Protein_aggregation_vs_subcellular_location

Initialization

Install all the needed libraries.

```
#install.packages("UniprotR")
#install.packages("BiocManager")
#BiocManager::install("Biostrings")
#BiocManager::install("GenomicAlignments")
#install.packages("sjmisc")
#install.packages("hash")
#install.packages("dplyr")
#install.packages("Peptides")
```

Load the needed libraries.

```
library(UniprotR)
library(sjmisc) # Used for str_contains
```

```
## Install package "strengejacke" from GitHub (`devtools::install_github("strengejacke/strengejacke")`)
```

```
library(hash) # Used to make a disctionary
```

```
## hash-2.2.6.1 provided by Decision Patterns
```

```
library(dplyr) # Used for aggregate (get avg tango score of protein)
```

```
##
## Caricamento pacchetto: 'dplyr'
```

```
## I seguenti oggetti sono mascherati da 'package:stats':
##
##     filter, lag
```

```
## I seguenti oggetti sono mascherati da 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(Peptides) # Used for charge (get the charge of a peptide sequence) ETC
library(ggplot2)
library(dplyr)
library(ggrepel)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.6     v purrr   0.3.4
## v tidyr   1.1.4     v stringr 1.4.0
```

```
## v readr   2.1.0     v forcats 0.5.1
## -- Conflicts ----------------------------------------- tidyverse_conflicts() --
## x tibble::add_case()  masks sjmisc::add_case()
## x dplyr::filter()     masks stats::filter()
## x purrr::is_empty()   masks sjmisc::is_empty()
## x dplyr::lag()        masks stats::lag()
## x tidyr::replace_na() masks sjmisc::replace_na()
```

Define the working directory.

```
directory = dirname(rstudioapi::getSourceEditorContext()$path) # Should work when data is placed in sam
setwd(directory)
```

Load the data.

```
source("load_data.R")
return_value = read_data() # See load_data.R
data = return_value$data
hashed_proteins = return_value$hashed_proteins
rm(return_value) # To get rid of the extra memory usage
```

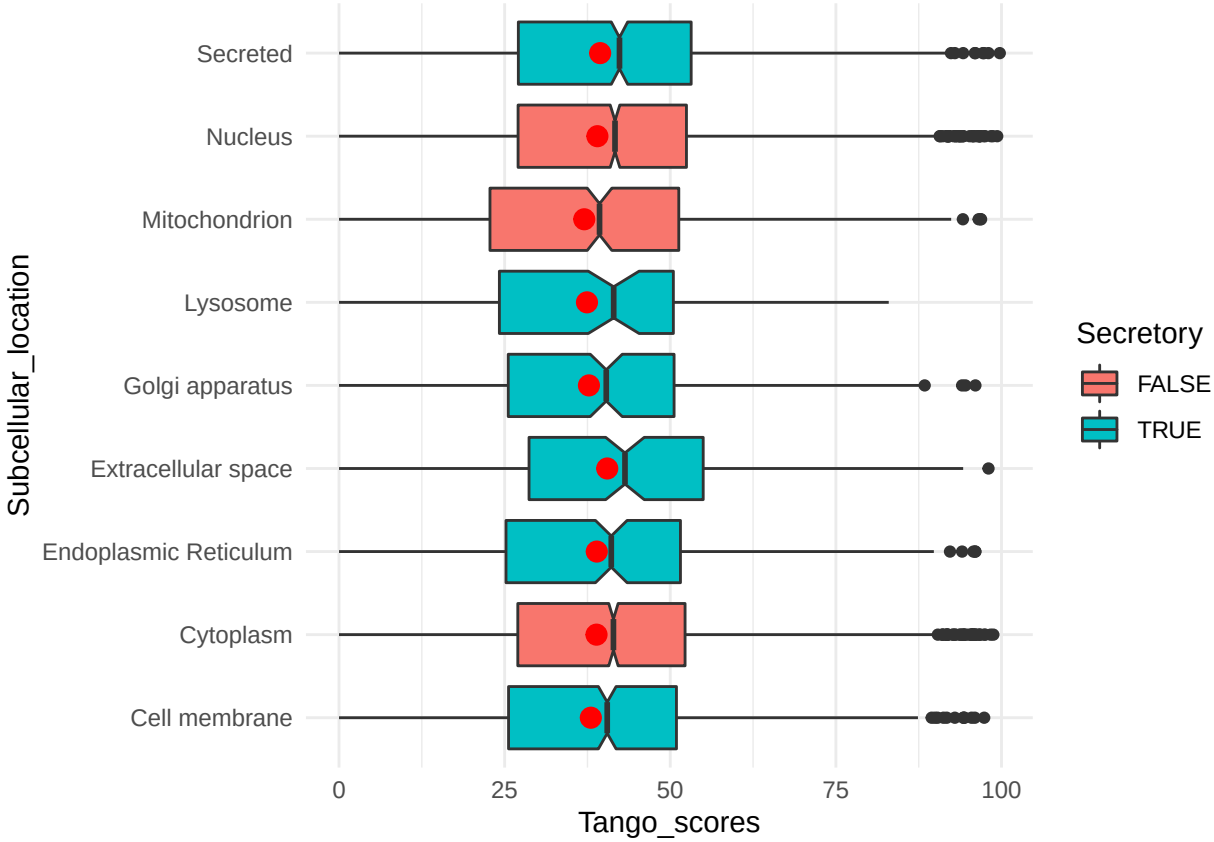
TANGO scores

```
source("tango_scores.R")
```

Average tango score distribution in all proteins

```
plot_average_tango_scores_complete_proteins()
```

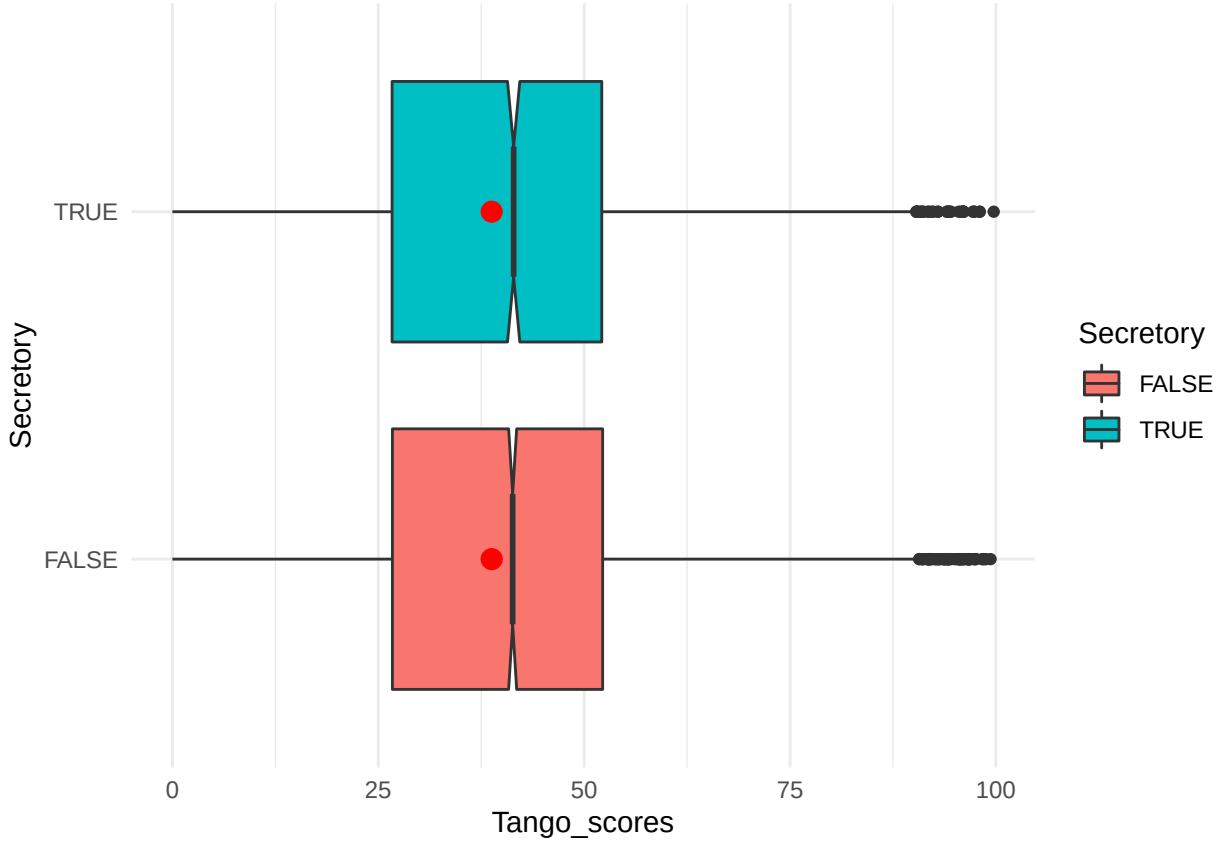
```
## Warning: Removed 3 rows containing non-finite values (stat_boxplot).
## Warning: Removed 3 rows containing non-finite values (stat_summary).
```




```
plot_average_tango_scores_complete_proteins_joined_secreted()
```

```
## Warning: Removed 3 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 3 rows containing non-finite values (stat_summary).
```

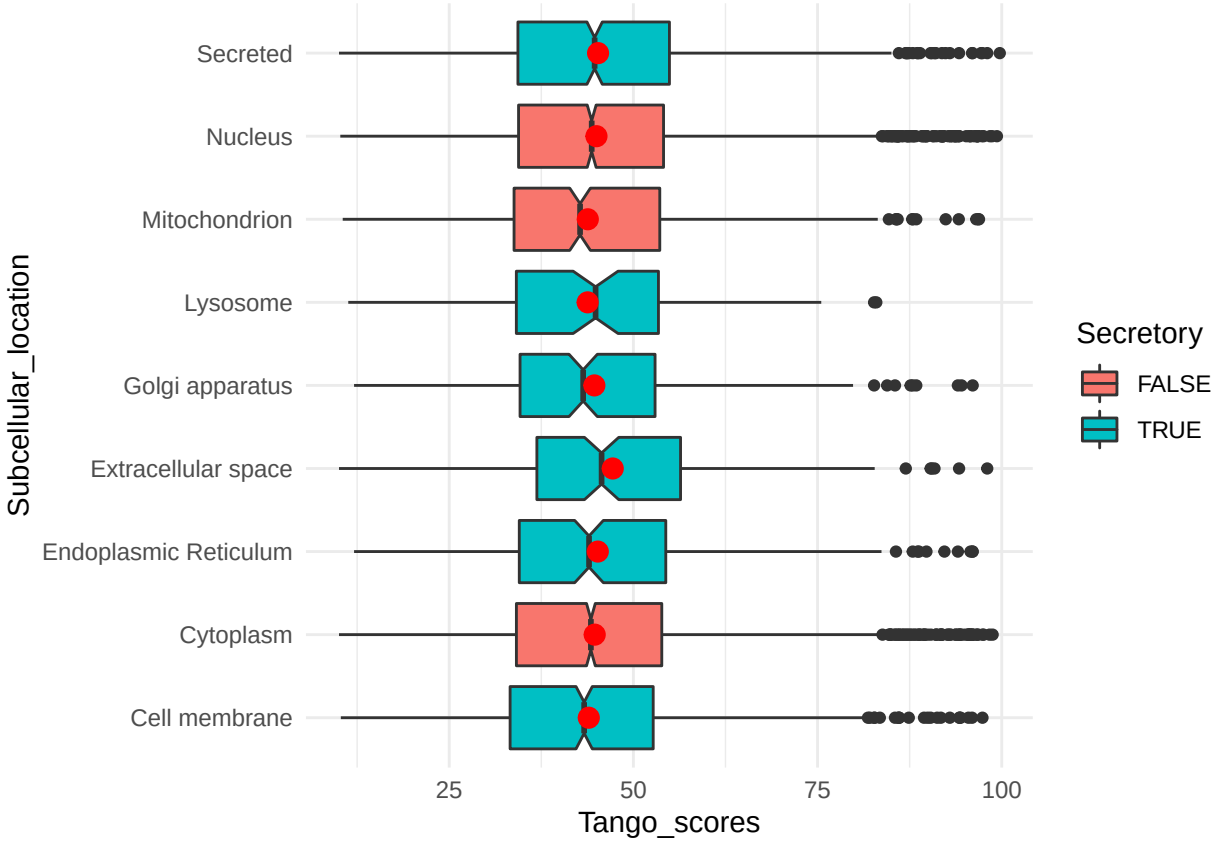


Average tango score distribution in proteins with an APR region

```
plot_average_tango_scores_APR_proteins()
```

```
## Warning: Removed 1317 rows containing non-finite values (stat_boxplot).
```

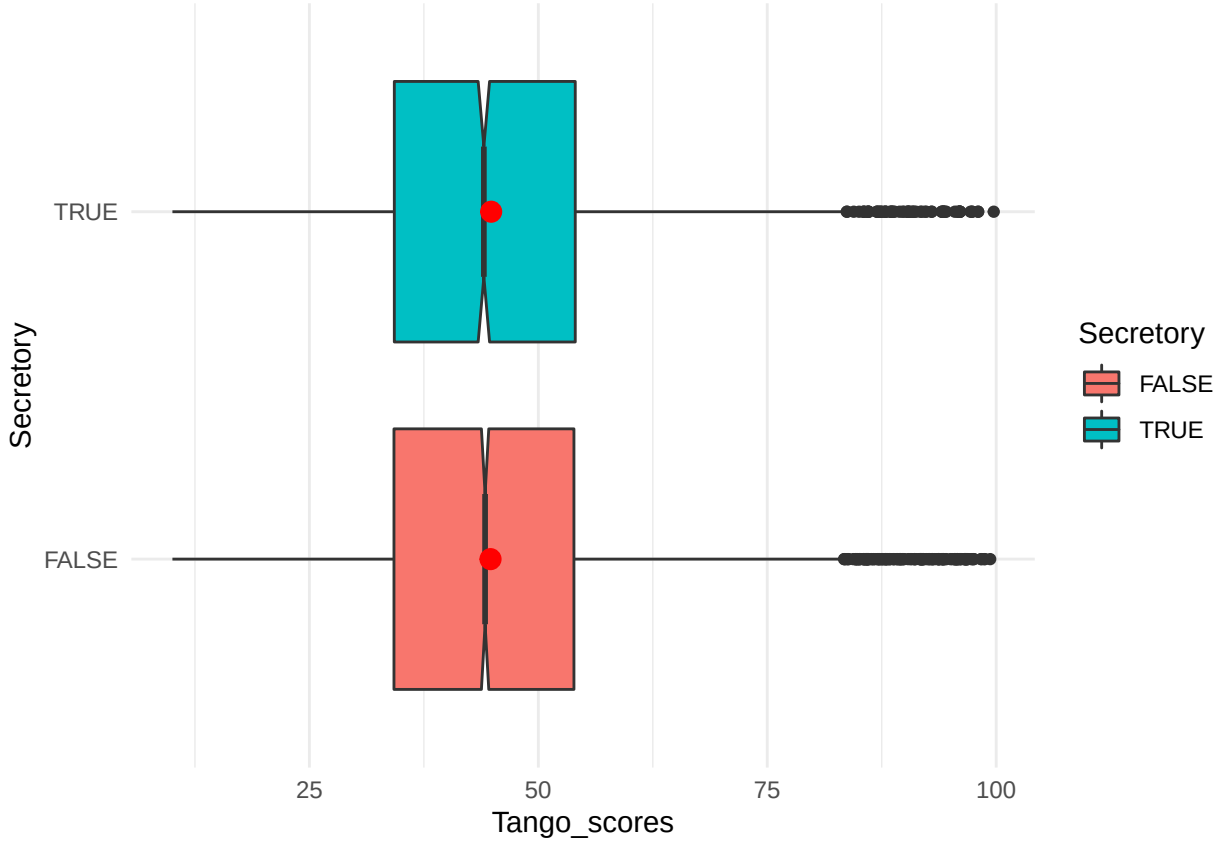
```
## Warning: Removed 1317 rows containing non-finite values (stat_summary).
```




```
plot_average_tango_scores_APR_proteins_joined_secreted()
```

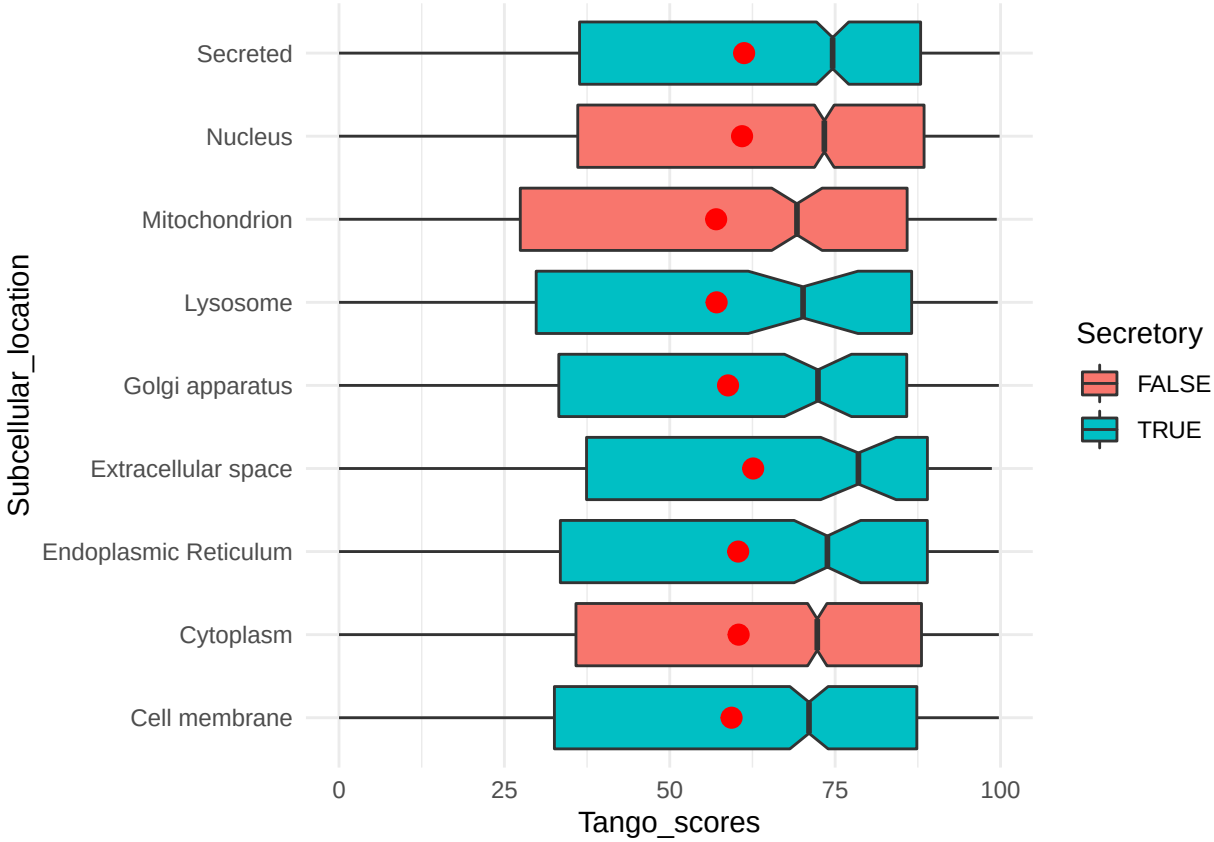
```
## Warning: Removed 1317 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 1317 rows containing non-finite values (stat_summary).
```



Distribution of maximal tango scores

```
plot_max_tango_scores()
```

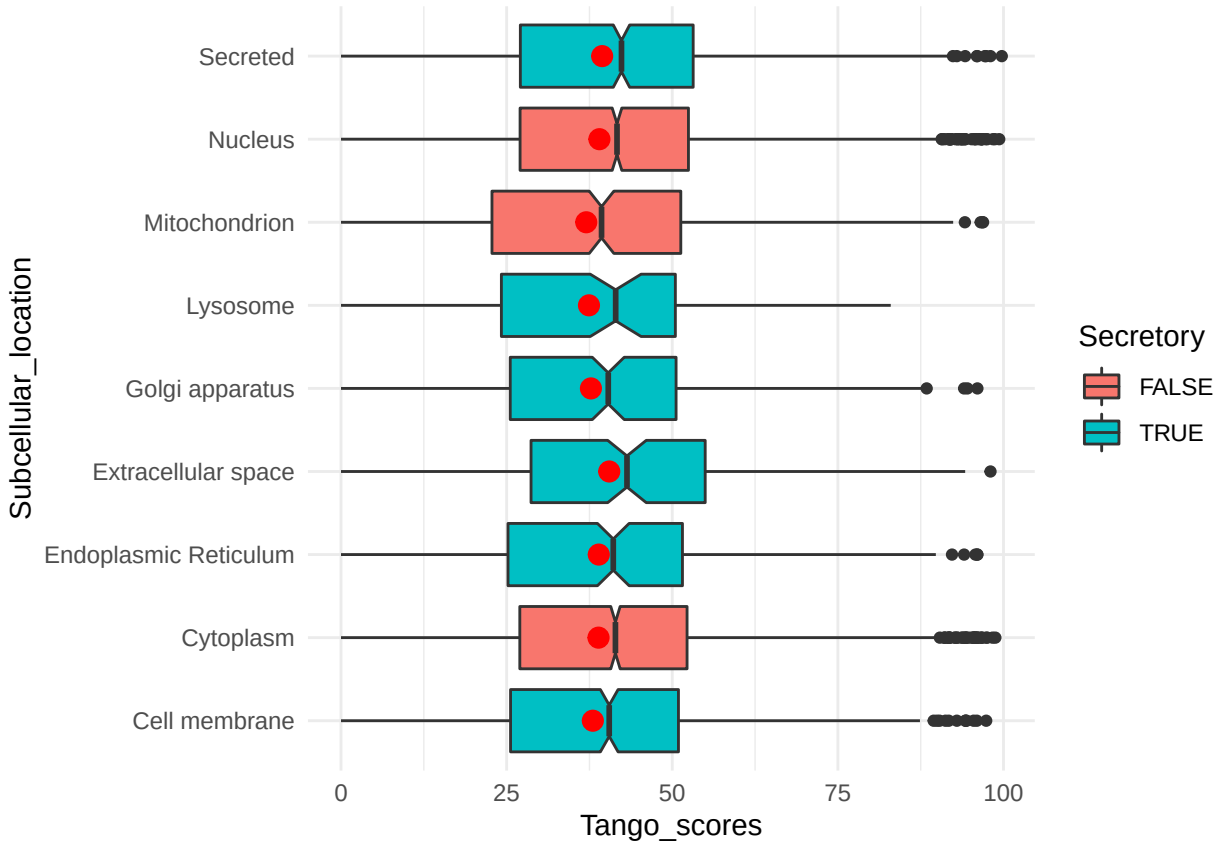


Distribution of number of APR regions per protein normalized for protein length

```
plot_average_tango_scores_complete_proteins()
```

```
## Warning: Removed 3 rows containing non-finite values (stat_boxplot).
```

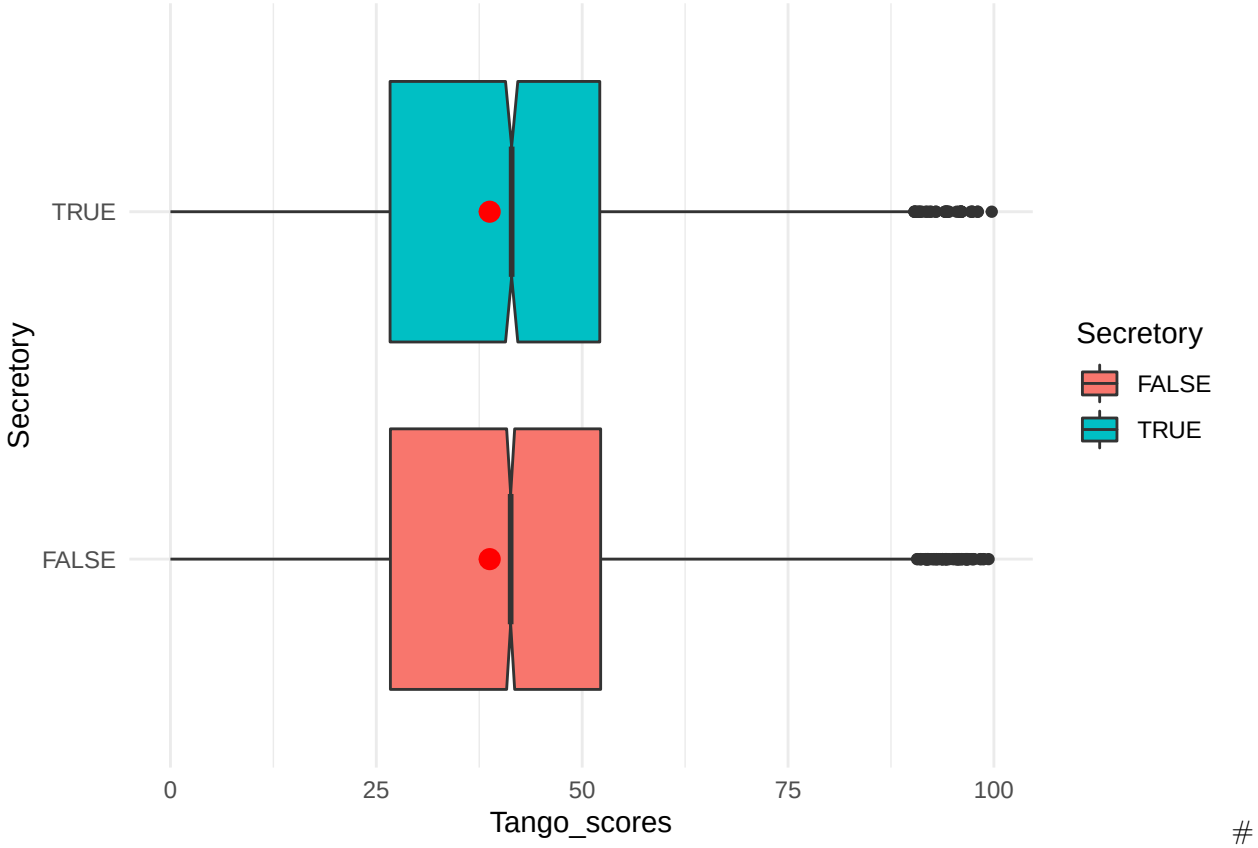
```
## Warning: Removed 3 rows containing non-finite values (stat_summary).
```




```
plot_average_tango_scores_complete_proteins_joined_secreted()
```

```
## Warning: Removed 3 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 3 rows containing non-finite values (stat_summary).
```



#

Peptide charges

```
source("charges.R")
classify_subsets()
```

```
# TODO: make plots showing the differences between APR, GK and FR based on charges
# TODO: see if differences when split into subcellular locations

# protein_sequences_APR_peptides = get_charge(APR_peptides)
# protein_sequences_GK_peptides = get_charge(GK_peptides)
# protein_sequences_FR_peptides = get_charge(FR_peptides)
```

Pie Plot for the residues

Lysine (Lys) –> K Arginine (Arg) –> R Aspartic acid (Asp) –> D Glutamic acid (Glu) –> E

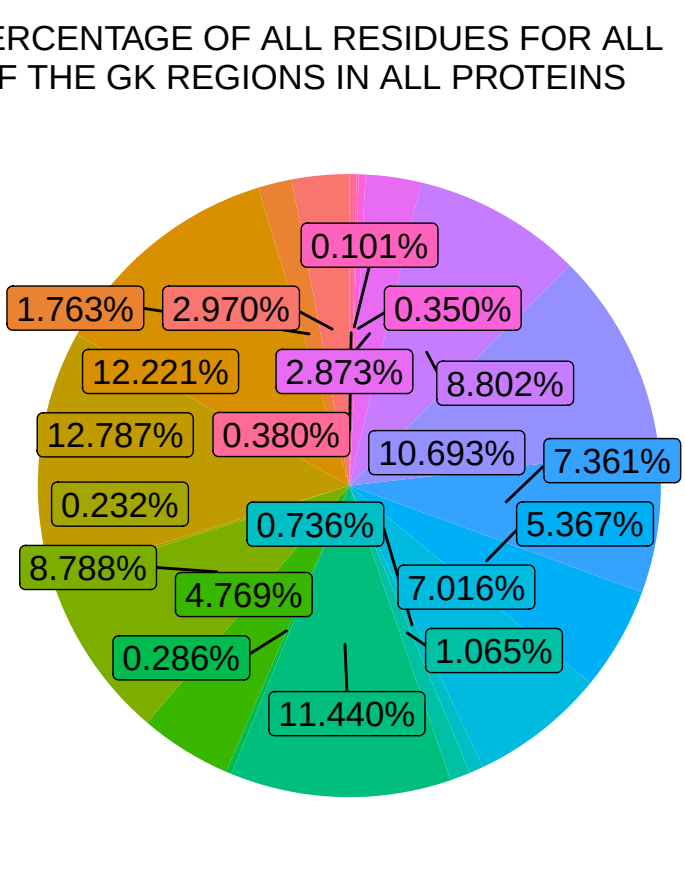
```
source("gatekeeper.R")
```

Only for the GK regions

```
GK_analysis = analyse_gate_keeper_regions()
```

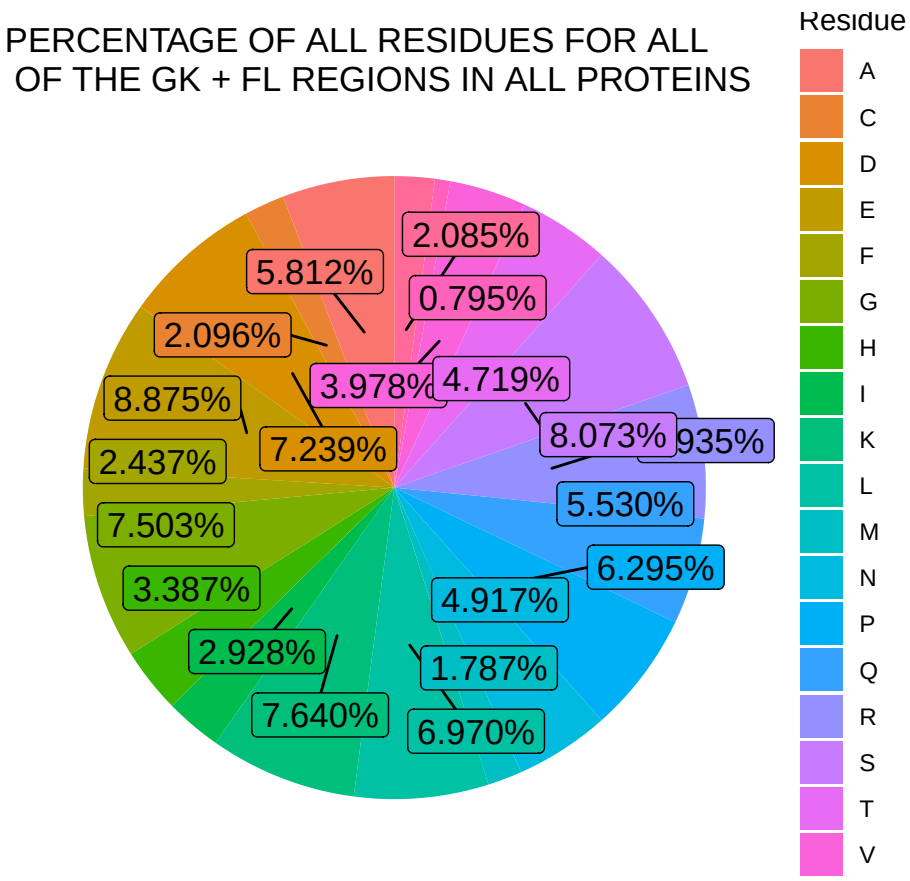
PERCENTAGE OF ALL RESIDUES FOR ALL OF THE GK REGIONS IN ALL PROTEINS

```
pie_plot_percentage_of_all_residues(GK_analysis$cts_gk, "GK REGIONS")
```



