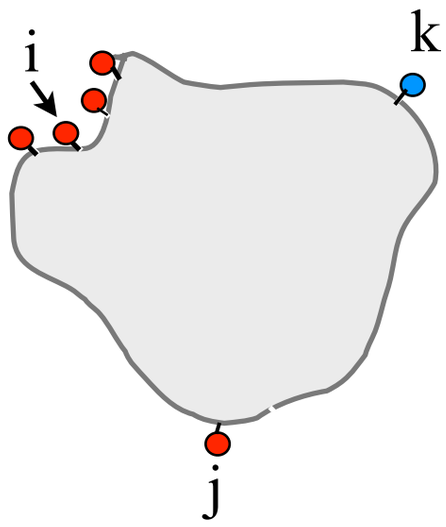
Course Project: Co-evolution as a tool for predicting protein-protein interactions.

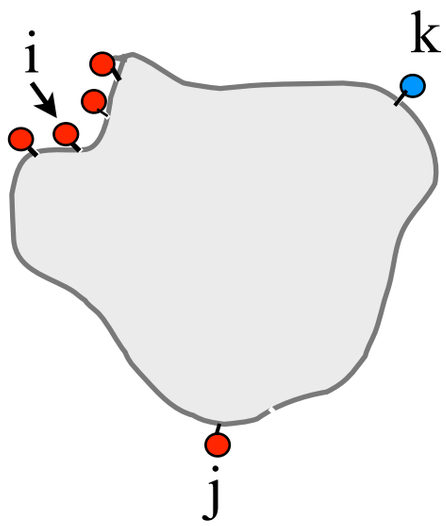


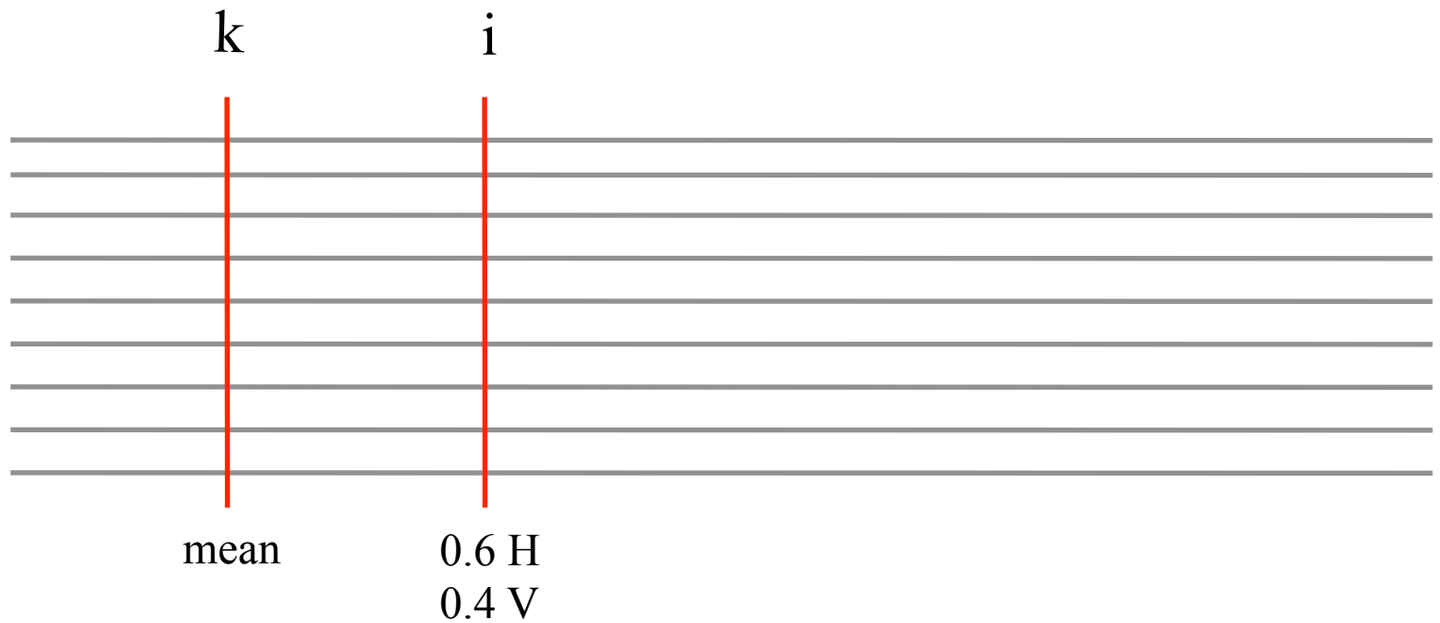
Protein **Multiple Sequence Alignment**

The basic approach:

Comparison of sequences across many species can be used to make a ***statistical model*** for the design of proteins.

<u>Course Project</u>: Co-evolution as a tool for predicting protein-protein interactions.

