Final Project - Mathematical Foundations of Quantitative Biology II

Background: Co-evolution has become a powerful tool for understanding protein function, structure, and regulation. The basic premise behind this approach is that correlated evolution between sequence elements – often individual amino acids – can be used as an indicator of a shared functional or structural constraint. For example, particular protocols for analyzing protein sequence co-evolution can be used to predict physical contacts in protein tertiary structure (e.g. Morcos et al. (2011) PNAS v.108(49):E1293), and the inclusion of co-evolutionary information is a key aspect of AlphaFold. More recently, co-evolution has been used to predict physically interacting protein domains.

In your project, you will extend co-evolution based methods to consider both physical and functional interactions between proteins. The goal is to examine if the pattern of co-evolution is distinct amongst physical binders, functional (but non-binding) interactors, and true non-interactors. You will be given a "test set" of four protein domains, where the protein family identifiers have been removed. While you might readily figure out the protein identity (with standard informatics tools), I encourage you to leave it a mystery till the end of class.

In the course of completing your project you will:

1. Curate alignments for the group of four proteins and learn to compute basic alignment statistics and measures of sequence identity.

2. Examine co-evolution through the lens of covariance and mutual information. What is the distribution of amino acid co-variance and mutual information within and between protein domains?

3. Use your co-evolutionary calculations to make some predictions. Given the computed patterns of co-variance and mutual information, can you statistically distinguish between interacting and non-interacting protein domains?

This project is intentionally somewhat open-ended, and we hope you will approach it with creativity. It is a true research project, in the sense that the answer is not already known. Instead, the goal is to learn something new about the pattern of coevolution between binders, biochemically interacting domains, and non-interactors. This is a team project, and we encourage you to form teams of 2-3 people (not more).

Please let Kim know your team composition by sending her an email by Wednesday Jan 10.

The central goal is to apply the mathematical measures learned in class to understand the proteins of interest… not to simply learn how to apply extant code packages. That said, you are welcome to look to the literature for inspiration (see: https://doi.org/10.1038/nrg3414 and https://doi.org/10.1101/cshperspect.a041463 for a couple of reviews). It is acceptable to use existing code or libraries within your own analysis. Your code can be in MATLAB, Python, Julia, C... using whichever language you feel most comfortable in and where you can work efficiently.

Check-ins, Presentations, and Review

Jan 17 – first project check-in. Data curation and calculation of basic alignment statistics. You have been given four alignments. For each alignment, please do the following:

- Convert the alignment from letters to numbers (i.e. create a one hot encoding)
- For each alignment, first remove highly gapped (more than 50%) positions and then highly gapped (more than 50%) sequences
- For each alignment, report the number of sequences and number of positions following gap removal
- For each alignment, plot a histogram of the average pairwise sequence identity. Which protein families are most conserved? Least conserved? Is there evidence of strong phylogenetic structure (sequence clades) in some families but not others?

Feb 14 – second project check-in. Analysis of covariance between amino acid positions.

- Decide how you will compute covariance. Between individual amino acid types, or physicochemical classes? How will you compress covariance between amino acid types to obtain a positional measure?
- Compute co-variance between all position pairs within each protein.
- Concatenate the alignments by species. You will need to make one concatenated alignment per protein pair.
- Compute co-variance between all *position* pairs across *protein* pairs.
- Plot the resulting covariance matrix as a heatmap.
- Plot histograms of the covariance within and between individual domains. Visually inspect these distributions – are they overlapping? Is covariance stronger within or between domains? How do the tails (strongest signal) of the distributions compare?

Feb 21 – third project check-in. Analysis of mutual information between amino acid positions.

- Decide how you will compute mutual information (MI). Between individual amino acid types, or classes? How will you compress covariance between amino acid types to obtain a positional measure?
- Compute MI between all position pairs within each protein.
- Using the concatenated alignments, compute MI between all protein pairs.
- Plot the resulting MI matrix as a heatmap.
- Plot histograms of MI within and between individual domains. Again, visually inspect and describe these distributions.

The final project presentations will be on February 28 and March 4. Final presentations will be 20 minutes long with 10 minutes for questions. The goal is now to examine if covariance and/or mutual information can be used to predict protein interactions. Choose an appropriate statistical testing strategy to decide whether the co-evolution between two proteins is significant. Consider carefully your null model.

To facilitate critical thinking, and invite constructive feedback, we will assess all projects through an in-class review on *March 4*. Each project will be assigned three or more peer reviewers. These reviewers will assign scores in the following categories:

- Technical soundness and execution. Are the methods used appropriately? Did the authors complete the work they set out to do? Are the results interesting?
- Innovation. Is the work creative? Does it make use of new ideas/concepts/approaches?
- Presentation. Was the work clearly presented? Was the use of plots/graphics appropriate and well-described?

Scores will be assigned on an NIH-like scale of 1-10. A score of 1 corresponds to "outstanding", 4-5 indicates "very good", and 10 indicates an area that needs substantial work or revision. The three reviewers will present their scores in class on March 4, along with their rationale – these presentations should be intellectually rigorous, provide specific feedback, and above all be constructive. After each reviewer presents their scores and rationale, all students in the class will submit a final score to the instructors. The instructors will use these peer scoring sheets alongside their own evaluations in assigning a final grade. The final project presentation and your participation in the review process and check-ins will be worth 30% of your grade.