1. Исследование датасета по продуктам процессинга белков протеасомой.

Смирнов Антон Сергеевич

Nov 2, 2022

Оглавление

Датасет взят из статьи(Specht et al. 2020)

Ссылка на датасет: https://doi.org/10.17632/nr7cs764rc.1

Датасет представляет из себя совокупность результатов процессинга пептидов разными протеасомами in vitro. Пептиды определяли с помощью масс-спектрометрии.

Присоединяю пакет: 'dplyr'

Следующие объекты скрыты от 'package:stats':

filter, lag

Следующие объекты скрыты от 'package:base':

intersect, setdiff, setequal, union

'data.frame': 147165 obs. of 23 variables:
$ sampleID : chr "010719FA6" "010719FA6" "010719FA6" "010719FA6" ...
$ sampleName : chr "FA6" "FA6" "FA6" "FA6" ...
$ runID : int 1 1 1 1 1 1 1 1 1 1 ...
$ protIsotype : chr "20S standard" "20S standard" "20S standard" "20S standard" ...
$ digestTime : int 20 20 20 20 20 20 20 20 20 20 ...
$ species : chr "human" "human" "human" "human" ...
$ sampleDate : chr "01/07/2019" "01/07/2019" "01/07/2019" "01/07/2019" ...
$ instrument : chr "QE_FT" "QE_FT" "QE_FT" "QE_FT" ...
$ fragmentation : chr "HCD" "HCD" "HCD" "HCD" ...
$ location : chr "MPI-BPC" "MPI-BPC" "MPI-BPC" "MPI-BPC" ...
$ substrateSeq : chr "KLSHKHLVLNYGVCVCGDENILVQEFVKFGSL" "KLSHKHLVLNYGVCVCGDENILVQEFVKFGSL" "KL
$ substrateOrigin: chr "O60674|JAK2[603-634]" "O60674|JAK2[603-634]" "O60674|JAK2[603-634]" "O60674|JAK2[603-634]"
$ substrateID : chr "TSN96" "TSN96" "TSN96" "TSN96" ...
$ pepSeq : chr "GDVKFGSL" "KLSHLNYGV" "KLSHKHLVLNYGVCVCGDENILVQEFVKFKFGSL" "KLLVLNY" ...
$ scanNum : int 8723 13051 26128 9651 6007 4480 23783 3846 24175 5455 ...

```
$ rankMS        : int  1 1 1 1 1 1 1 1 1 1 ...
$ ionScore      : num  20.1 20.1 20.2 20.4 20.6 ...
$ qValue        : num  0.0099 0.0097 0.0094 0.0092 0.0087 0.0085 0.03 0.0083 0.034 0.016 ...
$ productType   : chr  "PSP" "PSP" "PSP" "PSP" ...
$ spliceType    : chr  "cis" "cis" "trans" "cis" ...
$ positions     : chr  "17_18_27_32" "1_4_9_13" "1_29_28_32" "1_2_7_11" ...
$ charge        : int  2 2 4 2 1 1 2 1 3 1 ...
$ PTM           : chr  "" "" "2 Deamidated (NQ)" "" ...
```

|| || || ||

Посмотрим на основные интересующие нас моменты: тип протеасомы, видовое разнообразие, тип сплайсинга, разнообразие белков.

spliceType	species	protIsotype	n
	human	20S immuno	5066
	human	20S standard	102982
	human	26S immuno	14
	human	26S standard	11
cis	human	20S immuno	1373
cis	human	20S standard	10140
cis	human	26S immuno	7
cis	human	26S standard	8
revCis	human	20S immuno	799
revCis	human	20S standard	8680
revCis	human	26S immuno	23
revCis	human	26S standard	12
trans	human	20S immuno	2146
trans	human	20S standard	15879
trans	human	26S immuno	21
trans	human	26S standard	4

Как можно заметить, все сделано на человеческих белках, преобладают несплайсированные формы, крайне мало образцов, процессированных 26S протеосомой(20S + регуляторные субъединицы). Для дальнейшего рассмотрения возьмем не сплайсированные формы, процессированные 20S протеосомой без посттрансляционных модификаций.

substrateOrigin	n
O60674|JAK2[603-634]	709
O75376|NCOR1[414-439] P427S	6
P01116|KRAS[2-35]	3388
P01116|KRAS[2-35] G12V	1476
P06239-3|LCK[59-82] P74L	917
P0CG48|UBC[5-36]	737
P0CK49|BSRF1[47-79]	2903
P11210|IE1[162-186]	3158
P11802|CDK4[19-37]	2736
P12977|EBNA3[368-397]	1733
P14618|PKM[3-33]	1530
P15056|BRAF[577-602]	1887
P15056|BRAF[577-602] V590E	405
P15056|BRAF[584-615]	851
P33378|plcB[159-171|185-196]	1062
P33378|plcB[174-193]	2472
P33378|plcB[61-73|84 - 94]	1019
P35858-2|IGFALS[150-175] R169Q	1921
P36888|FLT3[829-850]	556
P40238|MPL[504-530]	320
P40967-2|PMEL[201-230] T210M	2731
P40967-2|PMEL[35-57]	5412
P40967-2|PMEL[40-52]	127
P40967|PMEL[201-230]	457
P42768|WASP[9-15|26-27]	1090
P42768|WAS[9-32]	1005
P63000-2|RAC1[20-44]	2037
P9WQP1|fbpB[175-198]	2723
P9WQP1|fbpB[175-198] L193E	2394
P9WQP1|fbpB[91-117]	927
P9WQP1|fbpB[91-117] V109E	1022
Q05925|EN1[9-39]	2808
Q12770|SCAP[749-773]	1402
Q14004|CDK13[870-892]	2961
Q14162-2|SCARF1[541-555] R544W	50