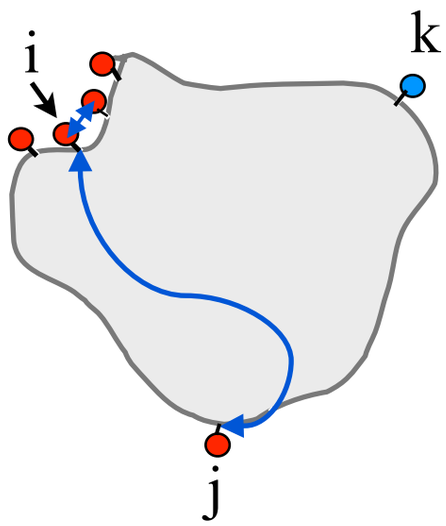


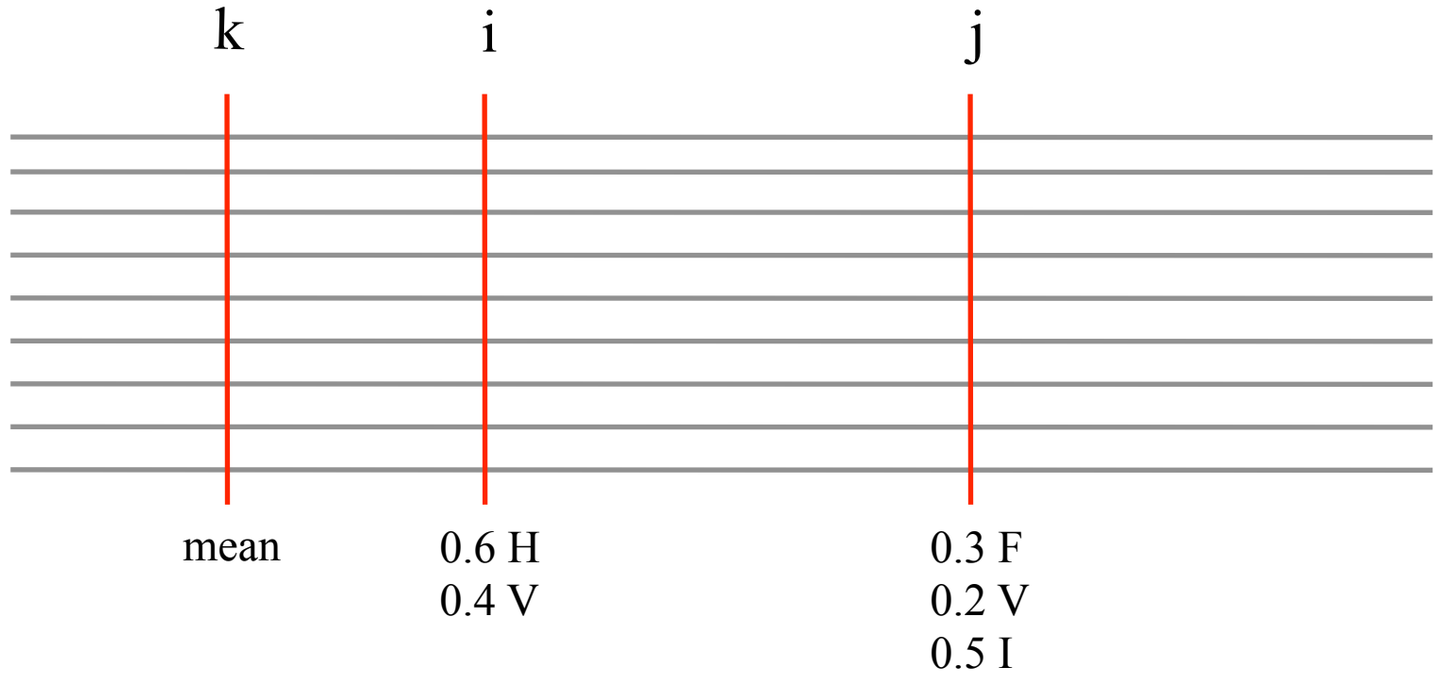


The basic approach:
Comparison of sequences across many species can be used to make a ***statistical model*** for the design of proteins.

1. Statistical invariance as a measure of importance (evolutionary **<u>conservation</u>**).

<u>Course Project</u>: Co-evolution as a tool for predicting protein-protein interactions.





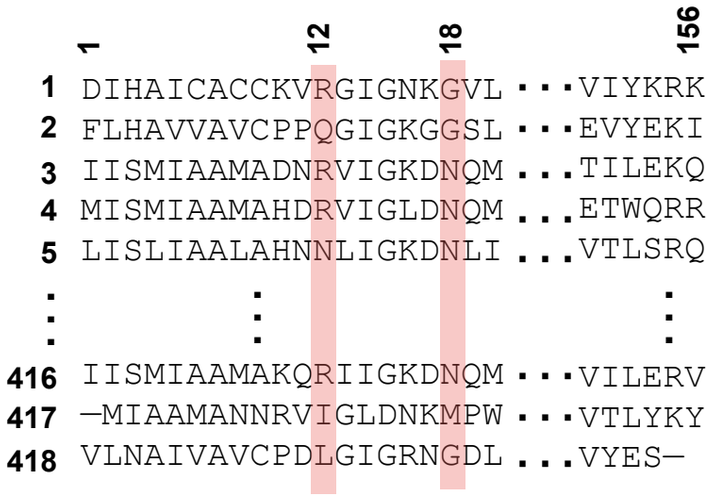
The basic approach:
Comparison of sequences across many species can be used to make a ***statistical model*** for the design of proteins.

