Results from UniProt

Dec 15th, 2020

UniProt Protein Filtering with Criteria

- 43,030,402 Proteins from UniRef50
  - Filter criterion: from one of three organisms & have sequence length in range [50, 2000]
  - Results:
    - Proteins from bacteria: 5,514,591 (12.81%)
    - Proteins from Fungi: 1,395,376 (3.24%)
    - Proteins from Plant: 725,086 (1.68%)
- 7,635,053 Proteins
  - Filter criterion: have sequence identity < 0.3 with positive proteins from each group
    - Group1 has 9 positive proteins, group2 has 4 positive proteins, group3 has 9 positive proteins
  - Results
    - Group1: 9745 (0.127%)
    - Group2: 392583 (5.141%)
    - Group3: 8235 (0.107%)

UniProt Protein Filtering with Model Predictions

- 8k~400k Proteins for each group
  - Filter criterion:
    - SCRFs: with ranking score higher than positive testing protein for each group
    - ProtCNN: with probability of being PFT higher than 0.5
    - Unique Proteins: resulted sequences should be different from each other (mutual sequence identity < 1 from blastp)
  - Results:
    - Group1: 34 (0.348%)
    - Group2: 114 (0.290%)
    - Group3: 33 (0.400%)
    - In total: 180/181 (remove repetitive proteins and proteins with 1 sequence identity)

Final Protein List

- Columns of protein list from each group (*results_summary_group1/2/3.csv*)
  - UniProt ID
  - PFT Probability: probability of being Pore-forming toxins estimated by UniProt model, the higher the better (rows sorted by PFT probability)
  - SCRFs Score: score of having similar structure with positive PFTs in each group, the higher the better
  - Organism: from one of bacteria, fungi, plant
  - Sequence
- Columns of protein list for all groups (*results_summary_3groups.csv*)
  - UniProt ID
  - PFT Probability: (rows sorted by PFT probability)
  - Organism
  - Sequence

Groups of Positive PFTs

Groups	I (9 proteins)
Proteins	**Endotoxin_C:** 1DLCA 1I5PA 1W99A 2C9KA 4ARXA 4D8MA 5ZI1A 6OWKA **Thiol_cytolysin:** 4ZGHA

Groups	II (4 proteins)
Proteins	**Clenterotox:** 2XH6A 2ZOEA 3WINE **BB_PF**: 4MJTA

Groups	III (9 proteins)
Proteins	**MACPF:** 2QP2A 2QQHA 3NSJA 3OJYB 5J67A 5OUPA 6CXOA 4OEJA **Thiol_cytolysin:** 2BK2A