Disorder in Cdk1/2 targets

updated: 2019-08-09

MobiDB

The MobiDB database of protein disorder and mobility annotations has been significantly updated and upgraded since its last major renewal in 2014. Several curated datasets for intrinsic disorder and folding upon binding have been integrated from specialized databases. The indirect evidence has also been expanded to better capture information available in the PDB, such as high-temperature residues in X-ray structures and overall conformational diversity.
The database is organized in three layers containing data of different nature: Disorder, Linear Interacting Peptides (LIPs) and Dynamic structure. The "Disorder" database bears data about protein regions without a defined tri-dimensional structure. The "LIPs" layer contains information about regions or residues that interact with other proteins or DNA/RNA preserving a linear (disordered) structure. At last, the data hosted in the "Dynamic structure" layer describes the propensity of a residue to assume a specific secondary structure conformation. When extracted from chemical shifts can be interpreted as a measure of the secondary structure populations in solution. Each one of three layers features three quality levels of annotation from high to low quality (Fig. 1), presenting different coverage of the universe of known proteins.

Disorder

As a first approach, I explored the Disorder database from Mobidb for the whole annotated proteome of Saccharomyces cerevisiae. The data set comprehends the information of disordered region over 6721 proteins. As explained before, each entry could have annotations of disorder in three different levels (DB, indirect, Predicted) which, at the same time, collect information from several sources in order to reach to a consensus of the regions with or without a defined tri-dimensional structure. I this way, the consensus of each level annotates the regions as structured "S", Disordered "D" or conflict "C", when there is more than one source of information over a protein region that contradicts each other.

In order to understand the data the first analysis to do is to evaluate the coverage of the different levels of annotations in the database. In this way, we can select a level or a combination of them for assigning the disorder regions to each protein considering the compromise between the quality of the annotation and the coverage of the protein set of interest.
DB is the level with less coverage, since its annotations are manually curated, and it has **56** entries covering a **0.83%** of the whole proteome. The data annotated in the "Derived" level holds **1372** entries and this covers **20.41%** of available proteins for S. cerevisiae. Since the predictors do not have any limitations in terms of which proteins could these methods be applied to, they cover the entire protein set.

Level	Method	N° entries	Level coverage (%)	Method coverage (%)
DB	full	56	100	0.83
Derived	full	1372	100	20.41
	bfactor	1223	89.14	18.2
	missing-residues	1372	100	20.41
	mobile	200	14.58	2.98
Predictors	simple	6721	100	100
	mobidb-lite	6105	90.83	90.83

Table 1: Percent of covered entries by each Level and method of the MobiDB database.

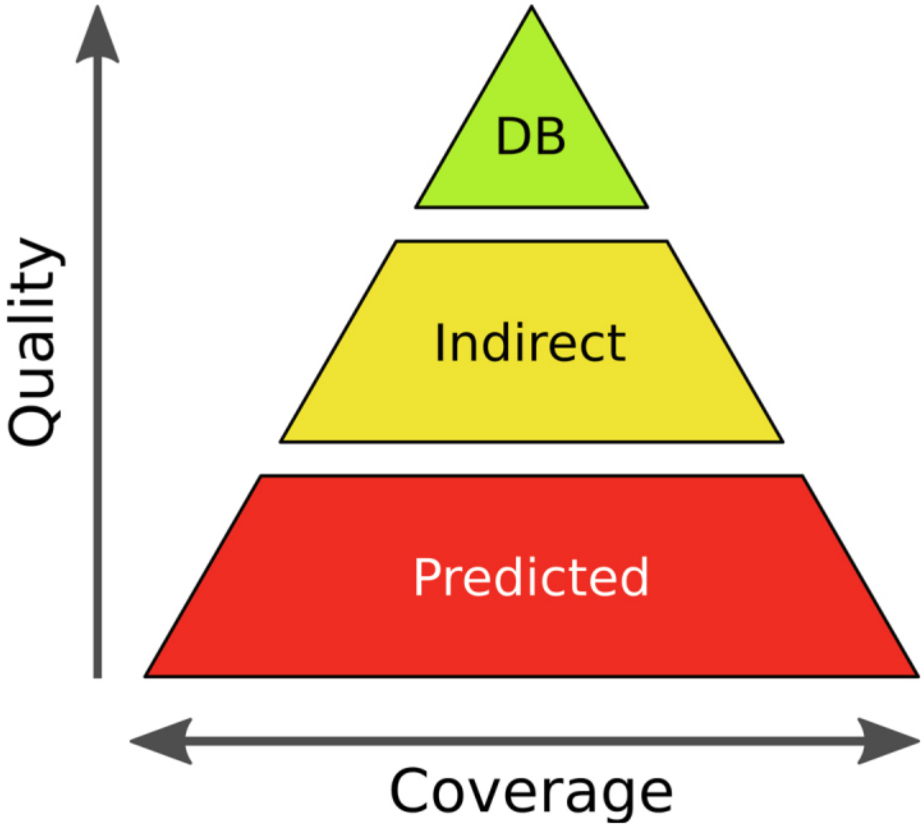


Figure 1: Figure 1: MobiDB confidence vs. coverage. DB refers to manually curated annotations from external databases. Indirect consist in annotations derived/calculated information from experimental data, i.e. PDB structures and/or chemical shifts. Predicted annotations are made using exclusively bioinformatic tools

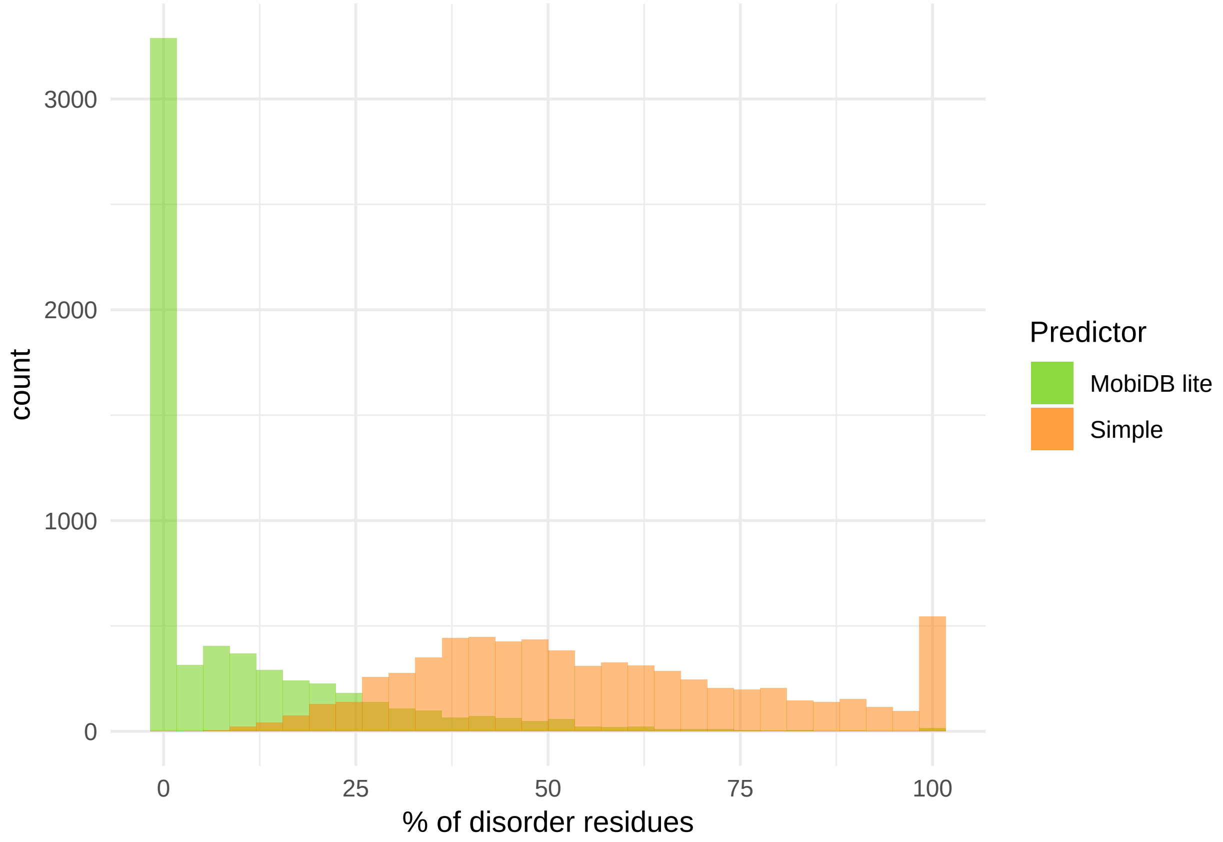


Figure 2: Figure 2: Histogram for the disorder percentages estimated by the two predictors

Percentage of disordered aminoacids

Once we have established the disorder regions for each protein it is useful to have a metric describing the global disorder status for each entry in order to summarise in one number the degree of disorder and to being able to make comparisons at the protein level. The percentage (or ratio) of disorder amino acids over the total number of amino acids of a protein constitute the most direct metric, providing one value that we can assign to each proteins describing its state of disorder.

Both of the predictors featured by the MobiDB database have a notorious discrepancy on the number of residues on which they assign the "Disorder" status. This could be due, in principle, to the differences in their conceptions. The "Simple" method is basically consensus derived from some of the already established disorder bioinformatic predictors and features region-wise annotations. In contrast, the method called MobiDB lite is, in fact, an *ad hoc* development that not only uses a consensus between these methods but also quantifies the agreement among different predictors, and it is optimized for detecting long intrinsically disordered regions (IDR) and the detection of short regions are filtered out. This could explain why Figure 1 presents a great proportion of proteins without any disorder residue for the MobiDB lite predictor. It is understandable then, that the ratio of the residues detected as intrinsically disorder over the total length of the proteins, by those different methods, are not in agreement and even they seem to be complementary between them.

For the "Derived" and "DB" levels, the identification of IDR has, allegedly, a higher degree of confidence. The comparison the percentage of disorder residues estimated by those levels (Figure 3) and the prediction methods could help to clarify which prediction is the better option if we want to use the percentage of disordered amino acids as metric to evaluate disorder at the protein level.

The use of the derived data for assigning IDRs seems to present a distribution of the percentage of disorder

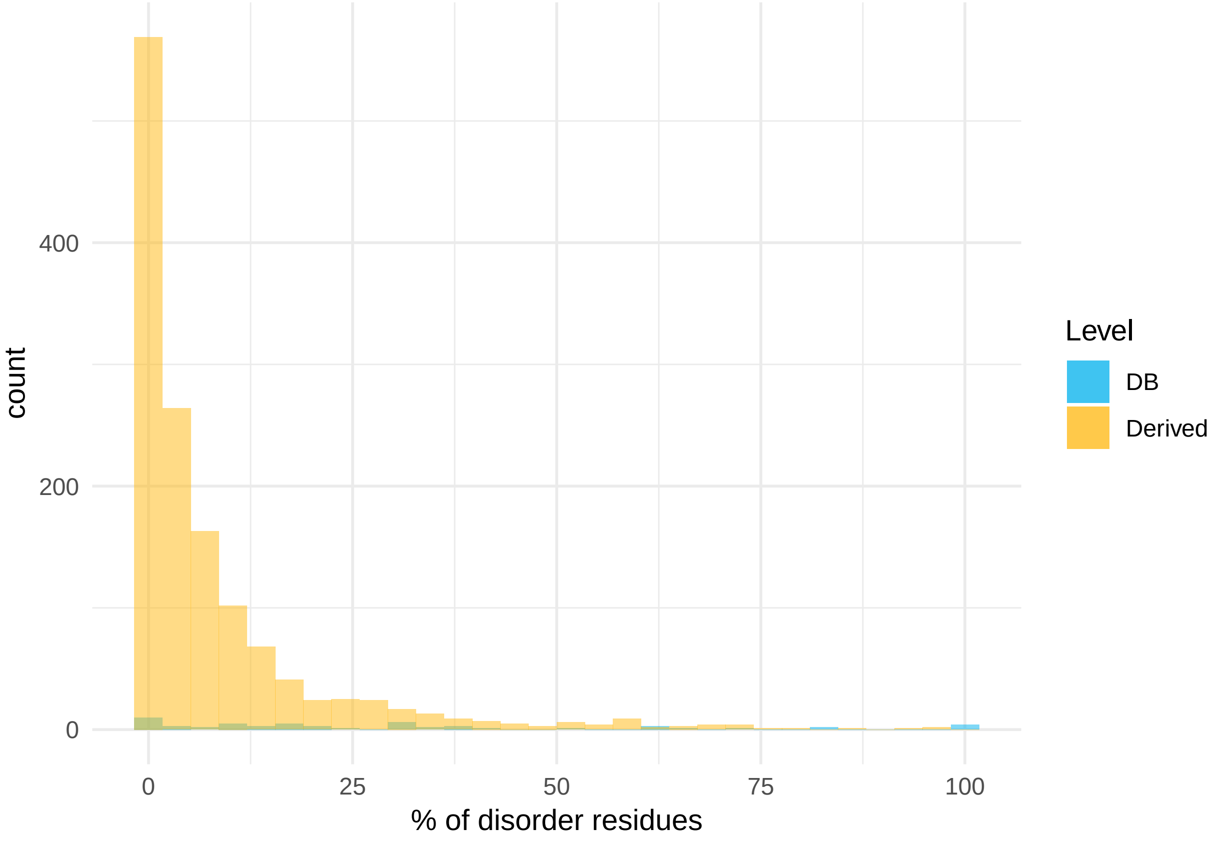


Figure 3: Figure 3: Histogram for the disorder percentages estimated for the ‘DB’ and ‘Derived’ levels

