CD-HIT vs. CD-HIT + graphpart

Katarzyna Sidorczuk & Michał

6/27/2022

General remarks

If we use graphpart we cannot predictn Sec proteins.

Moreover, PlastoGram performance drops for Tat, stromal proteins (mostly plastid-encoded) and nuclear-encoded transmembrane proteins.

Sec and TAT issue is mostly related to the number of sequences in the independent data set. Additional graphpart reduction removes 2 out of 6 sequences for Sec (33%) and for Tat 6 out of 12 (50%). Similarly, for plastid-encoded stromal proteins we are losing 21 proteins out of 63 (33%).

The unexplained phenomenon is a huge loss of accuracy for nuclear-encoded transmembrane proteins, when graphpart removes only few proteins in the independent dataset, b ut suddenly PlastoGram staerts to confuse this proteins with nuclear-encoded stromal proteins.

Datasets

CD-HIT + graphpart

1. CD-HIT reduction using 0.9 threshold
2. Graph-part partitioning to create independent and train-test datasets. Parameters: 0.4 threshold, 0.15 ratio of validation dataset, no moving between clusters.
3. N_IM and N_OM are reduced/partitioned separately and then grouped together as envelope.

Dataset	Before filtering	After CD-HIT	After partitioning(train-test)	After partitioning(independent)
N_E	118 (59 IM + 59 OM)	115 (59 IM + 56 OM)	96 (50 IM + 46 OM)	10 (6 IM + 4 OM)
N_TM	276	222	192	30
N_S	357	340	287	53
N_TL_SEC	49	43	37	4
N_TL_TAT	84	89	67	6
P_IM	187	128	106	11
P_TM	4456	1237	1073	156
P_S	1417	419	360	42

CD-HIT

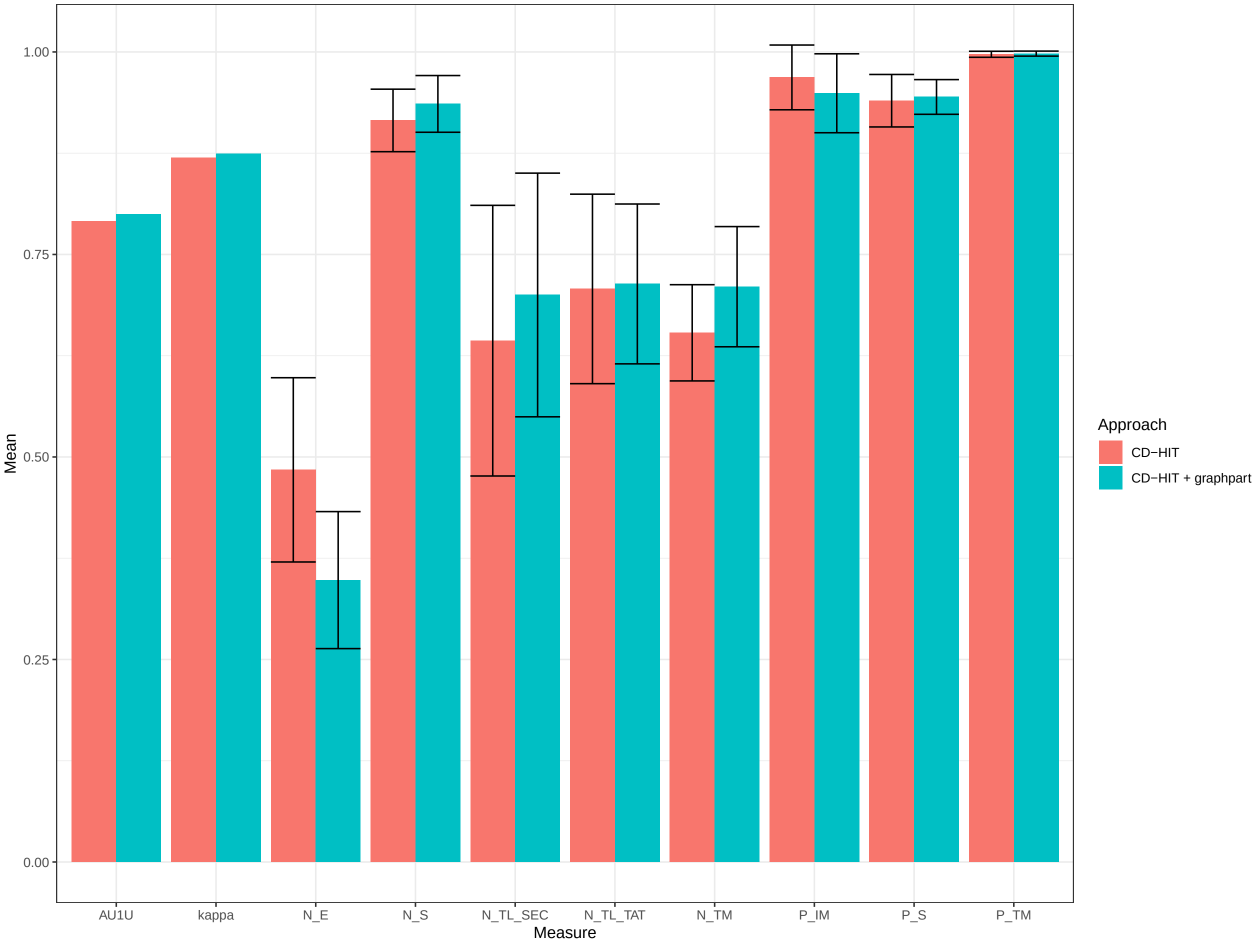
1. CD-HIT reduction using 0.9 threshold
2. 15% holdout to create independent dataset
3. N_IM and N_OM are reduced/partitioned separately and then grouped together as envelope.

