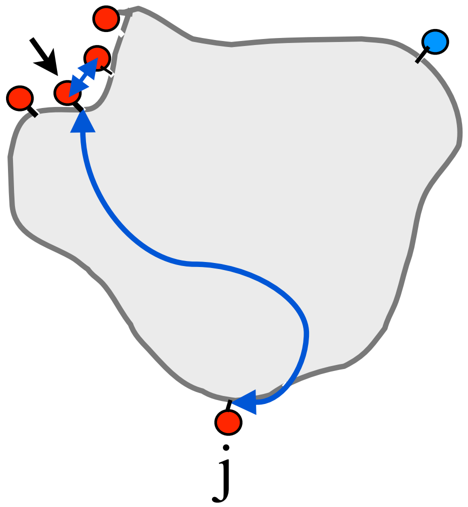
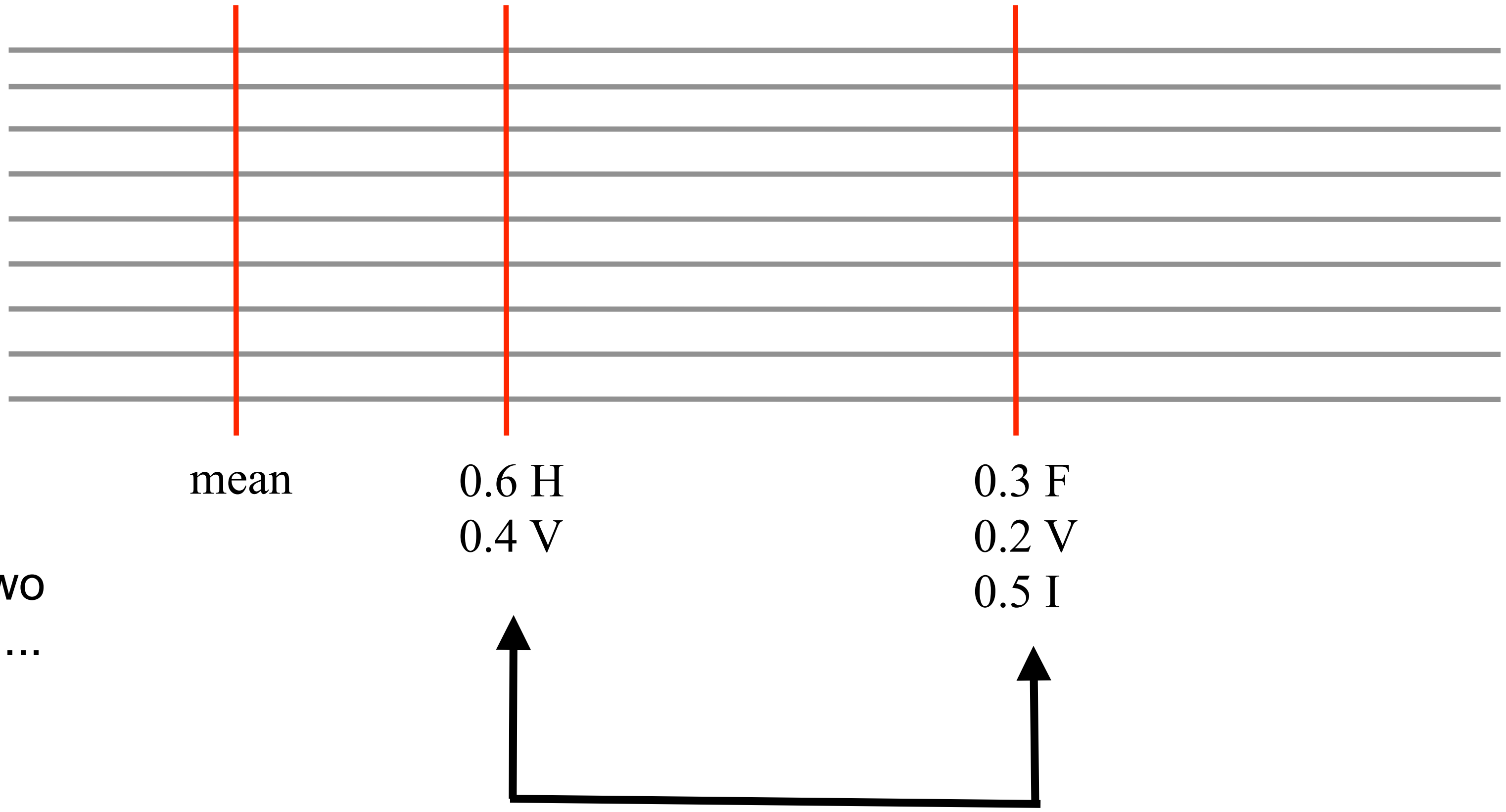
The next step… now that you have a curated alignment, it is time to start looking at positional covariance within and between proteins.

