**File Explanations**

* Training set files
  + train\_sequences.fasta - amino acid sequences for proteins in training set; each new sequence designated by “>” and then
    - entryID
    - Also includes:
      * db, ex) sp for UniProtKB
      * UniqueIdentifier, which is the primary accession number of the UniProtKB entry, ex) P20536
      * Entry name, which is the entry name of the UniProtKB entry ex) UNG\_VACCC
      * Protein name ex) Uracil-DNA glycosylase
      * Organism name, ex) OS=Vaccinia virus (strain Copenhagen)
      * Organism identifier, which is the unique identifier of the source organism, assigned by the NCBI, ex) OX=9606
      * GeneName, which is the first gene name of the UniProtKB entry, ex) GN=UNG
      * ProteinExistence, which is the numerical value describing the evidence for the existence of the protein, ex) PE=1
      * SequenceVersion (if 1, it means the first version, if revised 2 or higher), ex) SV=1
    - \*Purpose: contains amino acid sequences in training set
  + train\_terms.tsv - the training set of proteins and corresponding annotated GO terms with the following columns:
    - entry ID, ex) A0A009IHW8
      * each entry ID may have many GO terms
    - GO term, ex) GO:0008152
    - Aspect, ex) BPO
    - \*Purpose: links entryID of amino acid sequences in training set to its GO term and aspect
  + train\_taxonomy.tsv - taxon ID for proteins in training set with the following columns:
    - EntryID, ex) Q8IXT2
    - taxonomyID, ex) 9606
    - \*Purpose: links entryID of amino acid sequences in training set to its taxonomyID
  + go-basic.obo - ontology graph structure; for each term includes
    - GO term, ex) GO:0000001
    - Name, ex) mitochondrion inheritance
    - Namespace, ex) biological\_process
      * each term has only one subontology of the 3 (MF, BP or CC)
    - Def, ex) "The distribution of mitochondria, including the mitochondrial genome, into daughter cells after mitosis or meiosis, mediated by interactions between mitochondria and the cytoskeleton." [GOC:mcc, PMID:10873824, PMID:11389764] synonym: "mitochondrial inheritance" EXACT []
    - Ex) is\_a, ex) is\_a: GO:0048308 ! organelle inheritance
    - \*Purpose: ascribes function term to proteins according to its GO term
* Test set files
  + testsuperset.fasta - amino acid sequences for proteins on which the predictions should be made; each new sequence designated by “>” and then
    - entryID
    - ID, ex) 9606 … same as organism identifier?
    - \*Purpose: to test the amino acids sequences in place of train\_sequences.fasta
  + testsuperset-taxon-list.tsv - taxon ID for proteins in test superset (Note: you may need to use encoding="ISO-8859-1" to read this file in pandas), includes columns:
    - ID, ex) 9606 … same as organism identifier?
    - Species, ex) homo sapiens[All Names]
    - \*Purpose: to list the species for each ID
* Info files
  + IA.txt- “Information Accretion” for each term; contains the information accretion (weights) for each GO term. These weights are used to compute weighted precision and recall, as described in the Evaluation section, includes columns with no headers for:
    - GO term, ex) GO:0000001
    - IA weights (higher score indicates the source is more valuable by considering novelty and relevance)
  + sample\_submission.tsv- a sample submission file in the correct format, includes columns with no headers for:
    - protein targets (entry ID?), ex) A0A0A0MRZ7
    - GO term, ex) GO:0000001
    - probabilistic estimate of the relationship, ex) 0.123
      * the max F-measure based on weighted precision and recall
* concerns
  + some entryID in testsuperset-taxon-list.tsv overlaps with train\_sequences.fasta 🡪 doesn’t this violate what we learned about that the test terms should not be included in training!?
    - Ex) Q6P1M3
    - Ex) Q94FB9
    - Oh this is test set for competition, not necessarily for our model development

**Goal: predict Gene Ontology (GO) terms in each of the three subontologies: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). This set of sequences is referred to as test superset**

* MF, BP and CC are listed under namespace in go-basic.obo with specific functions described under name
* Combine MF, BP and CC in the prediction files, but evaluate independently and combine at the end

**Plan**

* Merge the training datasets into one dataframe
* Create training\_training (tt) set and validation set for model development with train\_test\_split
* Build neural network for tt set
* Evaluate network on validation set with keras
* Optimize the model
* Apply model to test data