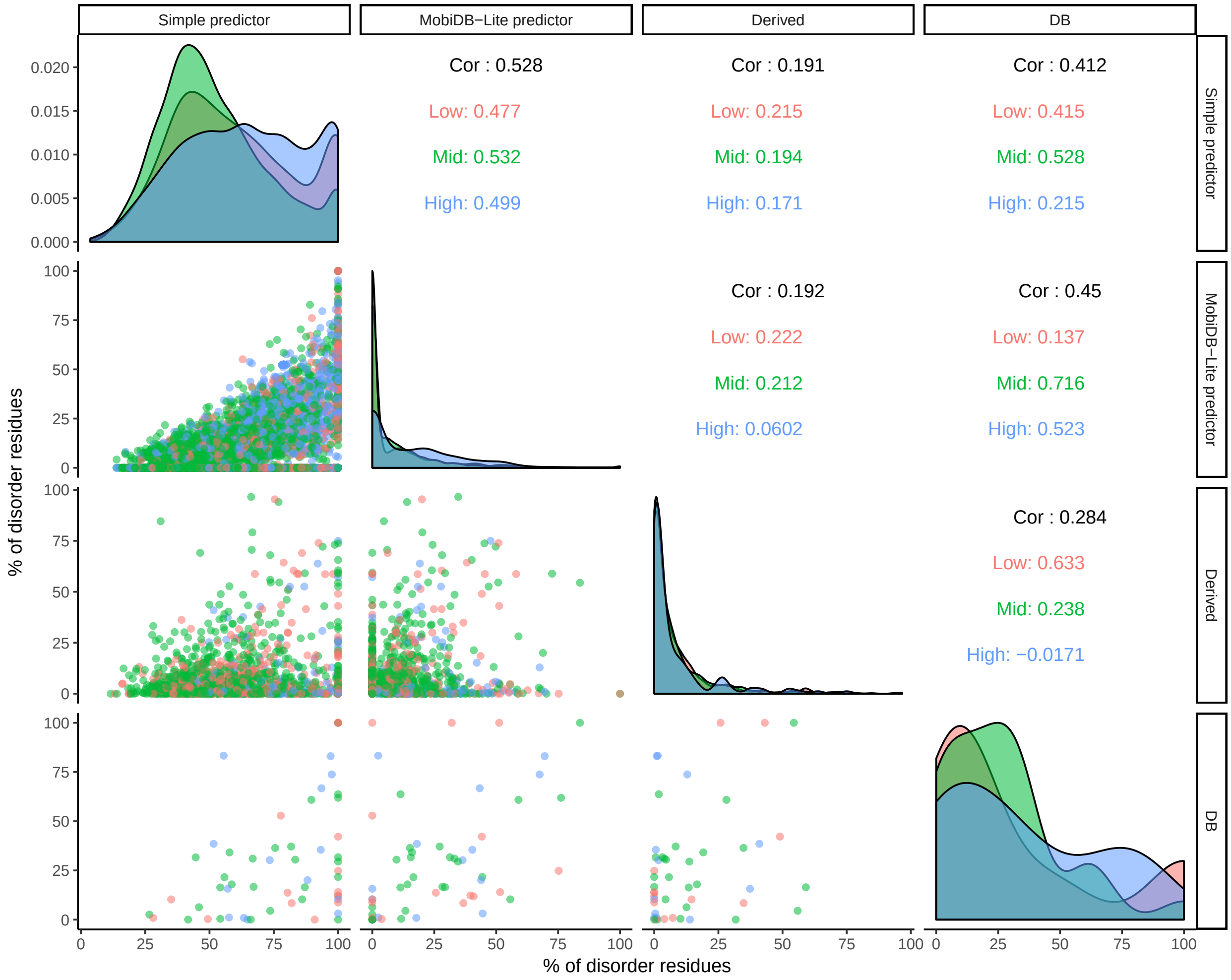


Figure 4: Figure 4: Pairwise comparison of the % of disorder residues determined by the different methods

residues congruent with the MobiDB lite predictor, presenting a considerable proportion of protein not having any disorder residues.