PERCENTAGE OF ALL RESIDUES FOR ALL OF THE GK + FL REGIONS IN ALL PROTEINS

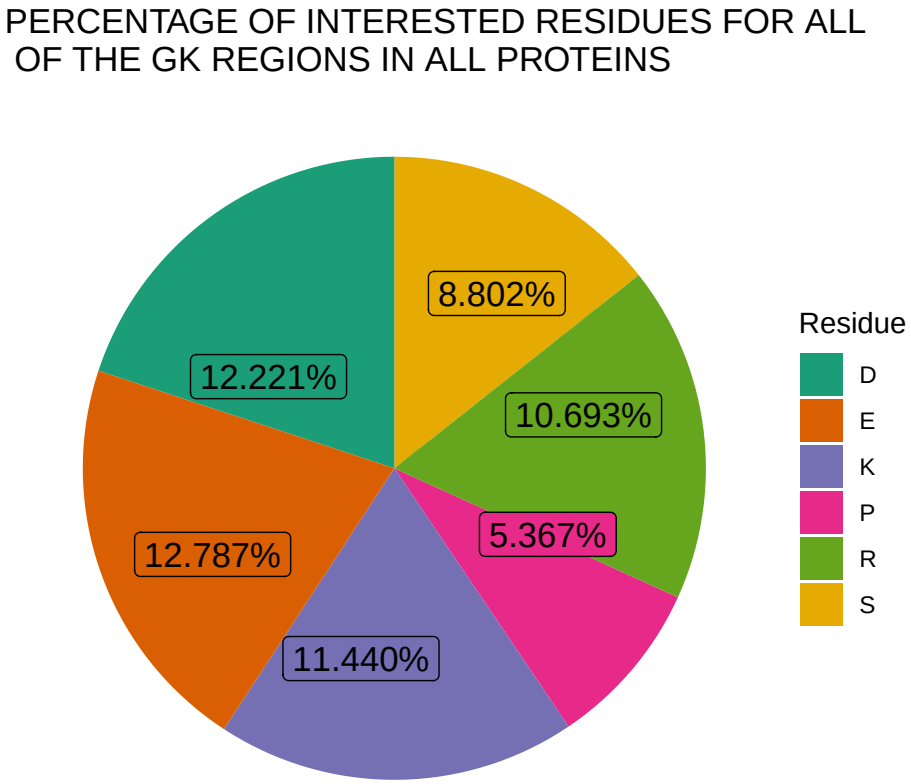
```
pie_plot_percentage_of_all_residues(GK_analysis$cts_gk_fl, "GK + FL REGIONS")
```



Interested residues Lysine (Lys) –> K Arginine (Arg) –> R Aspartic acid (Asp) –> D Glutamic acid (Glu) –> E Serine (Ser) –> S Proline (Pro) –> P PERCENTAGE OF INTERESTED RESIDUES FOR ALL OF THE GK REGIONS IN ALL PROTEINS

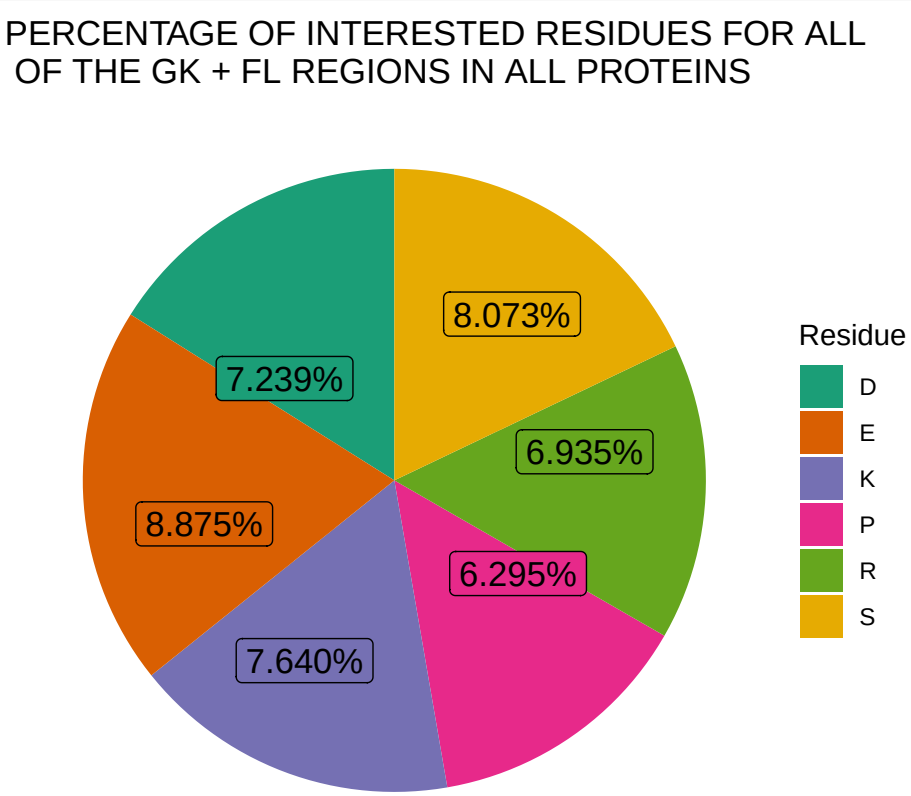
```
GK_analysis_interested_aa = analyse_interested_gate_keeper_regions(GK_analysis)
```

```
pie_plot_percentage_of_interested_residues(GK_analysis_interested_aa$cts_interest_gk, "GK REGIONS")
```



PERCENTAGE OF INTERESTED RESIDUES FOR ALL OF THE GK + FL REGIONS IN ALL PROTEINS

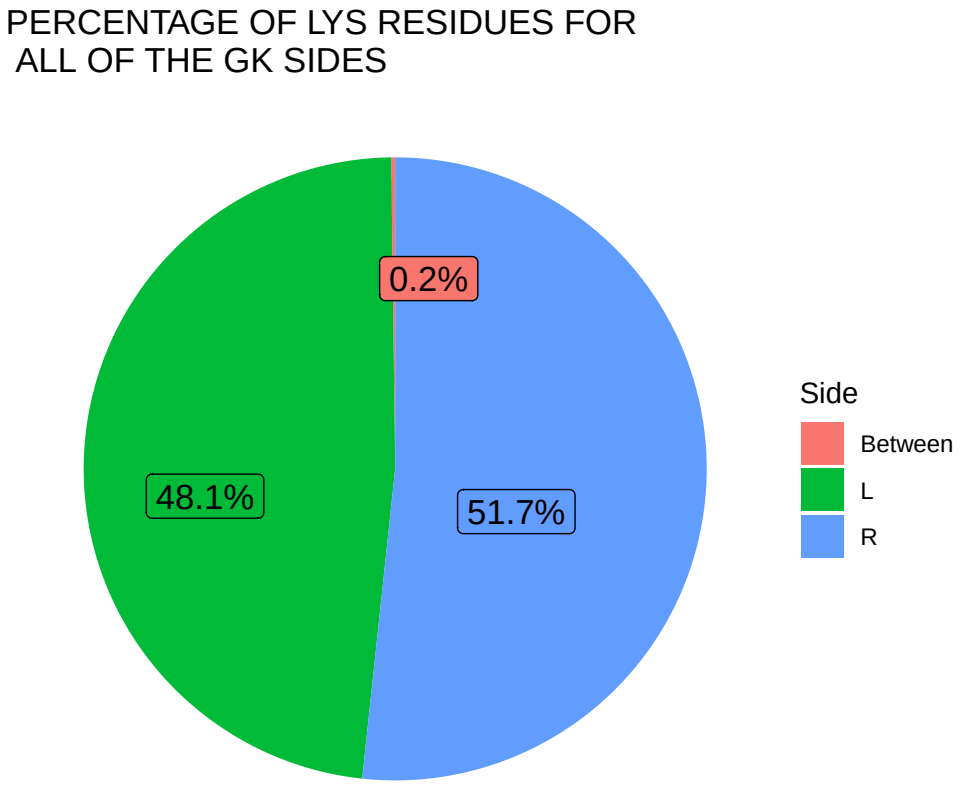
```
pie_plot_percentage_of_interested_residues(GK_analysis_interested_aa$cts_interest_gk_fl, "GK + FL REGIO
```



PERCENTAGE OF INTERESTED RESIDUES FOR ALL OF THE GK SIDES

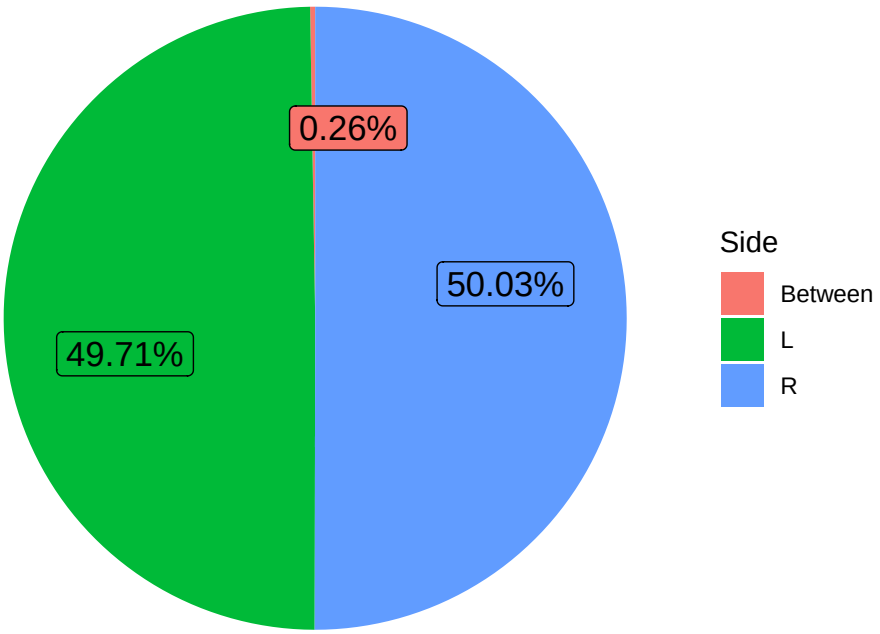
```
GK_analysis_GK_residues = analyse_gate_keeper_residues(GK_analysis$cts_gk_side)
side_string = "GK"
```

```
pie_plot_percentage_of_specific_residue(GK_analysis_GK_residues$lys_ct, "LYS", side_string)
```



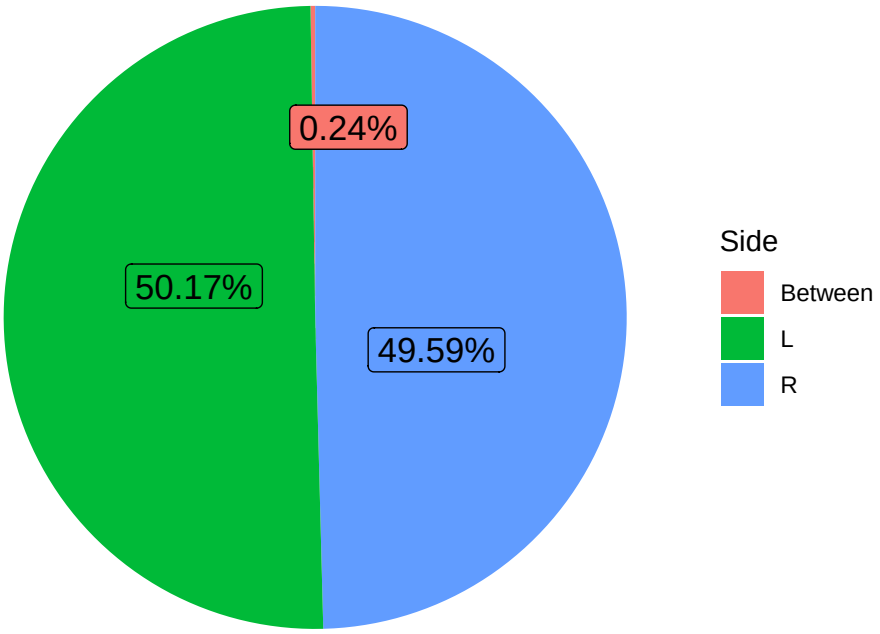

```
pie_plot_percentage_of_specific_residue(GK_analysis_GK_residues$arg_ct, "ARG", side_string)
```

PERCENTAGE OF ARG RESIDUES FOR
ALL OF THE GK SIDES

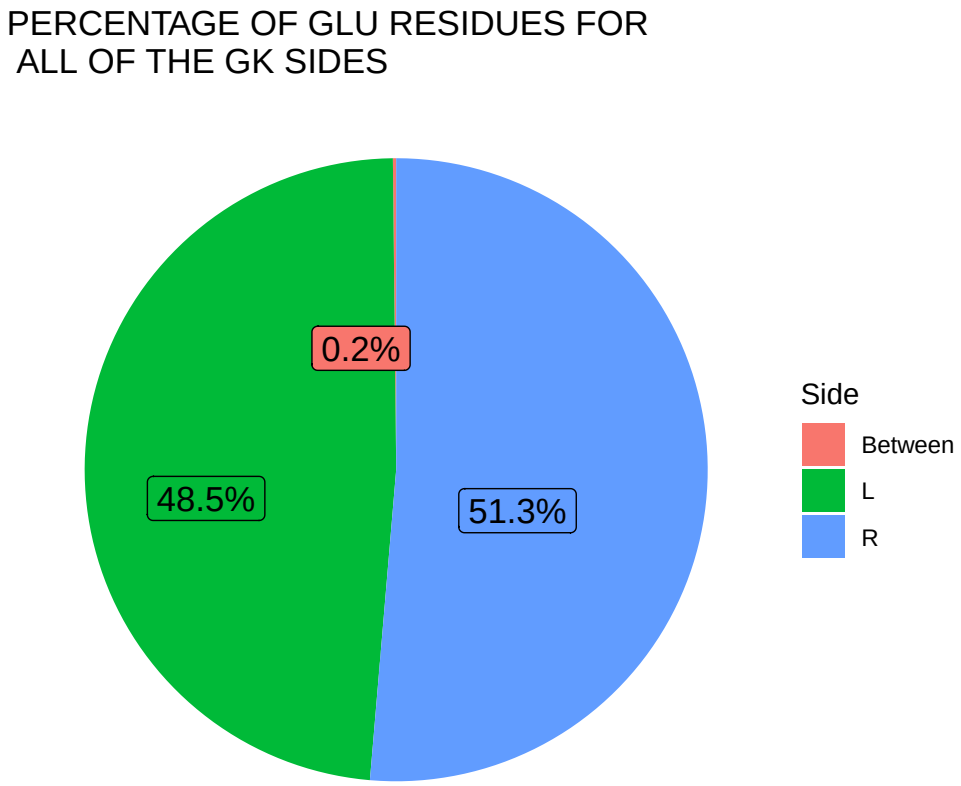



```
pie_plot_percentage_of_specific_residue(GK_analysis_GK_residues$asp_ct, "ASP", side_string)
```

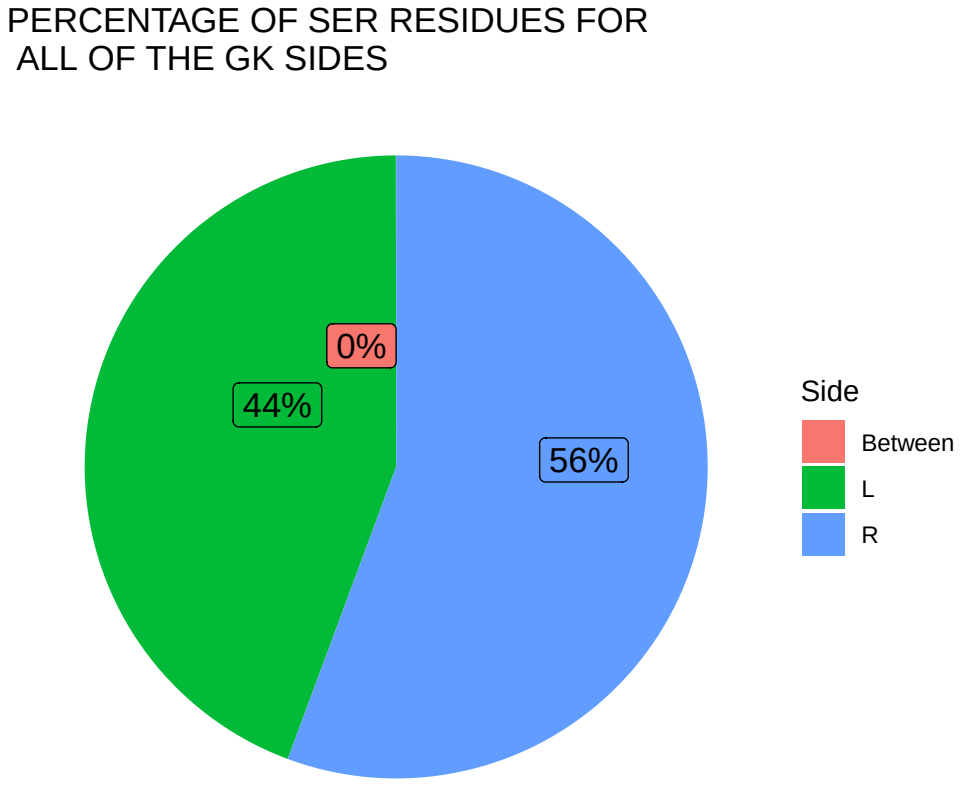
PERCENTAGE OF ASP RESIDUES FOR
ALL OF THE GK SIDES



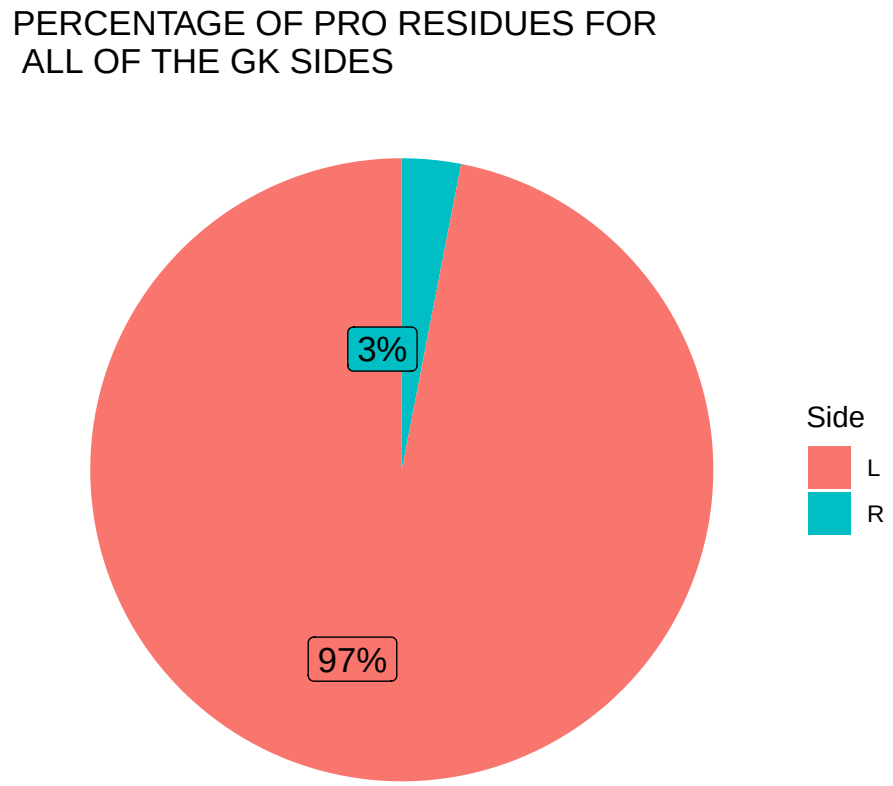
```
pie_plot_percentage_of_specific_residue(GK_analysis_GK_residues$glu_ct, "GLU", side_string)
```




```
pie_plot_percentage_of_specific_residue(GK_analysis_GK_residues$ser_ct, "SER", side_string)
```



```
pie_plot_percentage_of_specific_residue(GK_analysis_GK_residues$pro_ct, "PRO", side_string)
```

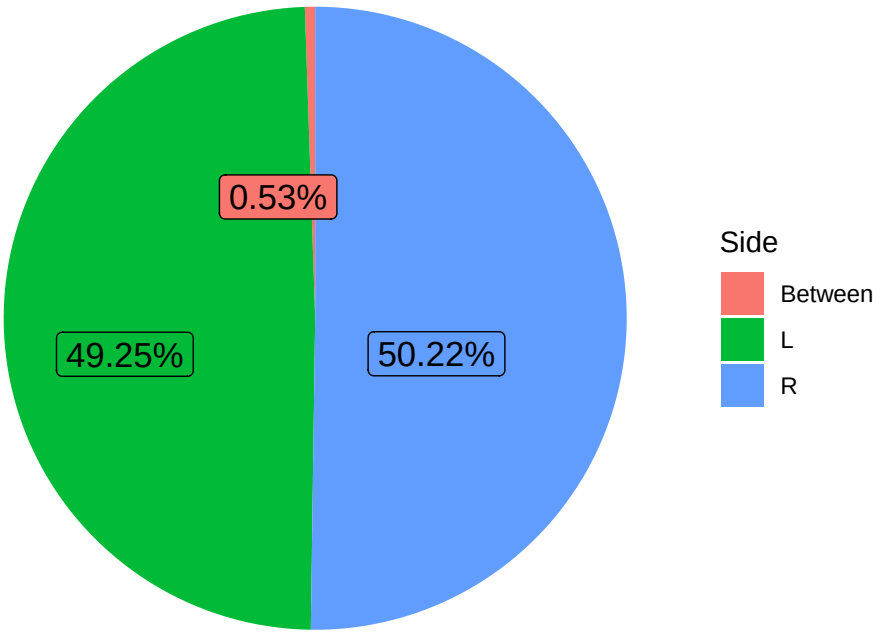


PERCENTAGE OF INTERESTED RESIDUES FOR ALL OF THE GK + FL SIDES

```
GK_analysis_GK_and_FL_residues = analyse_gate_keeper_residues(GK_analysis$cts_gk_fl_side)
side_string = "GK + FL"
```

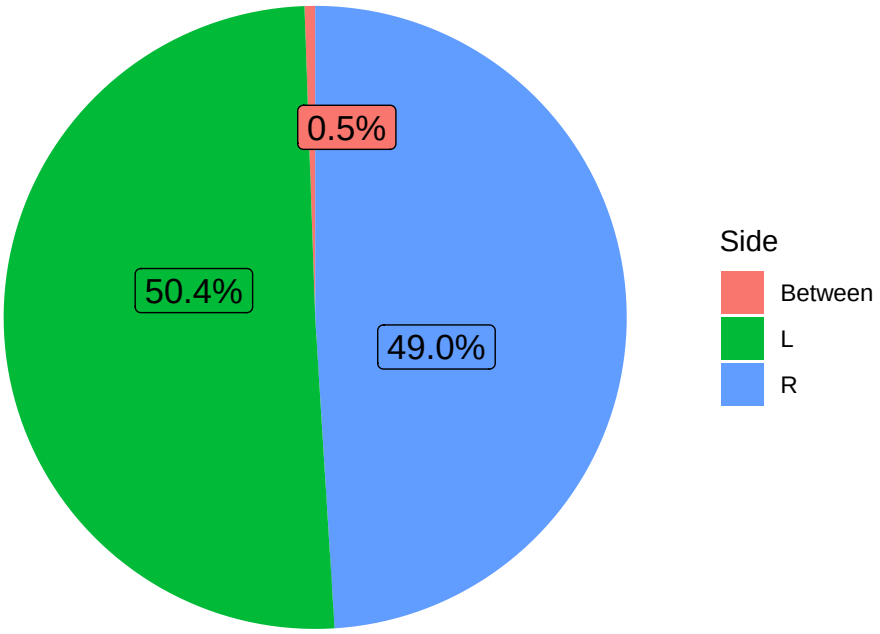
```
pie_plot_percentage_of_specific_residue(GK_analysis_GK_and_FL_residues$lys_ct, "LYS", side_string)
```

PERCENTAGE OF LYS RESIDUES FOR
ALL OF THE GK + FL SIDES

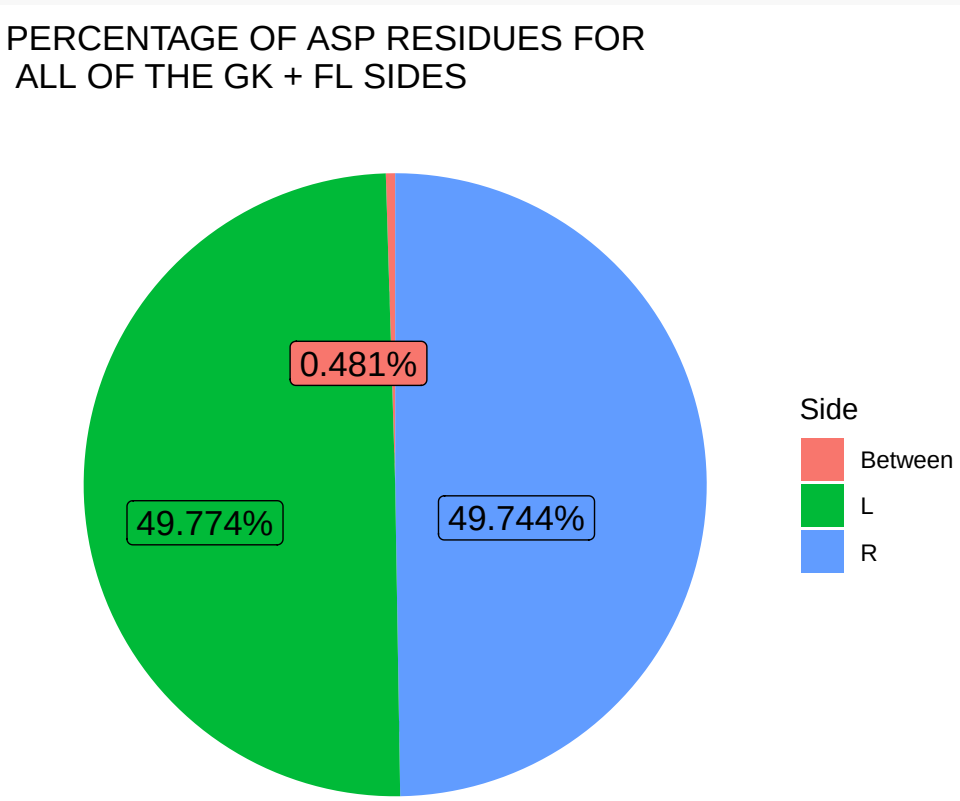



```
pie_plot_percentage_of_specific_residue(GK_analysis_GK_and_FL_residues$arg_ct, "ARG", side_string)
```

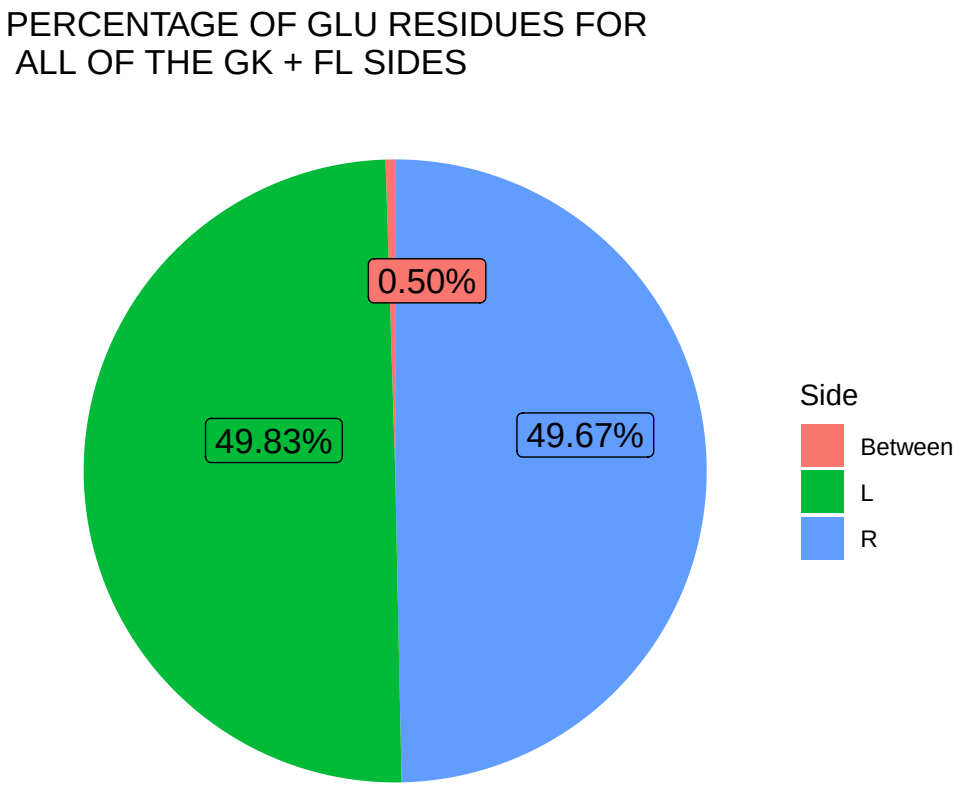
PERCENTAGE OF ARG RESIDUES FOR
ALL OF THE GK + FL SIDES



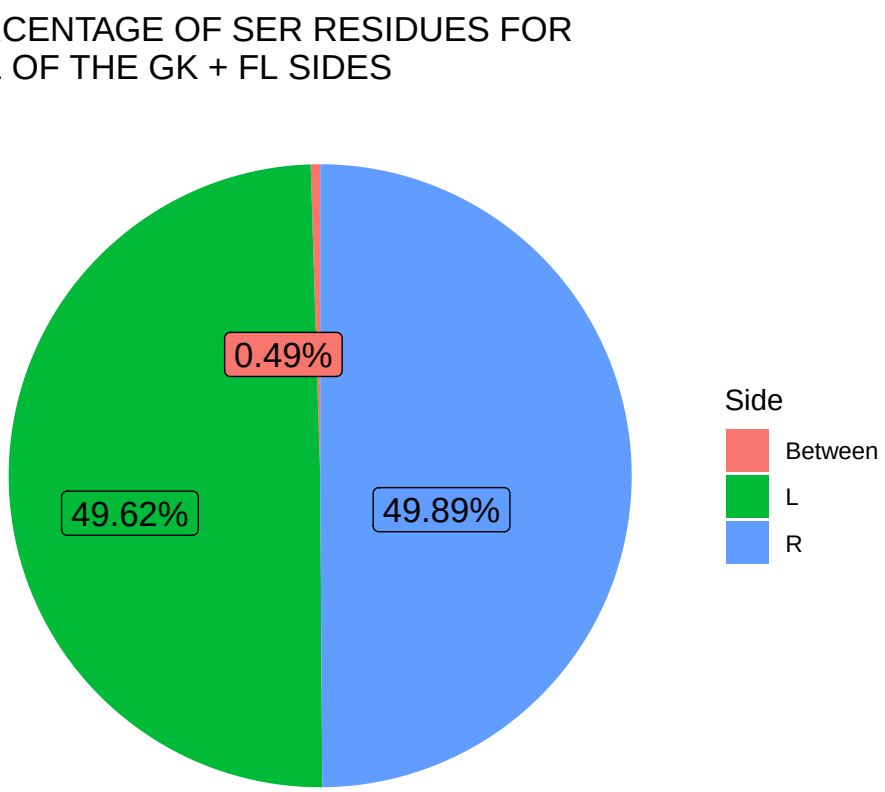
```
pie_plot_percentage_of_specific_residue(GK_analysis_GK_and_FL_residues$asp_ct, "ASP", side_string)
```



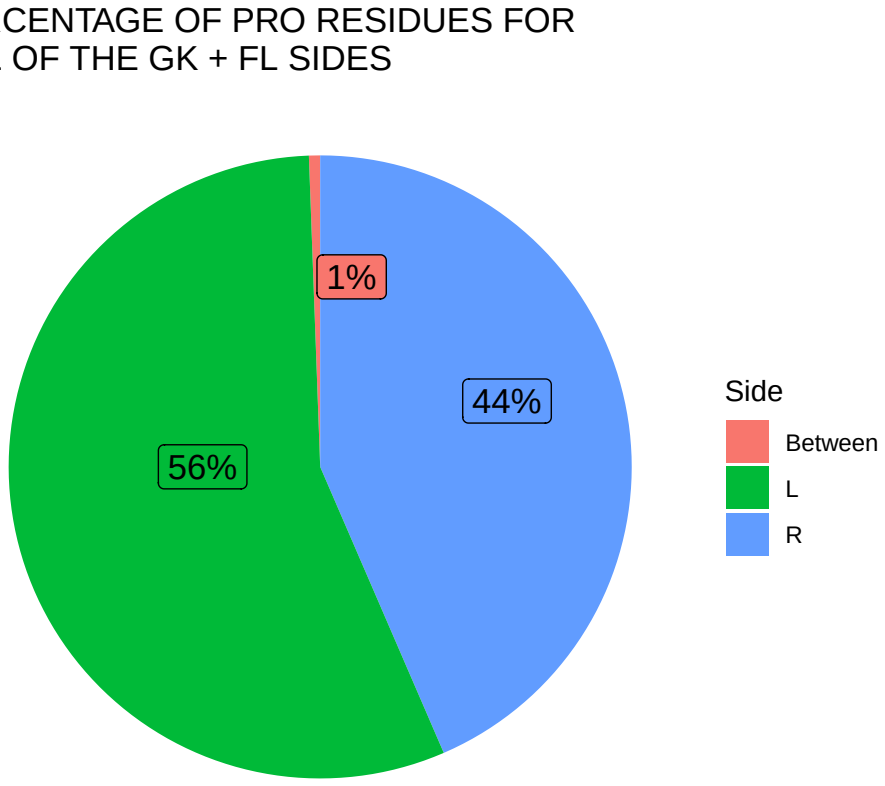

```
pie_plot_percentage_of_specific_residue(GK_analysis_GK_and_FL_residues$glu_ct, "GLU", side_string)
```



```
pie_plot_percentage_of_specific_residue(GK_analysis_GK_and_FL_residues$ser_ct, "SER", side_string)
```



