Seo et al., "DeepFam: deep learning based alignment-free method for protein family modeling and prediction", *Bioinformatics* 34:i254-i262, 2018.

Session Intro

The session will discuss the paper by Seo et al., which presents a deep learning-based method for protein function prediction. The results are exciting (circa 97% accuracy), seemingly implying that the protein function prediction problem is fully solved… and journals/reviewers can start throwing away submissions on protein function prediction without review.

I hope this catches your attention ☺, and read (and more importantly, think about) the Seo et al. paper carefully.

Session Plan

I am dividing the session into three parts as given below.

Although I have provided some possible topics/pointers for each aspect, I leave each presenting team to decide on what they want to talk about (i.e., it is perfectly ok to leave out some topics/points and/or include other topics/points.) Also, the presenting team need not just make presentations; they are encouraged to figure out how to engender more class interactions and lead discussions.

Part I, Background information:

This part deals with the background knowledge needed to understand the paper. I have highlighted some keywords below for you to look up background literature, Wikipedia, etc.

- Protein family
- Sequence similarity, homologs, orthologs, COG datasets
- Guilt by association approach to protein family/function prediction & modeling
- Alignment-free approach to protein family/function prediction
- Deep learning
- Accuracy, sensitivity, specificity, and other measures of prediction performance

Presentation team #1: AISHWARYA JAYAGOPAL, CAO XIAO, LIU ZHUANGHUA

.

The team members can decide who present what. Each presentation should be brief (say 5 minutes.) You can focus on any aspect of the topic, with the objective of making it easier for the class to understand the Seo et al. paper. **Total time limit: 15 minutes (presentation) + 5 minutes (audience questions.) Total slide count: 10 slides max.**

Part II, The paper by Seo et al.

This part presents the Seo et al. paper itself. We want to know some key technical details and key messages of the paper, such as:

Details of how Seo et al.'s DeepFam works

Details of how they validate DeepFam

What the main findings of Seo et al. are

Presentation team #2: MALAIKA AFRA TAJ, DAI YUHE, LEE JIANYI DAVID

I leave you to decide how you want to organize the presentation. **Total time limit: 15 minutes (presentation) + 5 minutes (audience questions.) Total slide count: 10 slides max.**

Part III, Possible points for discussion

This part discusses Seo et al. paper, hopefully in depth. We want to know whether there is any methodological issue, any doubt on the conclusions/key messages, any suggestion for improving the paper. Some pointers for discussion include:

- Any methodological issue? E.g., is a multi-class classifier (which assigns a query protein into one of the possible class labels) an appropriate framework for protein family/function prediction?

- Any issue in the validation? E.g., problems with the validation datasets, the performance metrics used, or how the validation is done.

- Any issue on the key messages? What can be done better?

Presentation team #3: CHENG YI, MD SALMAN SHAMIL, WANG JINGTAN

I leave you to decide how you want to organize the presentation. **Total time limit: 15 minutes (presentation) + 5 minutes (audience questions.) Total slide count: 10 slides max.**