Similar distributions of the percentage of disorder residues, for two different predictors, could be validated if the predictors are assigning similar percentage values to the same proteins. If this is the case, we should be able to find a correlation between the values calculated by different methods. In Figure 4, pairwise comparisons for both of the prediction methods and the Derived and DB annotations are shown, together with the density distributions and the corresponding correlation coefficients. Since we are interested in elucidating how CDK phosphorylation could be involved in the disorder states of its targets, another level of complexity is added to this plot. The percentage of phosphorylatable residues (Ser, Thr, Tyr) were calculated for all the proteins and Low, Mid and High levels of putative phosphorylations sites were defined with arbitrary breaks (0-15%, 15-20%,20-60%) of the percentage values, based on the quartiles of the distribution.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.333  15.487  17.541  18.138  20.036  57.379
```

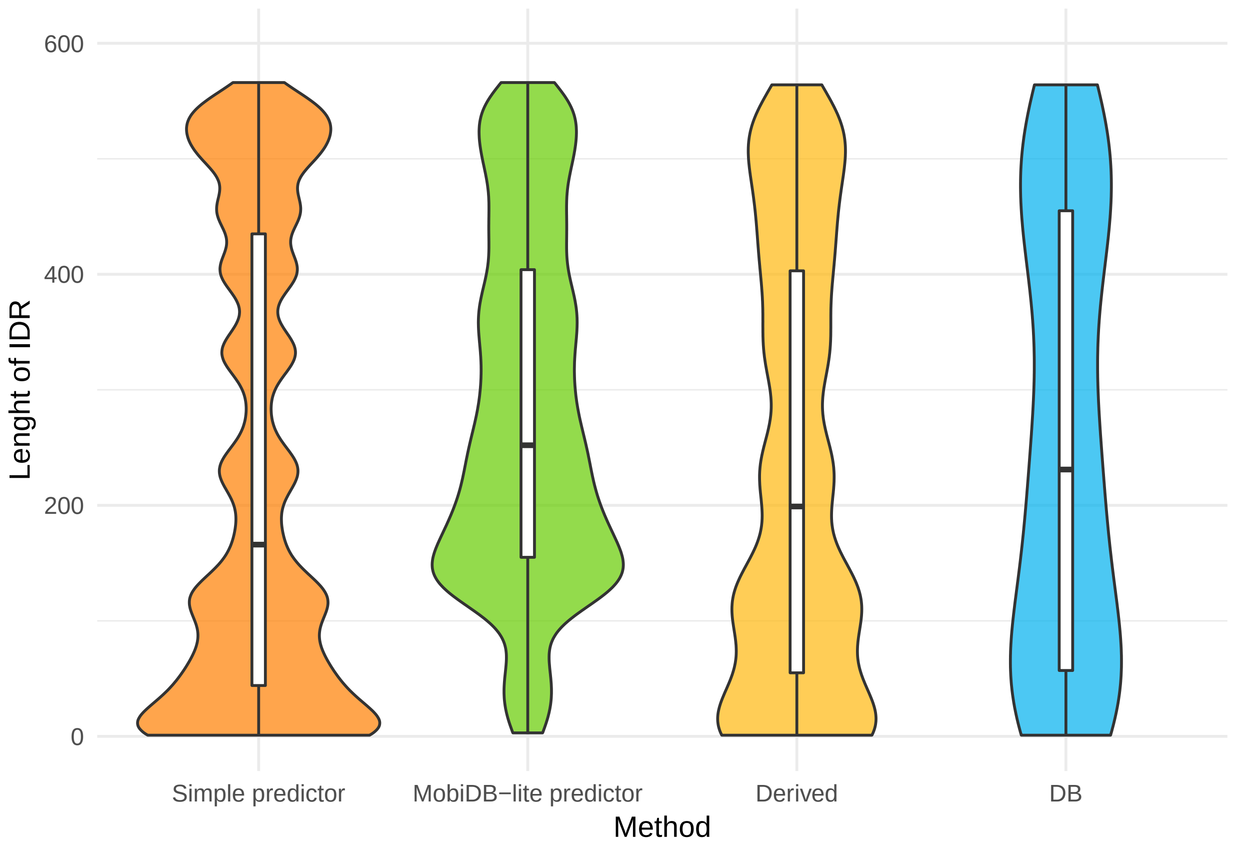


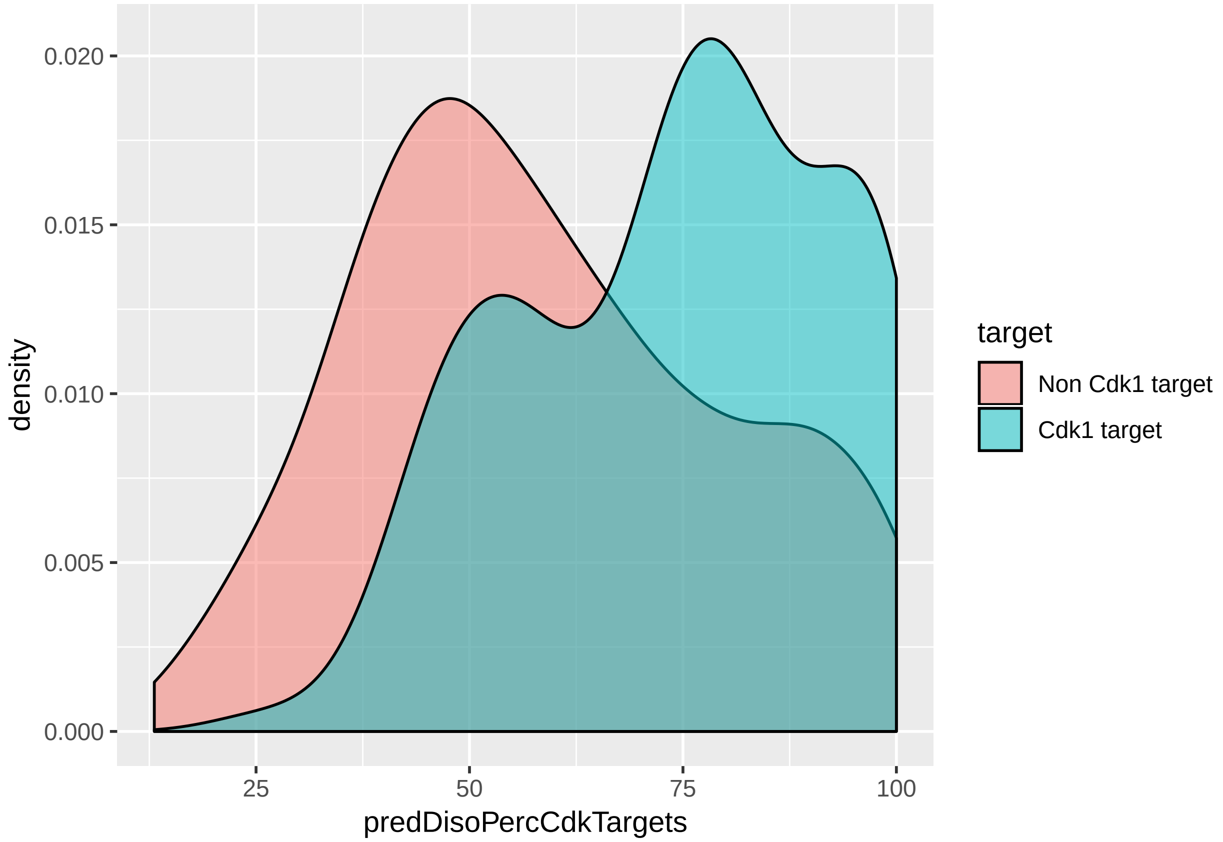