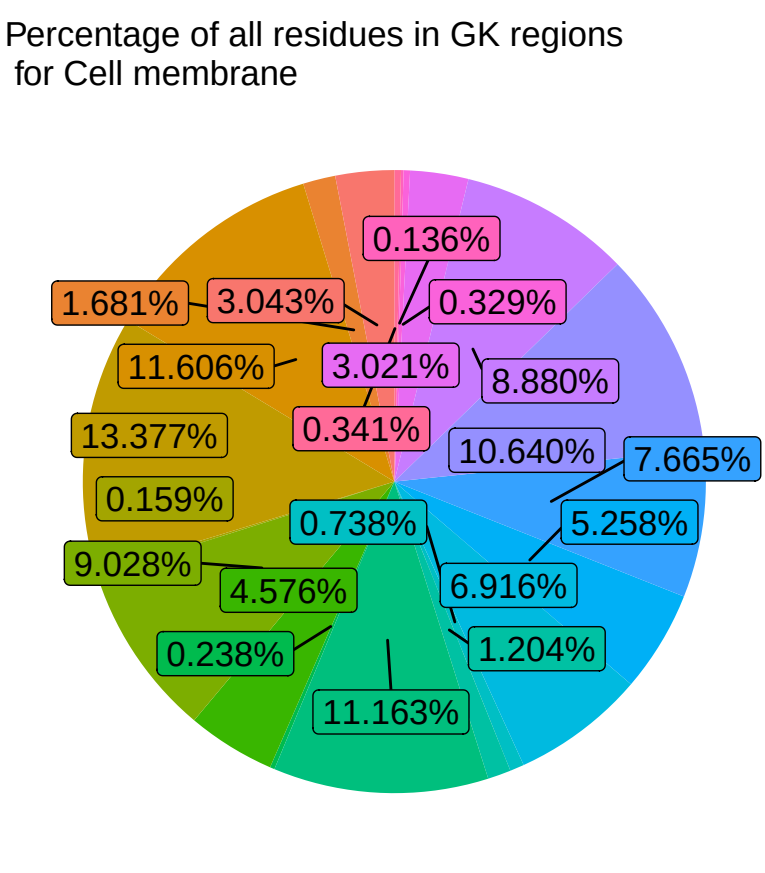

```
pie_plot_percentage_of_specific_residue(GK_analysis_GK_and_FL_residues$pro_ct, "PRO", side_string)
```



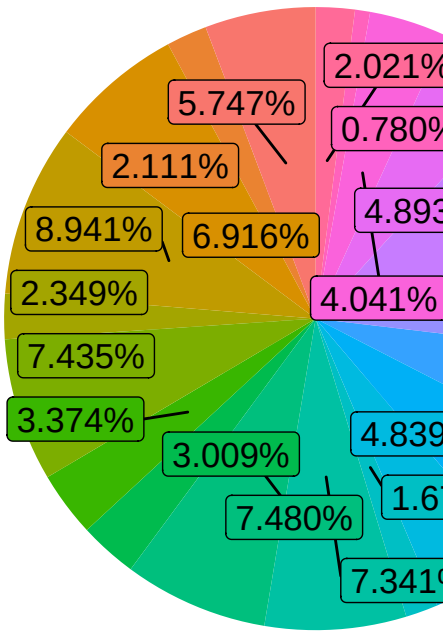
Subcellular location

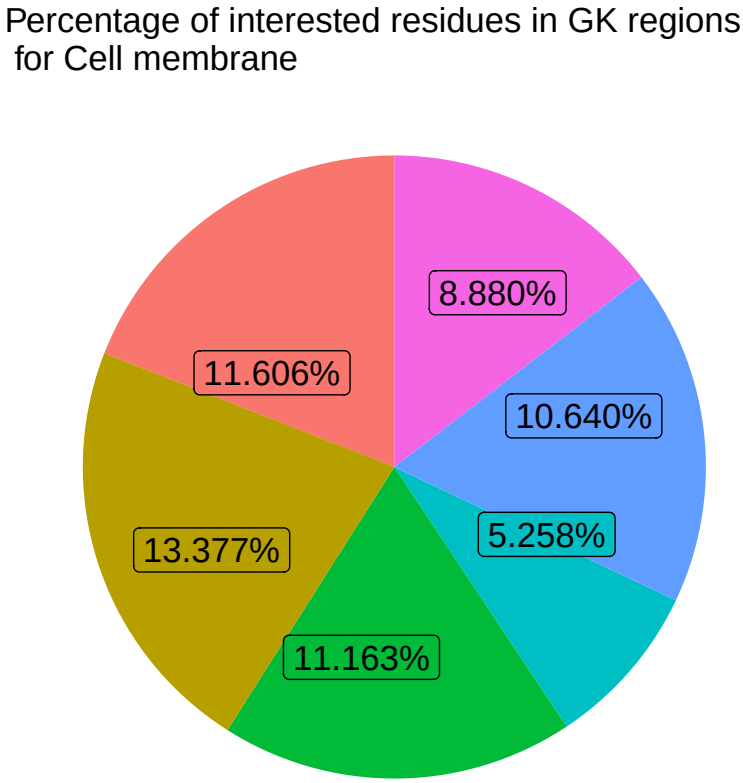
Previous analyses are done for all of the proteins, but now we will be working on all the proteins that belong to a specific subcellular location

```
# Statistics for each subcellular location
for (i in 1:length(search_terms)) {
  counts = get_counts_for_subcellular_location(search_terms[i], GK_analysis)

  residue_category = "all"
  pie_plot_subcellular_location(counts$cts_gk, residue_category, "GK", search_terms[i])
  pie_plot_subcellular_location(counts$cts_gk_fl, residue_category, "GK + FL", search_terms[i])

  # PERCENTAGE OF INTERESTED PROTEINS IN SUBCELLULAR LOCATIONS

  counts_interest <- analyse_interested_gate_keeper_regions(counts)
  residue_category = "interested"
  pie_plot_subcellular_location(counts_interest$cts_interest_gk, residue_category, "GK", search_terms[i]
  pie_plot_subcellular_location(counts_interest$cts_interest_gk_fl, residue_category, "GK + FL", search_

  # NOW THE SIDES PART
  counts_side_gk <- analyse_sides(counts$cts_gk_side)
  counts_side_gk_fl <- analyse_sides(counts$cts_gk_fl_side)

  region_string = "GK"
  pie_plot_sides(counts_side_gk$lys_ct, "LYS", region_string, search_terms[i])
  pie_plot_sides(counts_side_gk$arg_ct, "ARG", region_string, search_terms[i])
  pie_plot_sides(counts_side_gk$asp_ct, "ASP", region_string, search_terms[i])
  pie_plot_sides(counts_side_gk$glu_ct, "GLU", region_string, search_terms[i])
  pie_plot_sides(counts_side_gk$ser_ct, "SER", region_string, search_terms[i])
  pie_plot_sides(counts_side_gk$pro_ct, "PRO", region_string, search_terms[i])

  region_string = "GK + FL"
  pie_plot_sides(counts_side_gk_fl$lys_ct, "LYS", region_string, search_terms[i])
  pie_plot_sides(counts_side_gk_fl$arg_ct, "ARG", region_string, search_terms[i])
  pie_plot_sides(counts_side_gk_fl$asp_ct, "ASP", region_string, search_terms[i])
  pie_plot_sides(counts_side_gk_fl$glu_ct, "GLU", region_string, search_terms[i])
  pie_plot_sides(counts_side_gk_fl$ser_ct, "SER", region_string, search_terms[i])
  pie_plot_sides(counts_side_gk_fl$pro_ct, "PRO", region_string, search_terms[i])
}
```

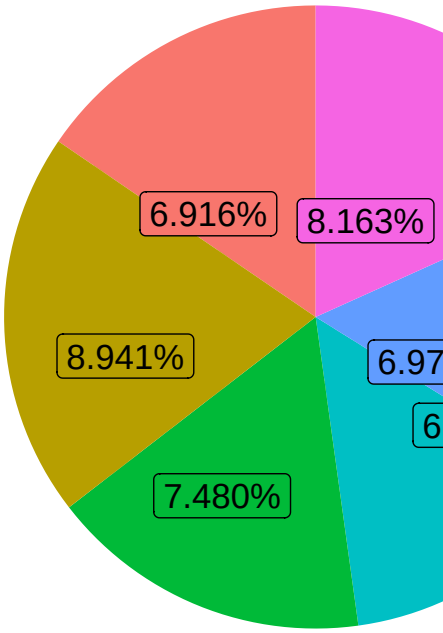

Percentage of all residues in GK regions
for Cell membrane
0.136%
1.681%
3.043%
0.329%
11.606%
3.021%
8.880%
13.377%
0.341%
10.640%
7.665%
0.159%
0.738%
5.258%
9.028%
4.576%
6.916%
1.204%
0.238%
11.163%

Residue
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V

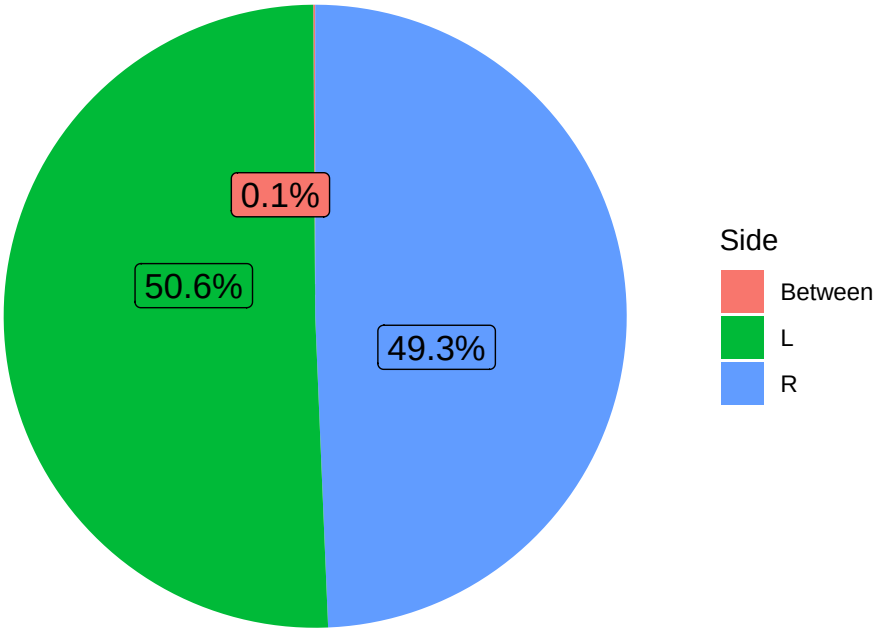

Percentage of all residues in GK +
for Cell membrane
2.021%
5.747%
0.780%
2.111%
4.893
8.941%
6.916%
2.349%
4.041%
7.435%
3.374%
4.839
3.009%
1.6
7.480%
7.341


Percentage of interested residues in GK regions
for Cell membrane
8.880%
11.606%
10.640%
5.258%
13.377%
11.163%

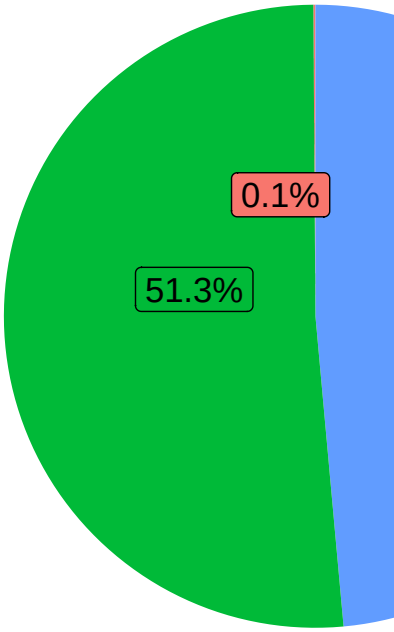
Residue
D
E
K
P
R
S


Percentage of interested residues
for Cell membrane
6.916%
8.163%
8.941%
6.97
6
7.480%

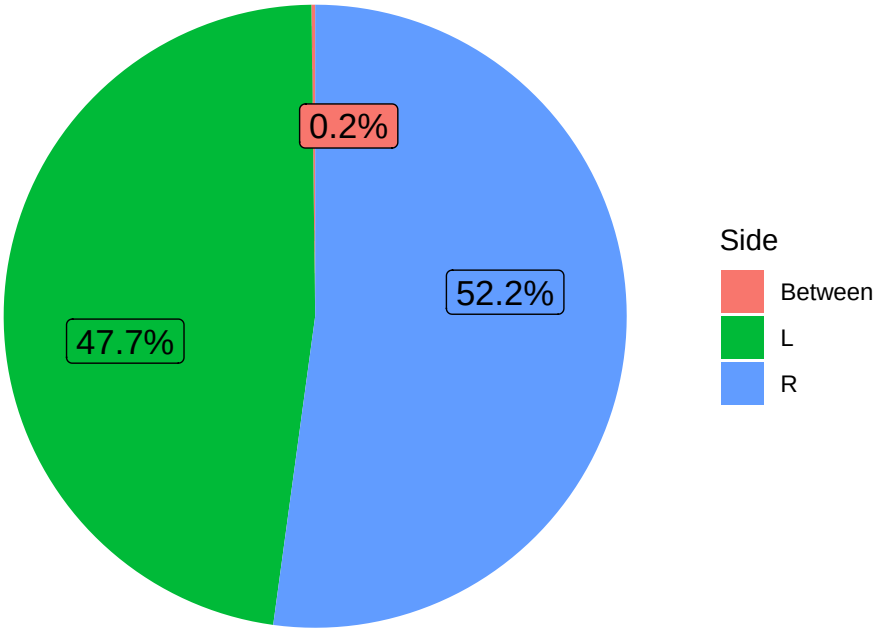
Percentage of LYS residue sides in GK regions for Cell membrane



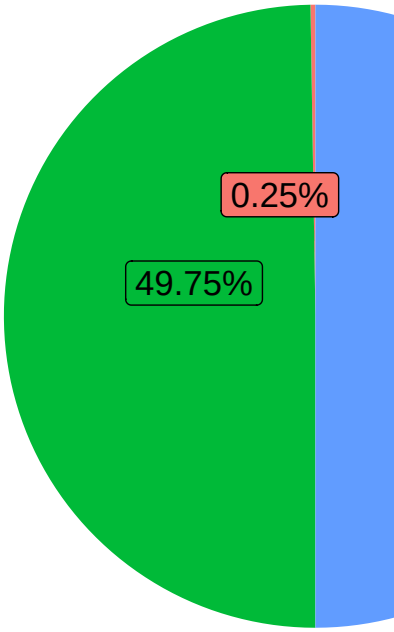
Percentage of ARG residue sid for Cell membrane



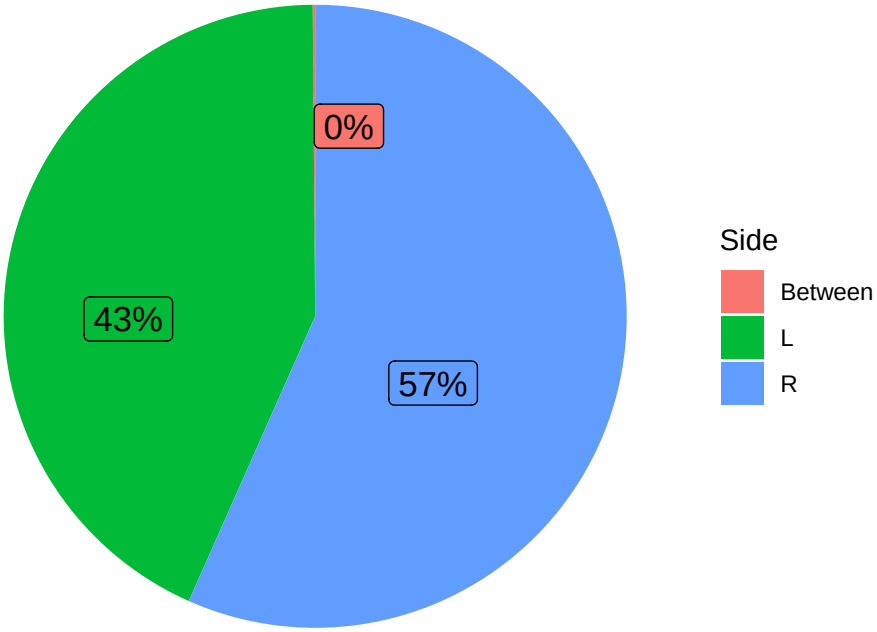
Percentage of ASP residue sides in GK regions for Cell membrane



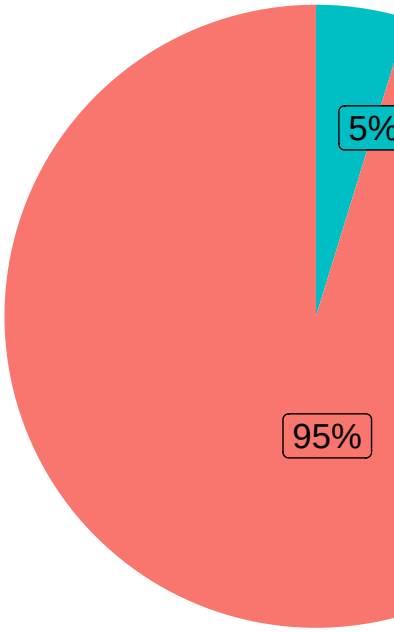
Percentage of GLU residue sid for Cell membrane



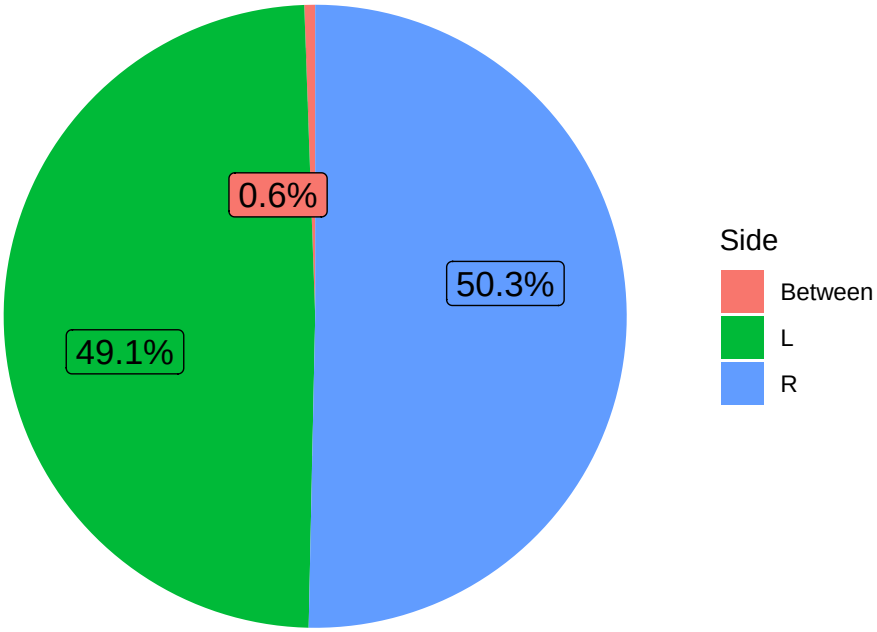
Percentage of SER residue sides in GK regions for Cell membrane



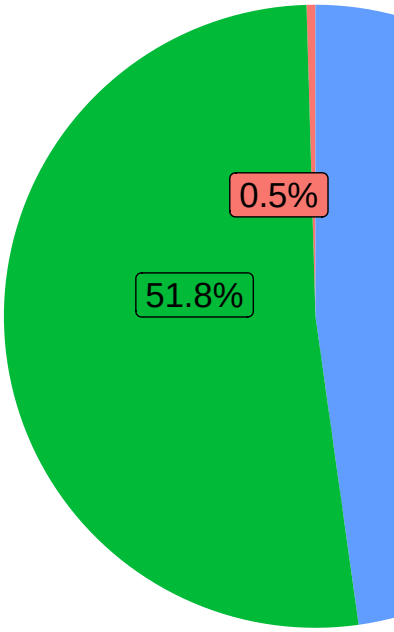
Percentage of PRO residue sid for Cell membrane



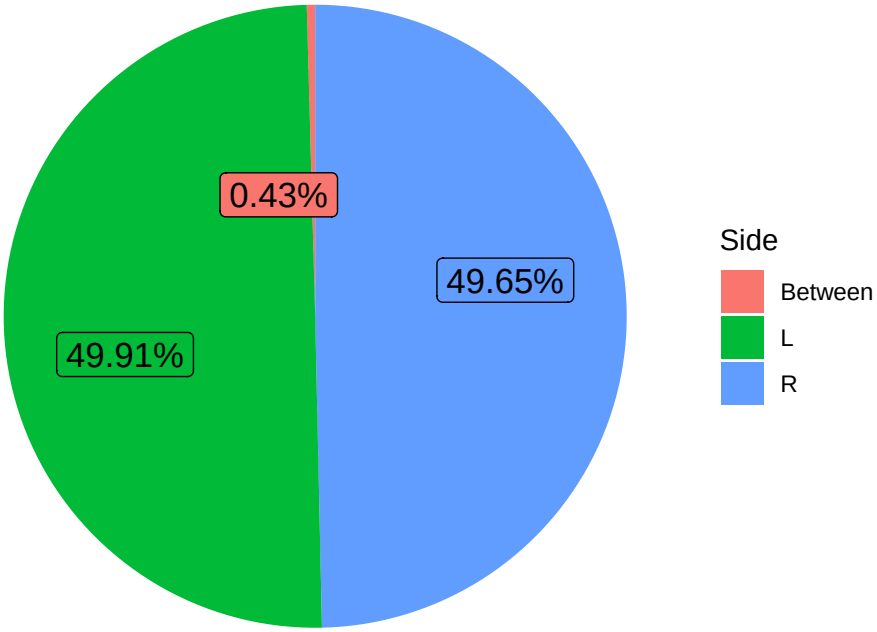
Percentage of LYS residue sides in GK + FL regions for Cell membrane



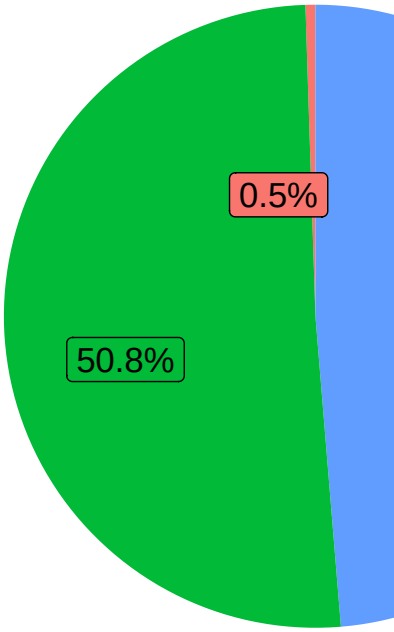
Percentage of ARG residue sid for Cell membrane



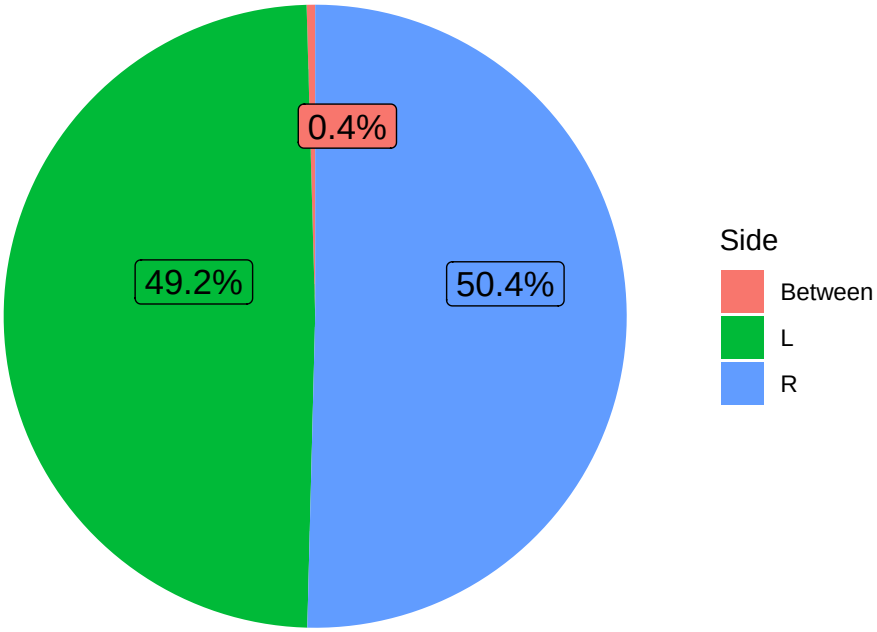
Percentage of ASP residue sides in GK + FL regions for Cell membrane



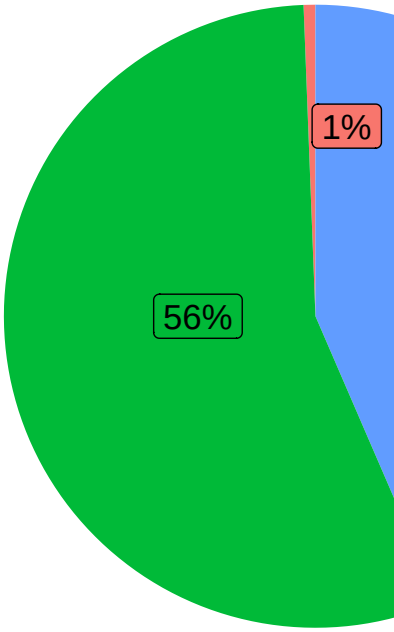
Percentage of GLU residue sid for Cell membrane

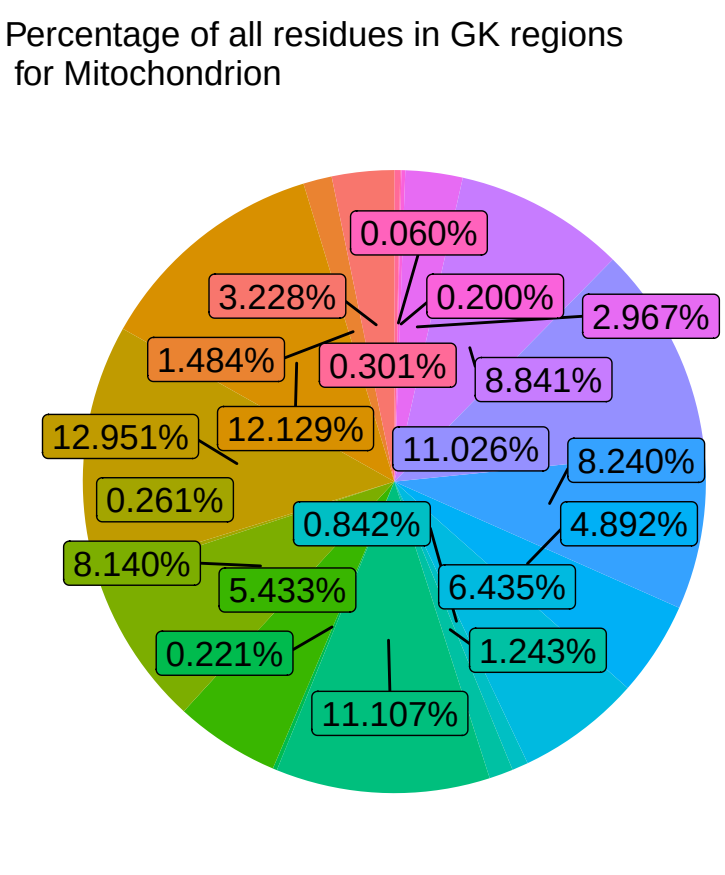


Percentage of SER residue sides in GK + FL regions for Cell membrane

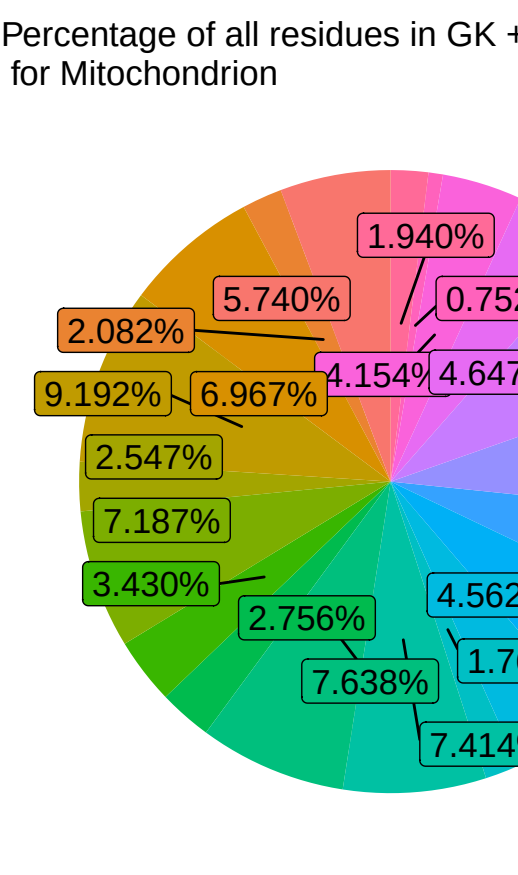


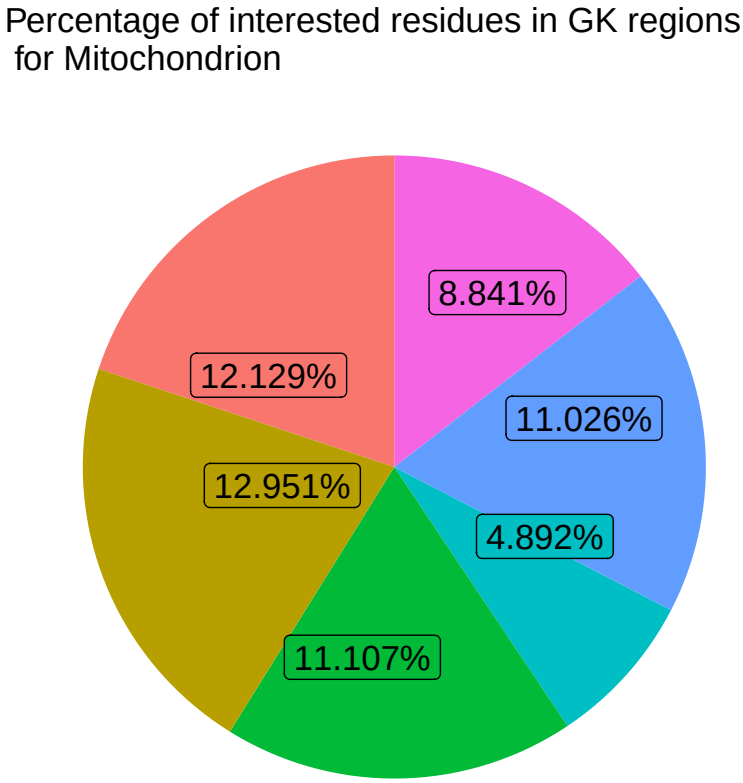
Percentage of PRO residue sid for Cell membrane



