Research Abbreviated Summary

* Compiled list of 335 bound antigen-antibody complexes from the researcher’s database (Jespersen et al.)
  + Heavy and light antibody chains named as “H” and “L,” respectively, in source pdbs
    - Antigen chains named beginning with A (alphabetical order)
  + Isolated antigen chains with corresponding annotated residues listed in researcher’s database
    - Limited complex (“bound”) number to 275 to ensure annotated residues are only exist on a single antigen chain per complex
  + Found corresponding researcher-RCSB chain letter identifiers by manual checking
    - This is not relevant for further calculations as only the source pdbs, and their systematic naming, were used
* From the list of 275, 195 unbound antigens were found in RCSB that had >95% similarity to the antigens in the complex.
  + Similarity searching was performed using the RCSB advanced search tool to find adequate structures with only one resolved protein entity
  + Mapped annotated residues from bound antigen chain onto corresponding unbound antigen chain
    - FASTA sequences of the bound and unbound proteins were generated and aligned
      * Mapping was used to generate the unbound protein annotated data
  + Further modified about 35 proteins from the set of 195 to remove extraneous protein segments
* Following alignments and curation of bound and unbound sets, we began predictor calculations
* Currently have:
  + VORFFIP, METAPPIPSP, ISPRED, SPPIDER, DOCKPRED, DIscoTope (bound results only)
    - For both bound/unbound
    - ISPRED gave us dynamic cutoff data
    - Will need to review this data to ensure completion
  + ISPIP
    - Completed for the bound
    - Unbound will require more analysis/work (ex: pymol)
* Clustering
  + Scripts developed
  + Capability to incorporate pymol
  + Have reviewed both pre-clustering and post-clustering F-score and MCC scores for the epitope predictions
* Planned future steps:
  + Review and fix up bound/unbound data
  + Review ISPIP unbound data
  + Better coordination through github

Github Naming:

GitHub link: <https://github.com/eved1018/Antigen_project/tree/carroll_antigen>

I'll bold the tabs that are of greater importance.

First, the "Alignment Work" tab contains the raw data that was used to align the bound and unbound pdbs. It is not so relevant at this point.

Second, the "Archived" tab contains previous work that we've done, but is again not so relevant to our current work. For example, the previous attempts at clustering are housed there.

Third, the "Researcher\_Data" tab contains the original pdbs of the bound complexes and the annotated residues.

Fourth, the "**Unbound\_predictor\_results**" tab contains the unbound predictions files and fscore results of the various methods. The "**bound\_predictor\_results"** tab contains this similar information for the bound predictions.

**Important: the "results" tab contains the summarized average f-score results for each method before and after clustering for both the bound and unbound proteins.**

Fifth, the "**bound\_pdbs\_and\_annotated\_residue\_data"** contains the pdbs, lists of the annotated residues, and the corresponding chains in RCSB. The "**unbound\_pdbs\_and\_annotated\_residue\_data**" contains this similar information for the unbound proteins.

Sixth, the "clustering\_analysis" tab contains the results and scripts to perform the clustering analysis.

Seventh, the "**cutoffs**" tab contains the bound and unbound cutoffs.

Eight, the "detailed\_individual\_method\_data" tab contains the raw data of each prediction method.

Ninth, the "meta-dpi\_work" tab contains the inputs for ISPIP. I will continue to develop this tab for the ISPIP results.

Tenth, the "protein\_size\_length" tab contains the lengths of the proteins.