1. Statistical invariance as a measure of importance (evolutionary **<u>conservation</u>**).

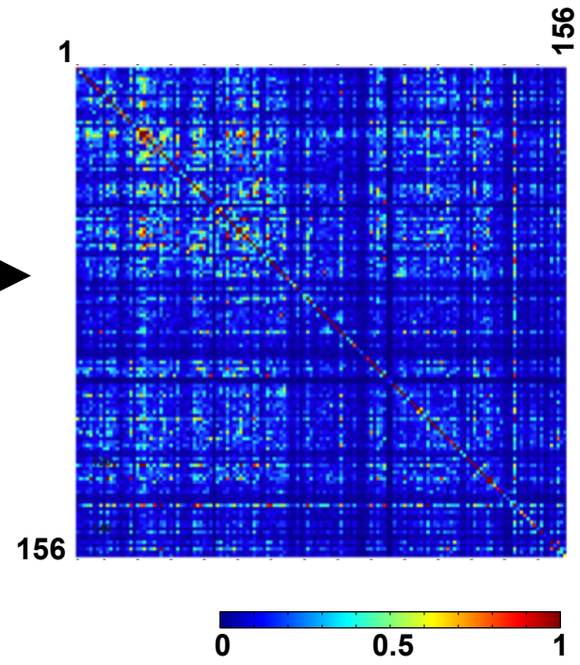
2. Correlation over sequences as a measure of interaction or shared functional constraint (**<u>co-evolution</u>**).

The basic analysis framework starts with constructing and analyzing an amino acid interaction matrix

1) Multiple Sequence Alignment



2) Interaction matrix



3) Analysis of the interaction matrix

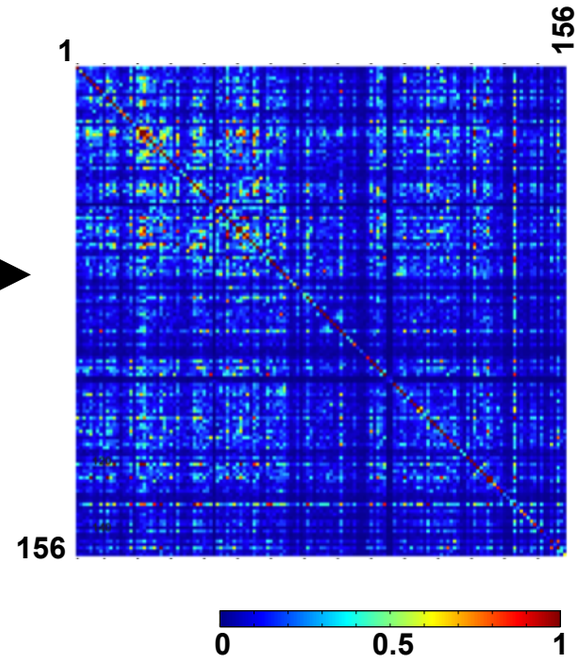
The basic analysis framework starts with constructing and analyzing an amino acid interaction matrix

… and can vary in three key ways.

1) Multiple Sequence Alignment

(1) Unit of co-evolution
(Individual amino acids, groups of AAs)

2) Interaction matrix

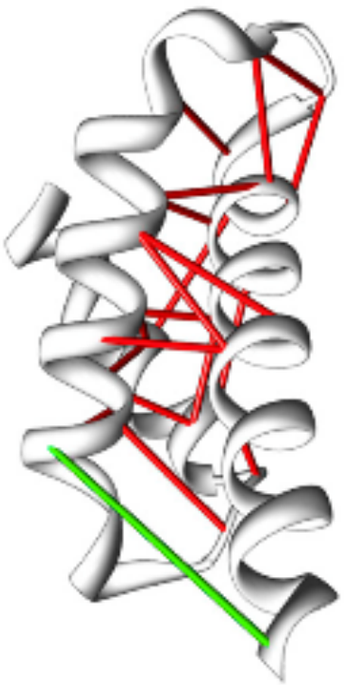


(2) Measure of co-evolution
(co-variance, mutual information, some weighted form of these, etc.)