Figure 5: Figure 5:distribution of the lenghts of disorder stretches

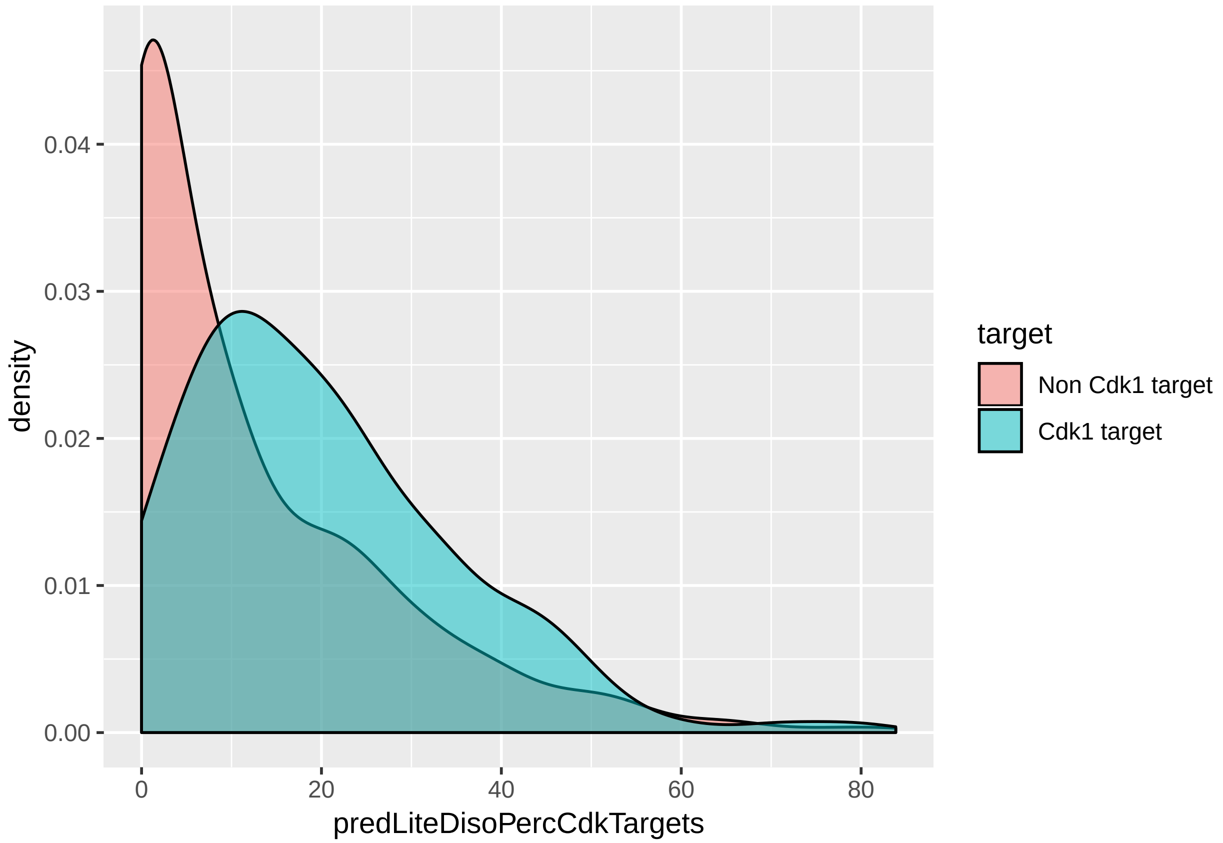
Lenght distribution in the detection of IDRs

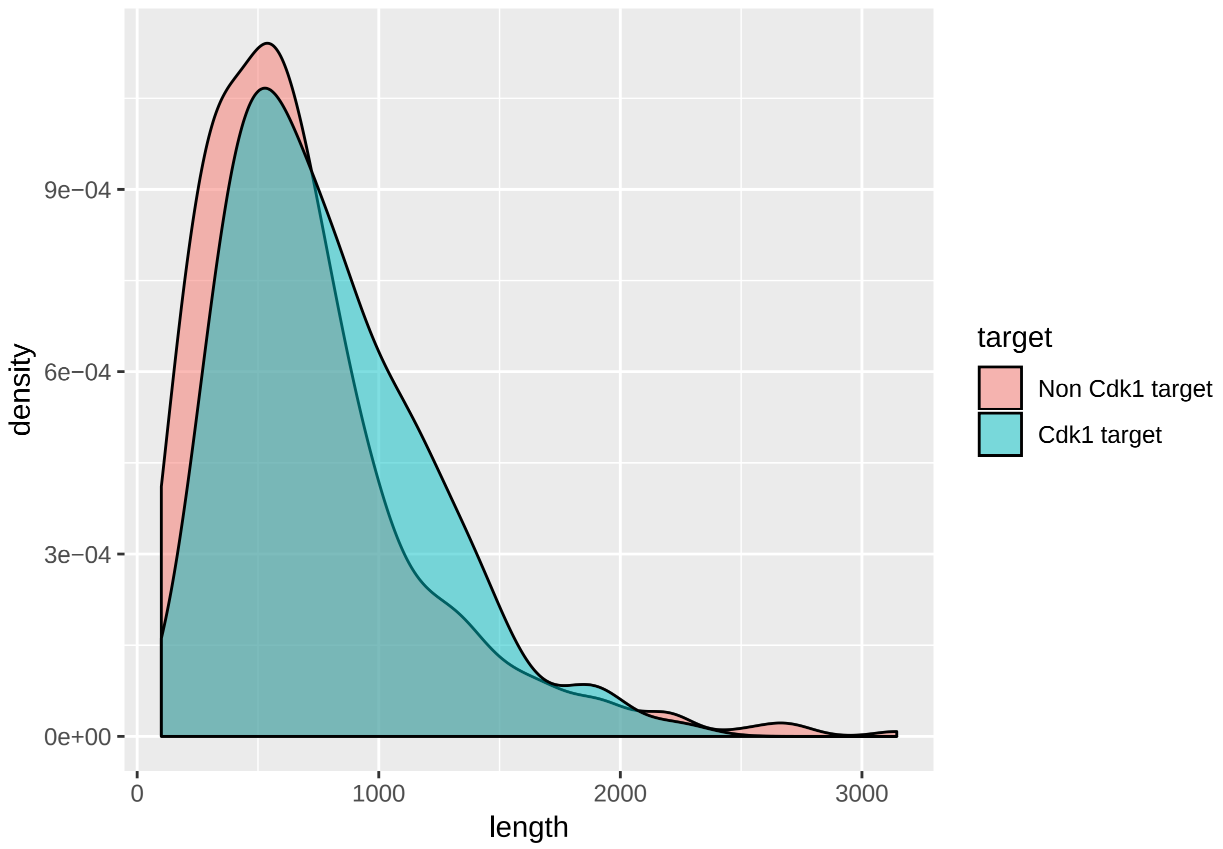
None of the methods for estimating the disorder regions of proteins seem to be in agreement with the other methods in the percentage of disorder residues for a given protein. The correlation coefficients are low, and the ratio of putative phosphorylation sites does not have any significant impact on these calculations. Moreover, the distribution of the percentage of phosphorylatable residues appears to be completely random in all the pairwise comparison scatterplots. The lack of correlation in this metric for different methods does not directly imply that the methods are not identifying the same regions. In fact, all of the aminoacids classified as disordered by the MobiDB lite method, for all the proteins, were also detected by the simple method. Disparities in the lengths of the disorder stretches lead to these observed effects.

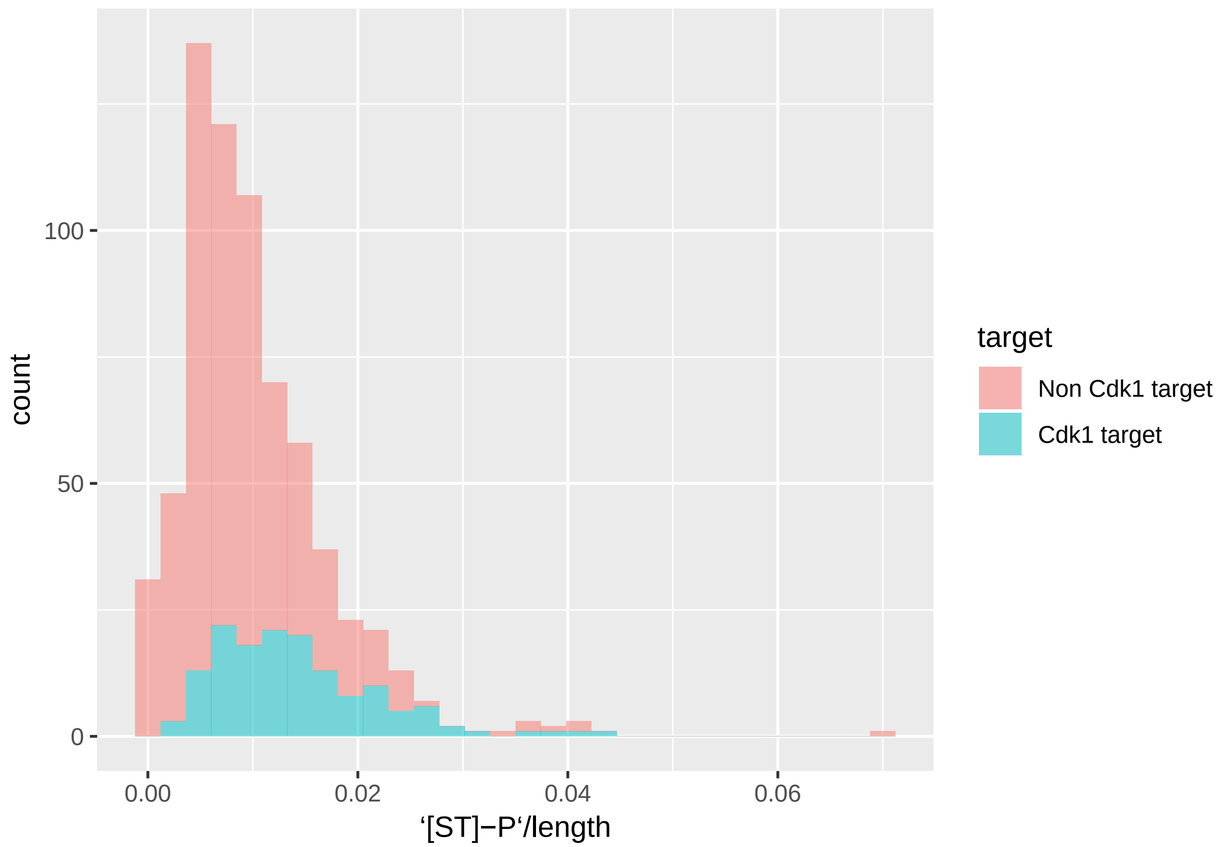