Computing covariance between amino acids at two positions:





mean

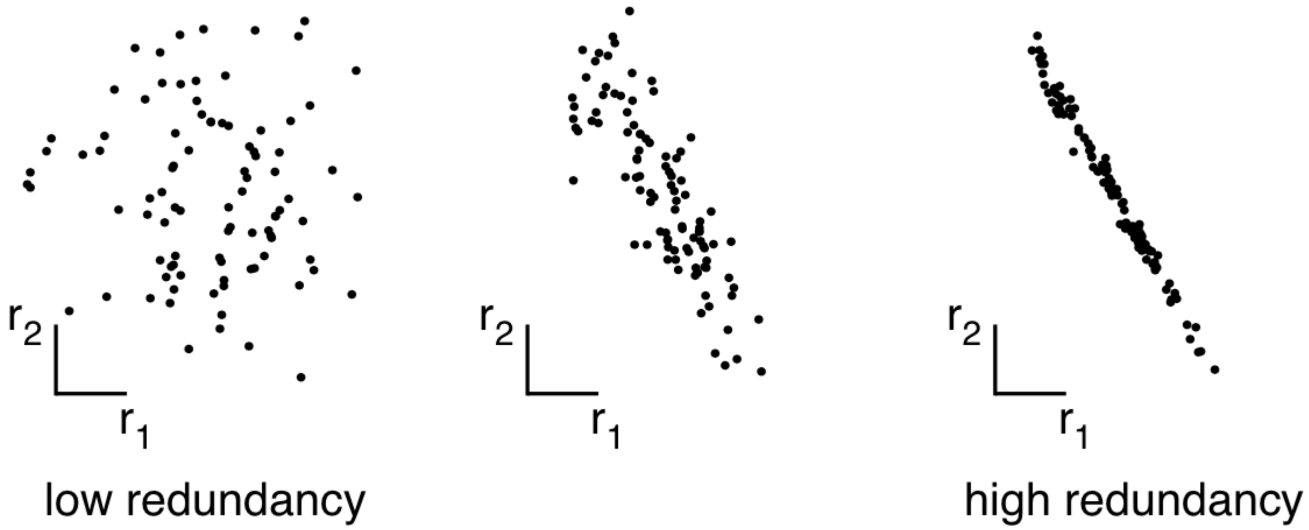
0.6 H
0.4 V

0.3 F
0.2 V
0.5 I

The **basic premise:** Functional Coupling of two amino acid positions should force co-evolution... provided that the interaction contributes to the fitness of the protein.

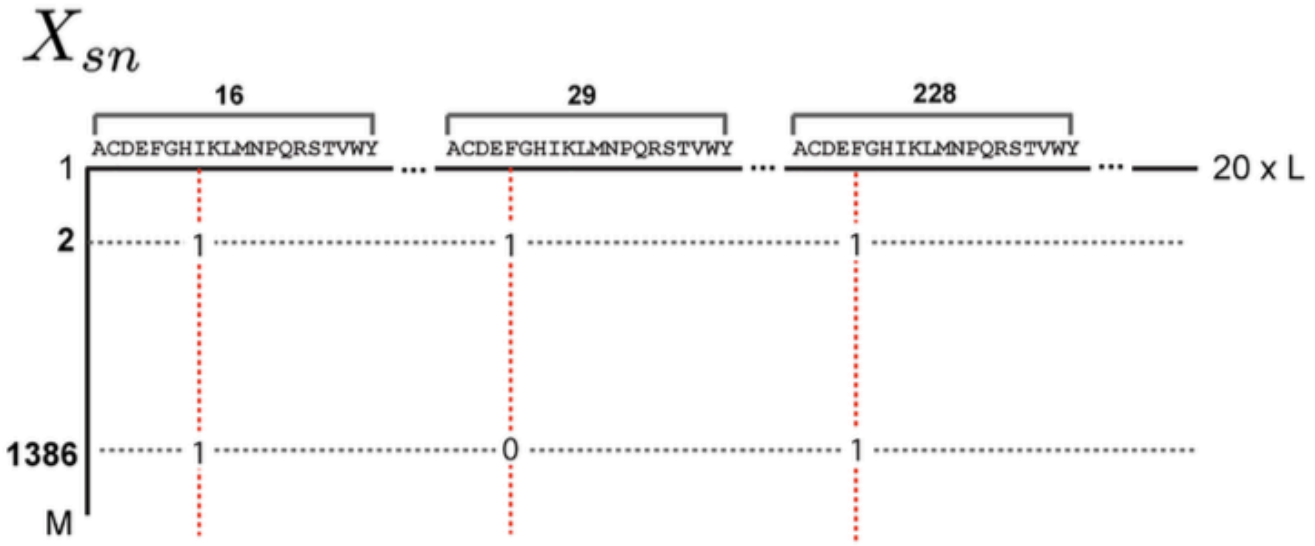
$$C_{ij}^{(ab)} = \left[f_{ij}^{(ab)} - f_i^{(a)} f_j^{(b)} \right]$$

