



Extraction of experimentally validated, eukaryotic and cytoplasmic proteins without a transit (mTP/cTP) or signal peptide from UniProt database.



Cutting of proteins according to the length distribution in sequences from the positive dataset.



Random selection of sequences for the negative dataset.