Data set	Before filtering	After filtering	Train-test	Independent
N_E	118 (59 OM + 59 IM)	115 (56 OM + 59 IM)	98 (48 OM + 50 IM)	17 (8 OM + 9 IM)
N_TM	276	222	189	33
N_S	357	340	289	51
N_TL_SEC	49	43	37	6
N_TL_TAT	84	79	67	12
P_IM	187	128	109	19
P_TM	4456	1237	1051	186
P_S	1417	419	356	63

Mean performance in 10-fold CV reapeated 5x

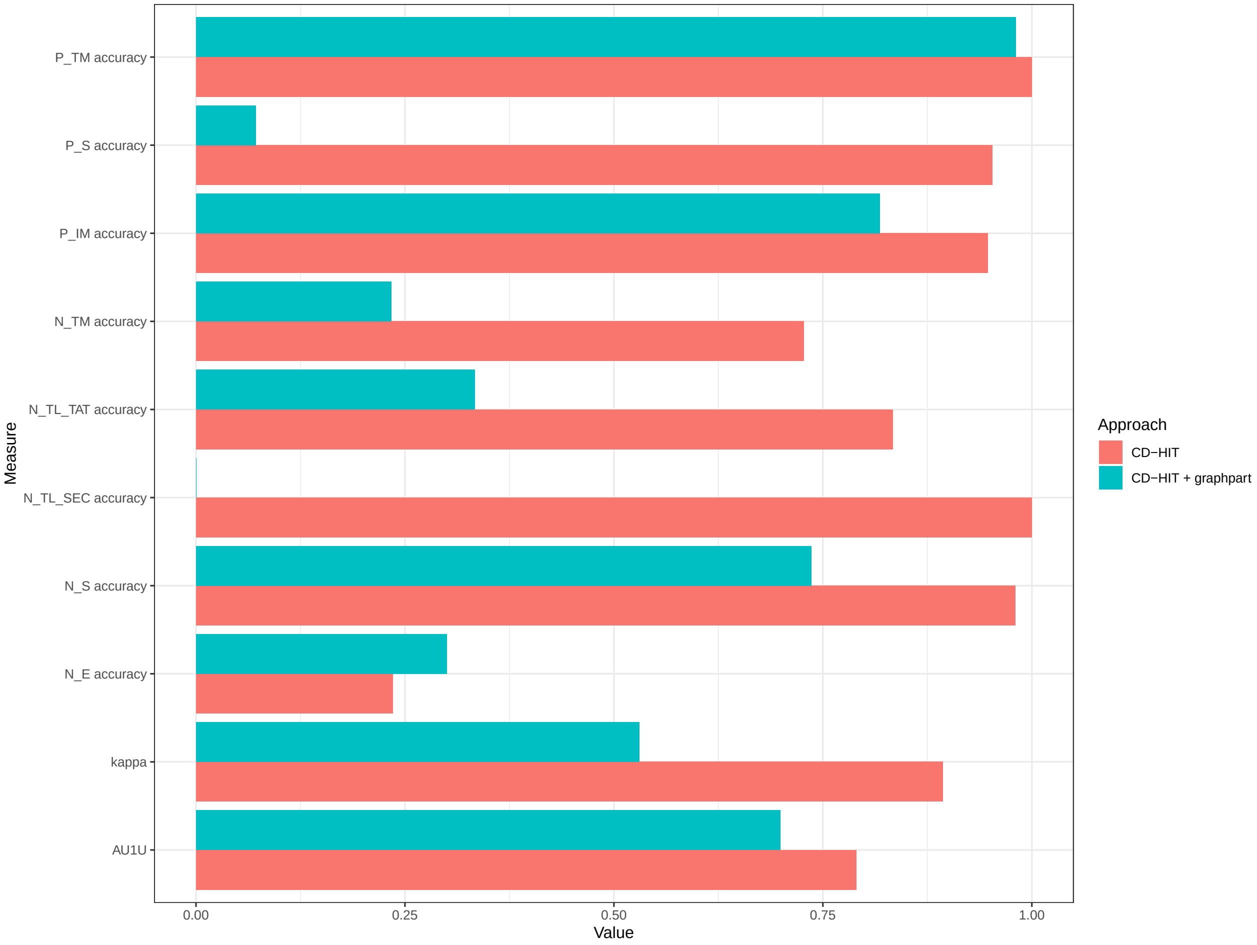
Both CD-HIT only and CD-HIT+graphpart approaches lead to the same best architecture: Architecture_v71_0-1_No_filtering_RF.