Remember that co-variance is effectively a measure of redundancy or statistical non-independence:

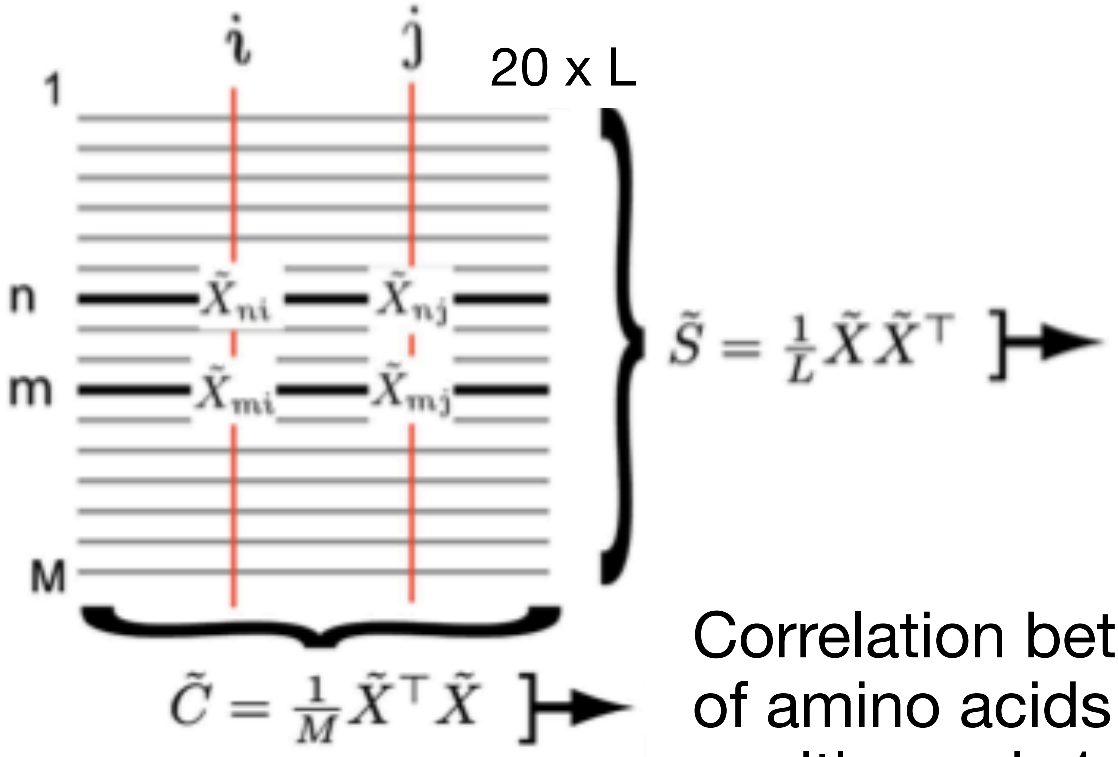


How well can you predict r2 given knowledge of r1?