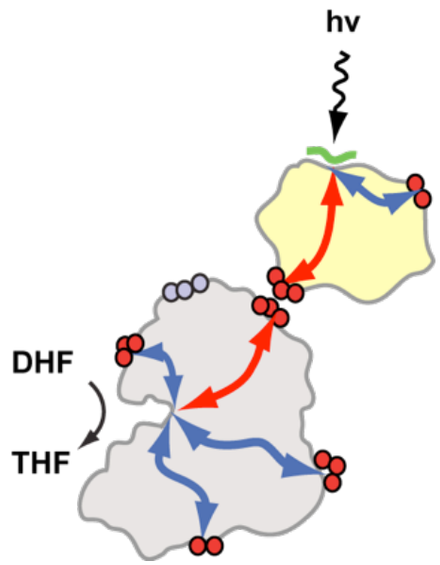
3) Analysis of the covariance matrix

(3) Strategy for identifying statistically significant interactions.
(Null model generation, PCA, maximum entropy models)

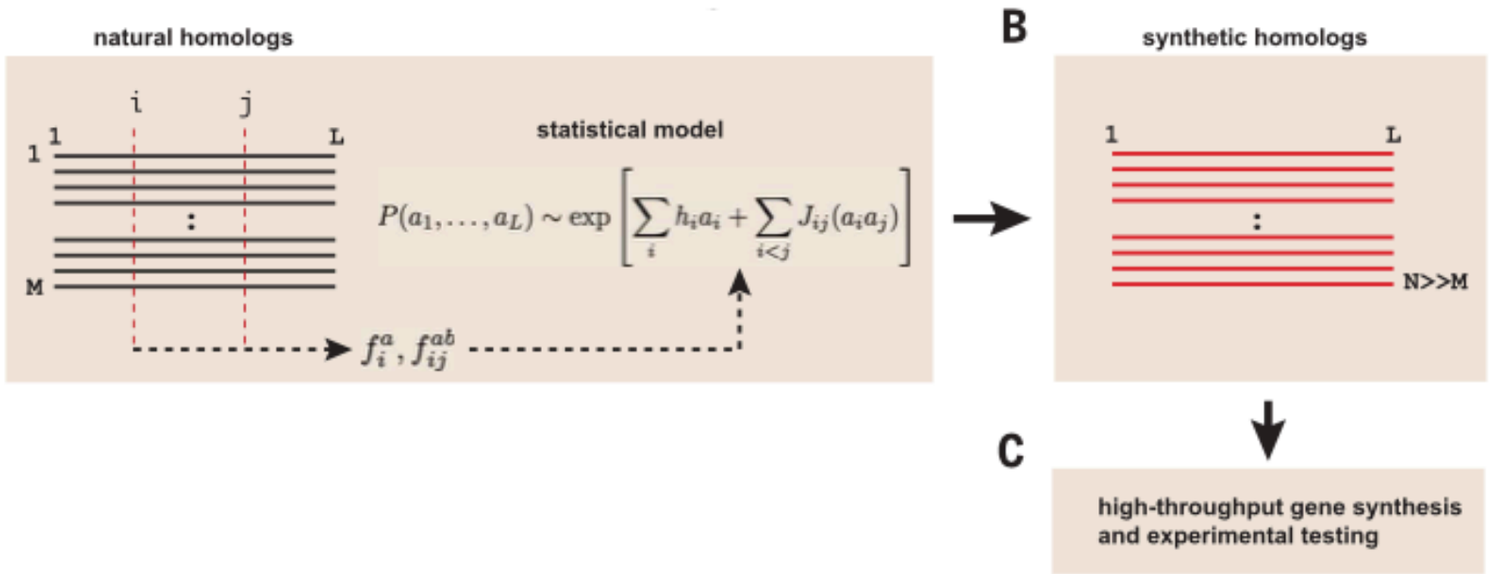
Recent work has used co-evolution to predict contacts within and between proteins, to understand (and engineer) allostery, and to design enzymes.



Contact prediction:
Morcos et al. (2011) PNAS. 108:E1293
Stein et al. (2015) PloS Comp Biol 11:e1004182
Cocco et al (2018) Rep Prog Phys 81:032601

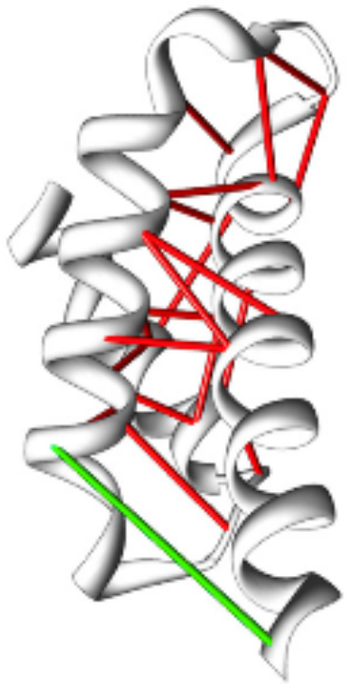


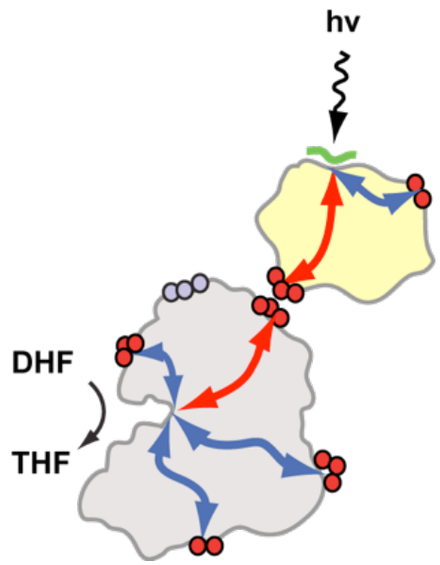
Allostery design:
Reynolds et al. (2011) Cell v.147: 564
Pincus et al. (2018) Science Signaling v11: 555



Enzyme design:
Russ et al. (2020) Science v. 369: 440

Recent work has used co-evolution to predict contacts within and between proteins, to understand (and engineer) allostery, and to design enzymes.