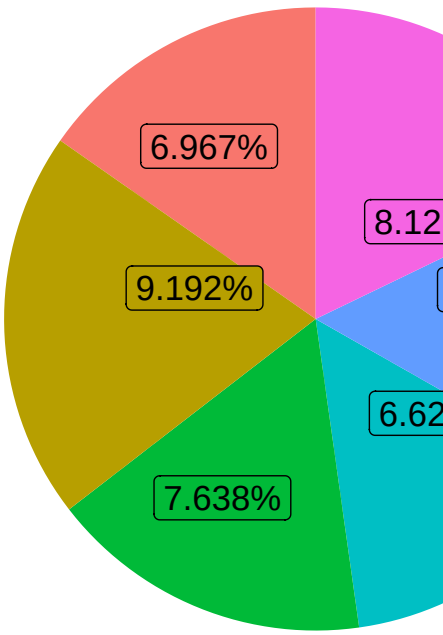

Percentage of all residues in GK regions
for Mitochondrion
0.060%
3.228%
0.200%
2.967%
1.484%
0.301%
8.841%
12.951%
12.129%
11.026%
8.240%
0.261%
0.842%
4.892%
8.140%
5.433%
6.435%
0.221%
1.243%
11.107%

Residue
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V

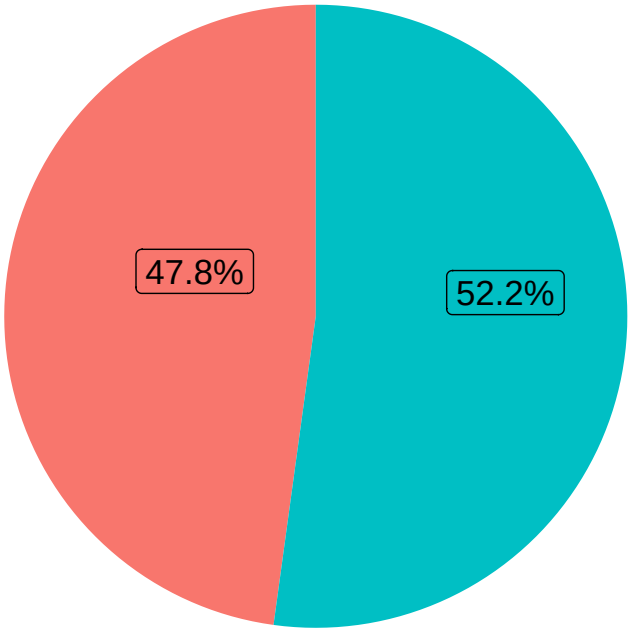

Percentage of all residues in GK +
for Mitochondrion
1.940%
5.740%
2.082%
9.192%
6.967%
2.547%
7.187%
3.430%
2.756%
7.638%


Percentage of interested residues in GK regions
for Mitochondrion
8.841%
12.129%
11.026%
12.951%
4.892%
11.107%

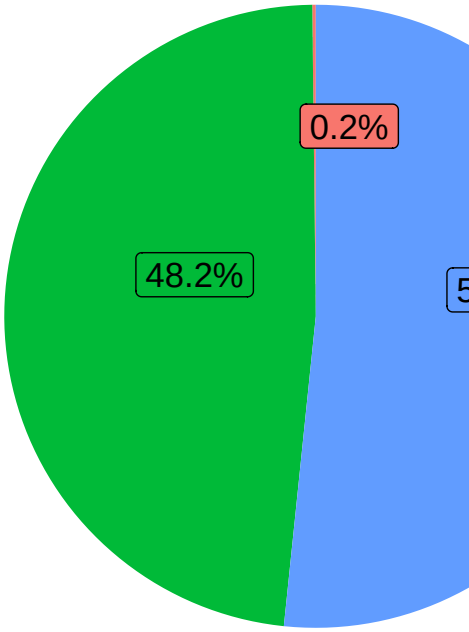
Residue
D
E
K
P
R
S


Percentage of interested residues
for Mitochondrion
6.967%
9.192%
7.638%

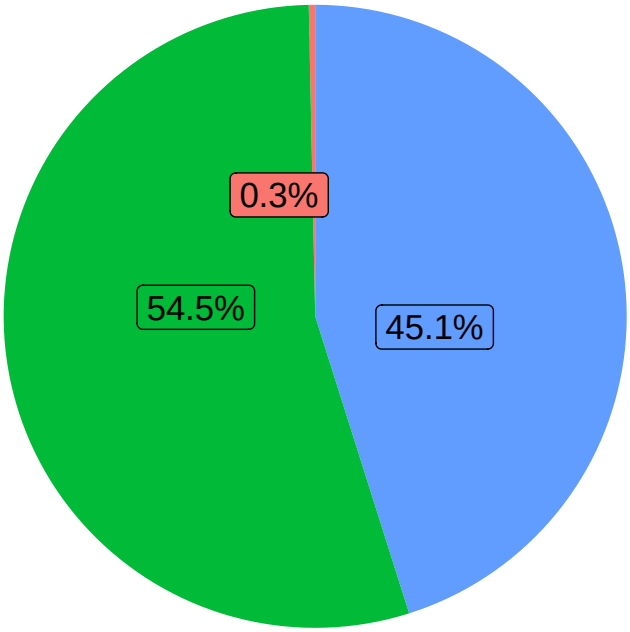
Percentage of LYS residue sides in GK regions for Mitochondrion



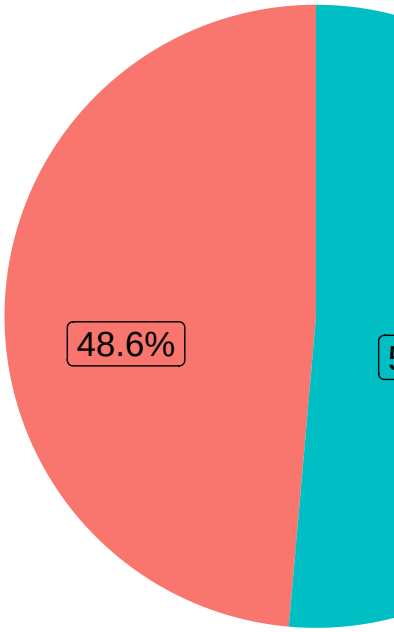
Percentage of ARG residue sides in GK regions for Mitochondrion



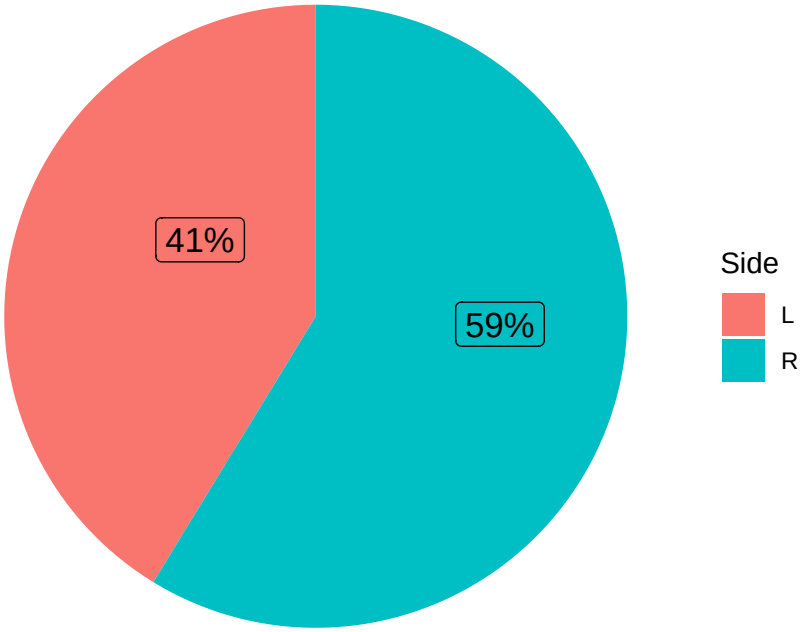
Percentage of ASP residue sides in GK regions for Mitochondrion



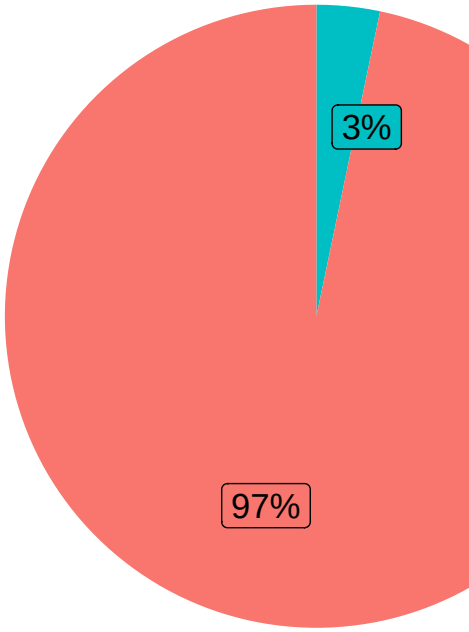
Percentage of GLU residue sides in GK regions for Mitochondrion



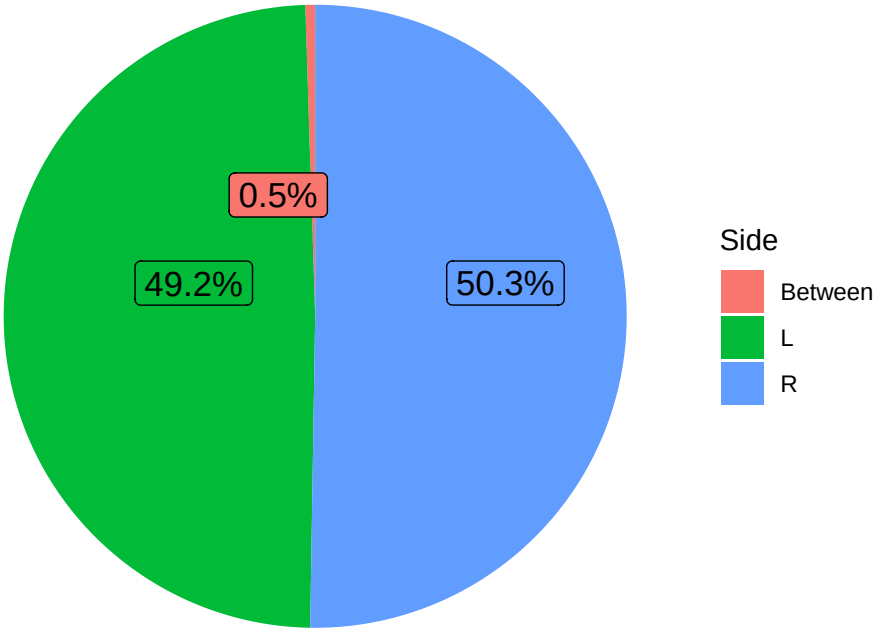
Percentage of SER residue sides in GK regions for Mitochondrion



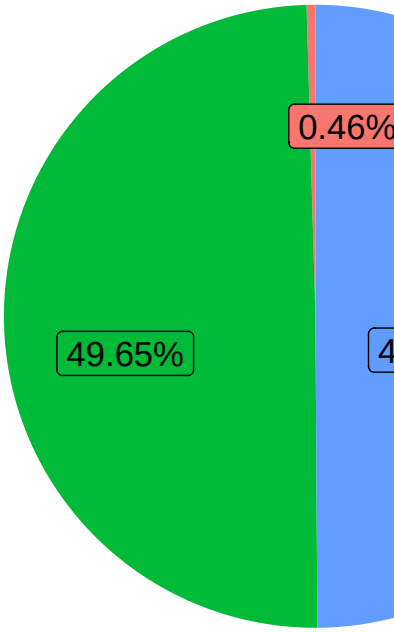
Percentage of PRO residue sides in for Mitochondrion



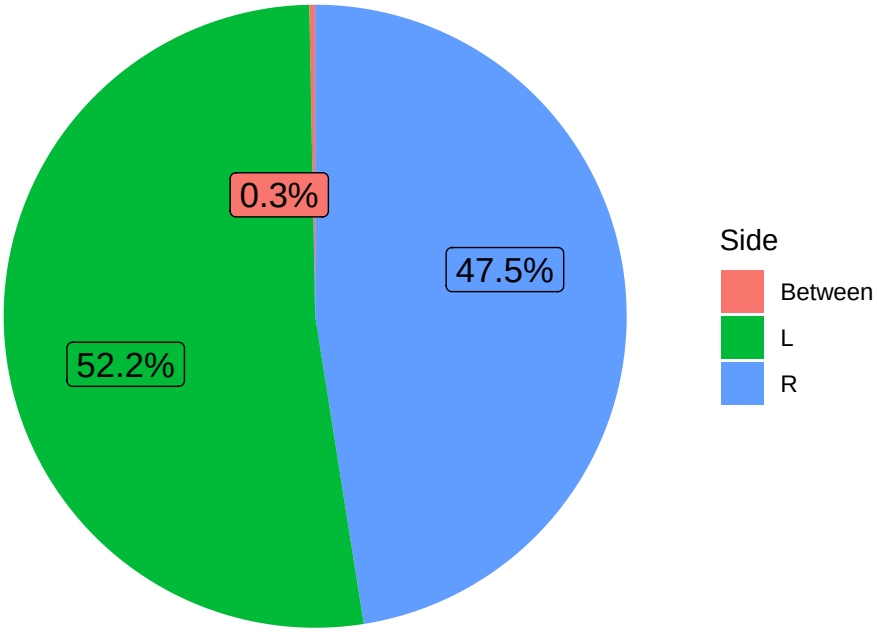
Percentage of LYS residue sides in GK + FL regions for Mitochondrion



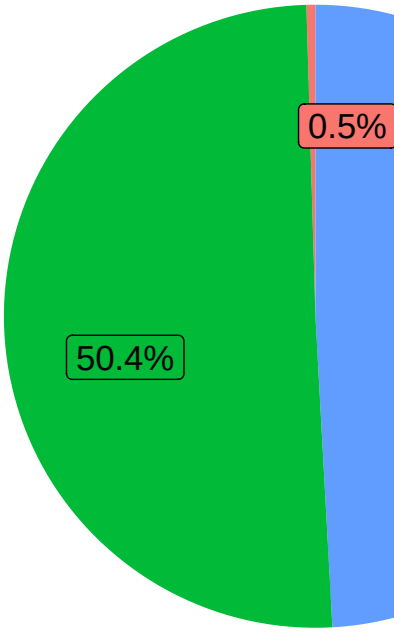
Percentage of ARG residue sid for Mitochondrion



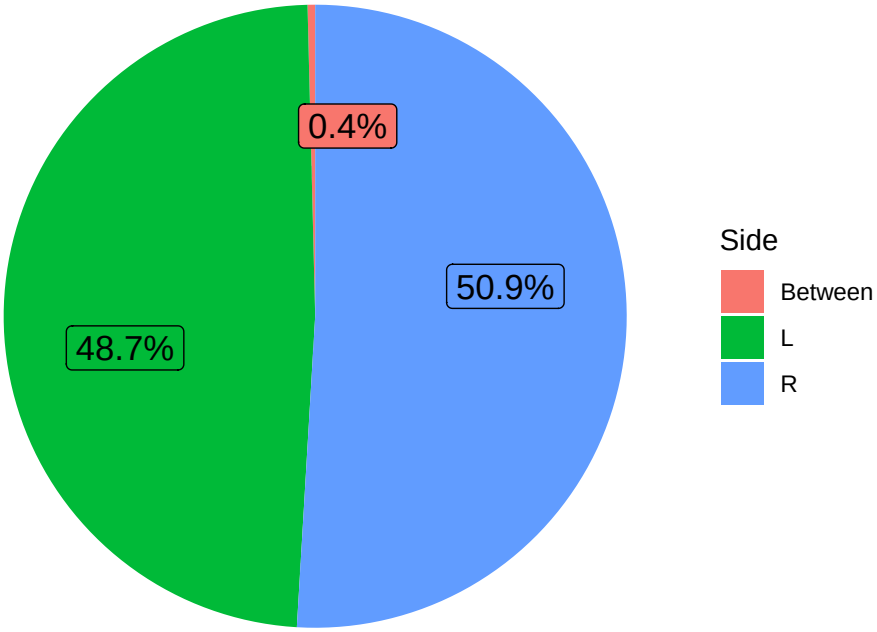
Percentage of ASP residue sides in GK + FL regions for Mitochondrion



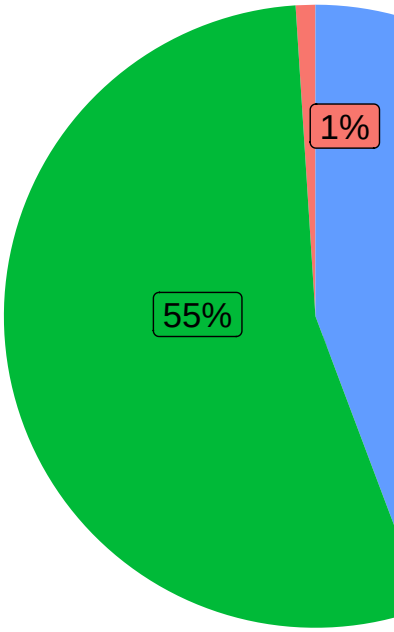
Percentage of GLU residue sid for Mitochondrion



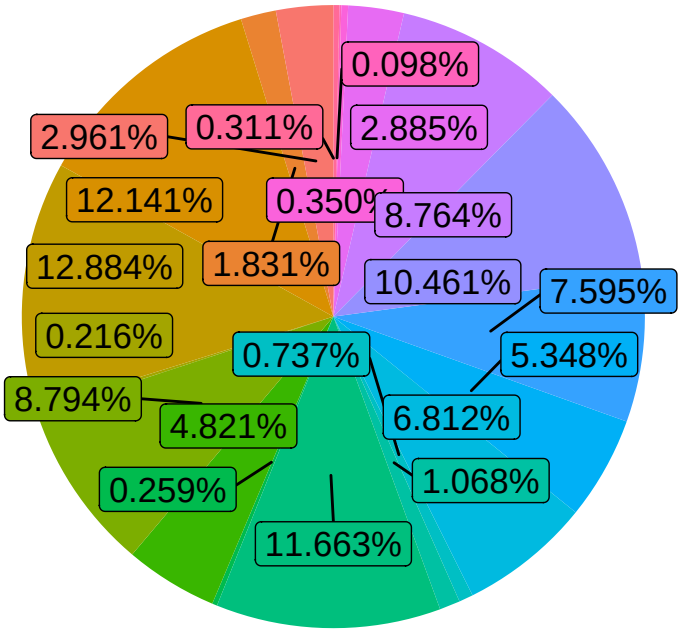
Percentage of SER residue sides in GK + FL regions for Mitochondrion



Percentage of PRO residue sid for Mitochondrion

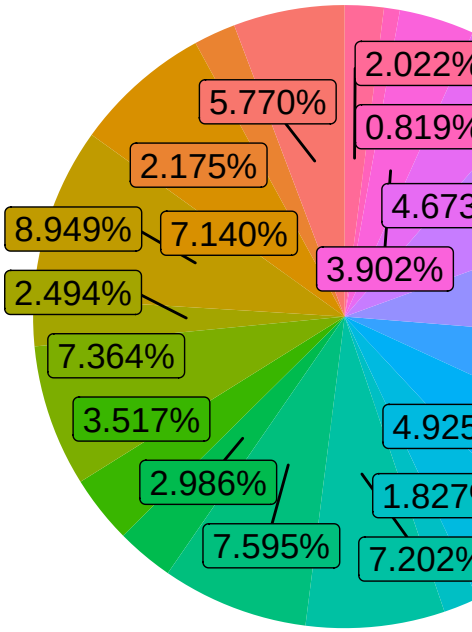


Percentage of all residues in GK regions for Nucleus

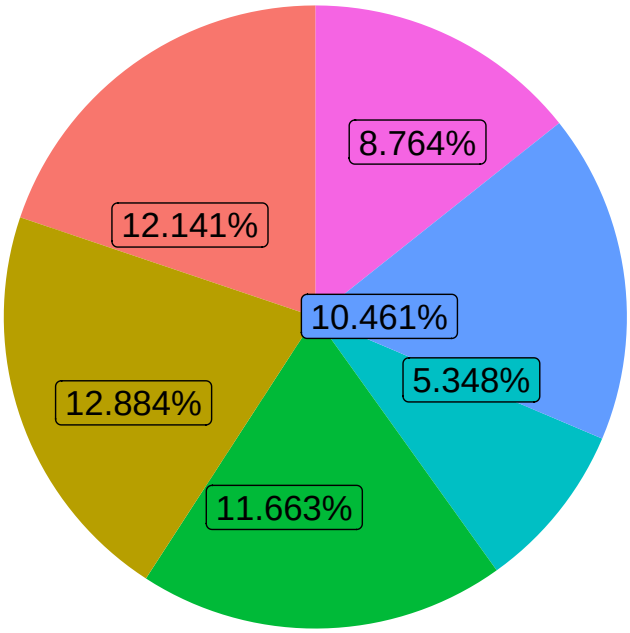


Residue
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V

Percentage of all residues in GK +
for Nucleus

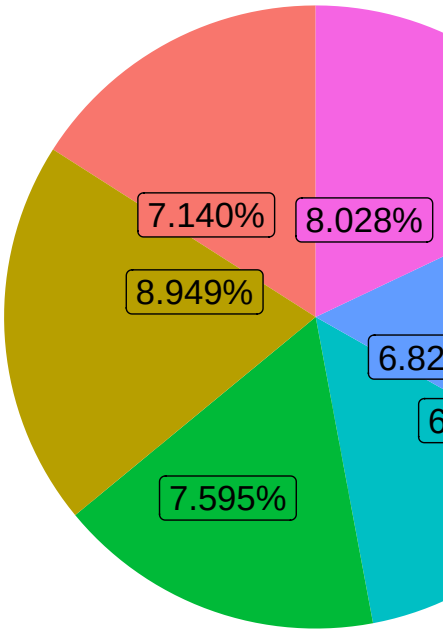


Percentage of interested residues in GK regions for Nucleus

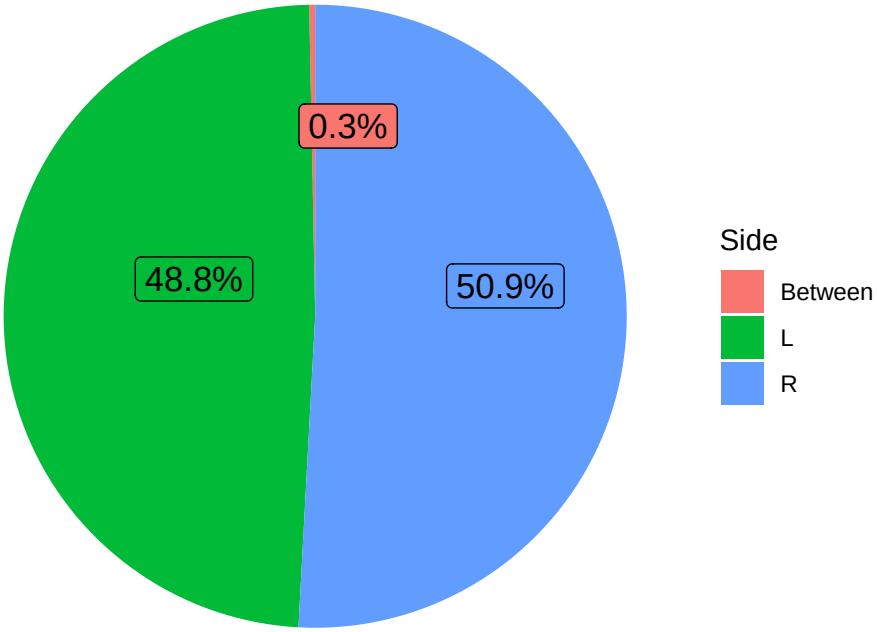


Residue
D
E
K
P
R
S

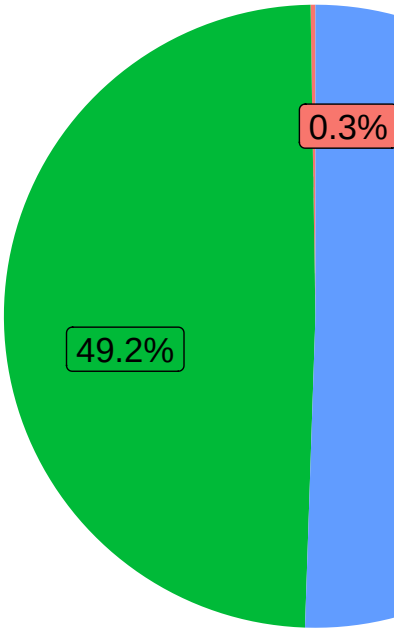
Percentage of interested residues
for Nucleus



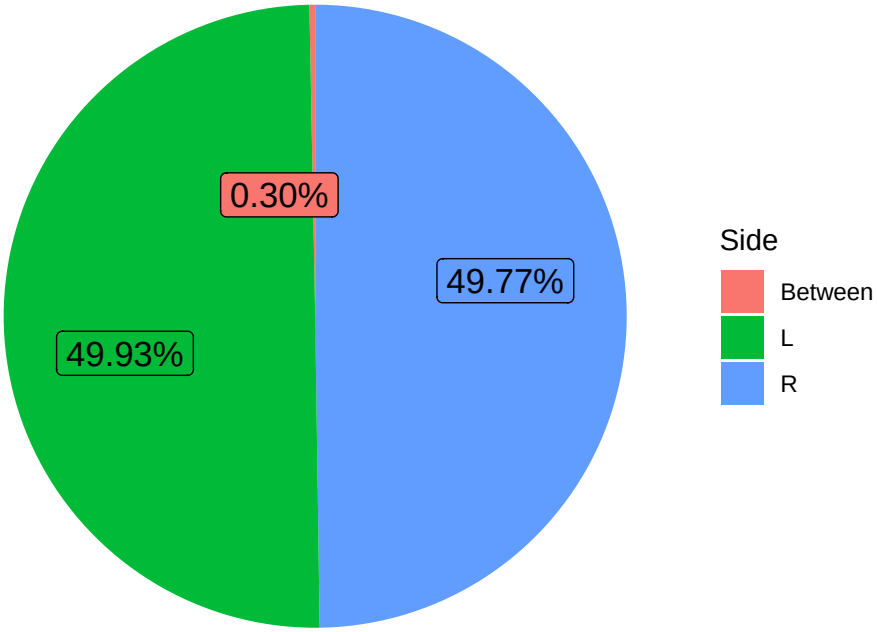
Percentage of LYS residue sides in GK regions for Nucleus



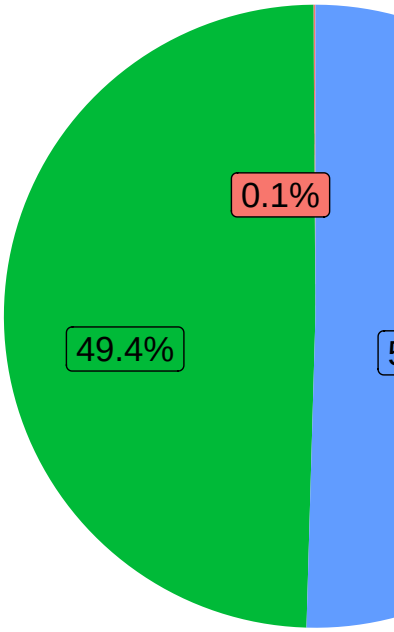
Percentage of ARG residue sid for Nucleus



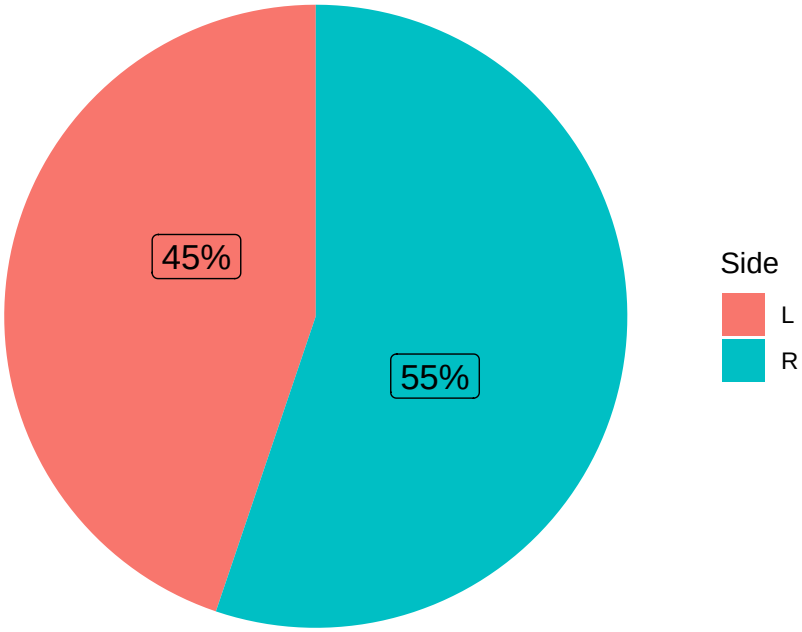
Percentage of ASP residue sides in GK regions for Nucleus



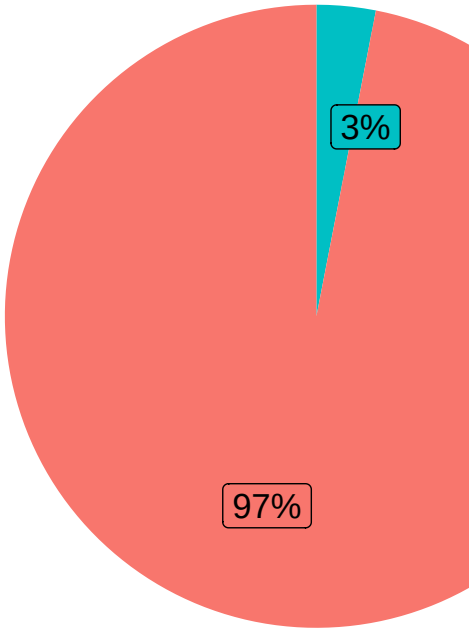
Percentage of GLU residue sid for Nucleus



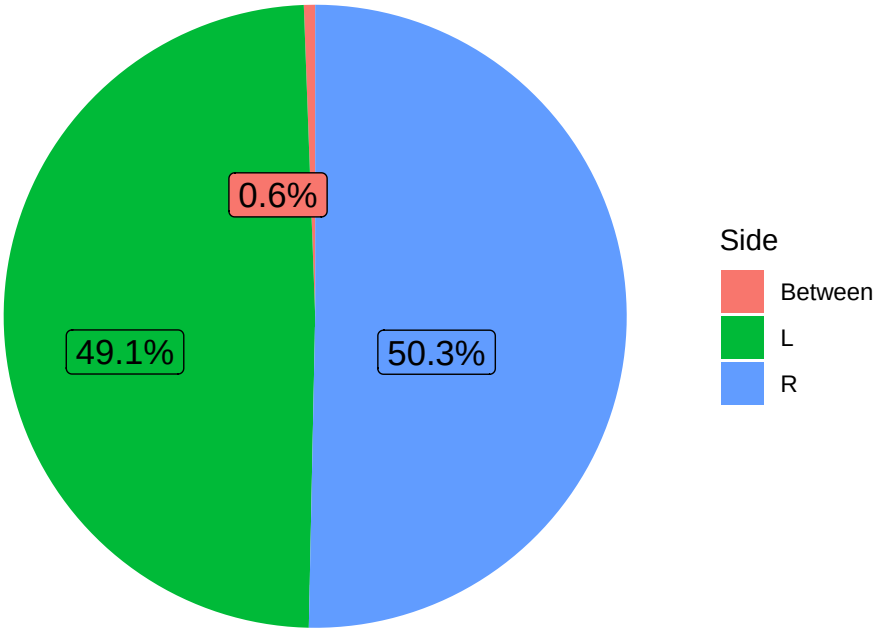
Percentage of SER residue sides in GK regions for Nucleus



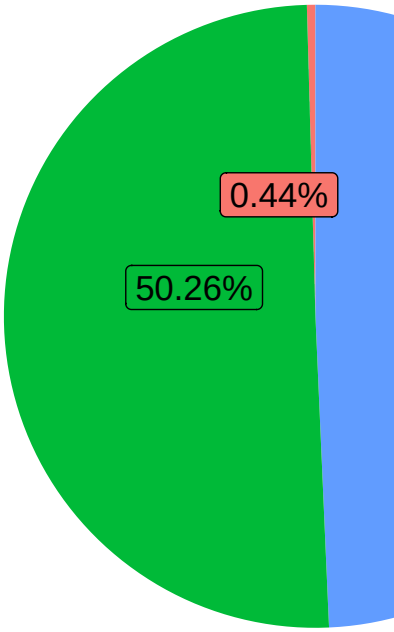
Percentage of PRO residue sides in for Nucleus