Nevertheless, inconsistencies in the measurements of the ratio of disordered amino acids between different methods do not indicate that the metric could not be used as an estimator of global disorder status in proteins. However, we should certainly be cautious about comparing values originated from different sources. If the percentage of disorder residues correlates well with one or more CDK phosphorylation target properties it could be informative regardless if the estimation is highly accurate.

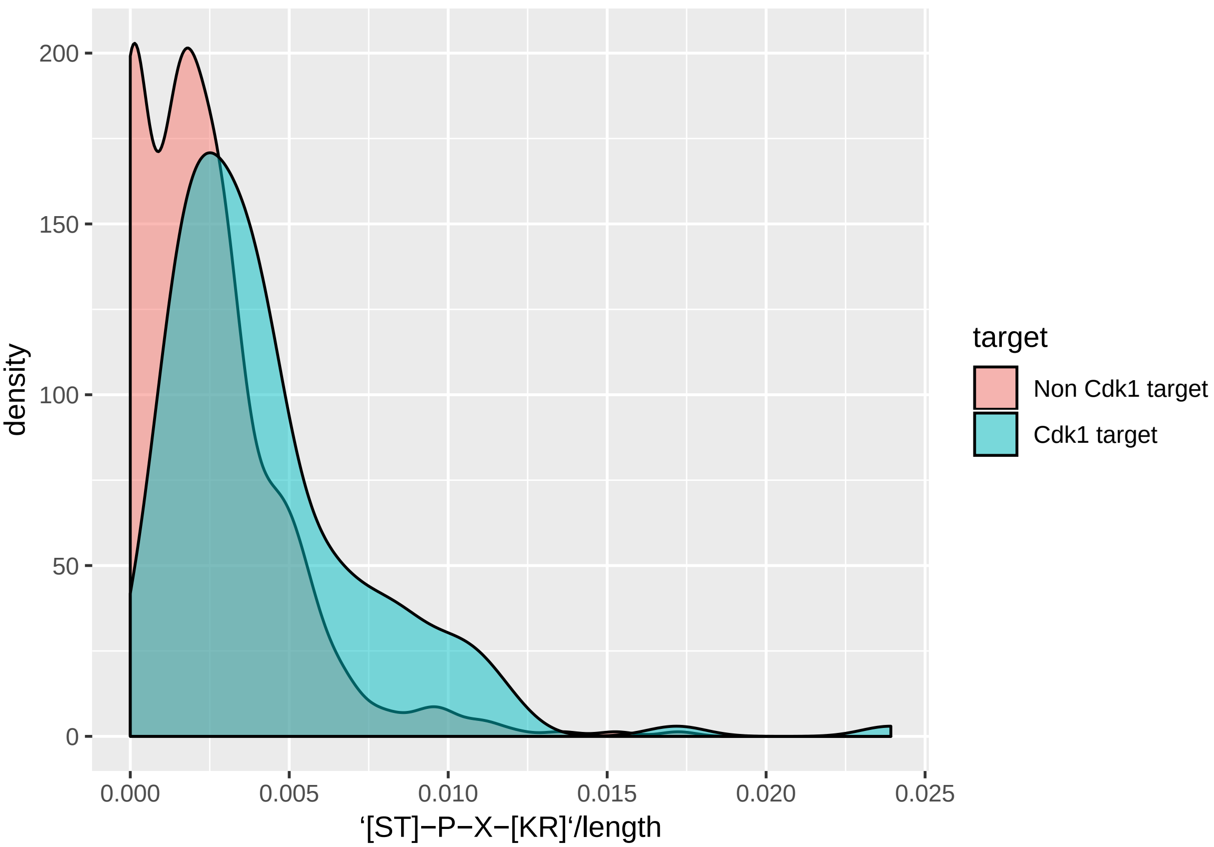
Cdk1 targets

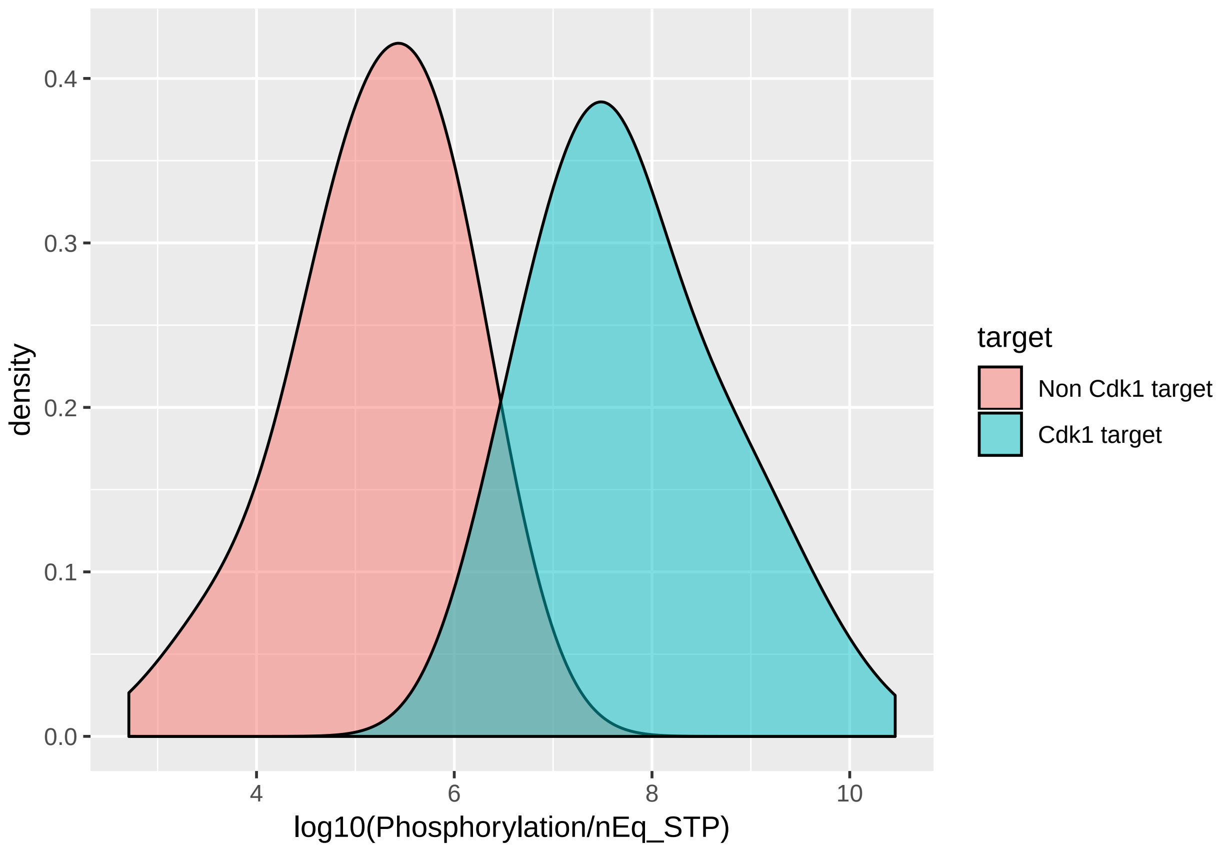


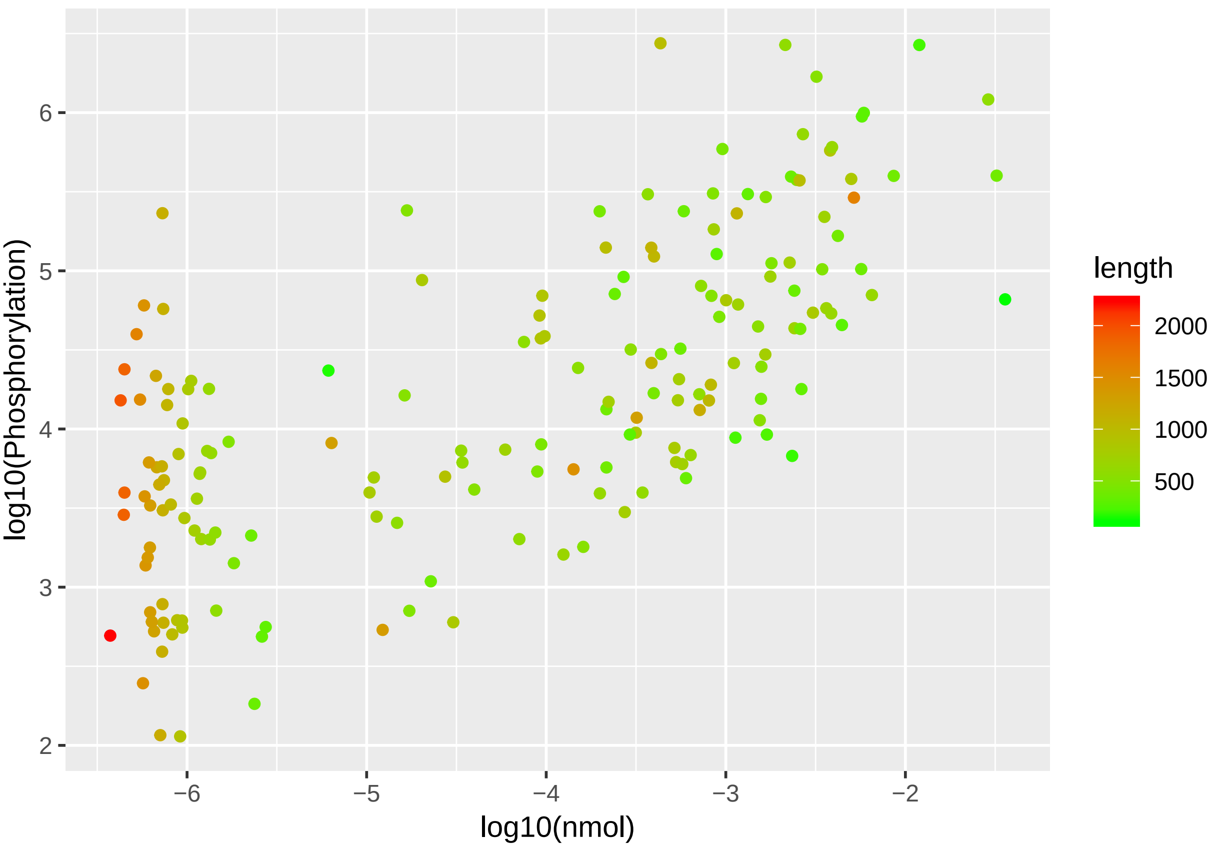



density
9e−04
6e−04
3e−04
0e+00
0
1000
2000
3000
length
target
Non Cdk1 target
Cdk1 target

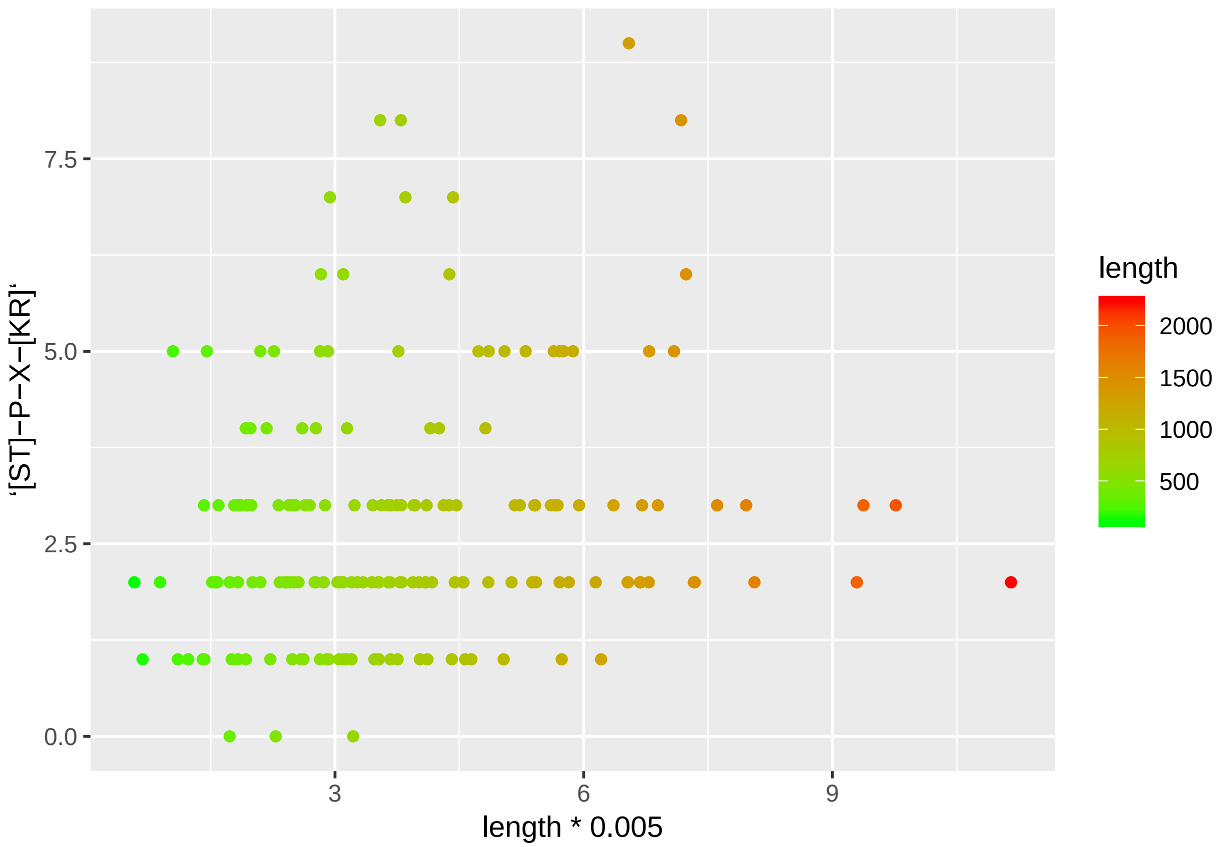

count
100
50
0
0.00
0.02
0.04
0.06
‘[ST]−P‘/length
target
Non Cdk1 target
Cdk1 target


density
200
150
100
50
0
0.000
0.005
0.010
0.015
0.020
0.025
‘[ST]−P−X−[KR]‘/length
target
Non Cdk1 target
Cdk1 target

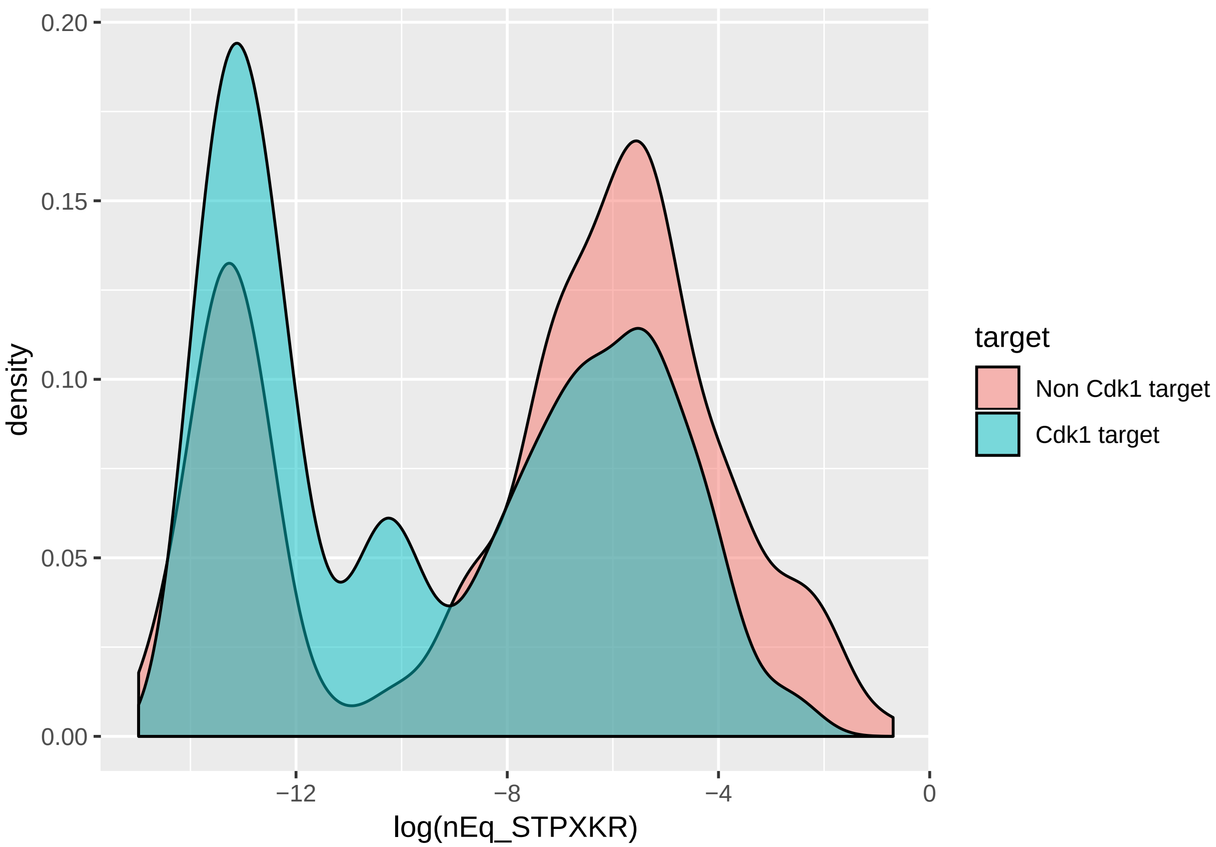

density
0.4
0.3
0.2
0.1
0.0
4
6
8
10
log10(Phosphorylation/nEq_STP)
target
Non Cdk1 target
Cdk1 target

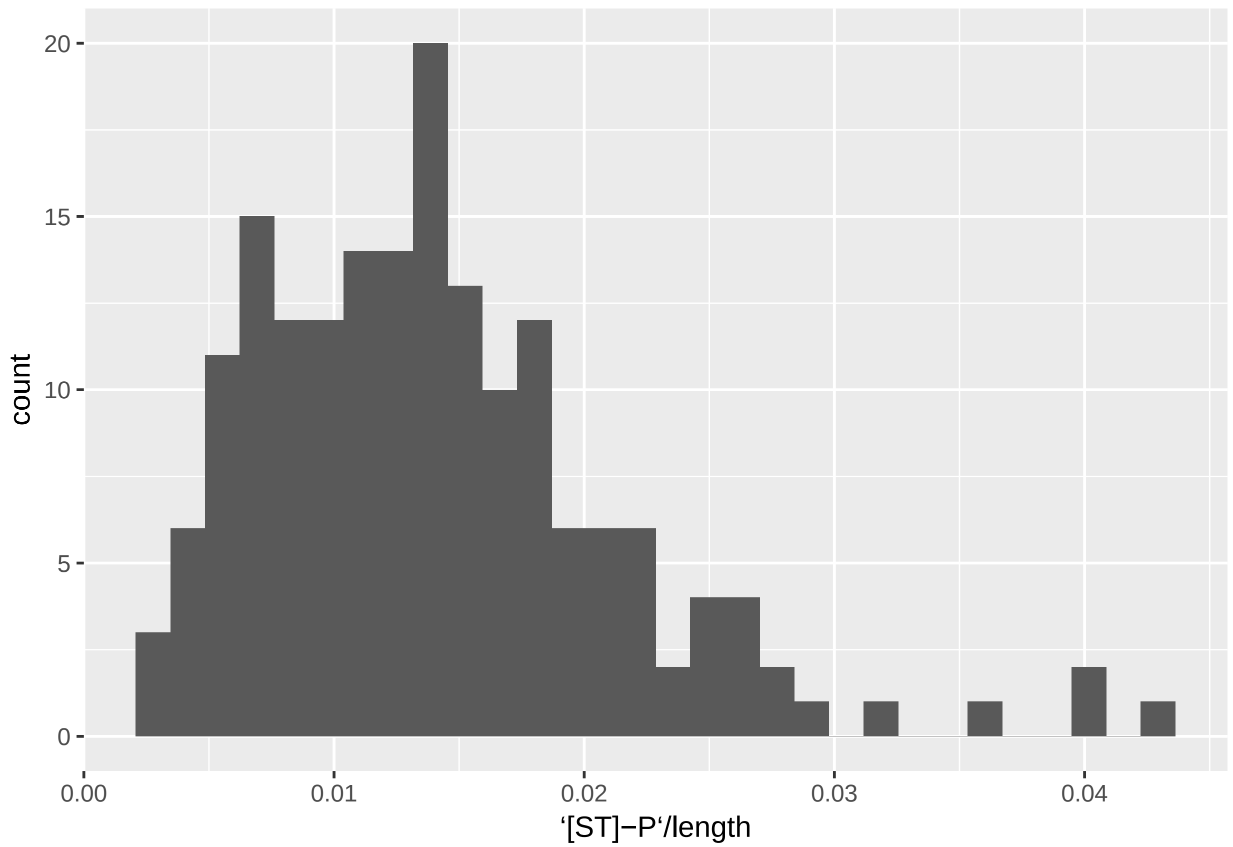