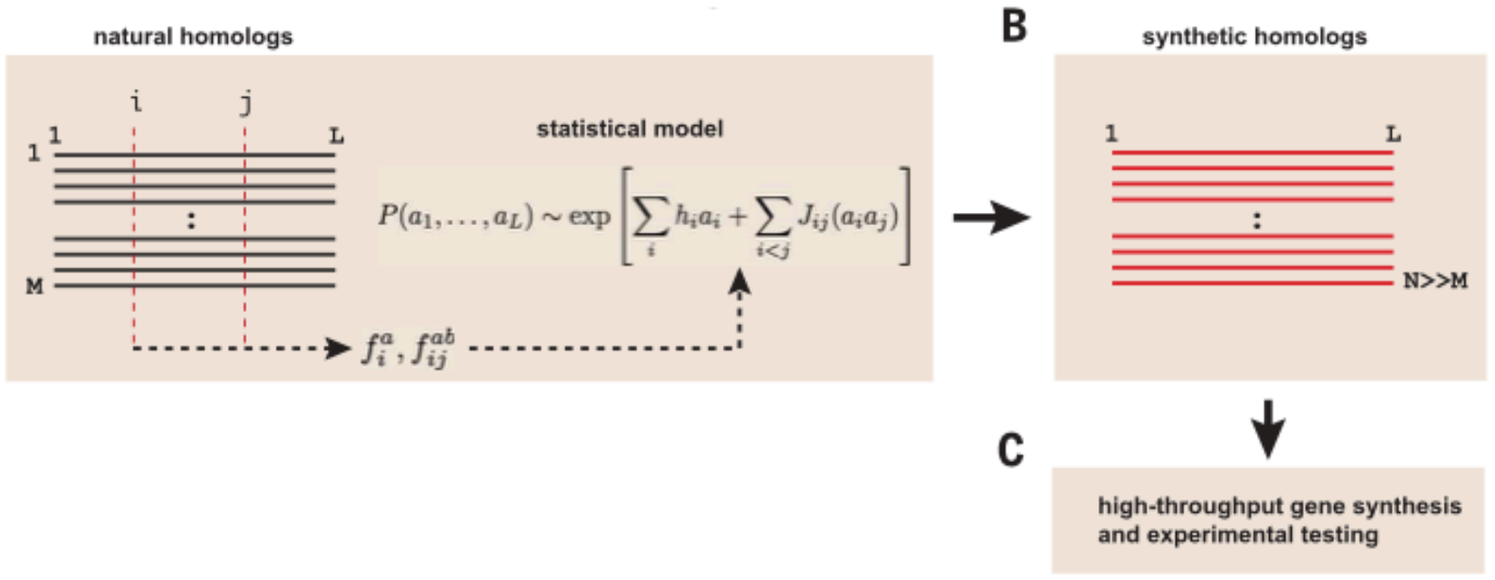




Allostery design:
Reynolds et al. (2011) Cell v.147: 564
Pincus et al. (2018) Science Signaling v11: 555

