinformatics* **2016**, *32* (1), 9–16. https://doi.org/10.1093/bioinformatics/btv522.
(2) Habib, N.; Kaplan, T.; Margalit, H.; Friedman, N. A Novel Bayesian DNA Motif Comparison Method for Clustering and Retrieval. *PLoS Comput Biol* **2008**, *4* (2), e1000010. https://doi.org/10.1371/journal.pcbi.1000010.
(3) Balaban, M.; Moshiri, N.; Mai, U.; Jia, X.; Mirarab, S. TreeCluster: Clustering Biological Sequences Using Phylogenetic Trees. *PLoS ONE* **2019**, *14* (8), e0221068. https://doi.org/10.1371/journal.pone.0221068.
(4) Bailey, T. L.; Elkan, C. Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization. *Machine Learning* **1995**, *21* (1), 51–80. https://doi.org/10.1007/BF00993379.