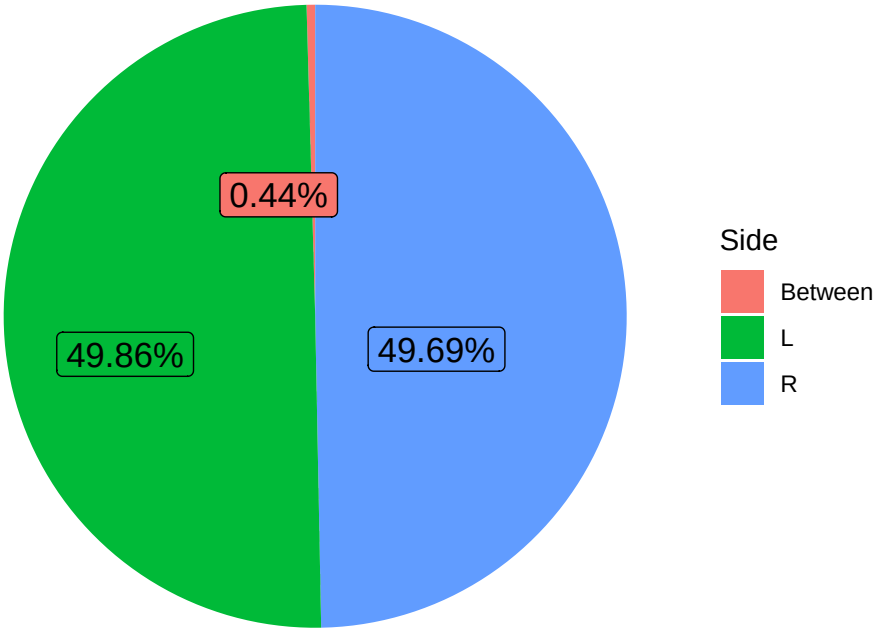
Percentage of LYS residue sides in GK + FL regions for Nucleus



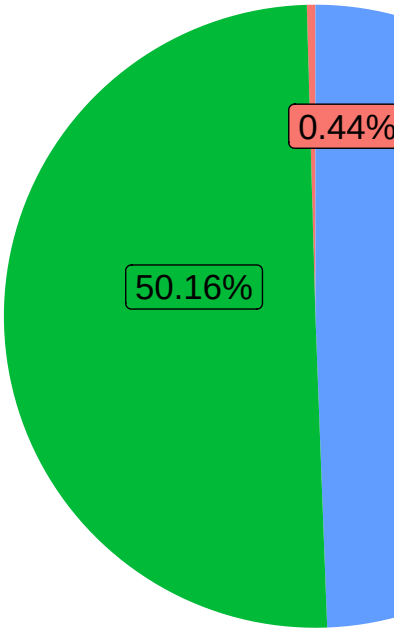
Percentage of ARG residue sid for Nucleus



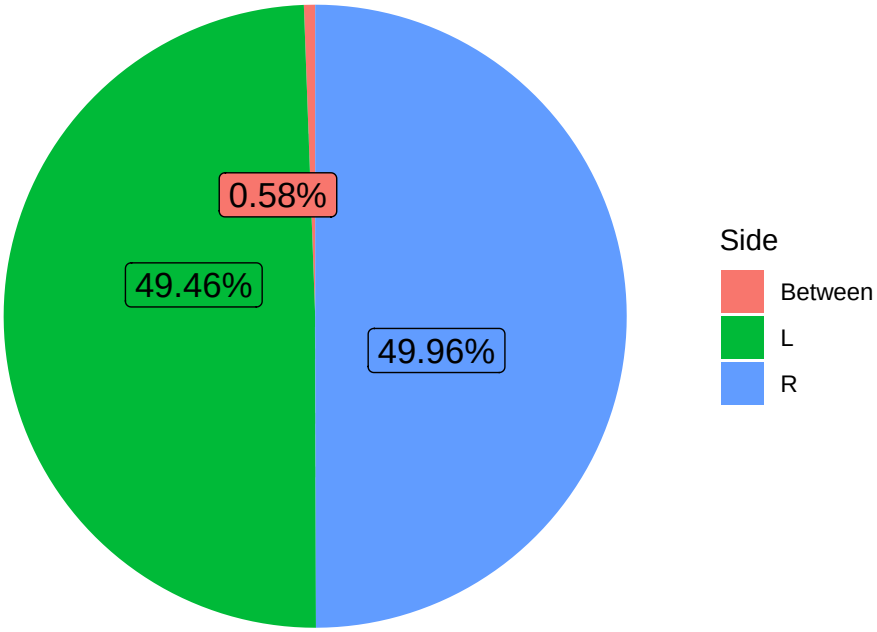
Percentage of ASP residue sides in GK + FL regions for Nucleus



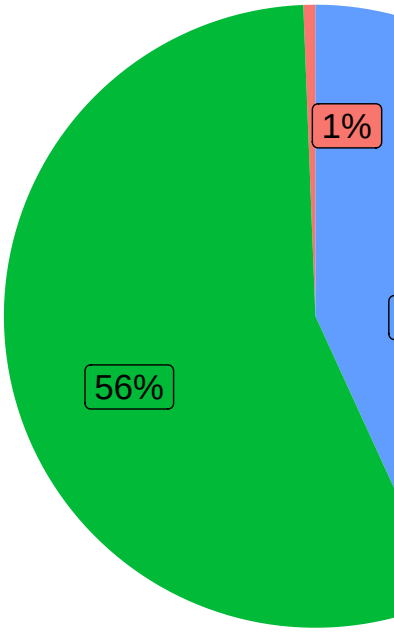
Percentage of GLU residue sid for Nucleus

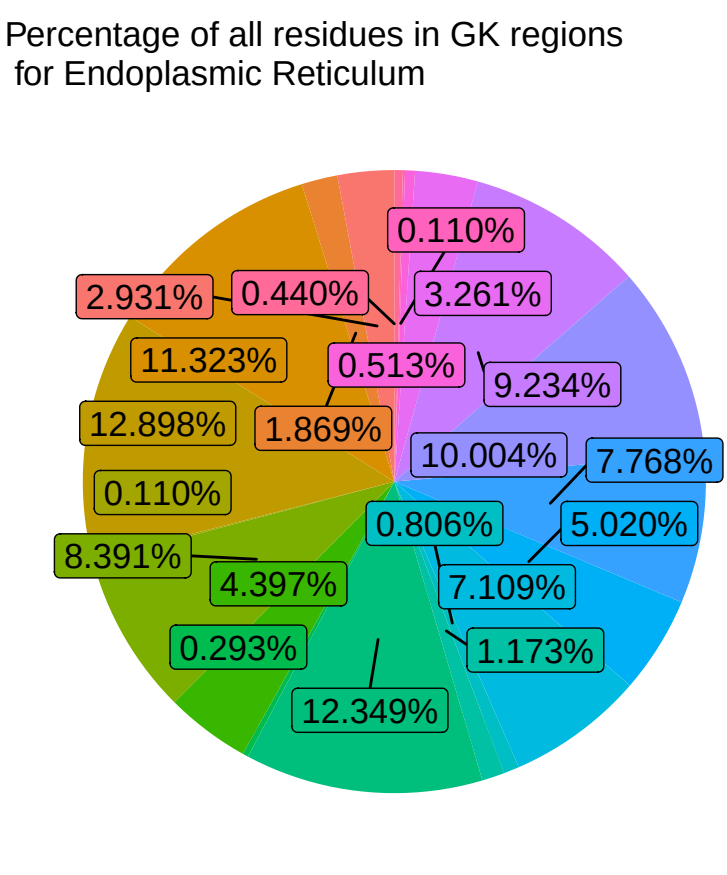


Percentage of SER residue sides in GK + FL regions for Nucleus

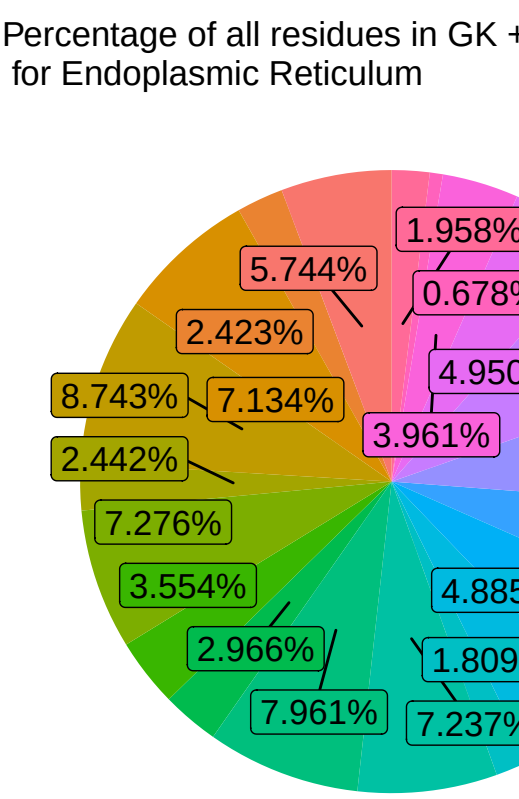


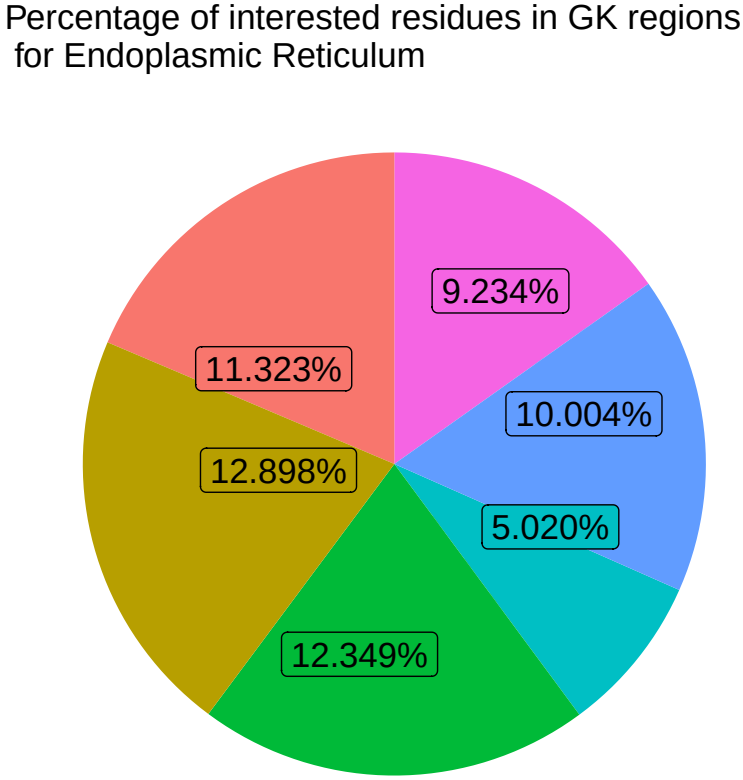
Percentage of PRO residue sid for Nucleus



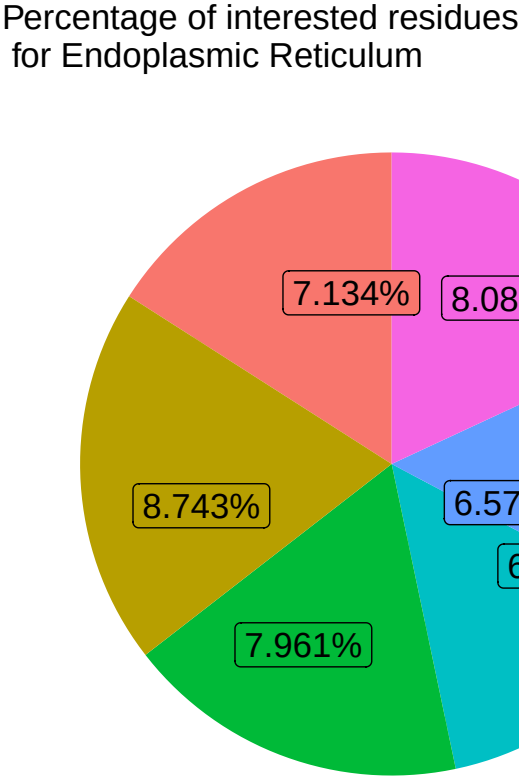

Percentage of all residues in GK regions
for Endoplasmic Reticulum
0.110%
2.931%
0.440%
3.261%
11.323%
0.513%
9.234%
12.898%
1.869%
10.004%
7.768%
0.110%
0.806%
5.020%
8.391%
4.397%
7.109%
0.293%
1.173%
12.349%

Residue
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V

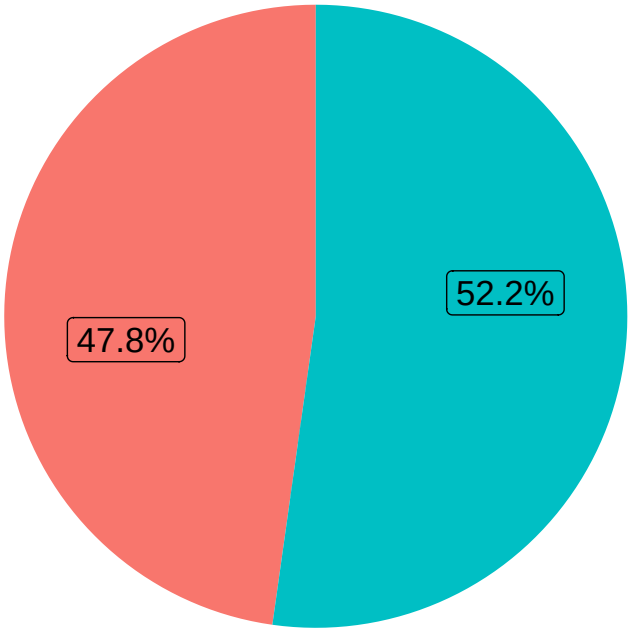

Percentage of all residues in GK +
for Endoplasmic Reticulum
1.958%
5.744%
0.678%
2.423%
4.950
8.743%
7.134%
3.961%
2.442%
7.276%
3.554%
4.885
2.966%
1.809
7.961%
7.237%


Percentage of interested residues in GK regions
for Endoplasmic Reticulum
9.234%
11.323%
10.004%
12.898%
5.020%
12.349%

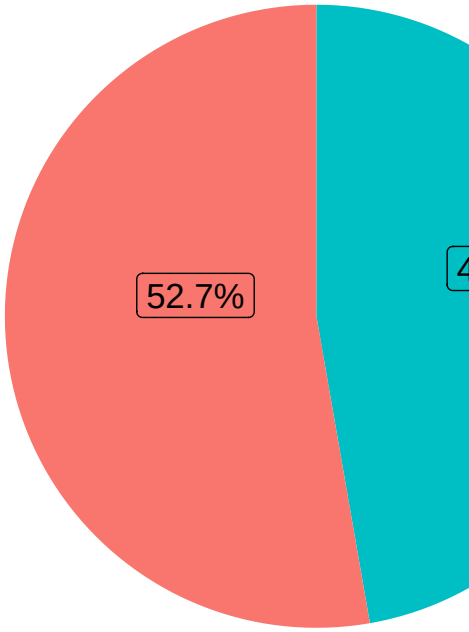
Residue
D
E
K
P
R
S


Percentage of interested residues
for Endoplasmic Reticulum
7.134%
8.08
8.743%
6.57
7.961%

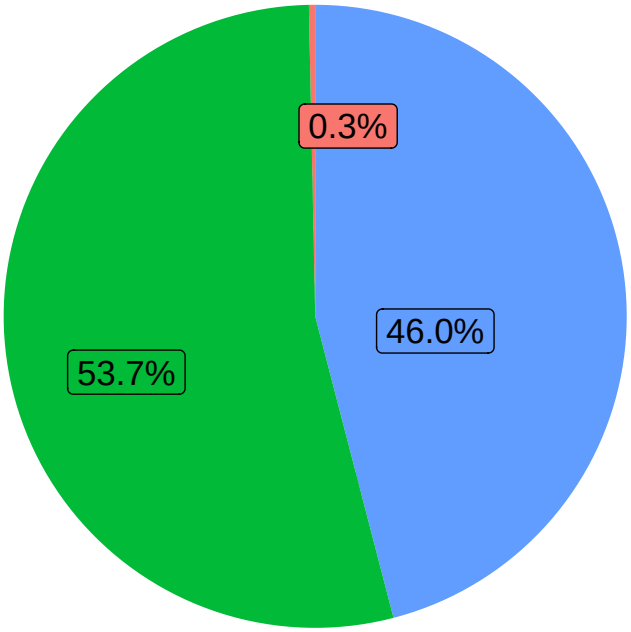
Percentage of LYS residue sides in GK regions
for Endoplasmic Reticulum



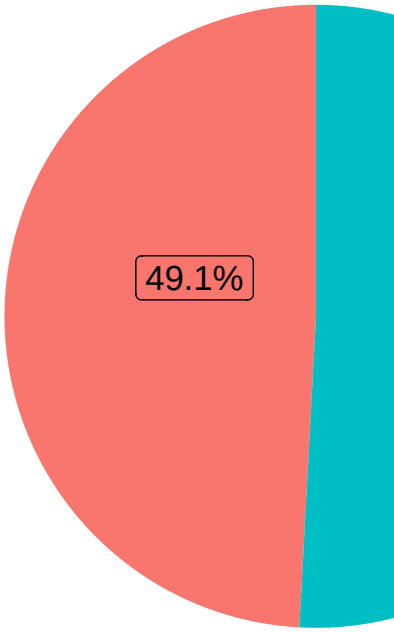
Percentage of ARG residue sides in
for Endoplasmic Reticulum



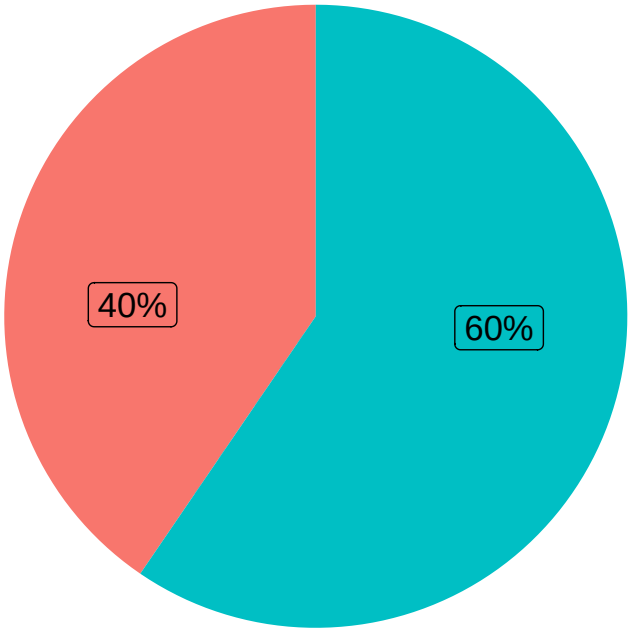
Percentage of ASP residue sides in GK regions
for Endoplasmic Reticulum



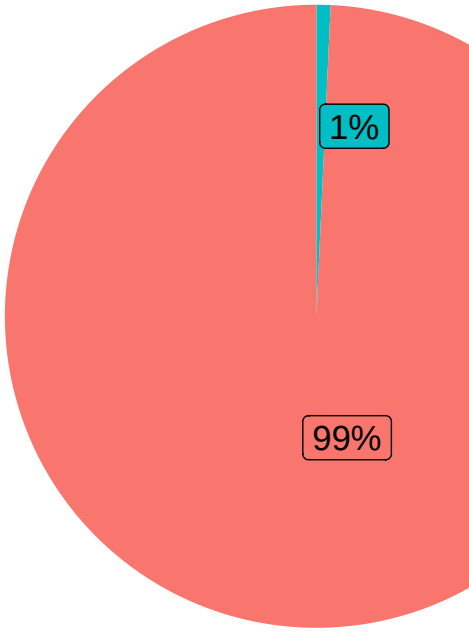
Percentage of GLU residue sid
for Endoplasmic Reticulum



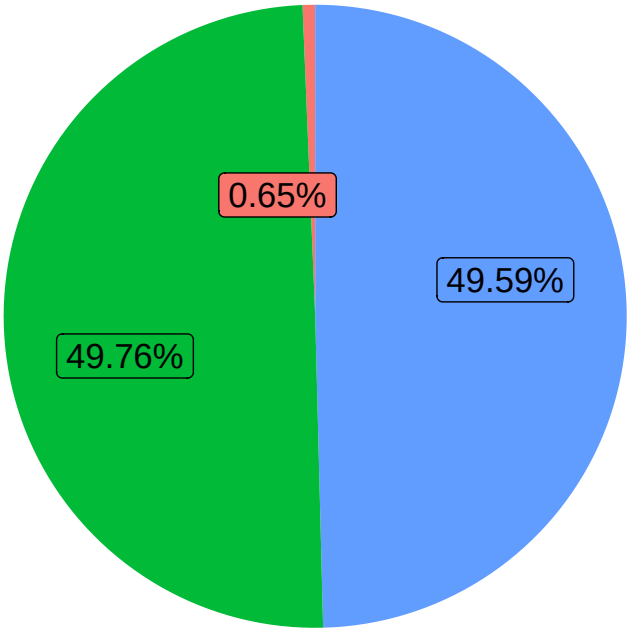
Percentage of SER residue sides in GK regions
for Endoplasmic Reticulum



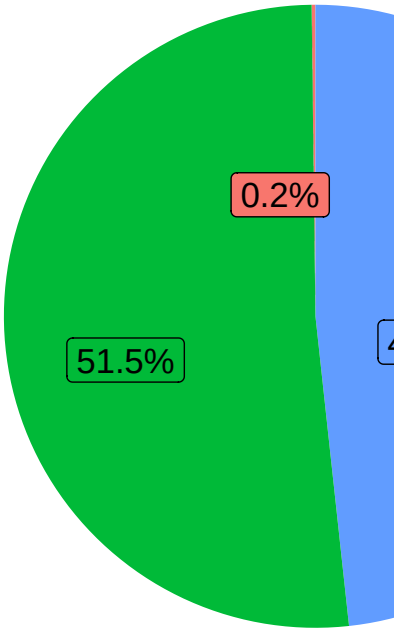
Percentage of PRO residue sides in
for Endoplasmic Reticulum



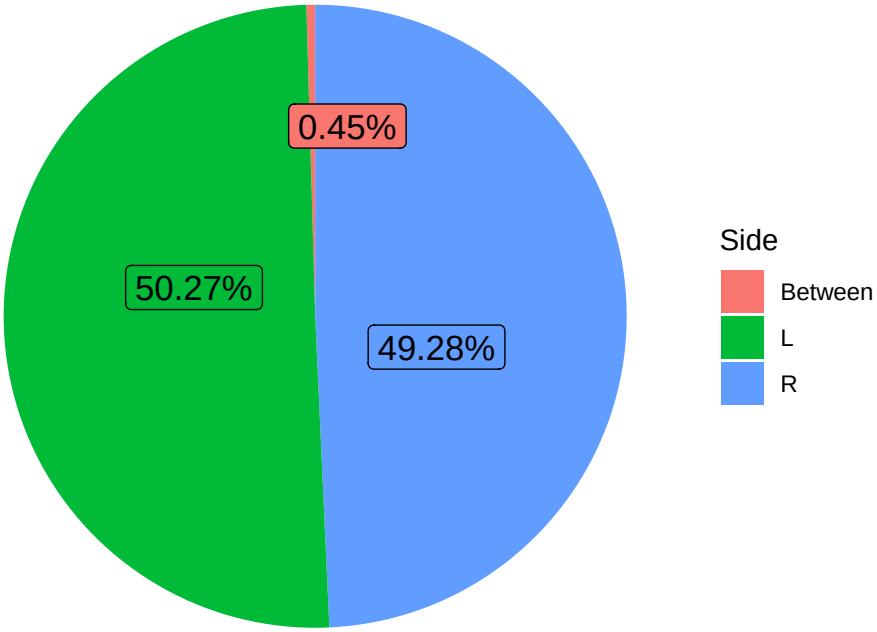
Percentage of LYS residue sides in GK + FL regions
for Endoplasmic Reticulum



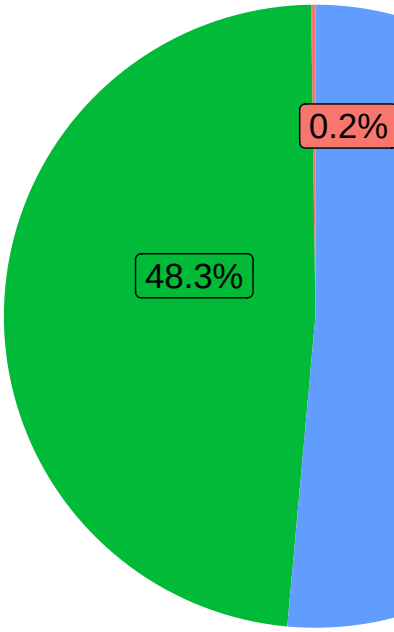
Percentage of ARG residue sid
for Endoplasmic Reticulum



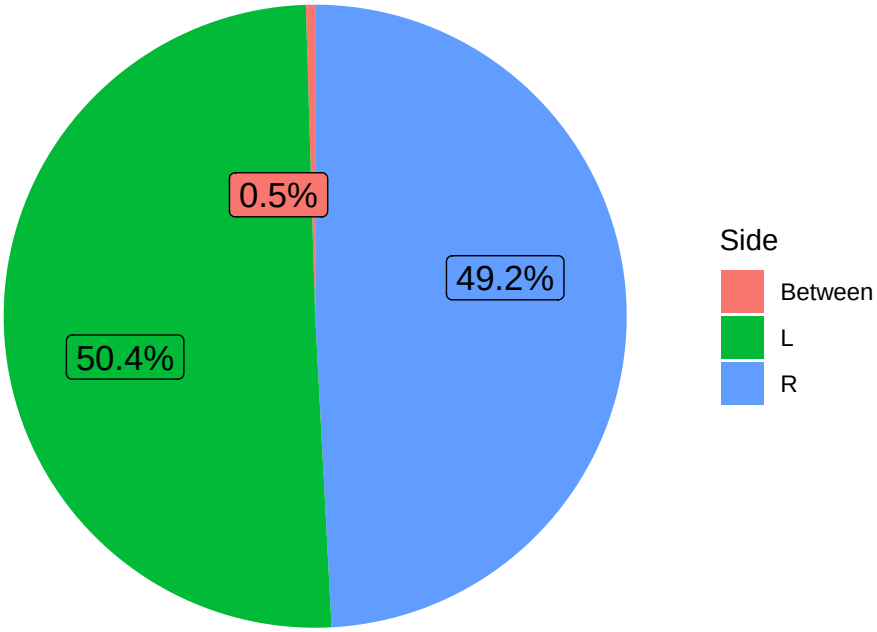
Percentage of ASP residue sides in GK + FL regions
for Endoplasmic Reticulum



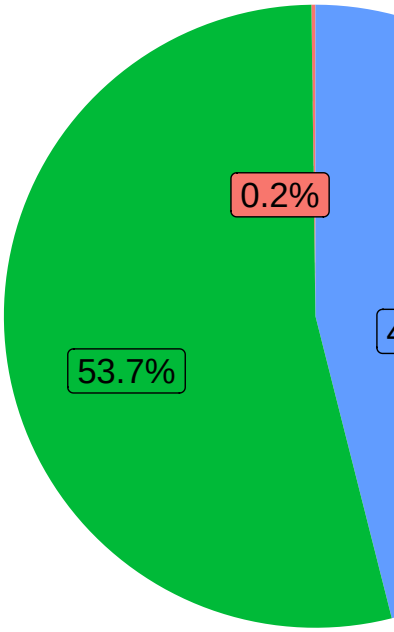
Percentage of GLU residue sid
for Endoplasmic Reticulum

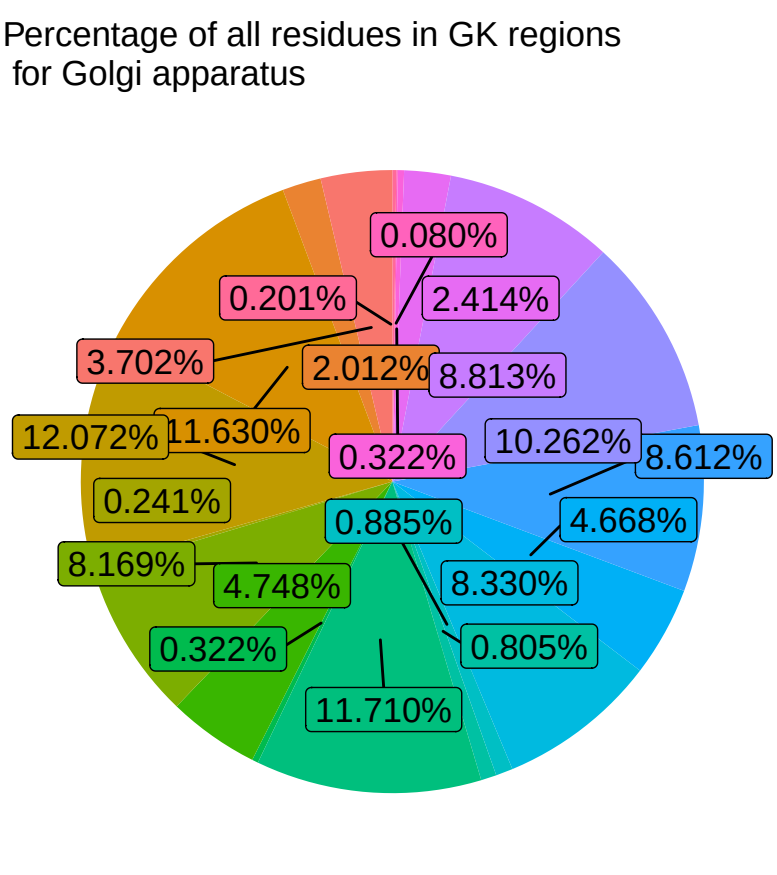


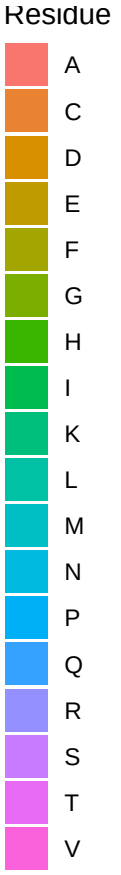
Percentage of SER residue sides in GK + FL regions
for Endoplasmic Reticulum

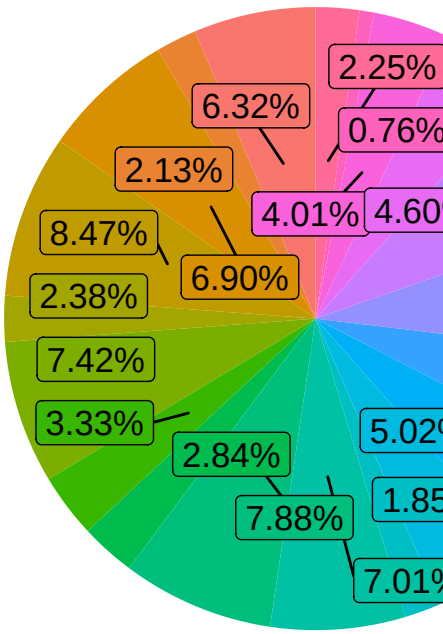


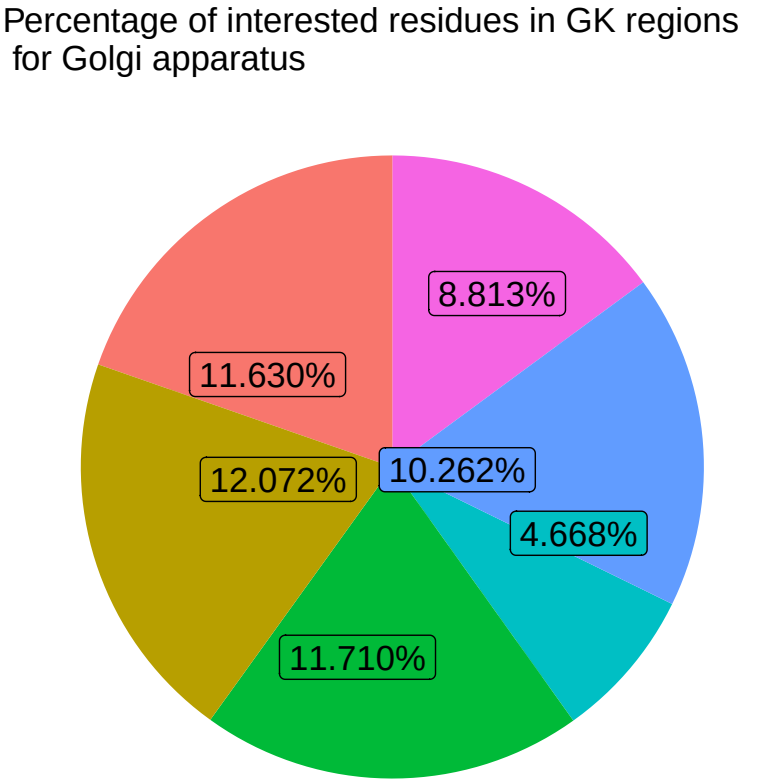
Percentage of PRO residue sid
for Endoplasmic Reticulum