A linear algebra shortcut to computing covariance $C_{ij}^{(ab)}$



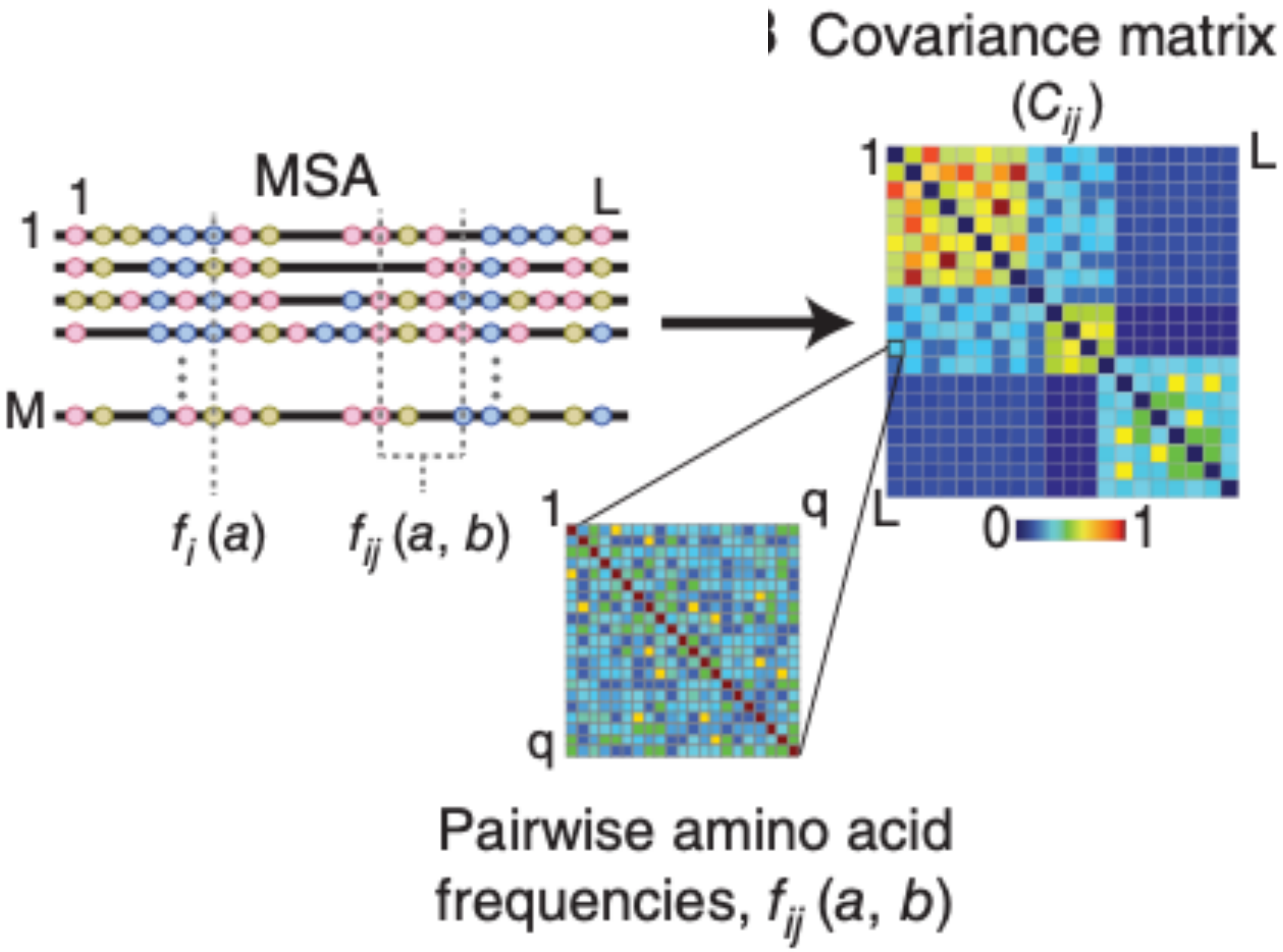
starting from the binarized (one-hot-encoding) matrix


i
j
20 x L
1
n
$\tilde{X}_{ni}$
$\tilde{X}_{nj}$
m
$\tilde{X}_{mi}$
$\tilde{X}_{mj}$
M
$\tilde{S} = \frac{1}{L}\tilde{X}\tilde{X}^\top$
$\tilde{C} = \frac{1}{M}\tilde{X}^\top\tilde{X}$

Correlation between all pairs of sequences n and m (rows)

Correlation between all pairs of amino acids at each position pair 1 x L x 20 (columns)

Now, this gives a 20 x 20 matrix of covariance for each amino acid position:

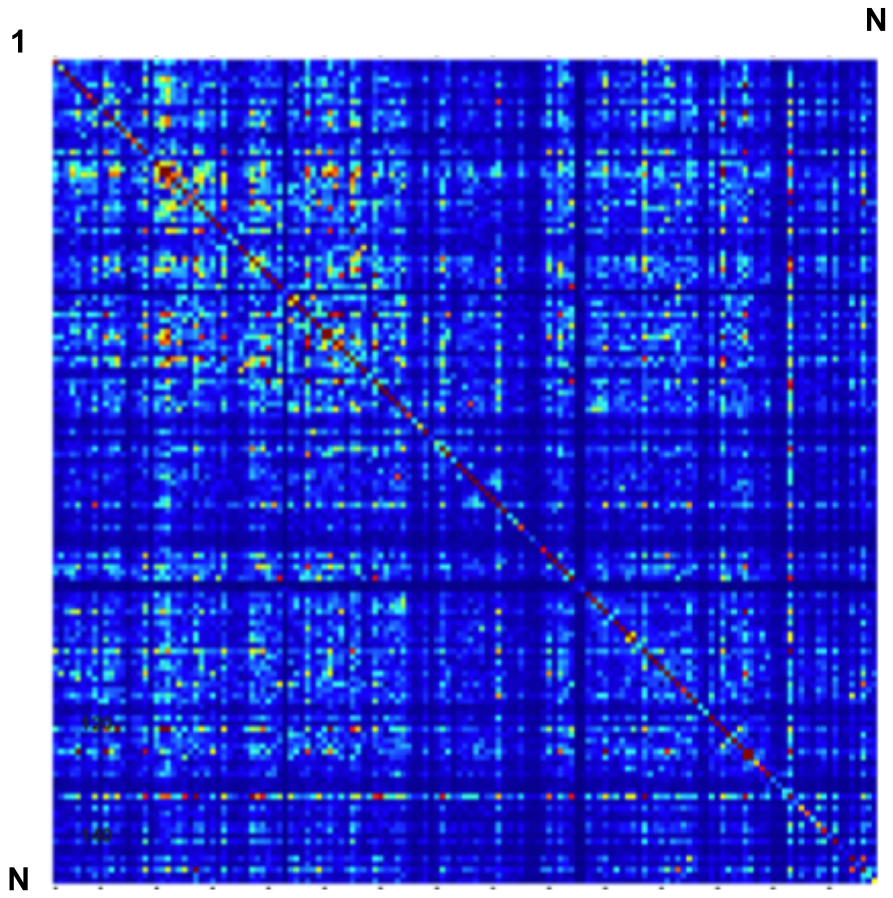


This should be collapsed to a pairwise positional measure.

One (but not the only) way:

$$C_{ij} = \sqrt{C_{ij}(a,b)^2}$$

The end product should be a N_{pos} x N_{pos} covariance matrix for each of your four proteins:

