# CS 482/682 Assignment 3

**Due**: March 15, 11:59pm

**Task**: This assignment requires you to use available public databases and software to study the RBD (receptor-binding domain) of the spike protein of SARS-Cov2 virus.

* Find the RBD domain sequences for the following variants:
  1. The earliest reported sequence: Wuhan-Hu-1 (data provided)
  2. Alpha (B.1.1.7)
  3. Beta (B.1.351)
  4. Gamma (P.1)
  5. Delta (B.1.617.2)
  6. Omicron (B.1.1.529)
* Conduct a multiple sequence alignment of the protein sequences of the RBD domains. In the multiple alignment, highlight the amino acids that are different from the earliest reported sequence, Wuhan-Hu -1.

**Submission**:

Your submission is a pdf file that includes the following:

* Your submission should include a list of sequences (one for each variant), together with a detailed description of the source where you get the sequence. Your sequence should only contain the RBD domains, but not the full sequence of the spike (S) protein. Describe in enough detail so that another person can follow the steps to access the sequences as well. For each variant, there may be slightly different sequences available depending which source you get the sequence. You can use any sequence you found from any reputable sources. No need to be too obsessed by the precise boundaries of the RBD domain - being off by a few amino acids is fine.
* You can use any existing multiple sequence alignment tool for this purpose. Your submission should include the multiple sequence alignment. It can be either a screenshot or in text format. The highlights of the different amino acids can be either produced by software or hand drawn. Describe what software you used.

**Hint**:

* For RBP in Wuhan-Hu-1, pay attention to the information in annotation\_report.jsonl and thereby locate the range of RBP in protein.faa from the given data. For example, in annotation\_report.jsonl, when you search for “receptor binding”, you can extract the range ("begin":"330","end":"583") of RBP information of which protein (“surface glycoprotein”).
* For protein sequences in other variants, use NCBI to search directly. For example, you can search Alpha by using <https://www.ncbi.nlm.nih.gov/search/all/?term=B.1.1.7>, and then look for the same protein.