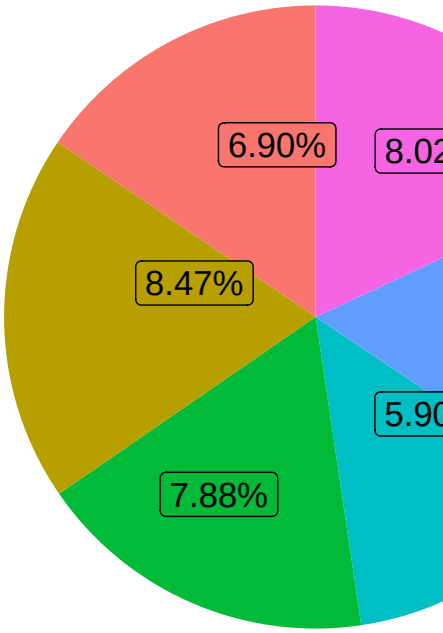

Percentage of all residues in GK regions
for Golgi apparatus
0.080%
2.414%
0.201%
3.702%
2.012%
8.813%
12.072%
11.630%
0.322%
10.262%
8.612%
0.241%
0.885%
4.668%
8.169%
4.748%
8.330%
0.322%
0.805%
11.710%


Residue
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V

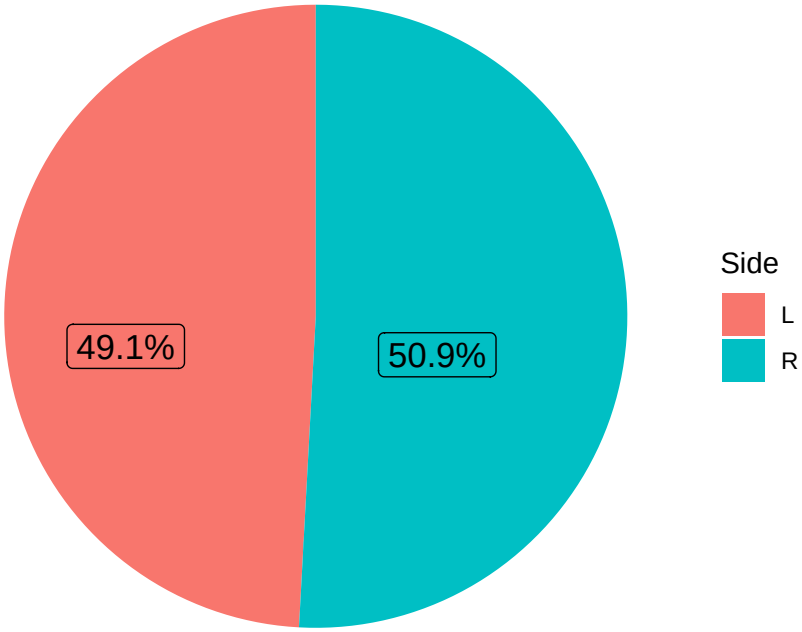

Percentage of all residues in GK +
for Golgi apparatus
2.25%
6.32%
0.76%
2.13%
4.01%
4.60
8.47%
6.90%
2.38%
7.42%
3.33%
5.02
2.84%
1.85
7.88%
7.01%


Percentage of interested residues in GK regions
for Golgi apparatus
8.813%
11.630%
12.072%
10.262%
4.668%
11.710%

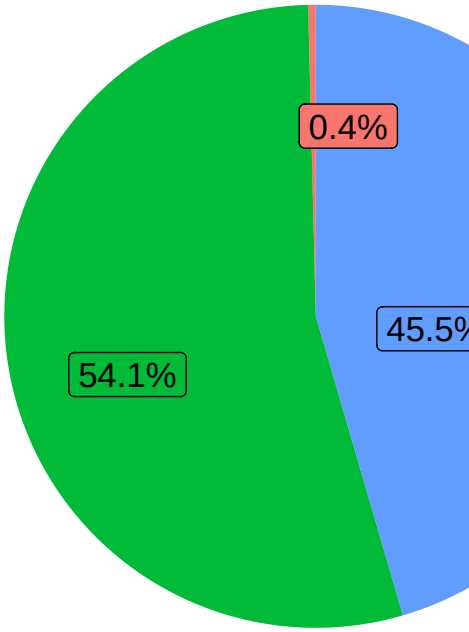
Residue
D
E
K
P
R
S


Percentage of interested residues
for Golgi apparatus
6.90%
8.02
8.47%
5.90
7.88%

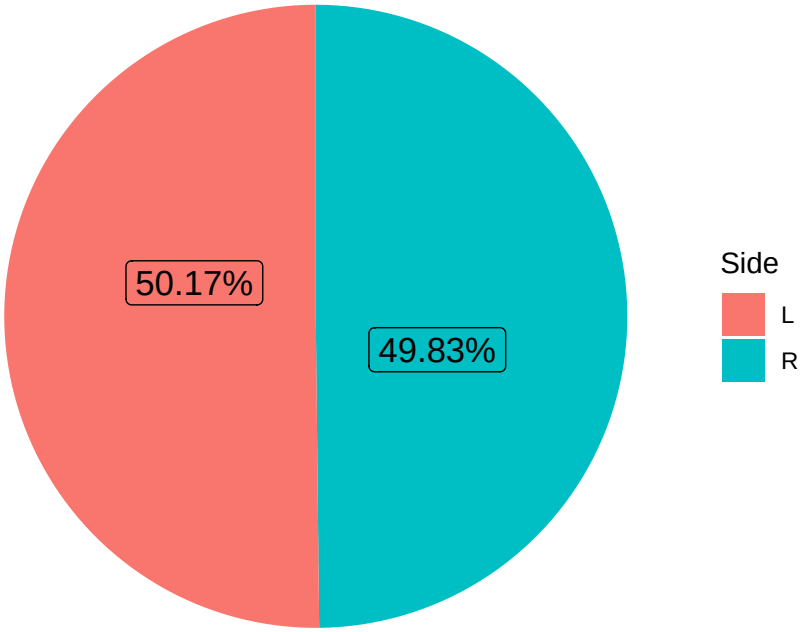
Percentage of LYS residue sides in GK regions for Golgi apparatus



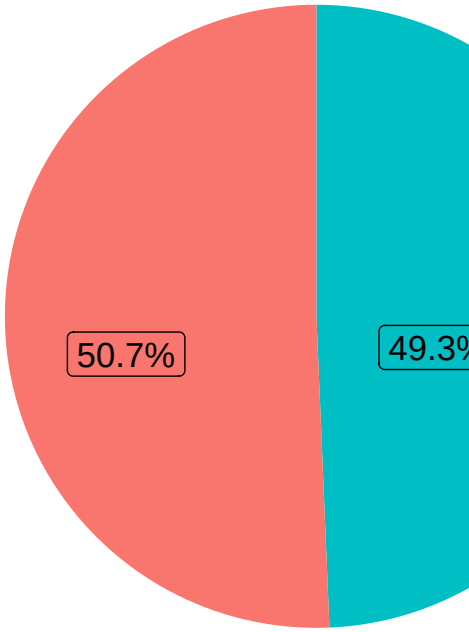
Percentage of ARG residue sides in for Golgi apparatus



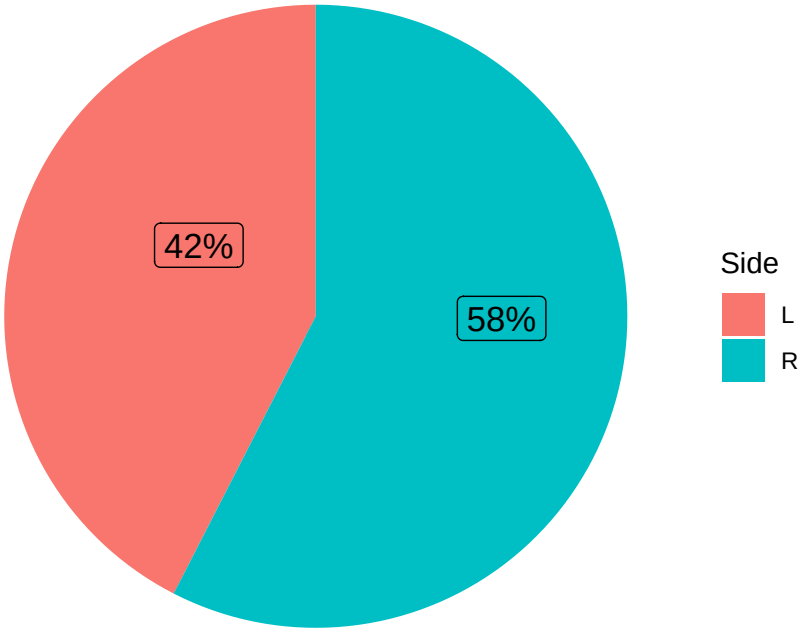
Percentage of ASP residue sides in GK regions for Golgi apparatus



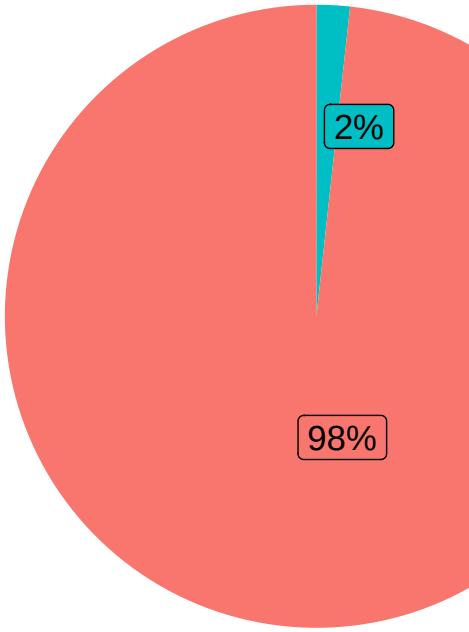
Percentage of GLU residue sides in for Golgi apparatus



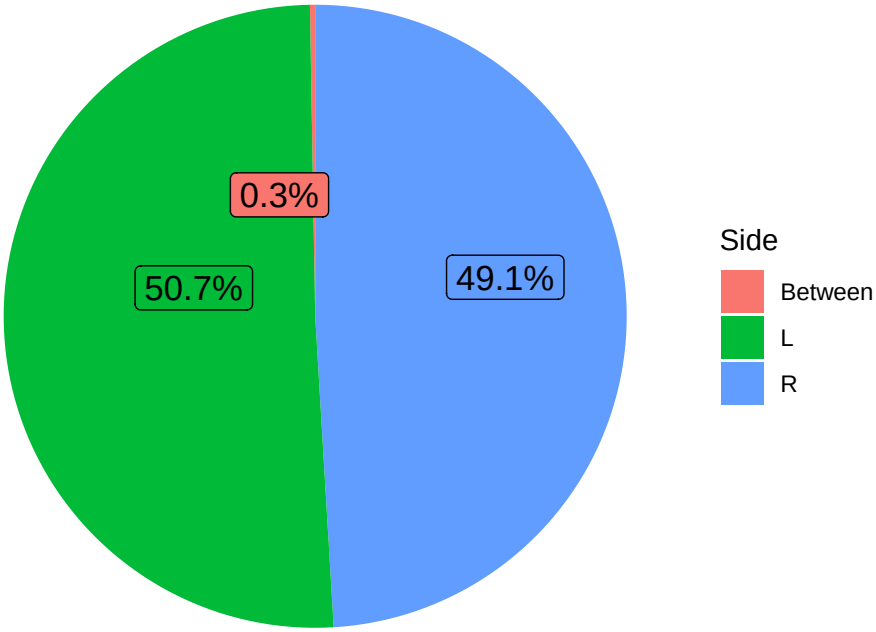
Percentage of SER residue sides in GK regions for Golgi apparatus



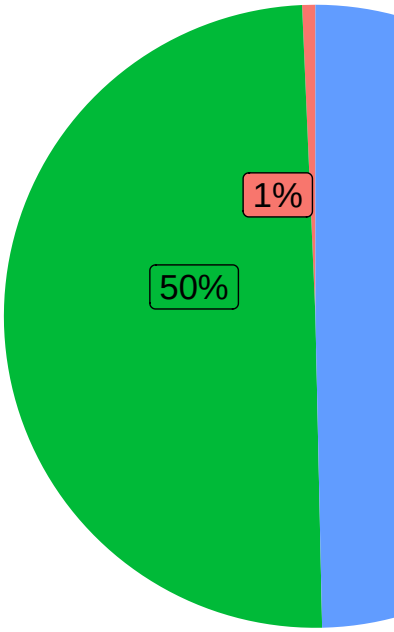
Percentage of PRO residue sides in for Golgi apparatus



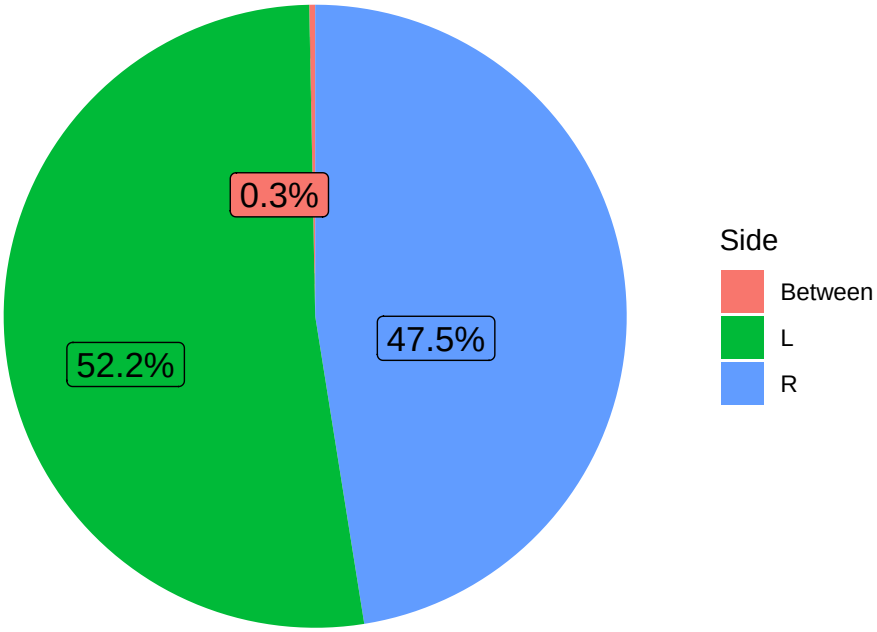
Percentage of LYS residue sides in GK + FL regions for Golgi apparatus



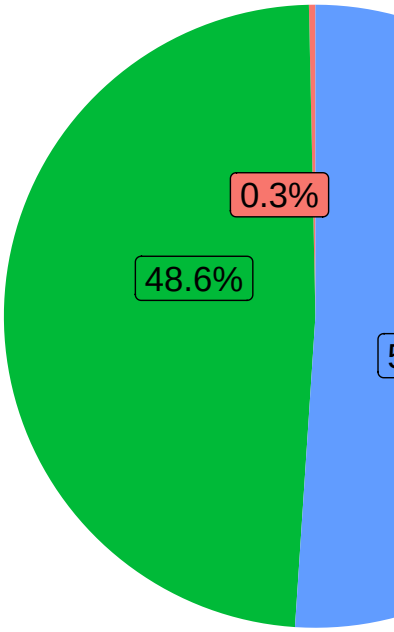
Percentage of ARG residue sid for Golgi apparatus



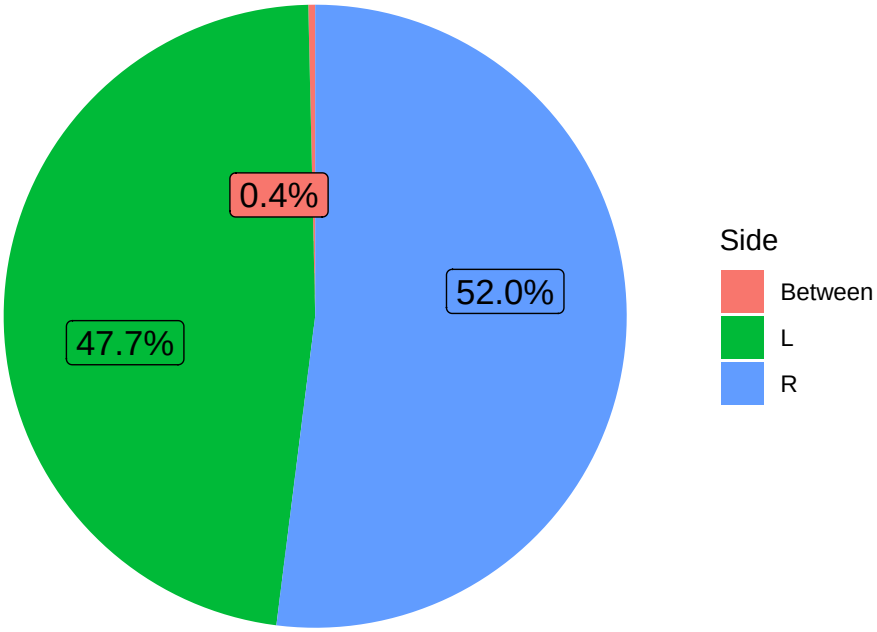
Percentage of ASP residue sides in GK + FL regions for Golgi apparatus



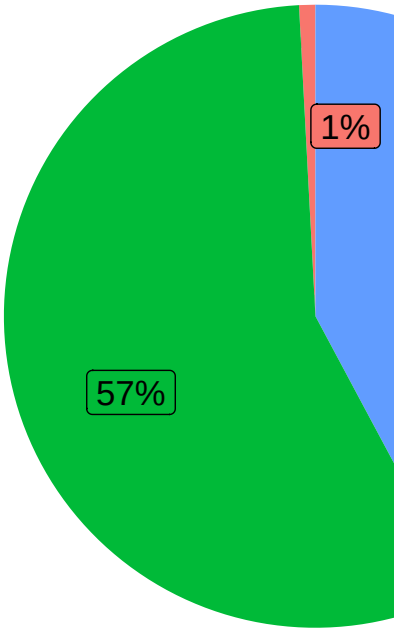
Percentage of GLU residue sid for Golgi apparatus



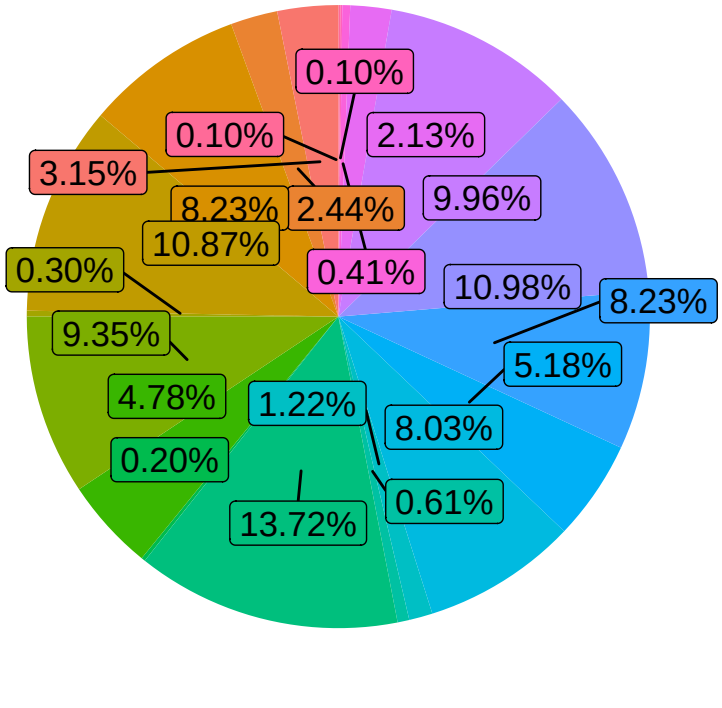
Percentage of SER residue sides in GK + FL regions for Golgi apparatus



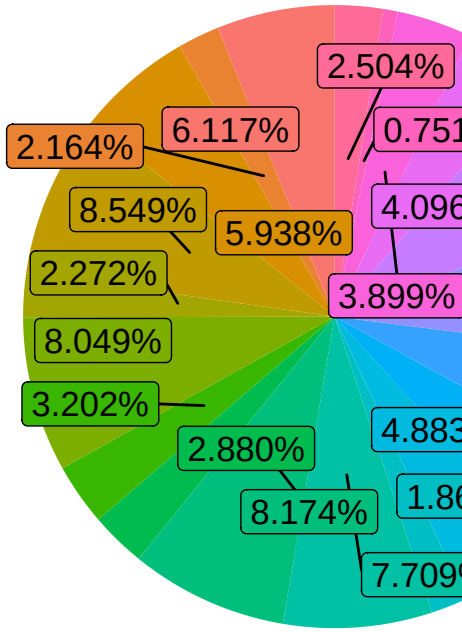
Percentage of PRO residue sid for Golgi apparatus



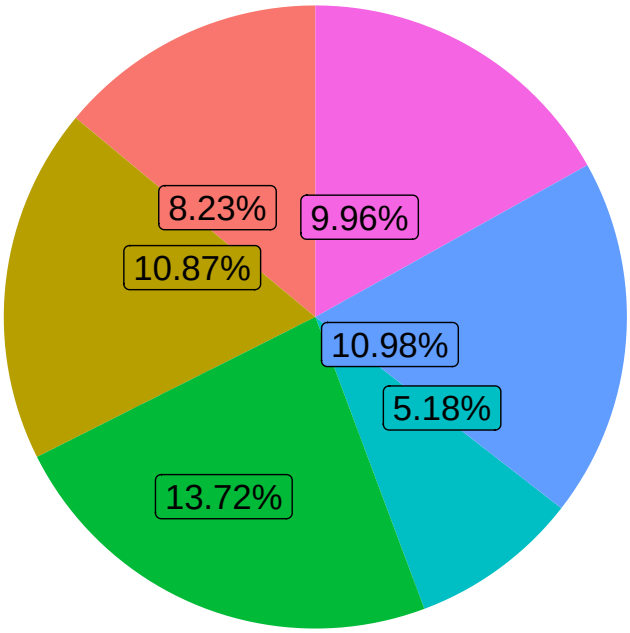
Percentage of all residues in GK regions for Lysosome



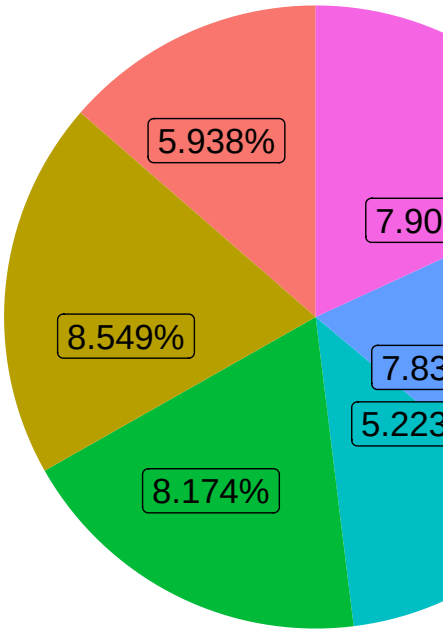
Percentage of all residues in GK + for Lysosome



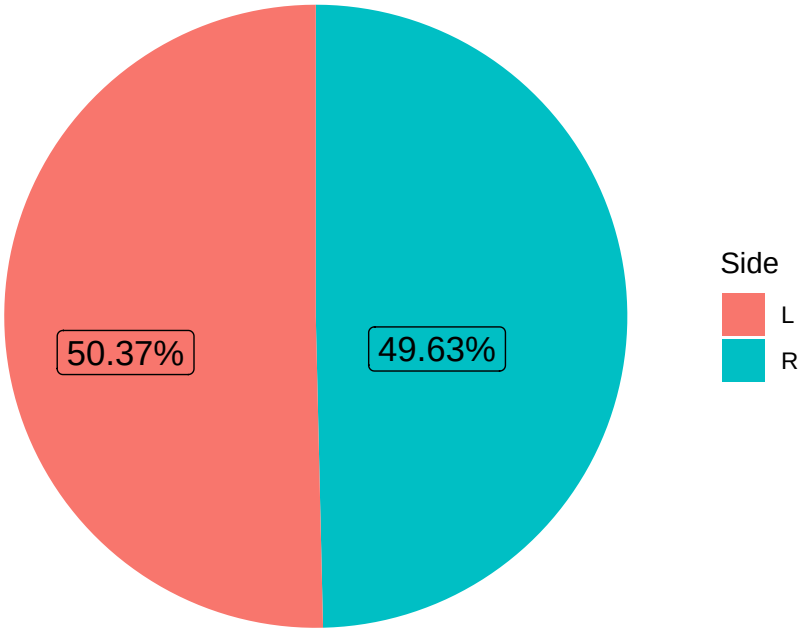
Percentage of interested residues in GK regions for Lysosome



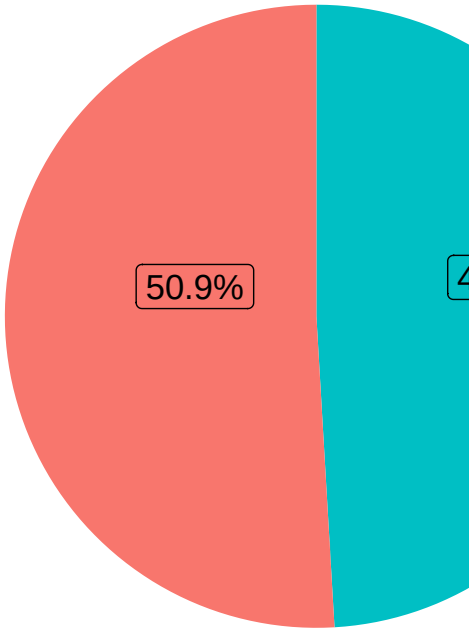
Percentage of interested residues for Lysosome



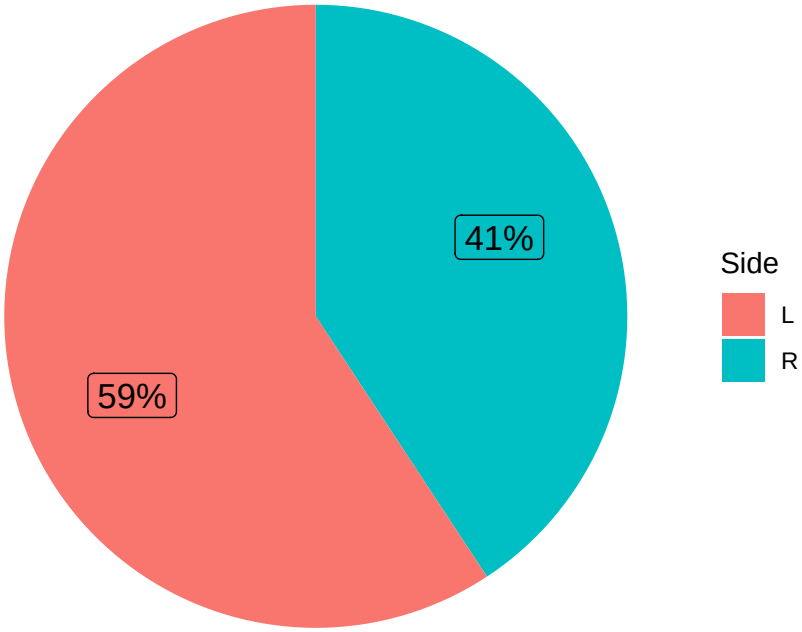
Percentage of LYS residue sides in GK regions for Lysosome



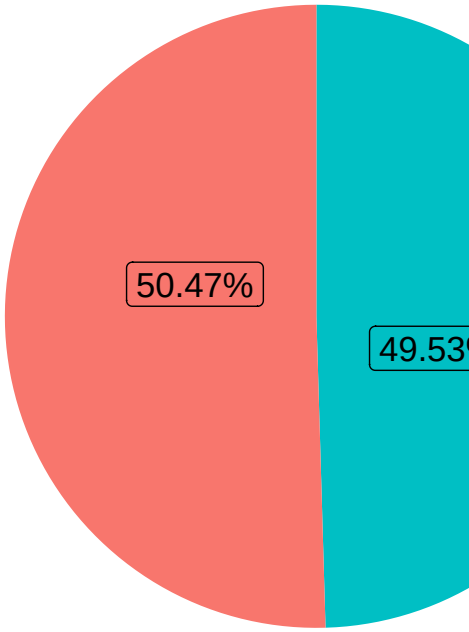
Percentage of ARG residue sides in GK regions for Lysosome



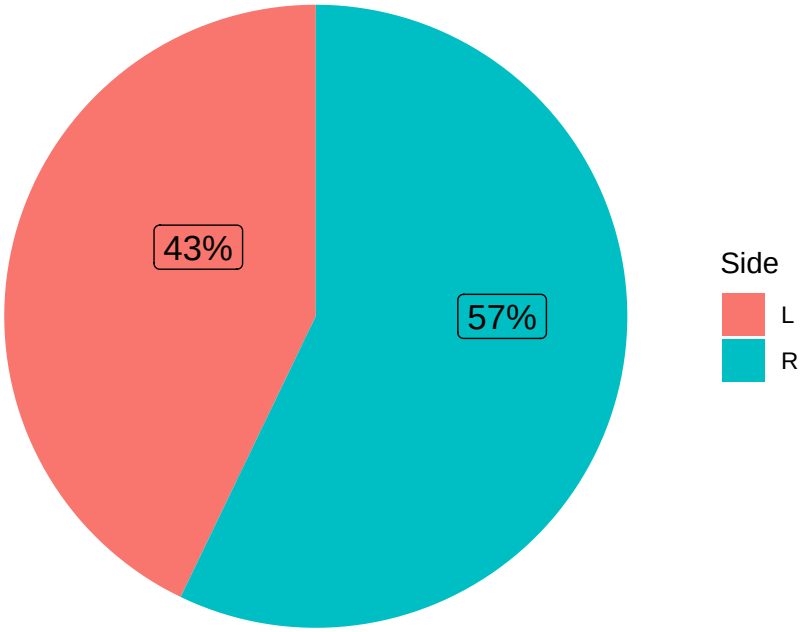
Percentage of ASP residue sides in GK regions for Lysosome



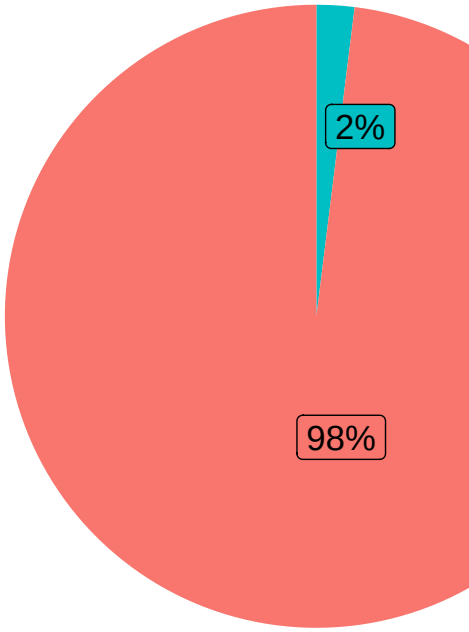
Percentage of GLU residue sides in GK regions for Lysosome



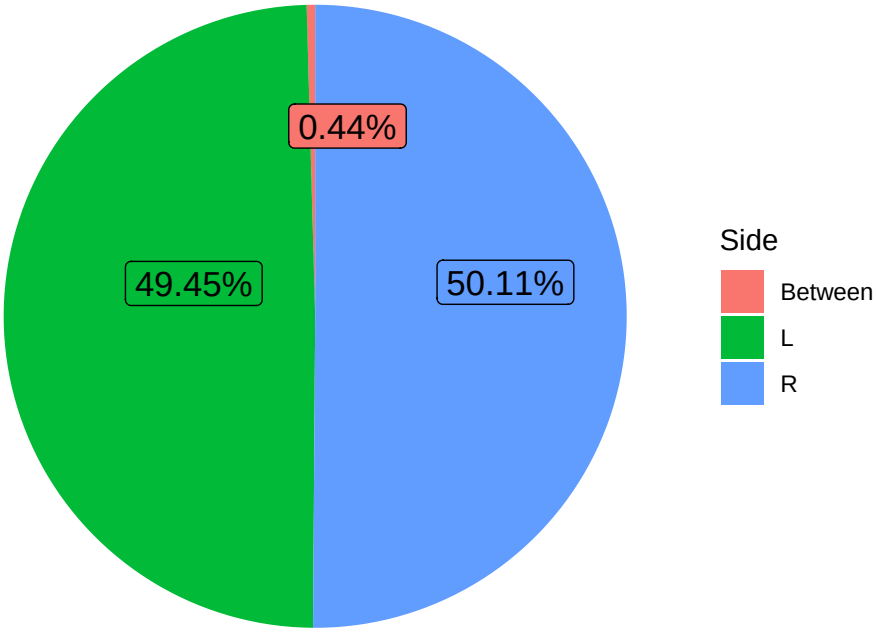
Percentage of SER residue sides in GK regions for Lysosome