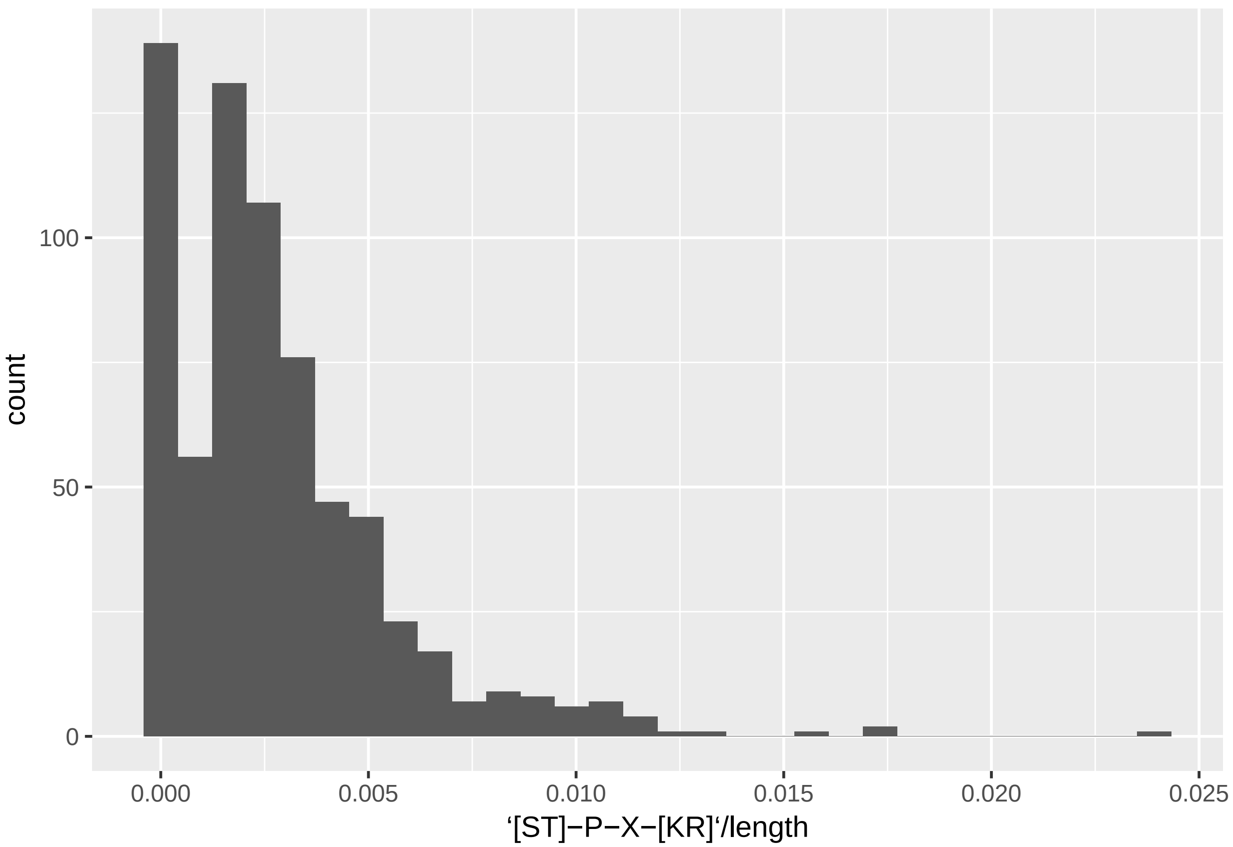
log10(Phosphorylation)
6
5
4
3
2
−6
−5
−4
−3
−2
log10(nmol)
length
2000
1500
1000
500

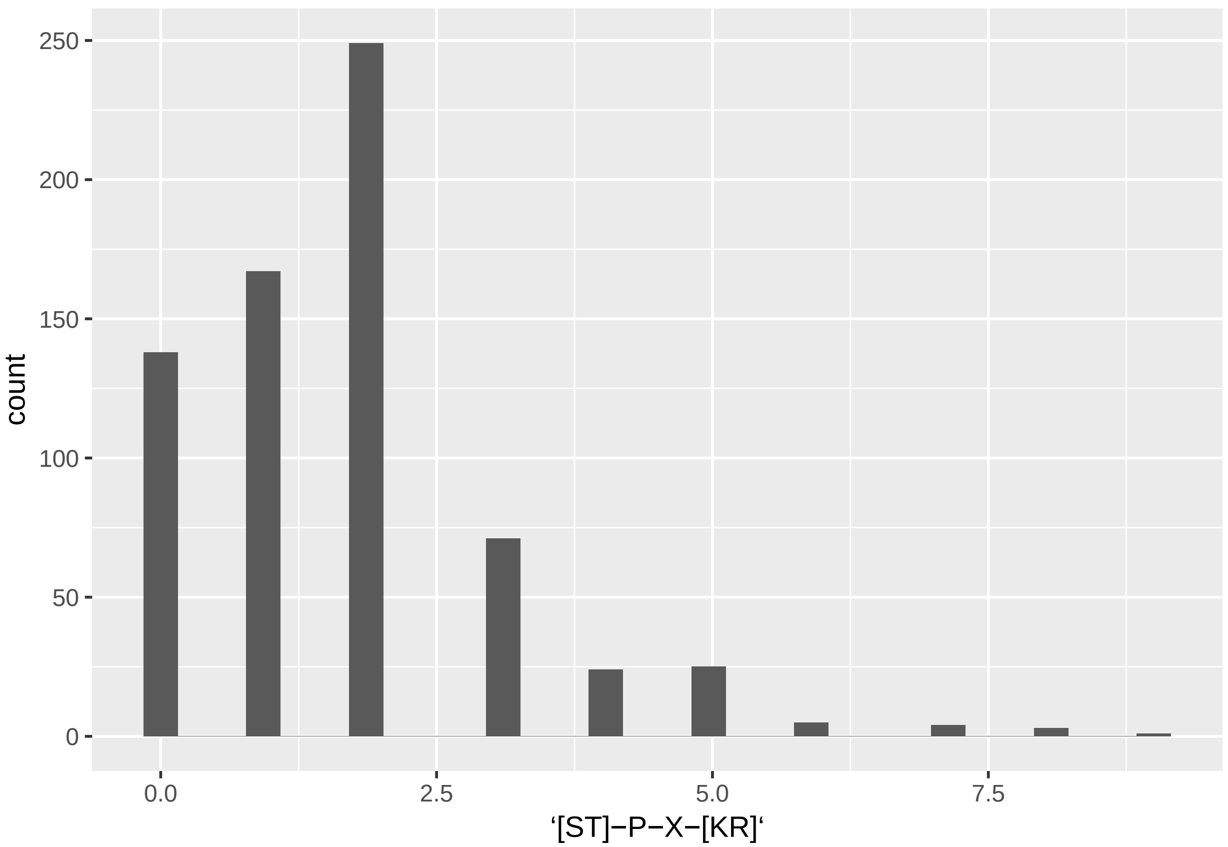


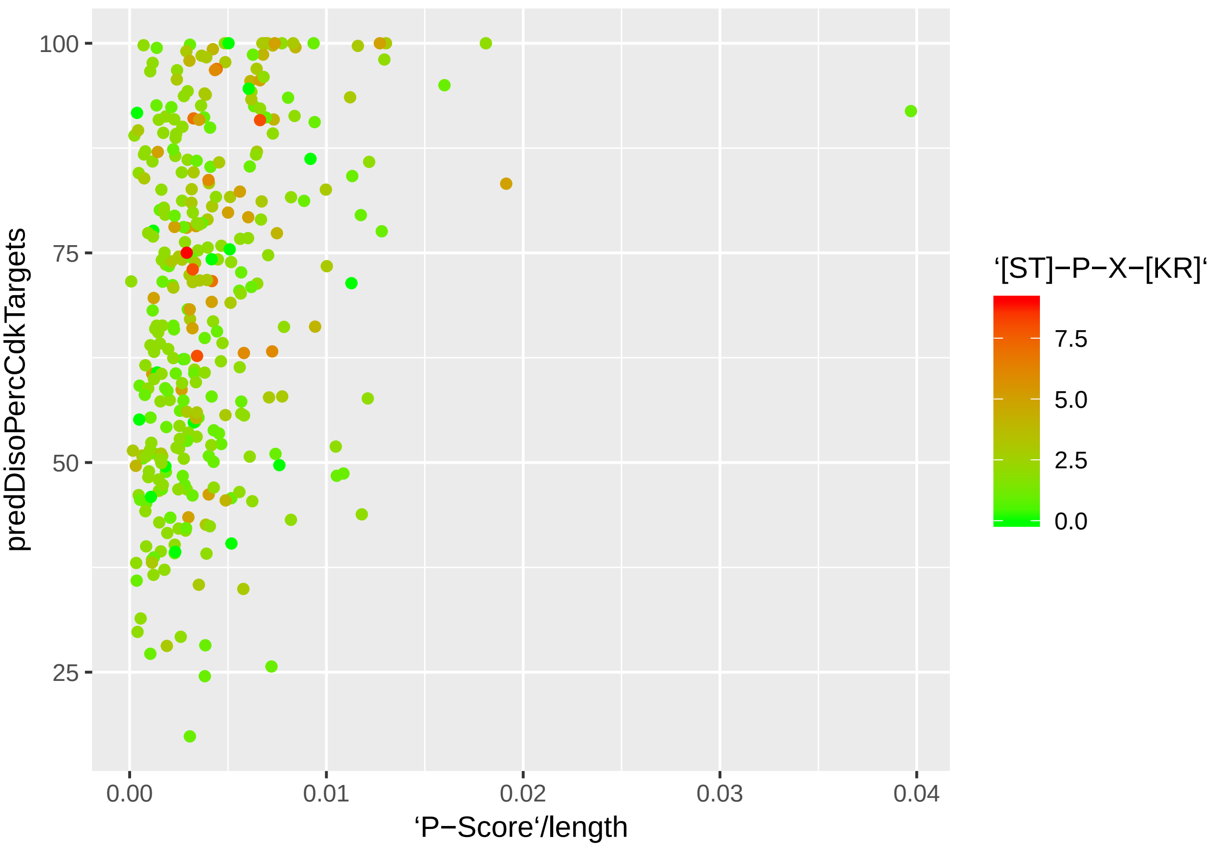
'[ST]−P−X−[KR]'
7.5
5.0
2.5
0.0
3
6
9
length * 0.005
length
2000
1500
1000
500

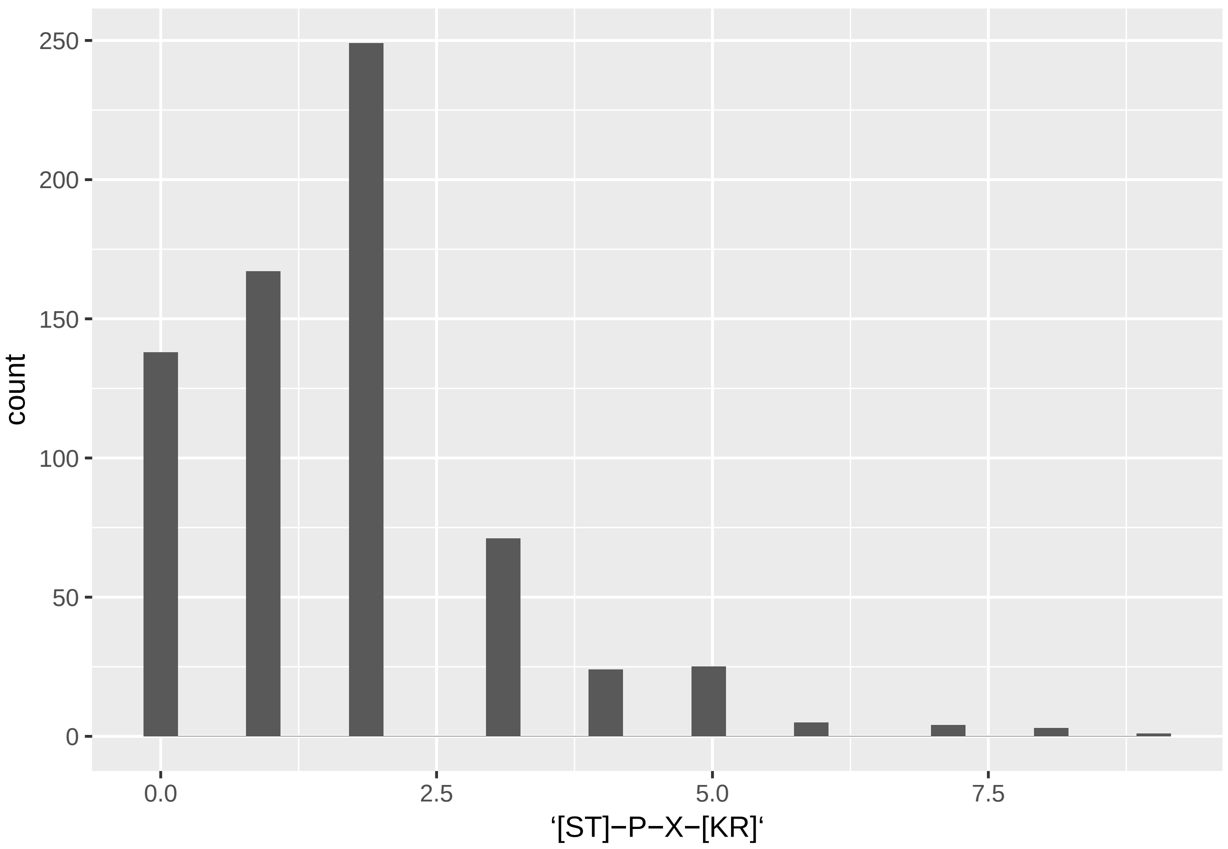

0.20
0.15
0.10
0.05
0.00
density
−12
−8
−4
0
log(nEq_STPXKR)
target
Non Cdk1 target
Cdk1 target

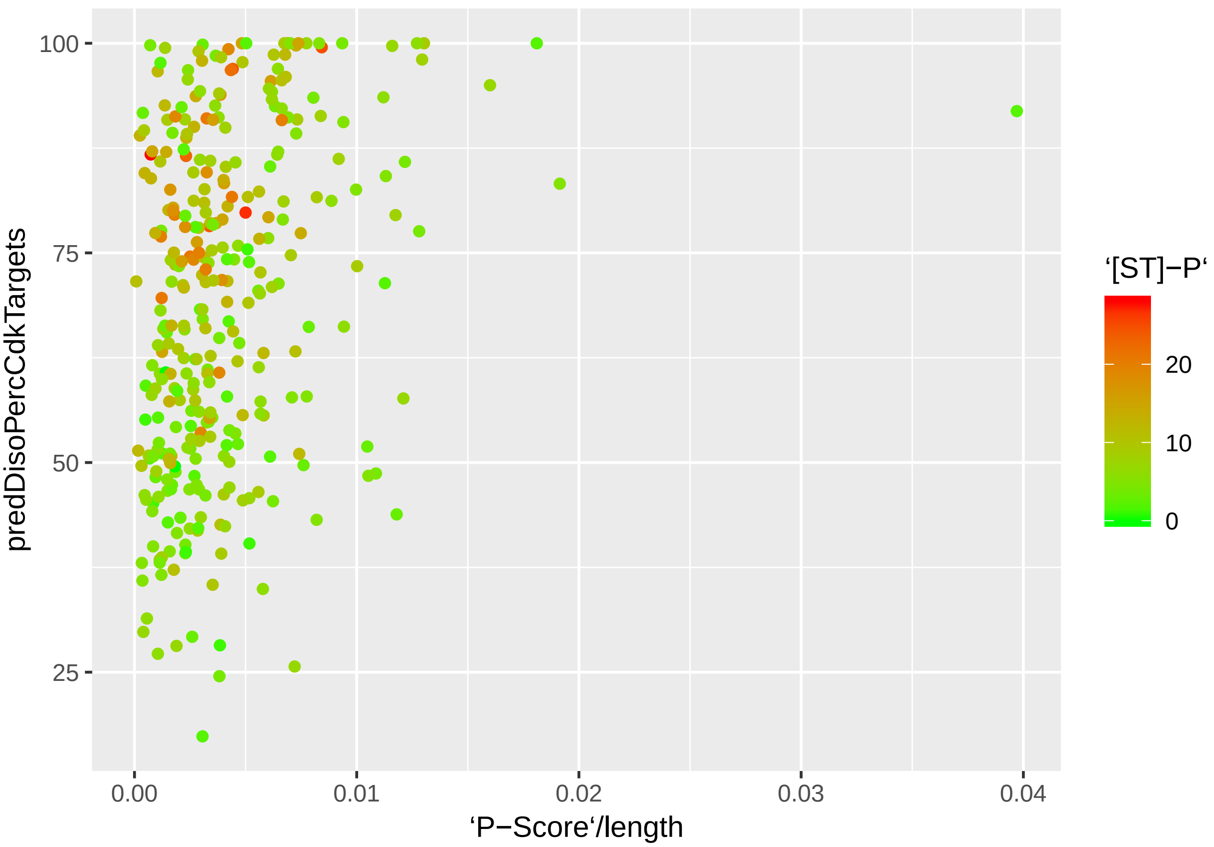

count
20
15
10
5
0
0.00
0.01
0.02
0.03
0.04
‘[ST]−P‘/length

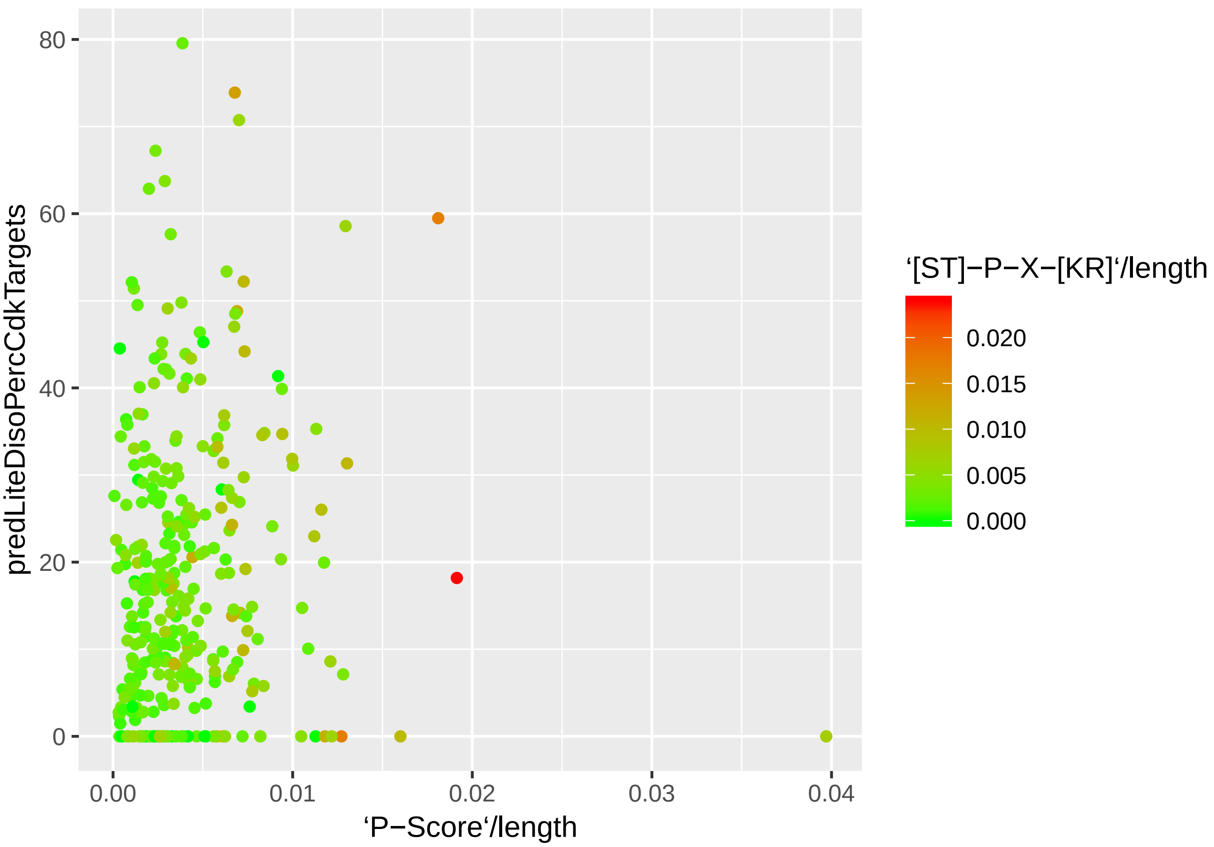

count
100
50
0
0.000
0.005
0.010
0.015
0.020
0.025
‘[ST]−P−X−[KR]‘/length

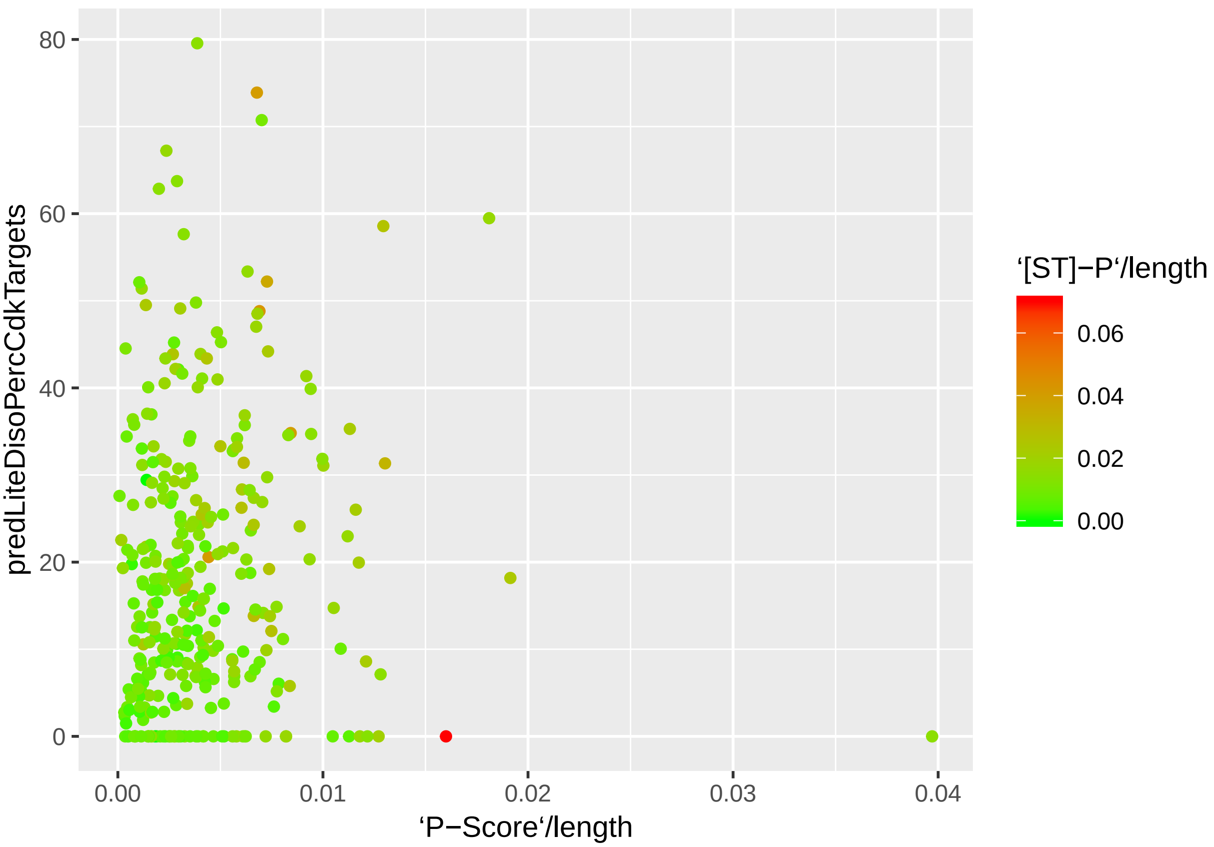

count
250
200
150
100
50
0
0.0
2.5
5.0
7.5
‘[ST]−P−X−[KR]‘

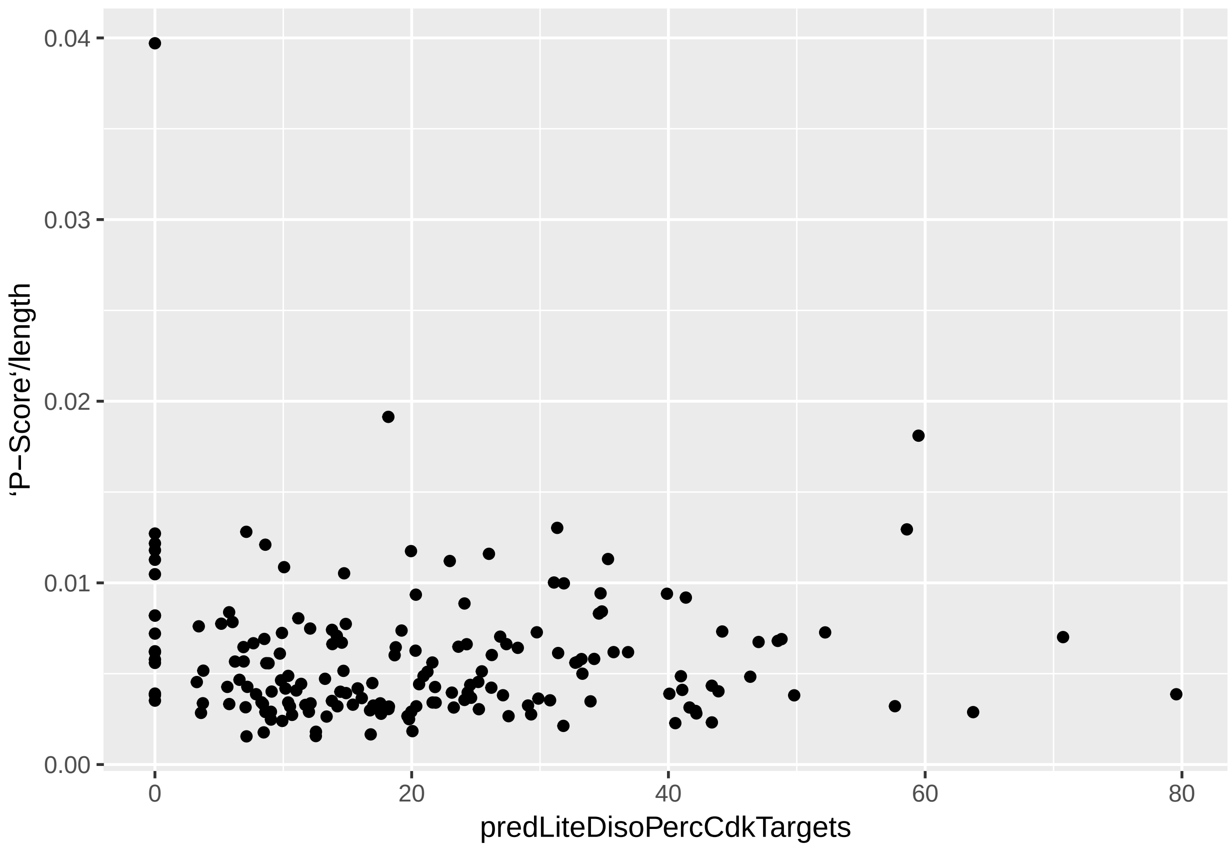

100
75
50
25
0.00
0.01
0.02
0.03
0.04
‘P−Score‘/length
predDisoPercCdkTargets
‘[ST]−P−X−[KR]‘
7.5
5.0
2.5
0.0

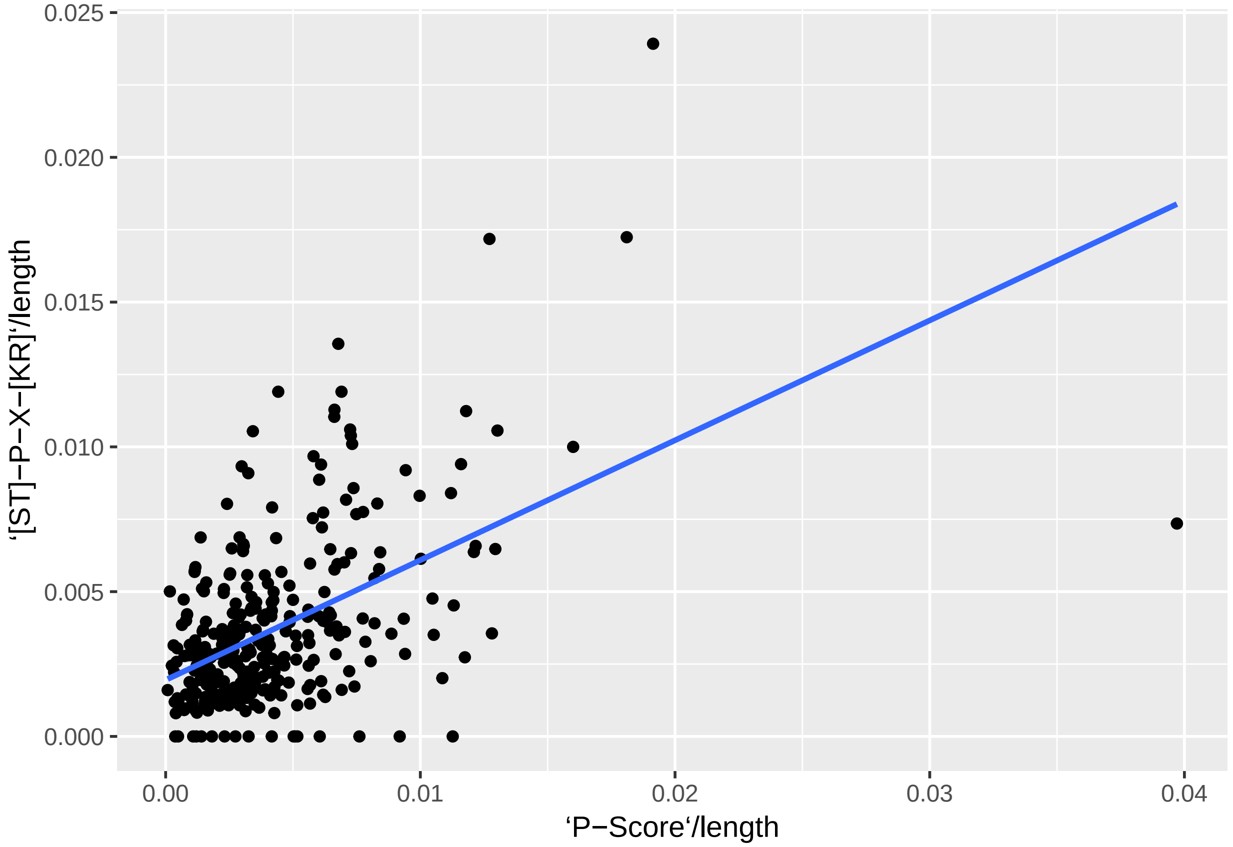

count
250
200
150
100
50
0
0.0
2.5
5.0
7.5
‘[ST]−P−X−[KR]‘

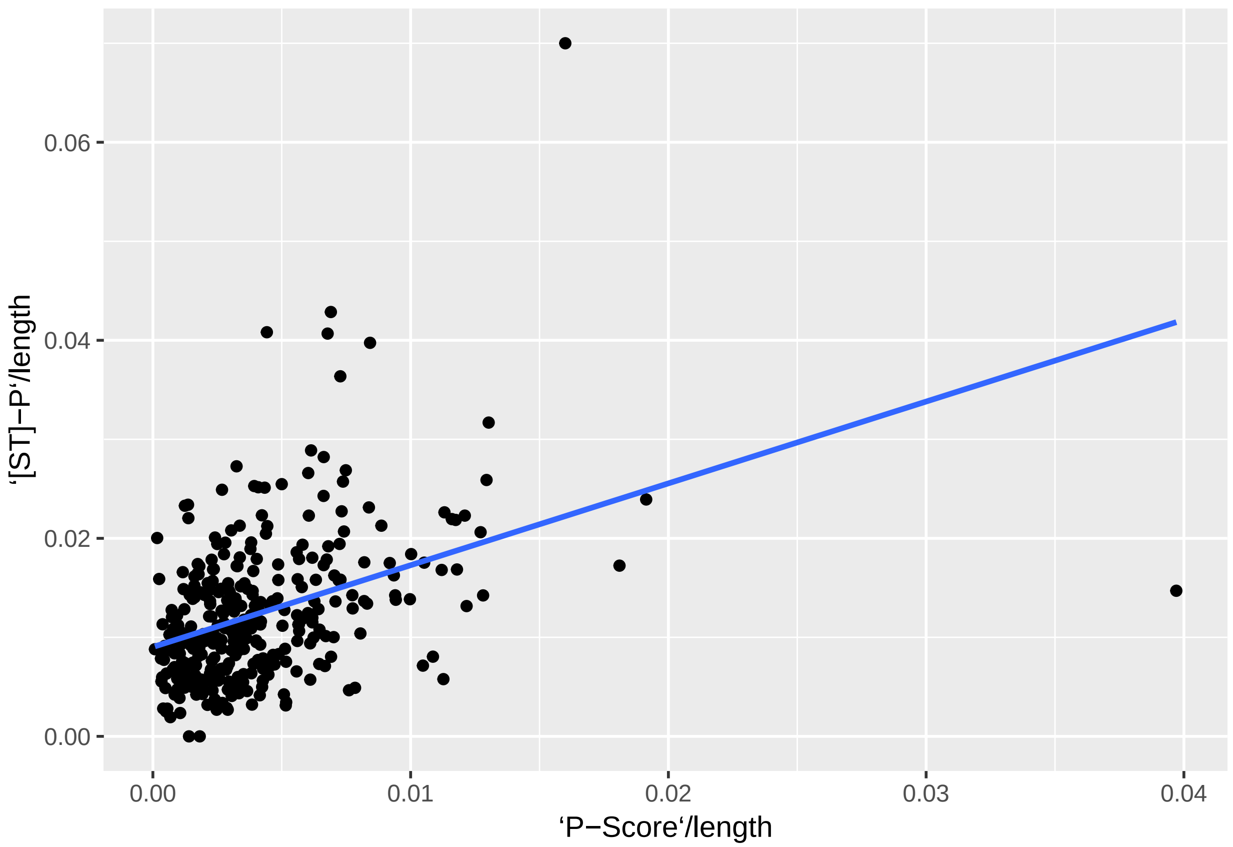

predDisoPercCdkTargets
100
75
50
25
0.00
0.01
0.02
0.03
0.04
‘P−Score‘/length
‘[ST]−P‘
20
10
0