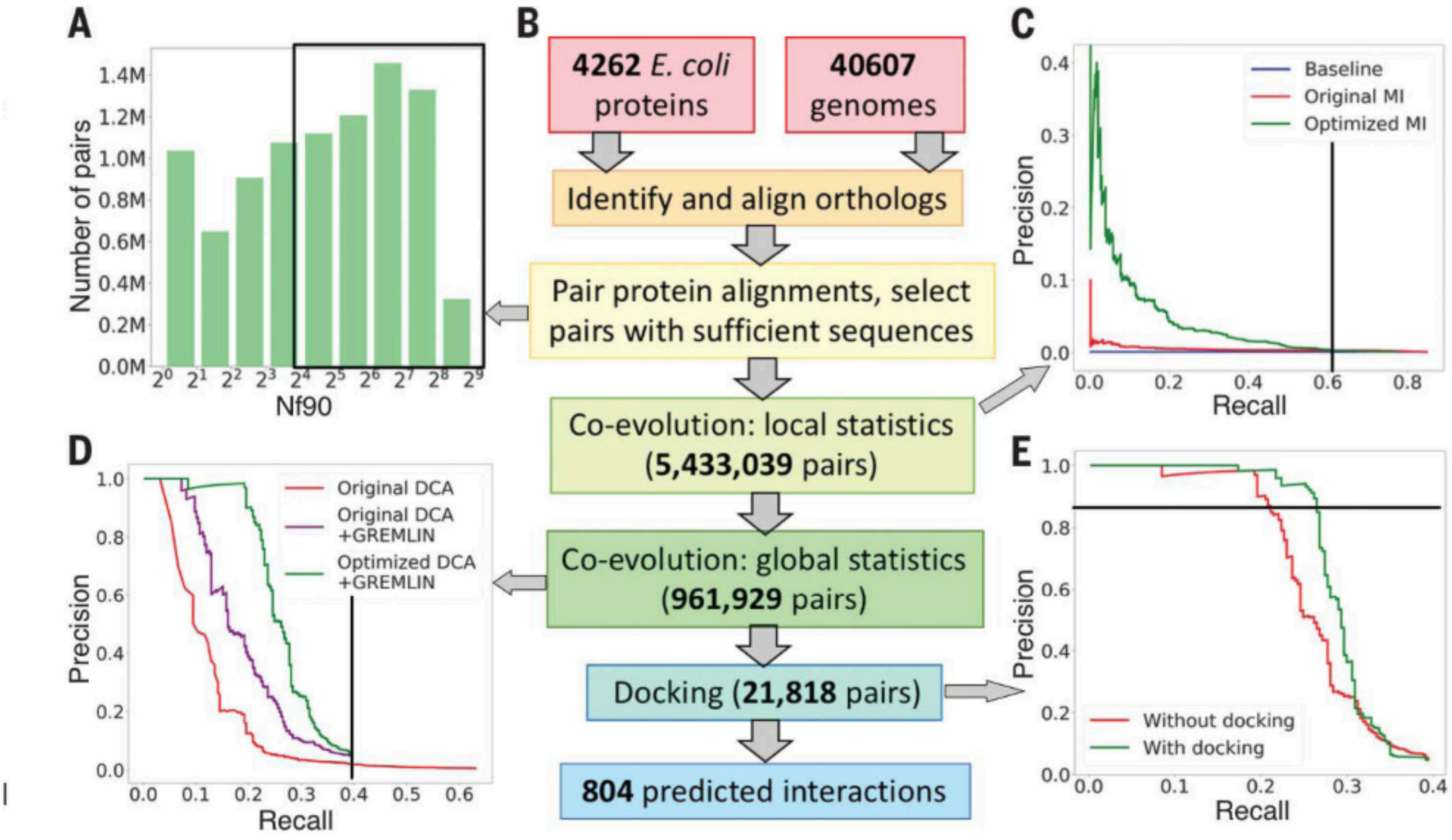
Upcoming seminar speaker!
Faruck Marcos, Jan 22 11AM. ND11.218 (this room)

Contact prediction:
Morcos et al. (2011) PNAS. 108:E1293
Stein et al. (2015) PloS Comp Biol 11:e1004182
Cocco et al (2018) Rep Prog Phys 81:032601



Enzyme design:
Russ et al. (2020) Science v. 369: 440

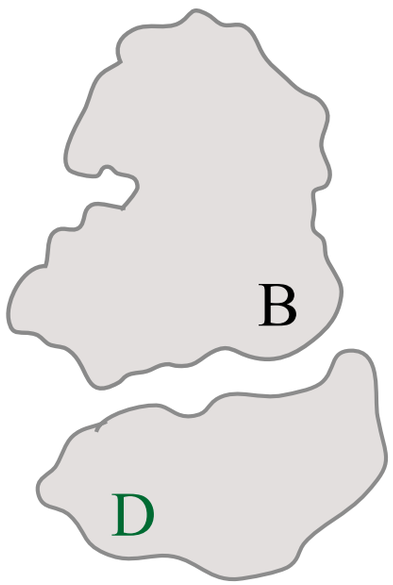
Co-evolution has also been used to predict physical protein interactions:



Cong et al. (2019) Science. 365:185

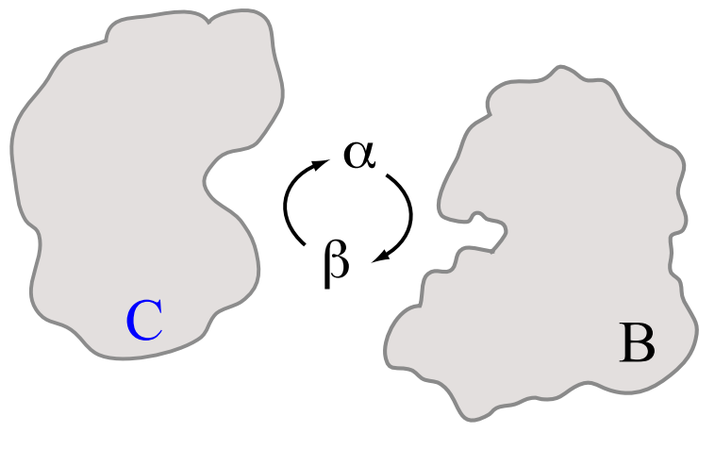
Thus far, the focus has been prediction of physical interactions, but what *biochemical* or *functional* interactions?