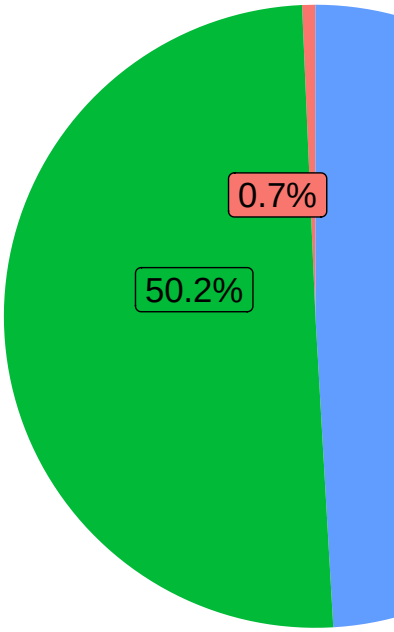
Percentage of PRO residue sides in for Lysosome



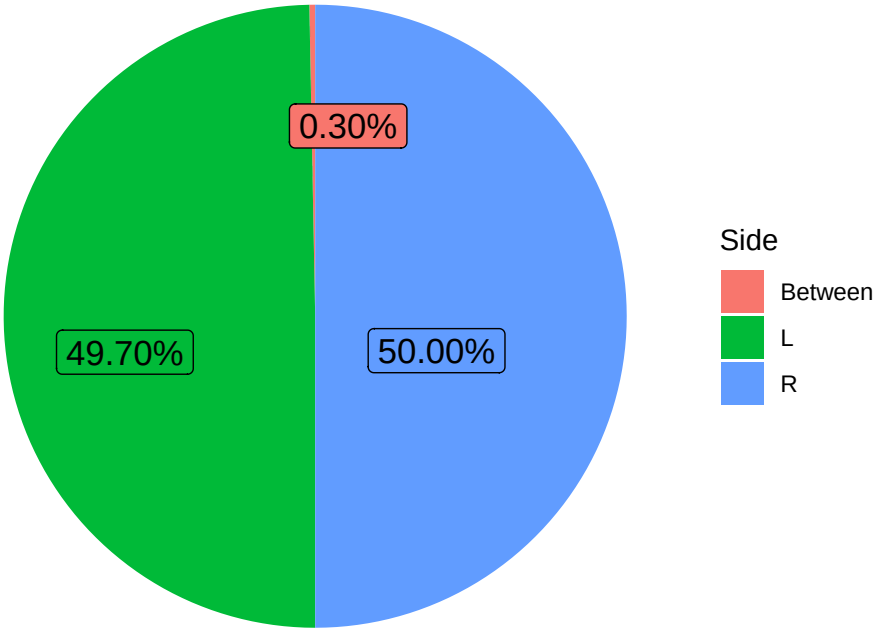
Percentage of LYS residue sides in GK + FL regions for Lysosome



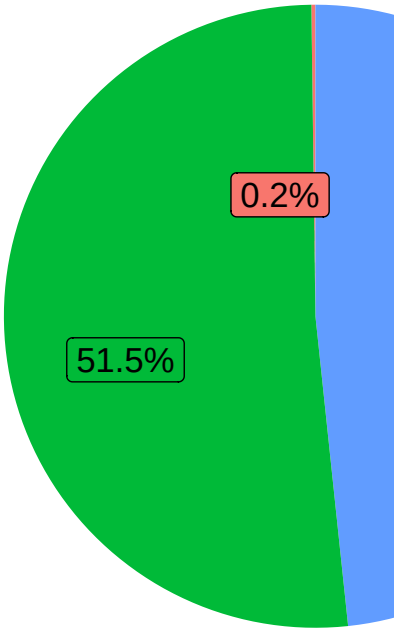
Percentage of ARG residue sid for Lysosome



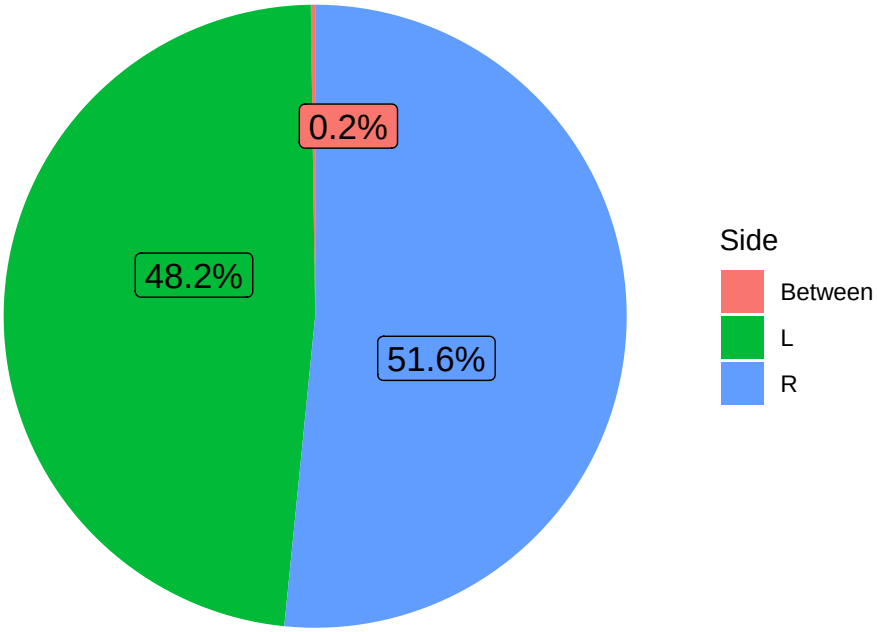
Percentage of ASP residue sides in GK + FL regions for Lysosome



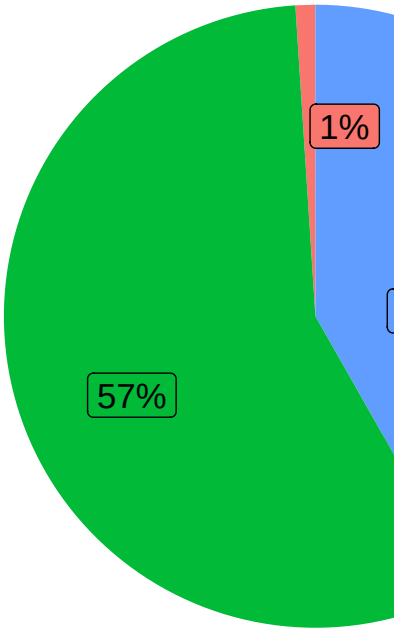
Percentage of GLU residue sid for Lysosome



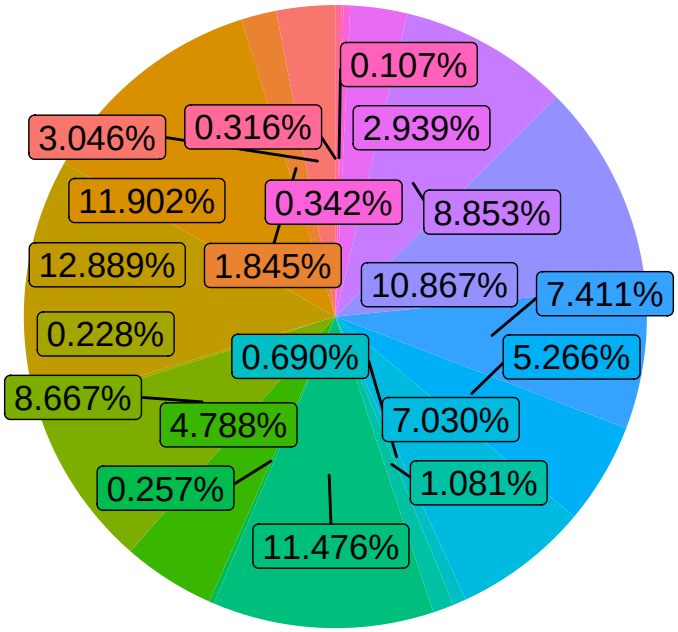
Percentage of SER residue sides in GK + FL regions for Lysosome



Percentage of PRO residue sid for Lysosome

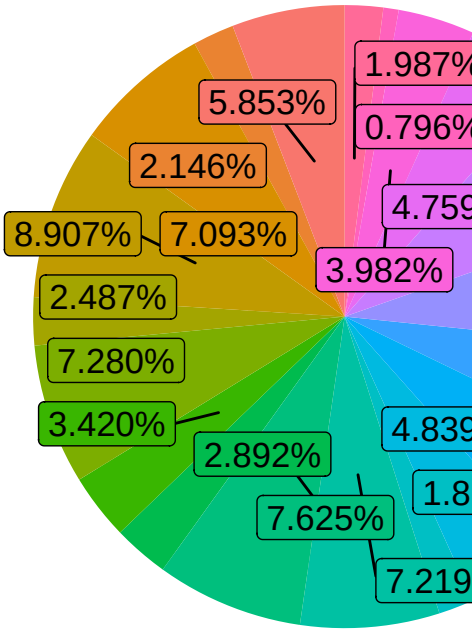


Percentage of all residues in GK regions for Cytoplasm

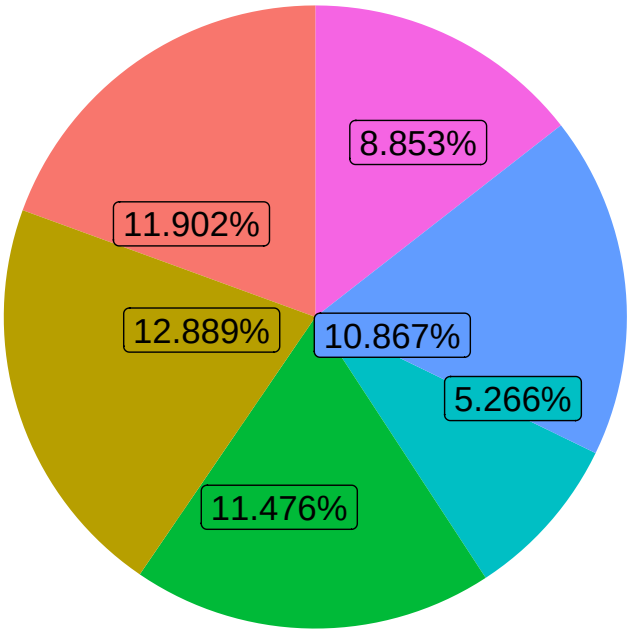


Residue
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V

Percentage of all residues in GK + for Cytoplasm

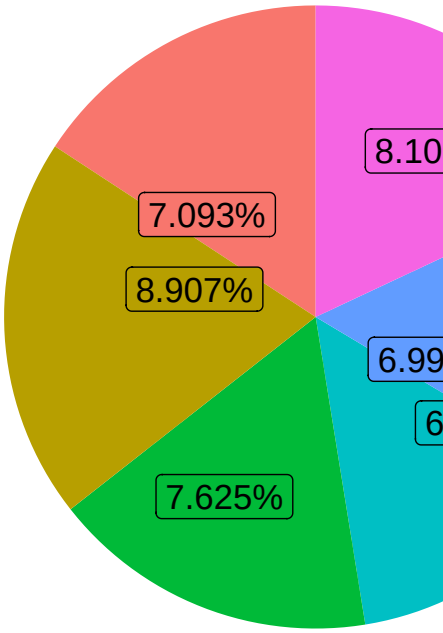


Percentage of interested residues in GK regions for Cytoplasm

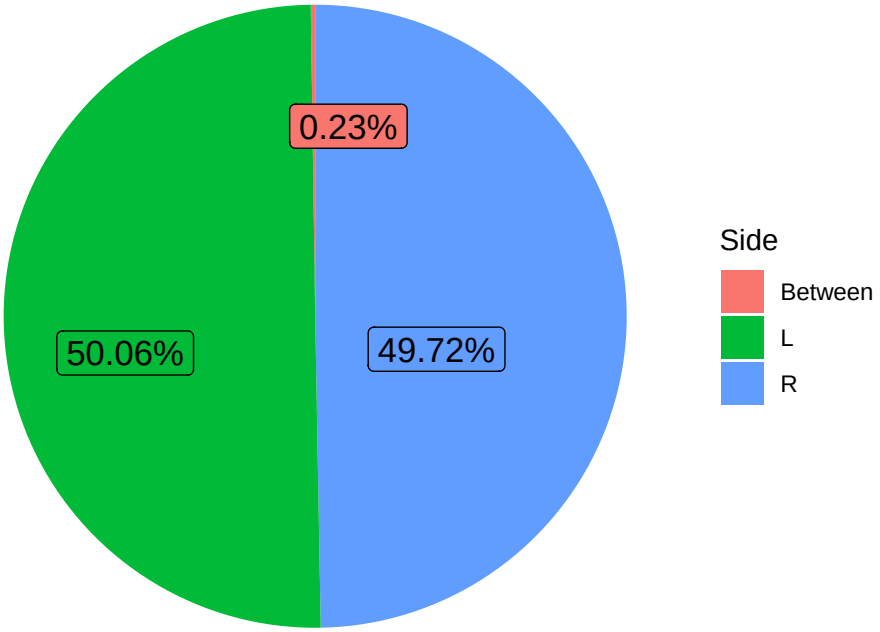


Residue
D
E
K
P
R
S

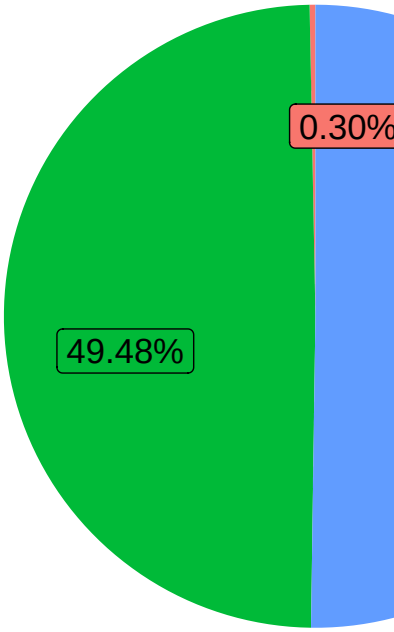
Percentage of interested residues for Cytoplasm



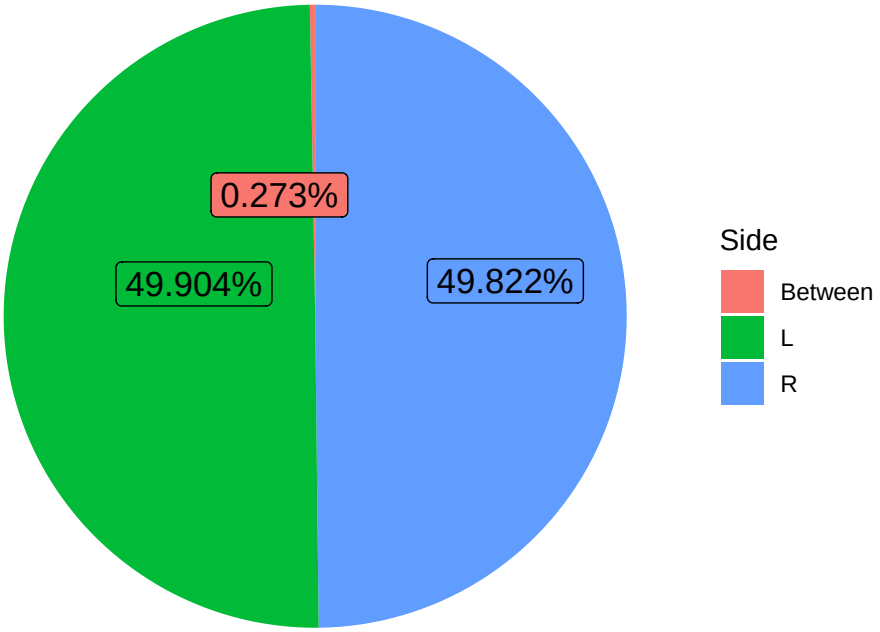
Percentage of LYS residue sides in GK regions for Cytoplasm



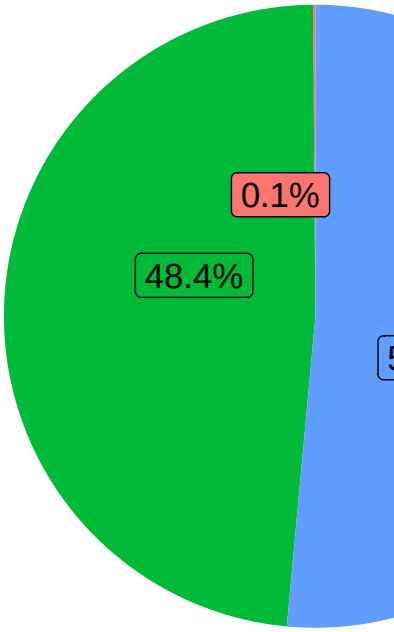
Percentage of ARG residue sid for Cytoplasm



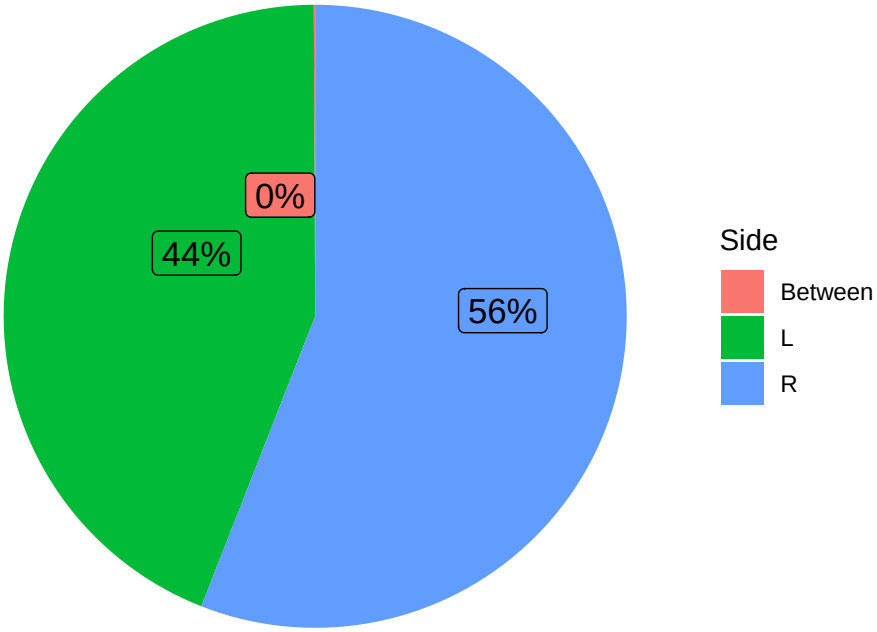
Percentage of ASP residue sides in GK regions for Cytoplasm



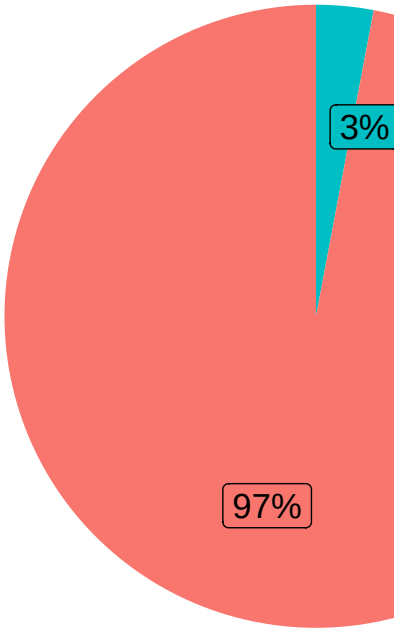
Percentage of GLU residue sid for Cytoplasm



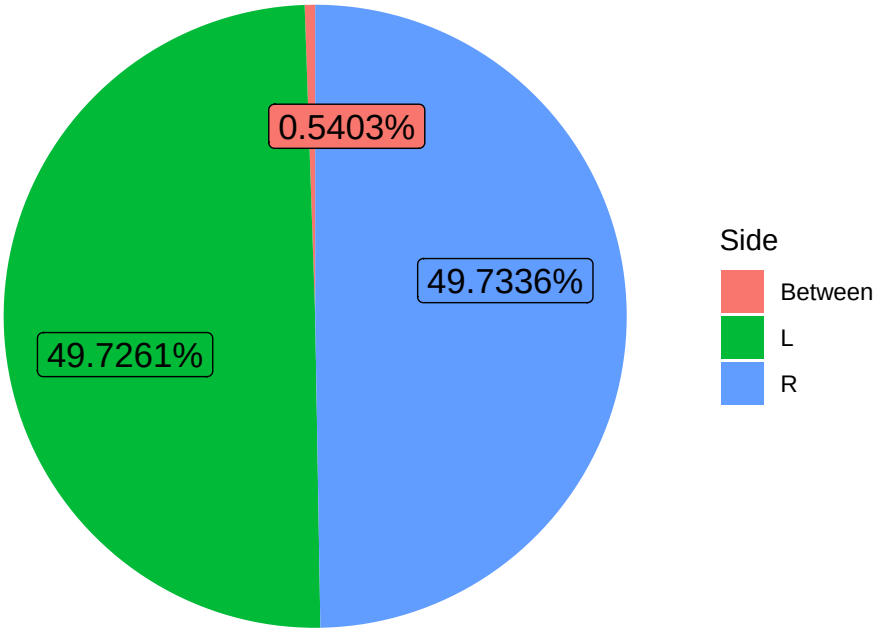
Percentage of SER residue sides in GK regions for Cytoplasm



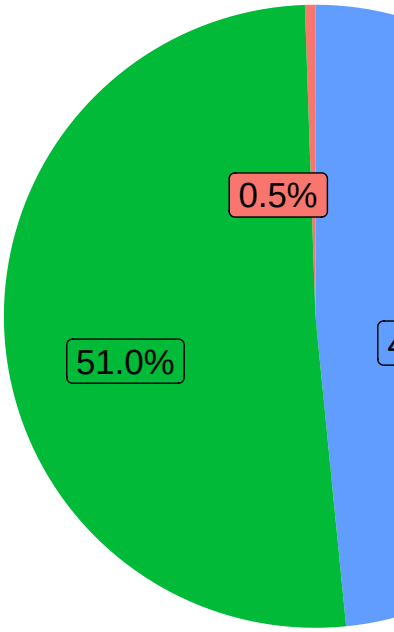
Percentage of PRO residue sid for Cytoplasm



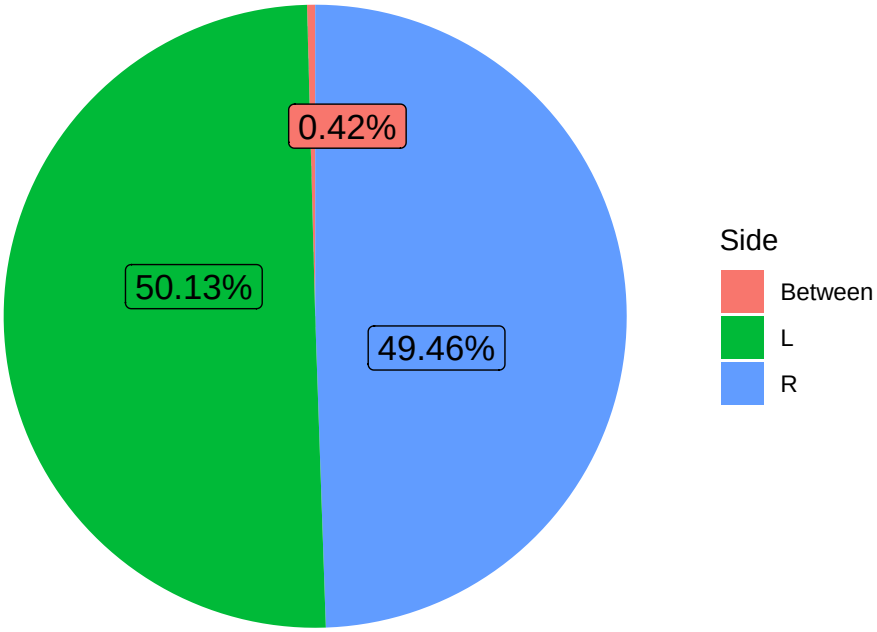
Percentage of LYS residue sides in GK + FL regions for Cytoplasm



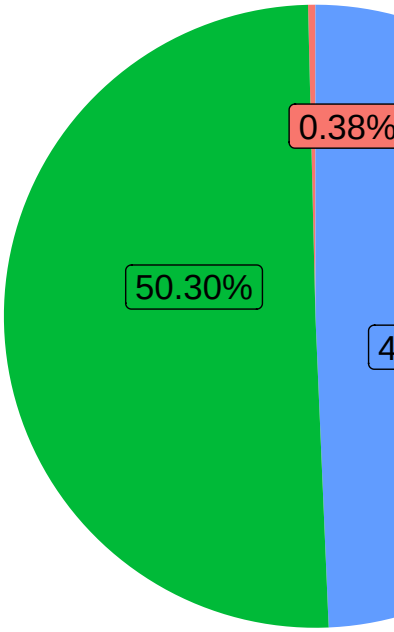
Percentage of ARG residue sid for Cytoplasm



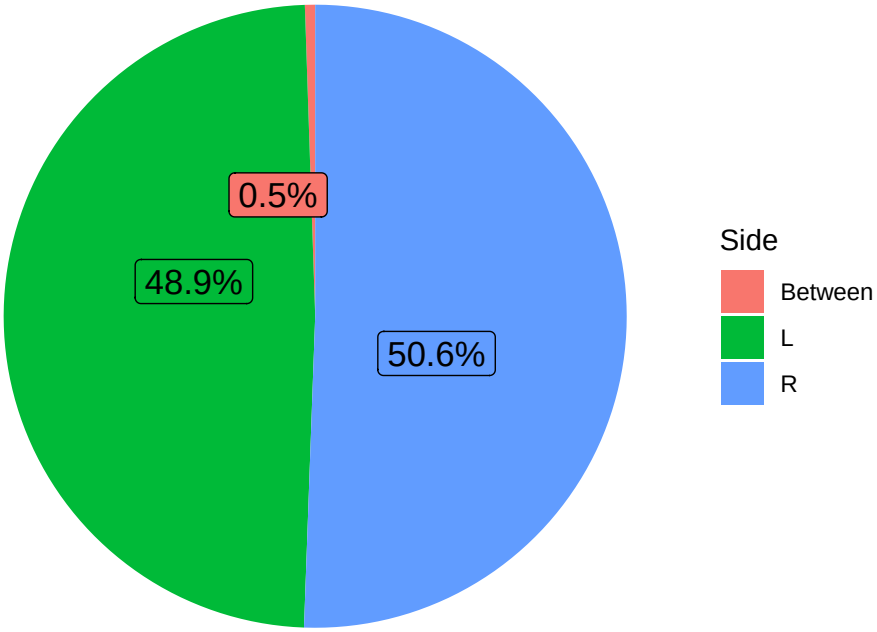
Percentage of ASP residue sides in GK + FL regions for Cytoplasm



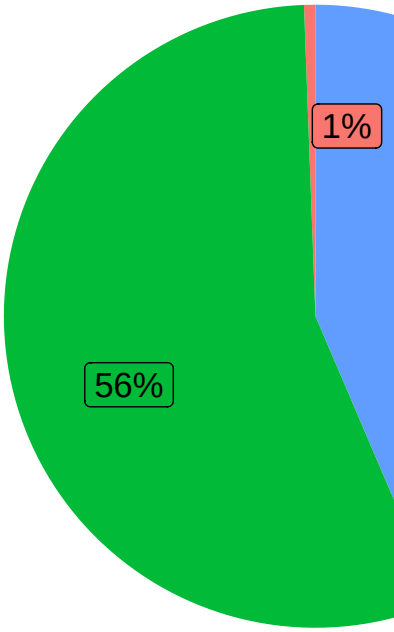
Percentage of GLU residue sid for Cytoplasm

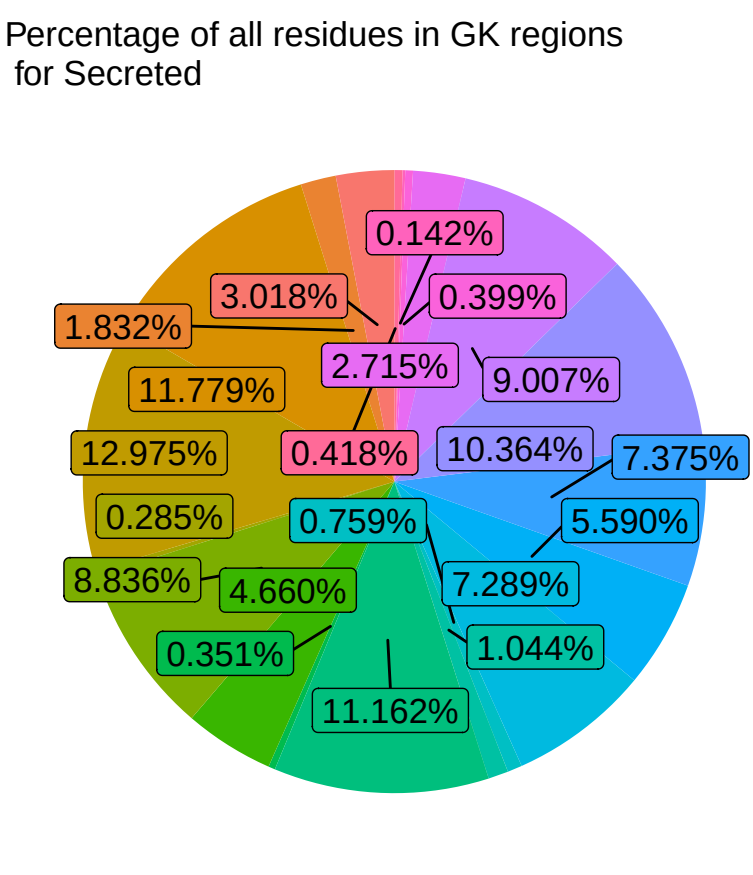


Percentage of SER residue sides in GK + FL regions for Cytoplasm

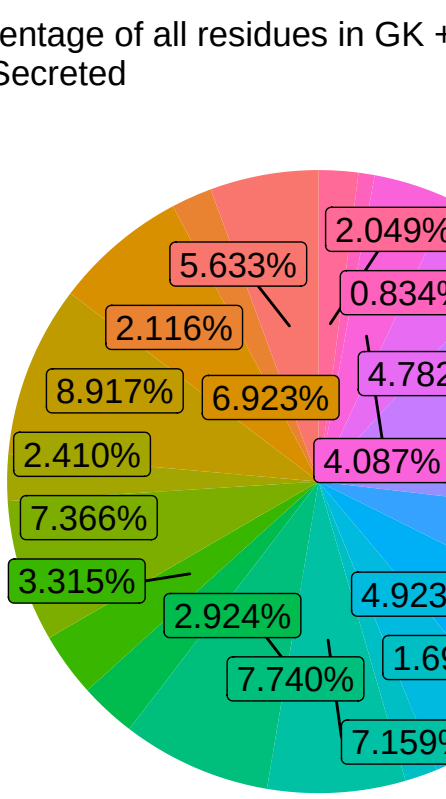


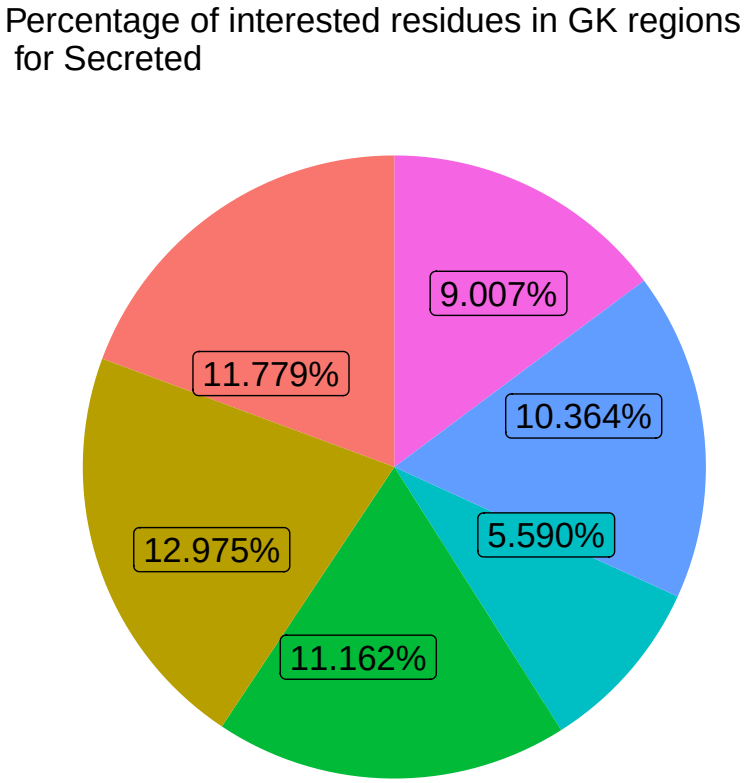
Percentage of PRO residue sid for Cytoplasm