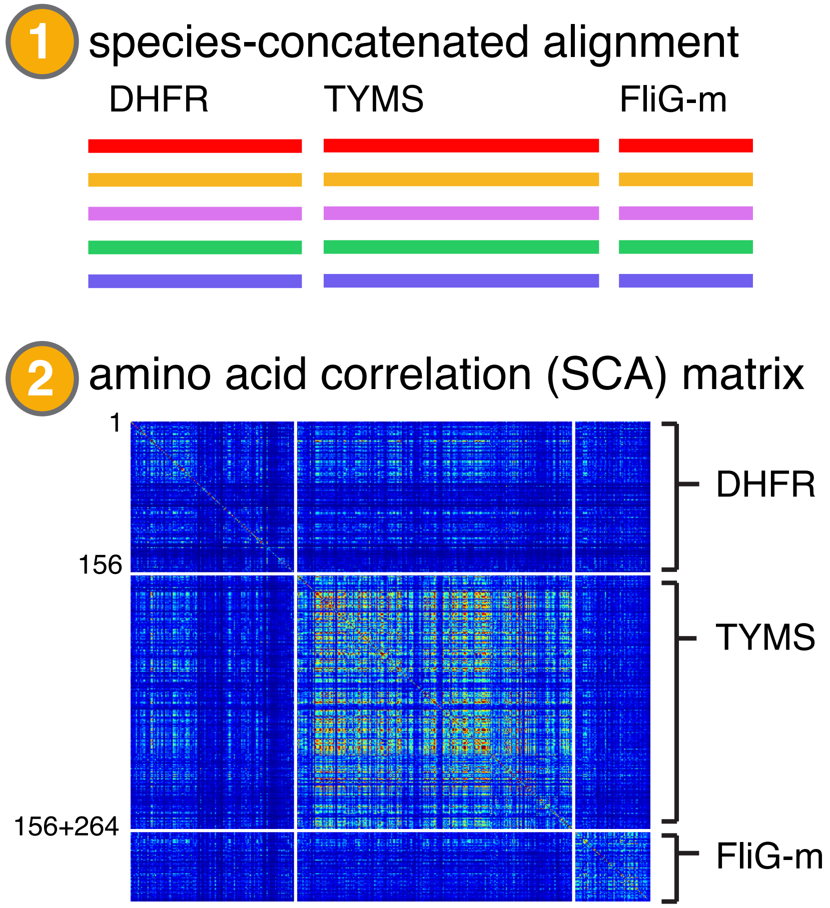


How is covariance organized in the matrix?
Is covariance sparse or abundant?
What is the distribution of covariance values?

Some analysis considerations:

- In calculating co-variance, do you want to examine frequency on a per amino acid basis (20 total possibilities) or group amino acids into more broader physicochemical classes in some way?
- How do you want to collapse covariance among amino acid types (or physicochemical class) into a position level measure?
- How will you handle absolutely conserved positions? Should these contribute to the co-evolutionary signal or no?

Now, to consider co-evolution between proteins, you need to concatenate sequences by species:


1 species-concatenated alignment
DHFR
TYMS
FliG-m
2 amino acid correlation (SCA) matrix
1
156
156+264
DHFR
TYMS
FliG-m

Practically you can make these matches using information in the fasta sequence headers:

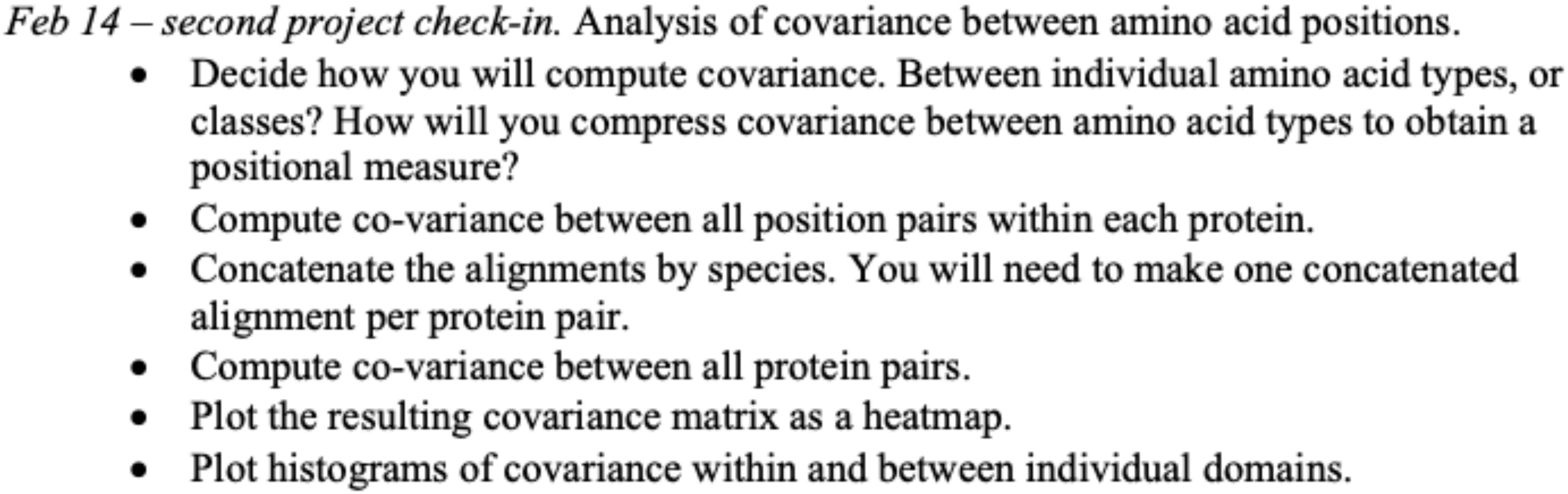
>Prot: A, level_taxid: 1236, organism taxid: "314275_0", organism name: "Alteromonas mediterranea"

>Prot: B, level_taxid: 1236, organism taxid: "314275_0", organism name: "Alteromonas mediterranea"

Some analysis considerations:

- Concatenating alignments will retain a different number and set of sequences. How will this impact the covariance signal between proteins?
- When you compare covariance between pairs of proteins (e.g. A/B vs A/C) how will you account for this difference?

At the next project check-in you will present (on Feb 14):



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.