Physical protein-protein interaction:



Coupling mediated by direct binding.

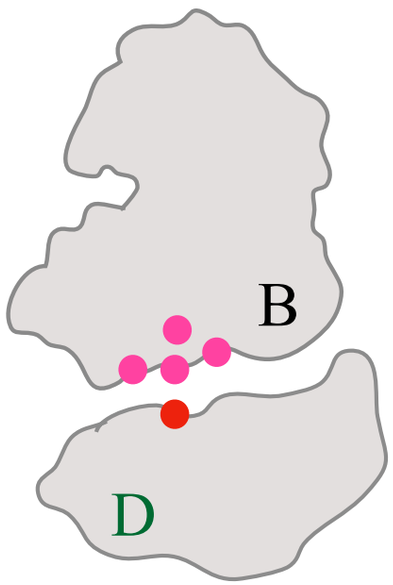
Biochemical protein-protein interaction:



Coupling mediated by a need to not accumulate intermediates and/or deplete key products

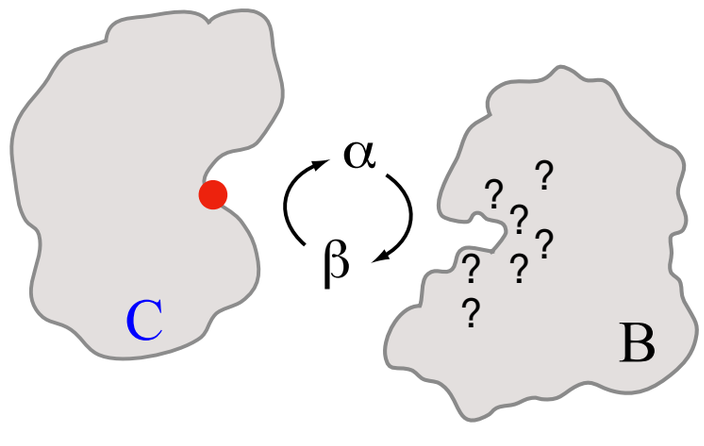
Thus far, the focus has been prediction of physical interactions, but what *biochemical* or *functional* interactions?

Physical protein-protein interaction:



Coupling mediated by direct binding.

Biochemical protein-protein interaction:



Coupling mediated by a need to not accumulate intermediates and/or deplete key products

This is the subject of our course project.

What does the co-evolutionary signal look like between a pair of physically interacting domains, vs. a pair of biochemically interacting domains?

<u>The dataset</u>: Multiple sequence alignments for four proteins across bacteria. The proteins have been renamed A, B, C, D.

While you might readily figure out the proteins (with standard informatics tools)…
I encourage you to leave it a mystery till the end of class.

Two of these bind and catalyze consecutive reactions in metabolism.

Two of these catalyze consecutive reactions in metabolism but do not bind.

The two interacting pairs (one physical and one biochemical) are in distinct cellular processes and not expected to interact.

What will the pattern of co-evolution look like?
Can you annotate the interacting and non-interacting pairs?

Some Guidelines:

The central goal is to apply the mathematical measures learned in 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 example 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 (one hot encoding)
- For each alignment, remove highly gapped (more than 50% gapped) positions and 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

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 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 protein pairs.
- Plot the resulting covariance matrix as a heatmap.
- Plot histograms of covariance within and between individual domains.

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.

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.

The review process:

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 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 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 opinions in assigning a final grade. The final project presentation and your participation in the review process and check-ins will be worth 20% of your grade.

