



Dear Dr. Francesco,

I have almost completed my midterm essays. Now I try to modify the uniprot database manually.

This e-mail is to report to you that the downloaded uniprot.tsv file is not compatible with NCBI data. You can see the detailed information by searching ‘7/3Question’ in the google docs. https://docs.google.com/document/d/1bOAe6lje0e1HV2Y88vFbeFtJsQwMR5qVErWyfsAOZxY/edit

I have manually labelled all of the proteins that I can retrieve from NCBI. But because of the variants of one protein and the protein sequences that cannot be retrieved, the ordered locus data is much less than 69. I also find that some retrieved ordered locus codes cannot be found in model.genes part. For instance, the locus codes of 10 and 12 row cannot be searched in model genes. That means these locus codes do not contribute to the construction of ecModel.

For these questions, I think maybe I should manually construct a uniprot.tsv from a blank tsv file rather than modify the downloaded uniprot file. If you think this is better than modifying the downloaded file, I have two questions: 1. Can I choose arbitrary proteins to construct the database? What is the suitable number of the selected proteins? 2. If I construct a blank uniprot file, the entry data (first column) in the downloaded file can not be used to label my uniprot file. Should I search for other entries? Where can I find the corresponding entries?