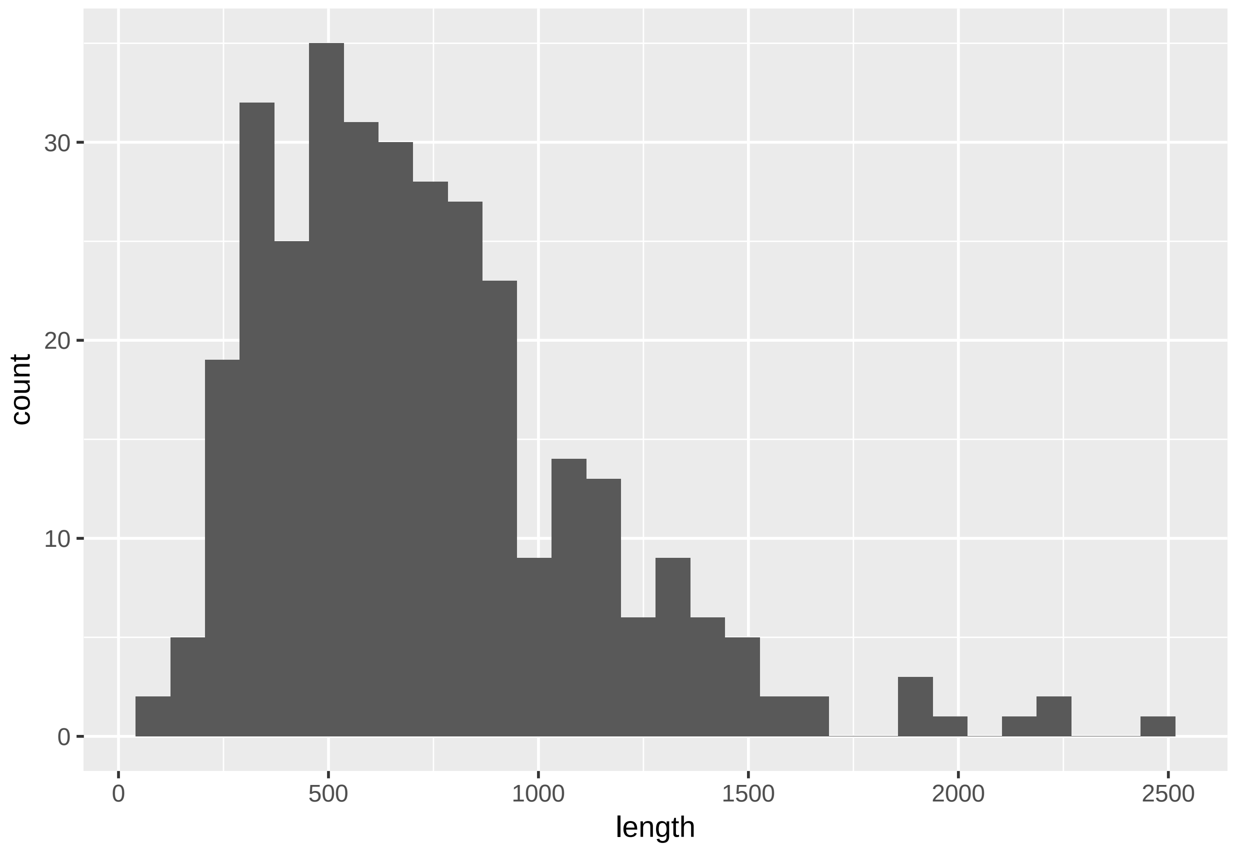

80
60
40
20
0
predLiteDisoPercCdkTargets
0.00
0.01
0.02
0.03
0.04
‘P−Score‘/length
‘[ST]−P−X−[KR]‘/length
0.020
0.015
0.010
0.005
0.000


predLiteDisoPercCdkTargets
80
60
40
20
0
0.00
0.01
0.02
0.03
0.04
‘P−Score‘/length
‘[ST]−P‘/length
0.06
0.04
0.02
0.00


‘P−Score‘/length
0.04
0.03
0.02
0.01
0.00
0
20
40
60
80
predLiteDisoPercCdkTargets


0.025
0.020
0.015
0.010
0.005
0.000
‘[ST]−P−X−[KR]‘/length
0.00
0.01
0.02
0.03
0.04
‘P−Score‘/length


0.06
0.04
0.02
0.00
‘[ST]−P‘/length
0.00
0.01
0.02
0.03
0.04
‘P−Score‘/length



P-score does not consider the size nor the number of consensus phospho sites for Cdk1. P-Score = log(Phosphorylation/ Protein level)

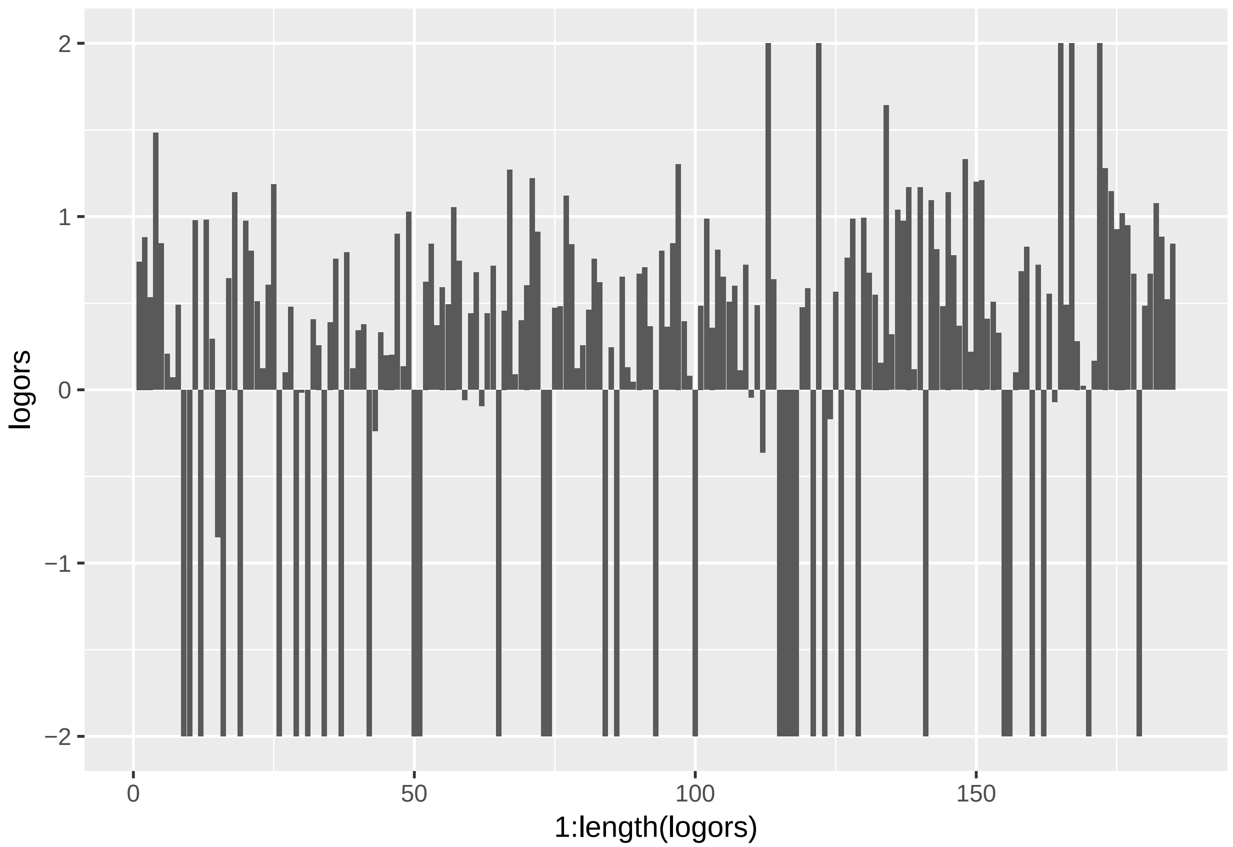
For correcting for protein lenght we can do P-Score Mol = log(Phosphorylation/ number of protein umols) = log(Phosphorylation/ (Protein Level / (length)*120))

Correcting for number of moles in the reaction or for number of Eq of [ST]-P sites or [ST]-P-X-[KR] doesn't change much, probably due to the fact that this correction is linear and the final effect in the logarithm scale of the Score is not relevant