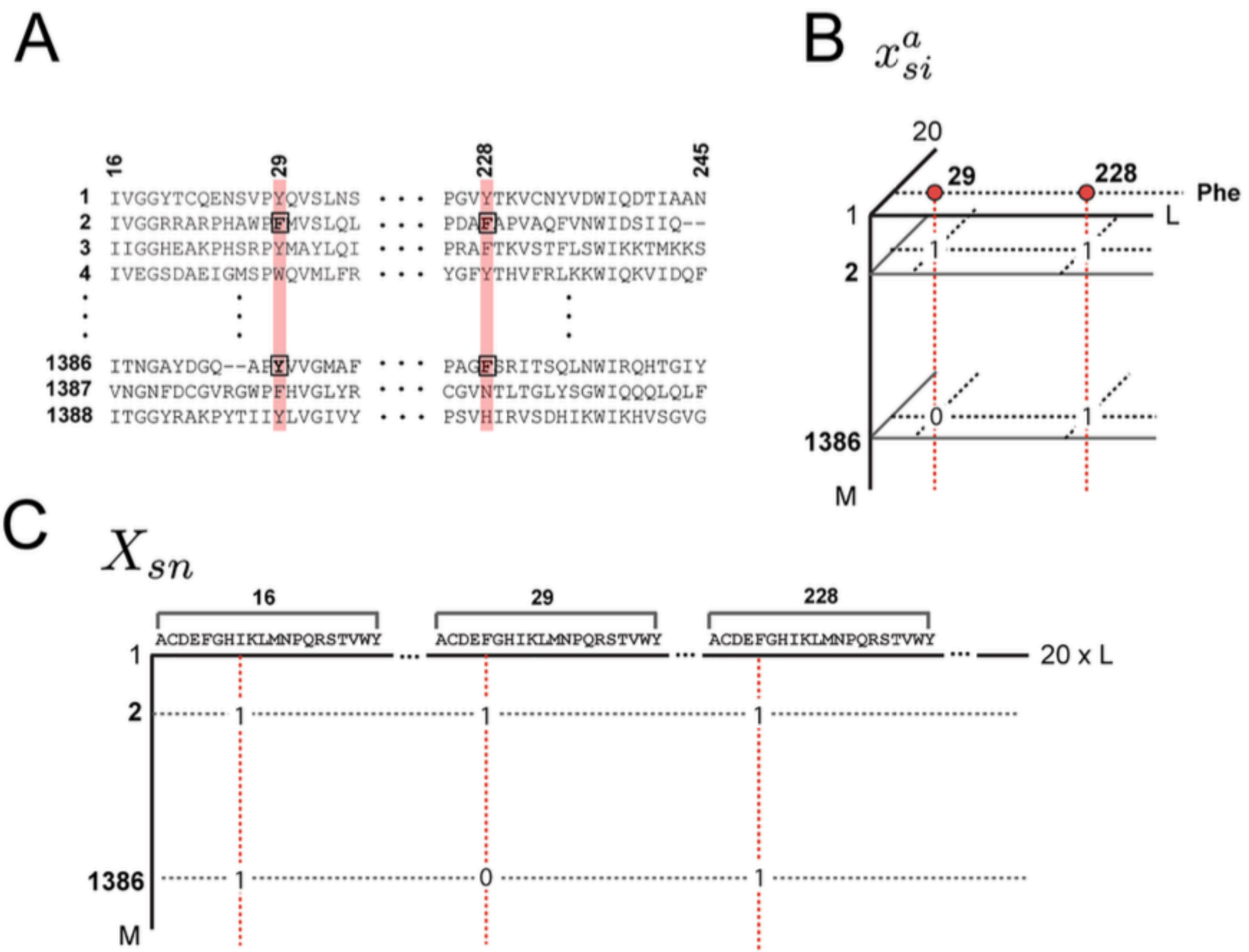
For the first check-in:

Check-ins, Presentations, and Review

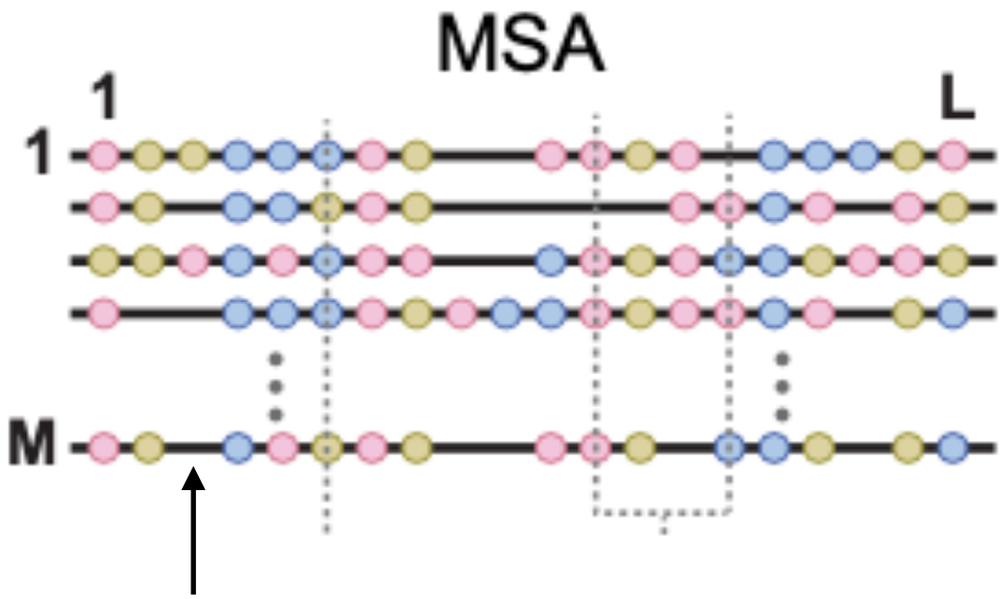
Jan 17 – first project check-in. Data curation and calculation of basic alignment statistics. You have been given four alignments. For each alignment, please do the following:

- Convert the alignment from letters to numbers (one hot encoding)
- For each alignment, remove highly gapped (more than 50% gapped) positions and 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

Encoding an alignment:



Filtering an alignment:



2. Is the sequence more than 50% gaps? ***Remove***

1. Is the position more than 50% gaps? ***Remove***

Computing average pairwise sequence identity (SID):