Measure	CD-HIT	CD-HIT + graphpart
mean_kappa	0.8694610	0.8745206
mean_AU1U	0.7912053	0.7996173
mean_N_E_sens	0.4841053	0.3480000
mean_N_TM_sens	0.6532859	0.7102294
mean_N_S_sens	0.9154628	0.9358742
mean_N_TL_SEC_sens	0.6435714	0.7000000
mean_N_TL_TAT_sens	0.7074725	0.7136264
mean_P_IM_sens	0.9685714	0.9490043
mean_P_TM_sens	0.9971456	0.9979483
mean_P_S_sens	0.9398200	0.9444444
sd_N_E_sens	0.1137329	0.0845621
sd_N_TM_sens	0.0593983	0.0741940
sd_N_S_sens	0.0385892	0.0350118
sd_N_TL_SEC_sens	0.1670598	0.1504075
sd_N_TL_TAT_sens	0.1169723	0.0986722
sd_P_IM_sens	0.0399679	0.0487234
sd_P_TM_sens	0.0037315	0.0031084
sd_P_S_sens	0.0323924	0.0214219



Performance on the validation dataset

Measure	CD-HIT	CD-HIT + graphpart
kappa	0.8931102	0.5304225
AU1U	0.7896798	0.6989366
N__E accuracy	0.2352941	0.3000000
N__TM accuracy	0.7272727	0.2333333
N__S accuracy	0.9803922	0.7358491
N__TL__SEC accuracy	1.0000000	0.0000000
N__TL__TAT accuracy	0.8333333	0.3333333
P__IM accuracy	0.9473684	0.8181818
P__TM accuracy	1.0000000	0.9807692
P__S accuracy	0.9523810	0.0714286



CD-HIT approach

True \ Predicted	N_E	N_S	N_TL_SEC	N_TL_TAT	N_TM	P_IM	P_S	P_TM
N_E	4	11	0	0	1	0	0	1
N_S	1	50	0	0	0	0	0	0
N_TL_SEC	0	0	6	0	0	0	0	0
N_TL_TAT	0	1	0	10	1	0	0	0
N_TM	0	5	0	0	24	0	0	4
P_IM	0	0	0	0	0	18	0	1
P_S	0	1	0	0	0	0	60	2
P_TM	0	0	0	0	0	0	0	186

CD-HIT + graphpart approach

True \ Predicted	N_E	N_S	N_TL_SEC	N_TL_TAT	N_TM	P_IM	P_S	P_TM
N_E	3	6	0	0	0	0	0	1
N_S	1	39	0	0	8	0	4	1
N_TL_SEC	0	2	0	1	1	0	0	0
N_TL_TAT	0	3	0	2	0	0	0	1
N_TM	1	15	2	0	7	0	0	5
P_IM	0	0	0	0	0	9	0	2
P_S	0	24	0	0	0	0	3	15

True \ Predicted	N_E	N_S	N_TL_SEC	N_TL_TAT	N_TM	P_IM	P_S	P_TM
P_TM	0	1	0	0	1	0	1	153