```
##
##  Woolf-test on Homogeneity of Odds Ratios (no 3-Way assoc.)
##
## data:  stratContingencyArray
## X-squared = 171.13, df = 184, p-value = 0.7428
```

```
##    [1] 0.2560122523 1.0000000000 1.0000000000 0.3445599225 0.2459207757
##    [6] 1.0000000000 1.0000000000 0.5922847936 1.0000000000 1.0000000000
##   [11] 0.4016857625 0.5545432737 0.2097117287 0.0416281287 0.5272820744
##   [16] 1.0000000000 1.0000000000 0.2169287711 0.2244391776 1.0000000000
##   [21] 0.0511842388 0.5979574342 1.0000000000 1.0000000000 0.0012334189
##   [26] 1.0000000000 1.0000000000 0.1956012048 1.0000000000 0.3118847582
##   [31] 0.0989783205 0.1267951888 1.0000000000 1.0000000000 1.0000000000
##   [36] 0.0890168367 0.6698201781 0.0023870925 1.0000000000 0.0029716692
##   [41] 0.0062127646 0.3251046576 0.0510919002 0.6993794389 1.0000000000
##   [46] 1.0000000000 0.0398078171 0.4608077291 1.0000000000 1.0000000000
##   [51] 1.0000000000 0.0690275253 0.1224396075 1.0000000000 1.0000000000
##   [56] 0.0918615889 0.0090176482 0.1609956131 0.3439383432 0.2223088150
##   [61] 0.1951122421 0.3279087814 1.0000000000 1.0000000000 1.0000000000
##  [66] 0.0210912360 0.4560385565 0.3354970917 0.2457090823 0.3072515448
##  [71] 0.7006359698 1.0000000000 0.1005439710 0.1130325664 0.0030131149
##  [76] 0.0613943332 0.0009159634 1.0000000000 1.0000000000 0.3701044220
##  [81] 0.1849923924 0.6281675880 1.0000000000 0.7003761028 0.7558743328
##  [86] 1.0000000000 0.0574614889 1.0000000000 0.0575045352 0.7372455902
##  [91] 1.0000000000 0.1368173965 1.0000000000 0.0032267293 0.0466429858
##  [96] 0.0002229540 0.1503696232 0.0933610887 0.6186725076 1.0000000000
## [101] 0.0001352781 0.1647046367 1.0000000000 0.7005605239 0.2453632495
## [106] 0.0697415508 0.2014878011 0.0170758505 0.0853247588 1.0000000000
## [111] 0.0207811221 1.0000000000 0.1010455747 0.0040820959 1.0000000000
## [116] 0.0398988319 0.0212816326 0.6698201781 0.0445549189 0.3010504901
## [121] 1.0000000000 1.0000000000 0.2578788813 1.0000000000 0.6787626067
## [126] 0.5822224338 0.2027914306 1.0000000000 0.3252545464 0.4998231554
## [131] 0.0765295441 0.0088932185 0.5917355169 0.5780537136 1.0000000000
## [136] 1.0000000000 0.7000621675 0.0304913776 1.0000000000 0.4107612396
## [141] 1.0000000000 0.1100015237 0.1341417725 0.0473322840 1.0000000000
## [146] 0.2162902312 0.0186695864 1.0000000000 1.0000000000 1.0000000000
## [151] 0.3878500938 1.0000000000 0.0066922970 0.2439133076 1.0000000000
## [156] 1.0000000000 1.0000000000 1.0000000000 0.1446771475 0.5896511353
## [161] 0.0049462159 0.3394409373 0.3639568508 0.6064009131 1.0000000000
## [166] 0.0887197667 0.1098837769 0.7472272065 1.0000000000 0.3766806983
## [171] 0.0524161358 0.3022987330 1.0000000000 0.0610707147 1.0000000000
## [176] 0.0647439830 0.0266130908 0.3651737576 0.0416692004 1.0000000000
## [181] 0.0059767641 0.0761225204 0.6007763263 0.1248043248 0.1878710282

## 
##  Woolf-test on Homogeneity of Odds Ratios (no 3-Way assoc.)
## 
## data:  liteStratContingencyArray
## X-squared = 281.82, df = 184, p-value = 4.602e-06
```



REFERENCES