Either filter the protein sequences by protein names [DE] or by sequences and apply either Rule 1 or Rule 2.

Rule 1: For any two protein sequences with the same protein names [DE] or sequences, if the length difference is less than 20%, remove the shorter one, else, keep them both.

Rule 2: For protein sequences with the same protein names [DE] or sequences, if the difference of length between the shortest and second shortest is less than 20%, only remove the shorter and keep the rest, else, keep them both and the rest of the protein sequences with the same protein name or sequences.

This results in:

**Filtered by protein name**

Rule 1: 2237

Rule 2: 16197

**Filtered by sequence**

Rule 1: 3440

Rule 2: 15103