substrateOrigin	n
Q15021|NCAPD2[407-426]	1213
Q15021|NCAPD2[411-426]	2684
Q15051|IQCB1[406-438]	2169
Q15051|IQCB1[411-433]	2790
Q16576-2|RBBP7[51-69]	2217
Q16576-2|RBBP7[51-69] N61D	495
Q16653-13|MOG[93-122]	2228
Q16787|LAMA3[3189-3195|3203-3204]	2150
Q16787|LAMA3[3189-3209]	2933
Q6E949|plcA[19-28|46-58]	2161
Q8N819|PPM1N[178-201]	937
Q8WUX9|CHMP7[311-330]	818
Q8WUX9|CHMP7[311-330] A324T	304
Q92621|NUP205[1263-1291]	2743
Q92621|NUP205[1268-1290]	3822
Q9H7F0|ATP13A3[112-139]	549
Q9HCG8|CWC22[229-258]	2784
Q9NRK6|ABCB10[518-540]	4345
Q9NRK6|ABCB10[518-540] P518L	218
Q9Y6D9|MAD1L1[150-181]	1193

Определим сайты протеолиза иммунопротеосомы.

i.sites	Freq	N
A	0.05100	A
AF	0.00256	A
AH	0.00341	A
AL	0.00049	A
AN	0.00158	A
AQ	0.07278	A
AW	0.02300	A
DG	0.00986	D
DN	0.00024	D
DQ	0.00024	D

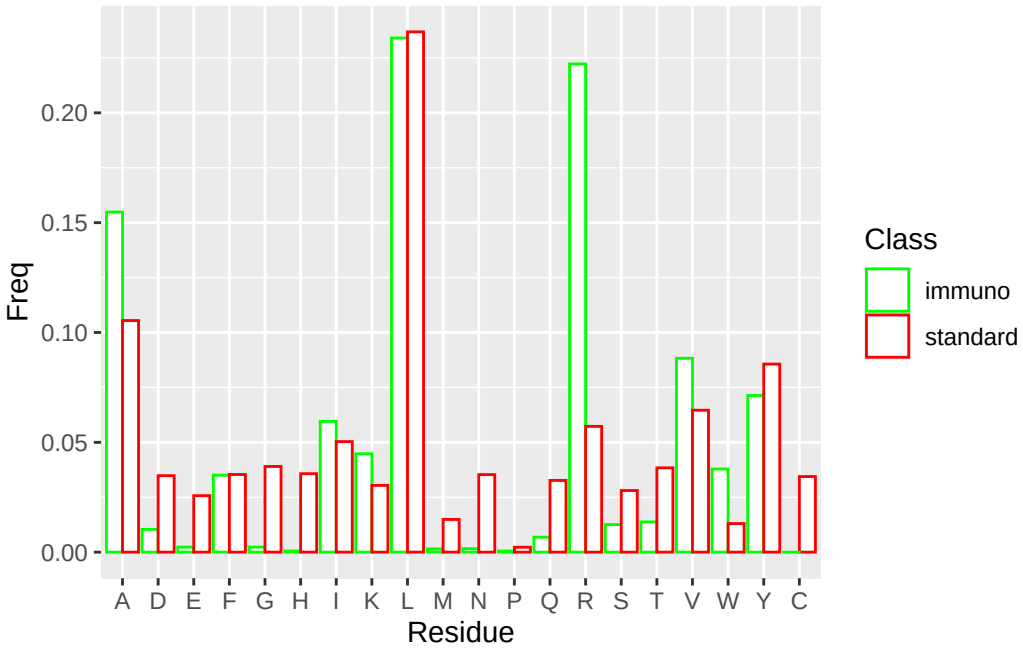
Var1	Freq
A	7
D	3
E	2
F	4
G	3
H	1
I	2
K	2
L	4
M	1
N	5
P	2
Q	3
R	4
S	4
T	4
V	4
W	2
Y	4

Определим сайты разрезания стандартной протеосомы

s.sites	Freq	N
A	0.00507	A
AA	0.02696	A
AD	0.00112	A
AF	0.00250	A
AG	0.00668	A
AH	0.00037	A
AI	0.00250	A
AK	0.00794	A
AL	0.01122	A
AM	0.00070	A

Var1	Freq
A	19
C	10
D	17
E	19
F	16
G	19
H	16
I	18
K	18
L	21
M	11
N	18
P	14
Q	18
R	16
S	18
T	16
V	19
W	11
Y	17

Как можно видеть, протеосомы могут разрезать белок по всем аминокислотам, но с разной частотой.



Specht, Gerd, Hanna P. Roetschke, Artem Mansurkhodzhaev, Petra Henklein, Kathrin Textoris-Taube, Henning Urlaub, Michele Mishto, and Juliane Liepe. 2020. “Large Database for the Analysis and Prediction of Spliced and Non-Spliced Peptide Generation by Proteasomes.” *Scientific Data* 7 (1): 146. https://doi.org/10.1038/s41597-020-0487-6.