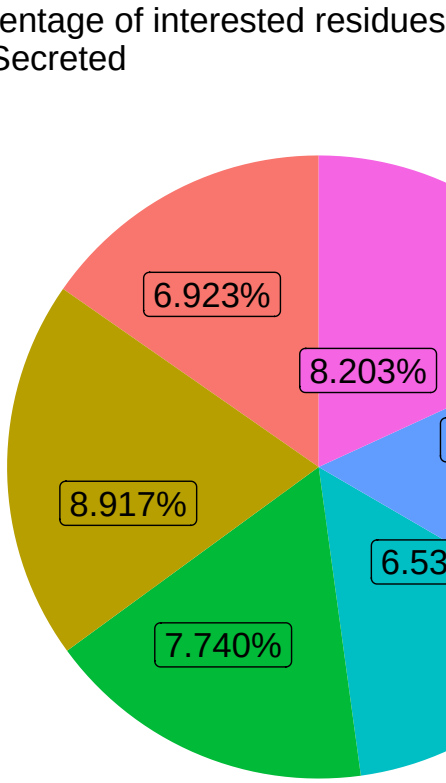

Percentage of all residues in GK regions
for Secreted
0.142%
3.018%
0.399%
1.832%
2.715%
9.007%
11.779%
12.975%
0.418%
10.364%
7.375%
0.285%
0.759%
5.590%
8.836%
4.660%
7.289%
0.351%
1.044%
11.162%

Residue
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V

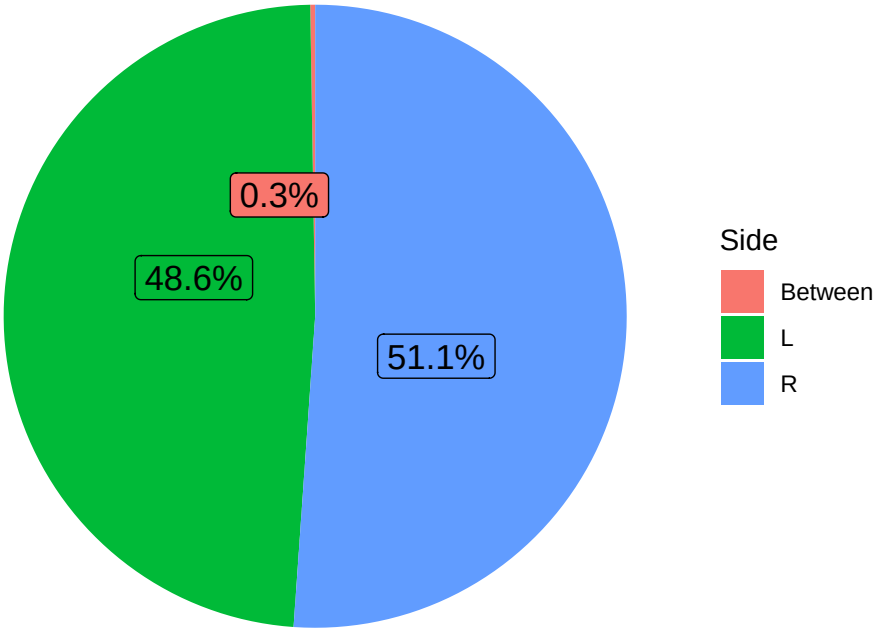

Percentage of all residues in GK +
for Secreted
2.049%
5.633%
0.834%
2.116%
4.782
8.917%
6.923%
2.410%
4.087%
7.366%
3.315%
4.923
2.924%
1.6
7.740%
7.159%


Percentage of interested residues in GK regions
for Secreted
9.007%
11.779%
10.364%
5.590%
12.975%
11.162%

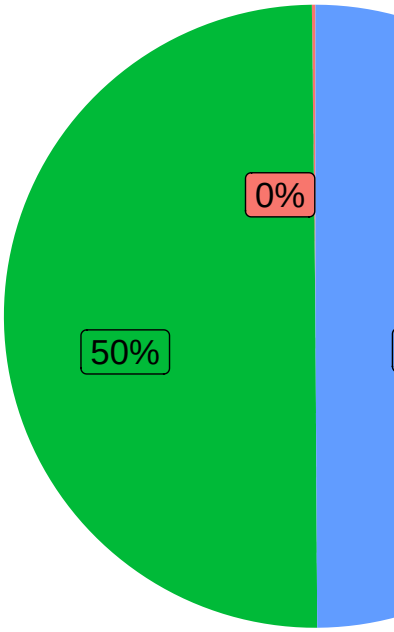
Residue
D
E
K
P
R
S


Percentage of interested residues
for Secreted
6.923%
8.203%
8.917%
6.53
7.740%

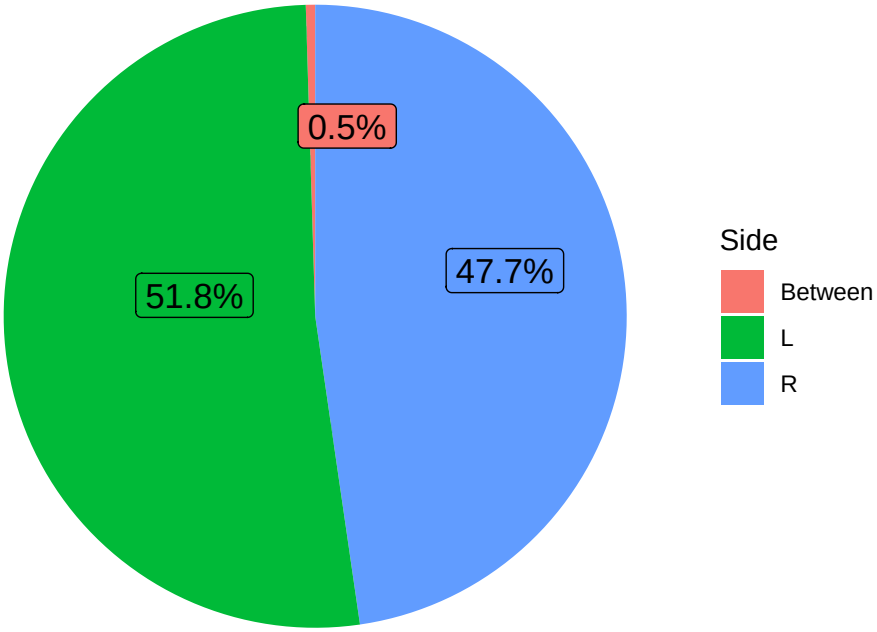
Percentage of LYS residue sides in GK regions for Secreted



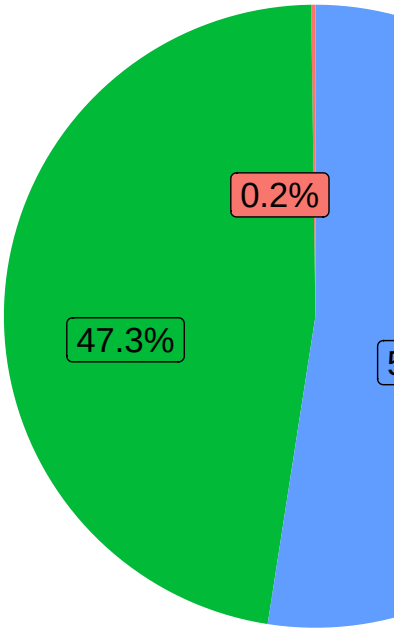
Percentage of ARG residue sid for Secreted



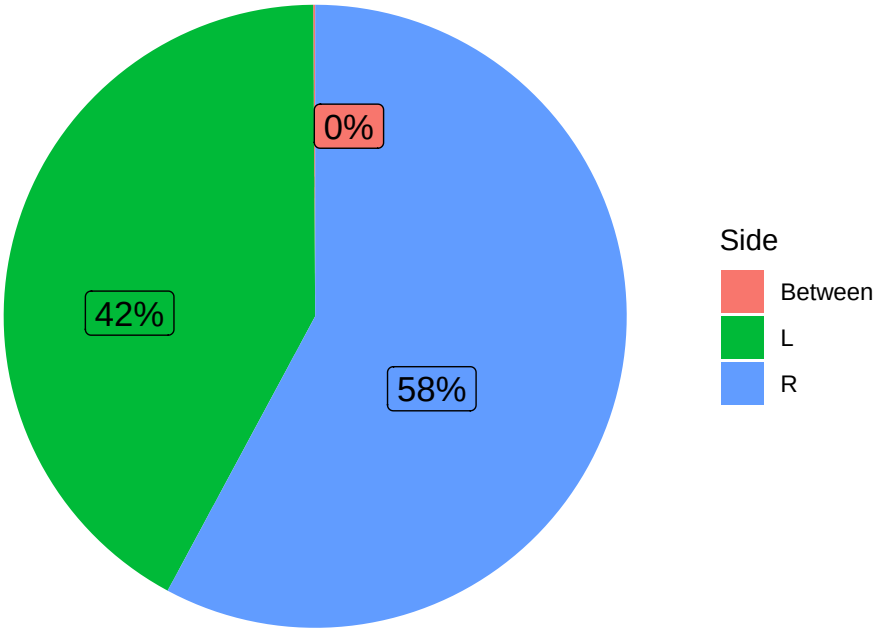
Percentage of ASP residue sides in GK regions for Secreted



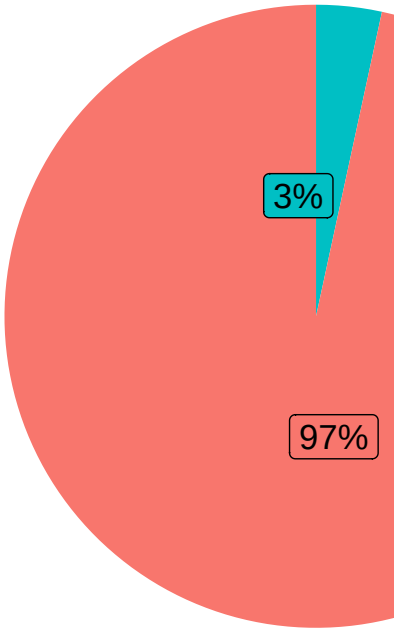
Percentage of GLU residue sid for Secreted



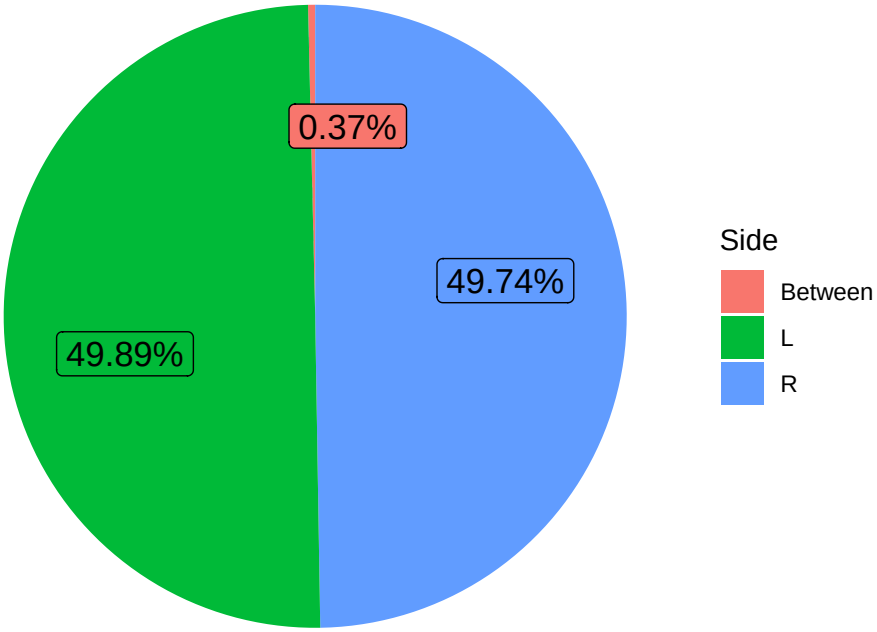
Percentage of SER residue sides in GK regions for Secreted



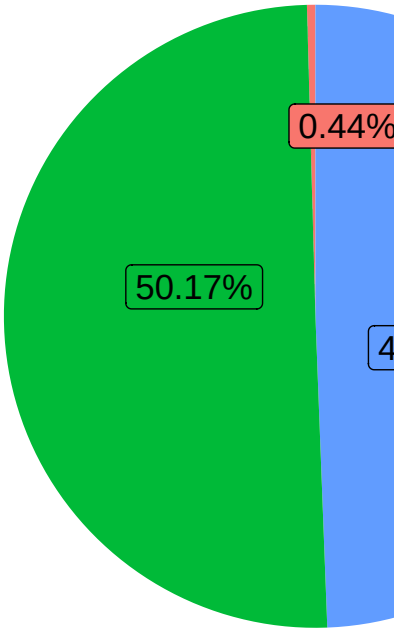
Percentage of PRO residue sid for Secreted



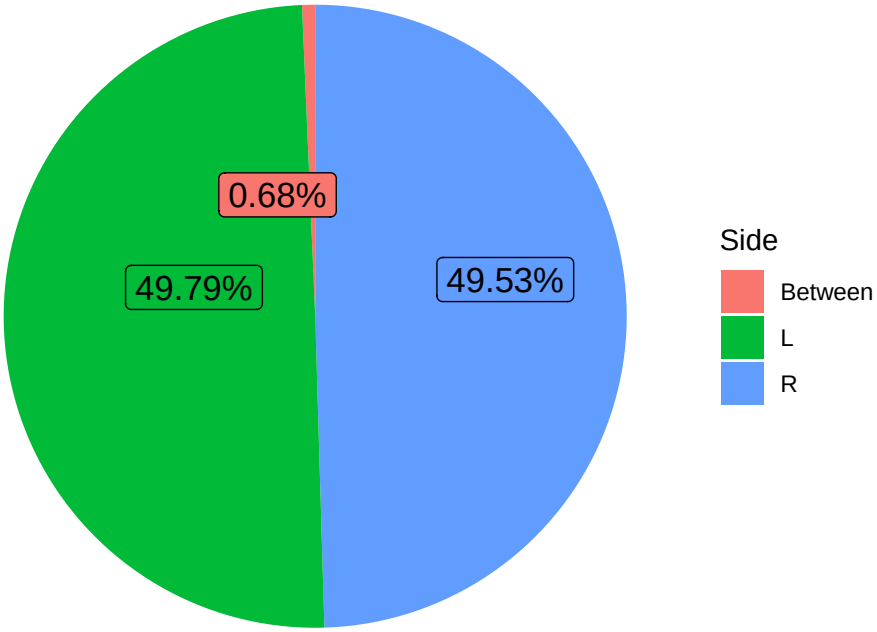
Percentage of LYS residue sides in GK + FL regions for Secreted



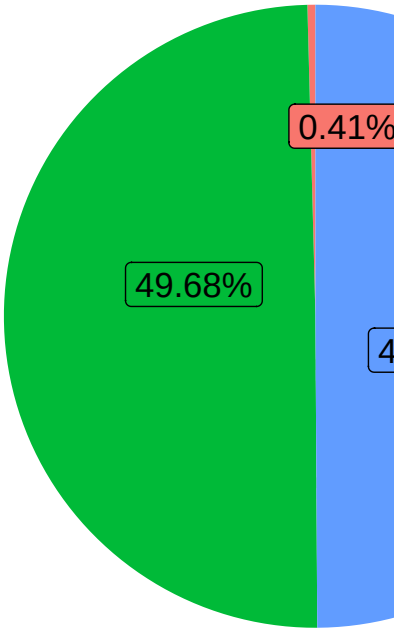
Percentage of ARG residue sid for Secreted



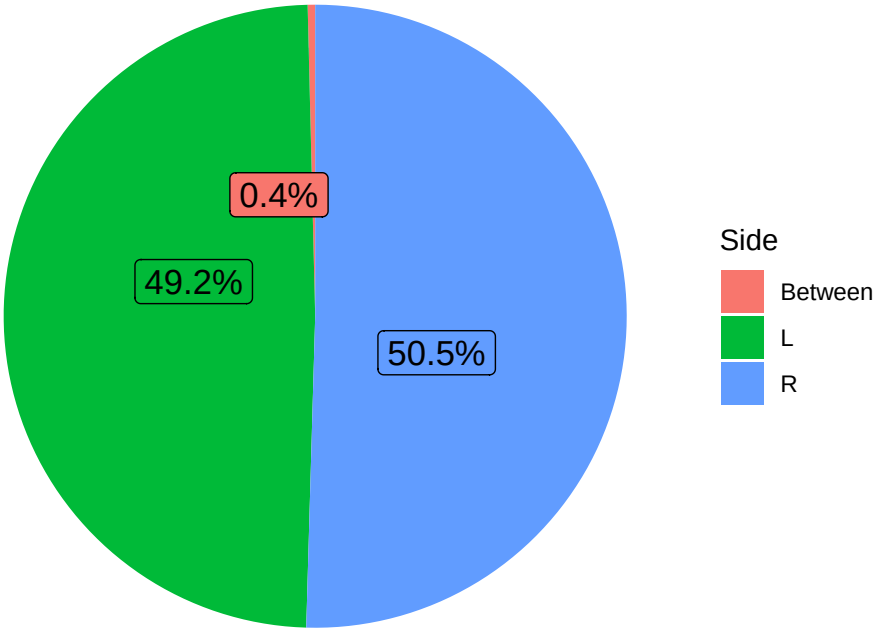
Percentage of ASP residue sides in GK + FL regions for Secreted



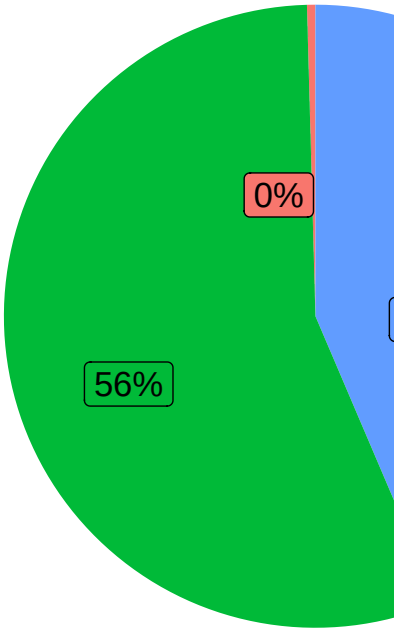
Percentage of GLU residue sid for Secreted

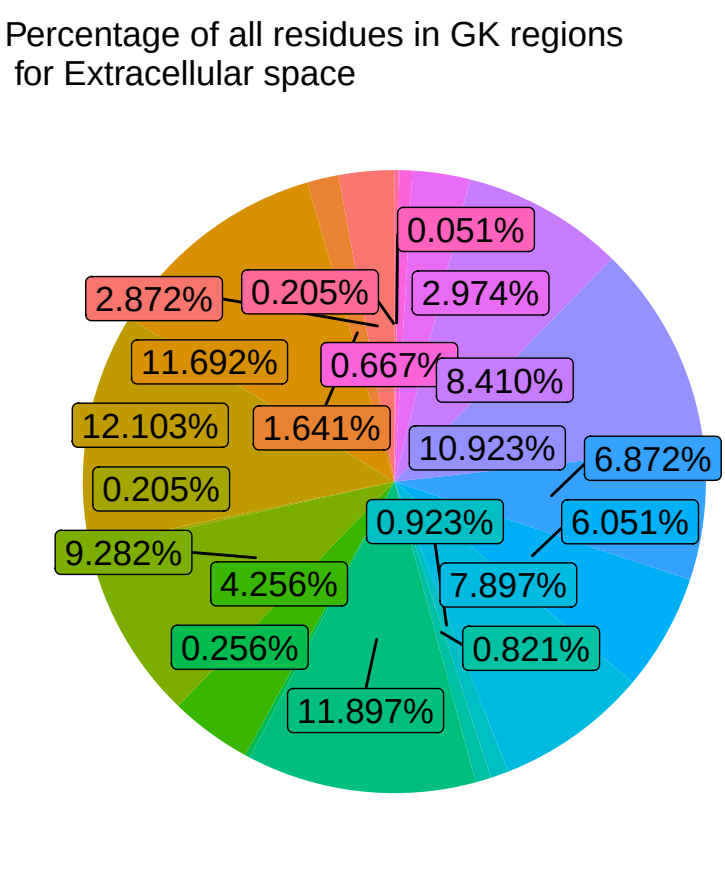


Percentage of SER residue sides in GK + FL regions for Secreted

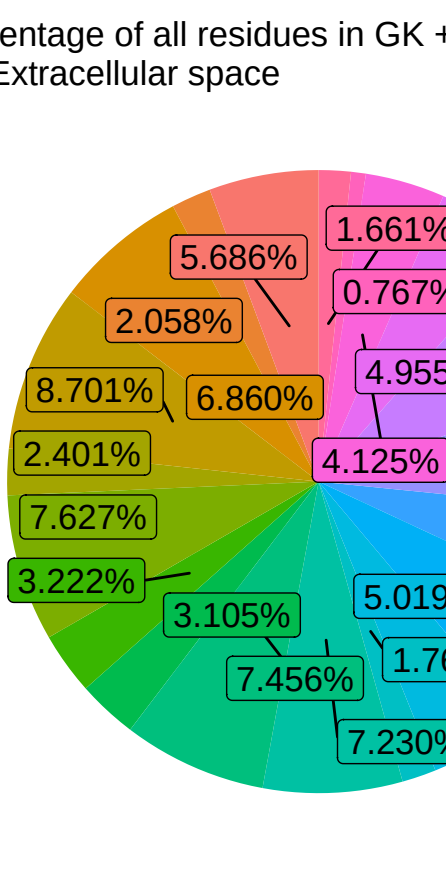


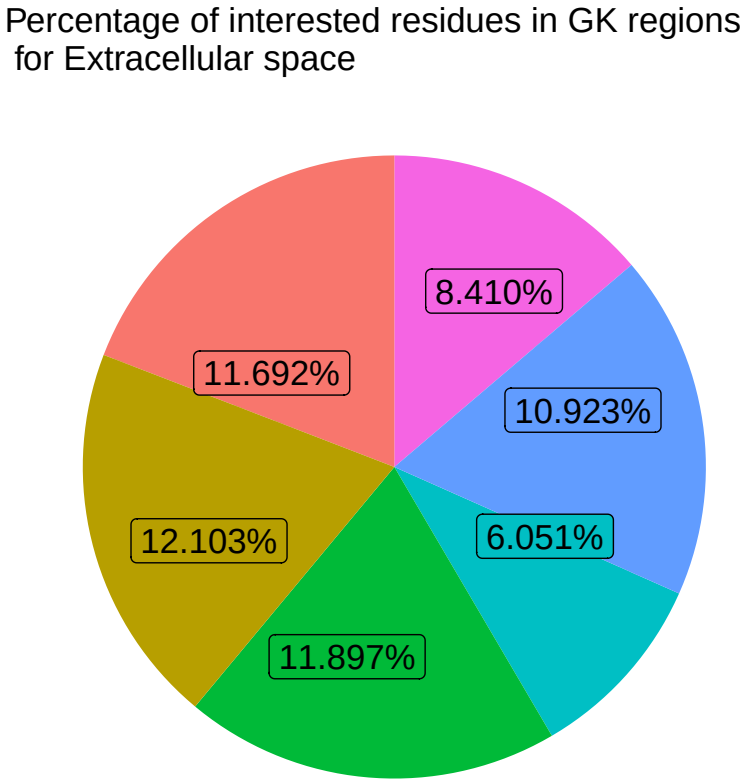
Percentage of PRO residue sid for Secreted



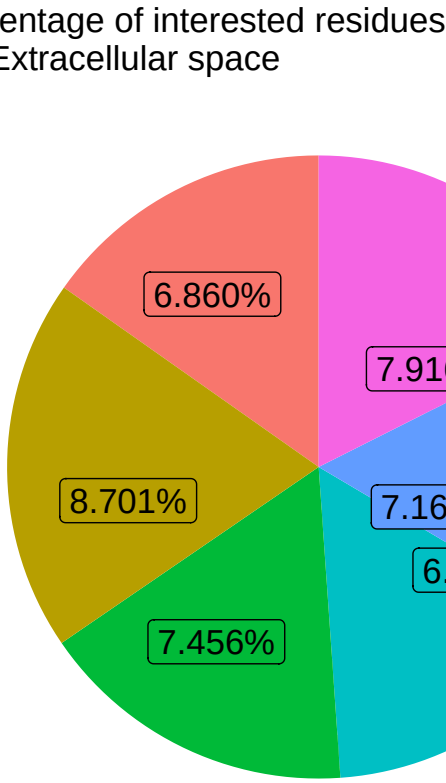

Percentage of all residues in GK regions
for Extracellular space
0.051%
2.872%
0.205%
2.974%
11.692%
0.667%
8.410%
12.103%
1.641%
10.923%
6.872%
0.205%
0.923%
6.051%
9.282%
4.256%
7.897%
0.256%
0.821%
11.897%

Residue
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V

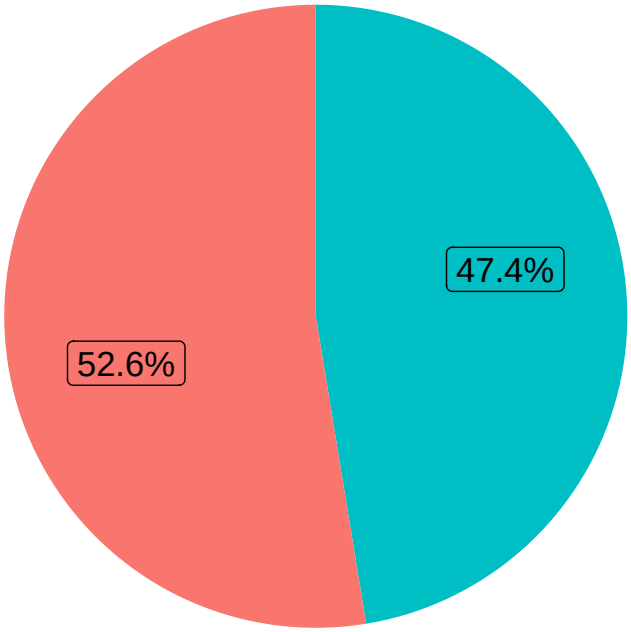

Percentage of all residues in GK +
for Extracellular space
1.661%
5.686%
0.767%
2.058%
4.955
8.701%
6.860%
2.401%
4.125%
7.627%
3.222%
5.019
3.105%
1.76
7.456%
7.230%


Percentage of interested residues in GK regions
for Extracellular space
8.410%
11.692%
10.923%
12.103%
6.051%
11.897%

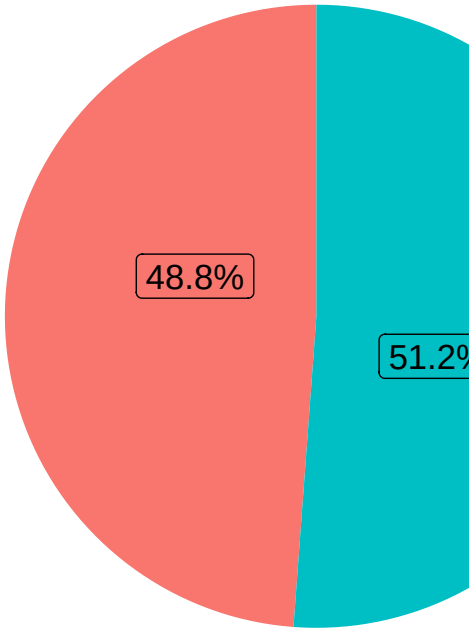
Residue
D
E
K
P
R
S


Percentage of interested residues
for Extracellular space
6.860%
7.91
8.701%
7.16
6
7.456%

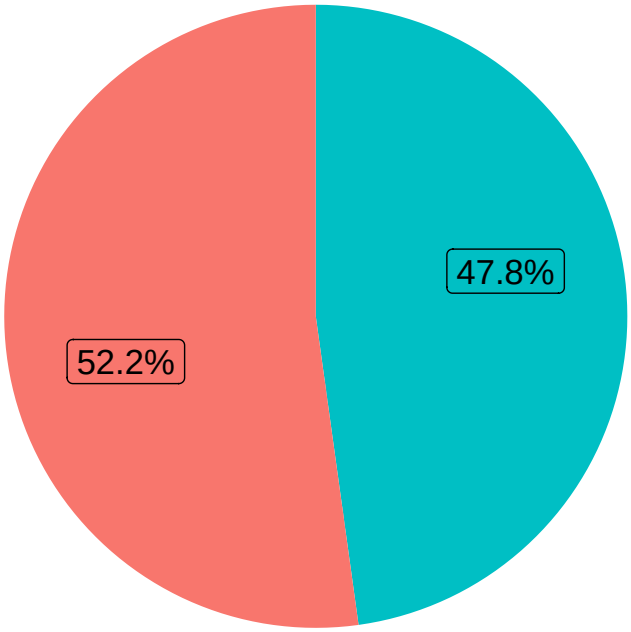
Percentage of LYS residue sides in GK regions for Extracellular space



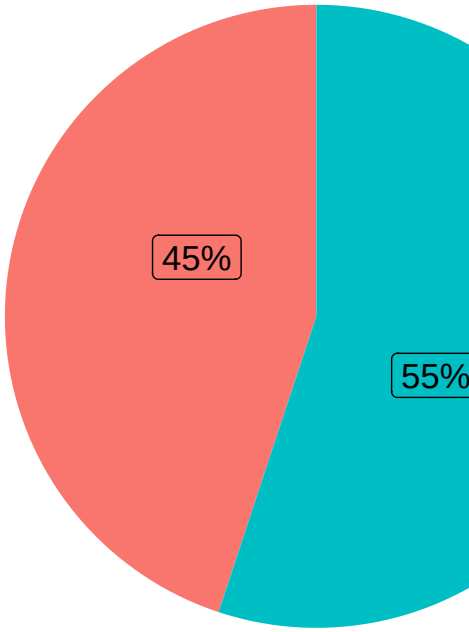
Percentage of ARG residue sides in GK regions for Extracellular space



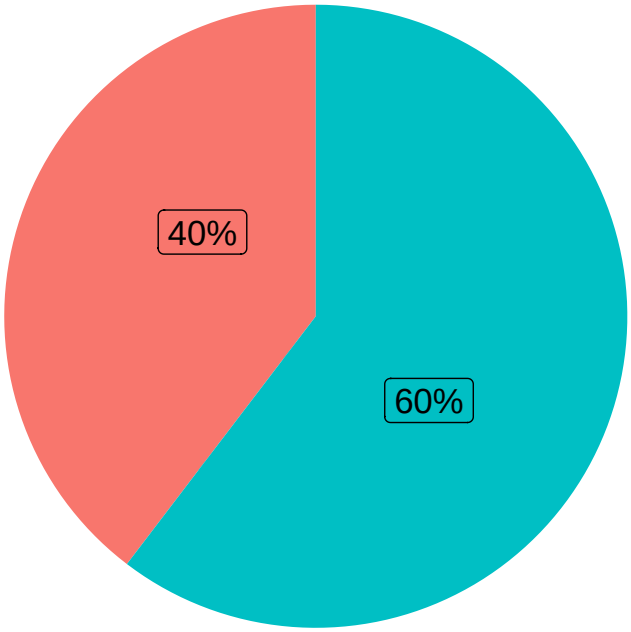
Percentage of ASP residue sides in GK regions for Extracellular space



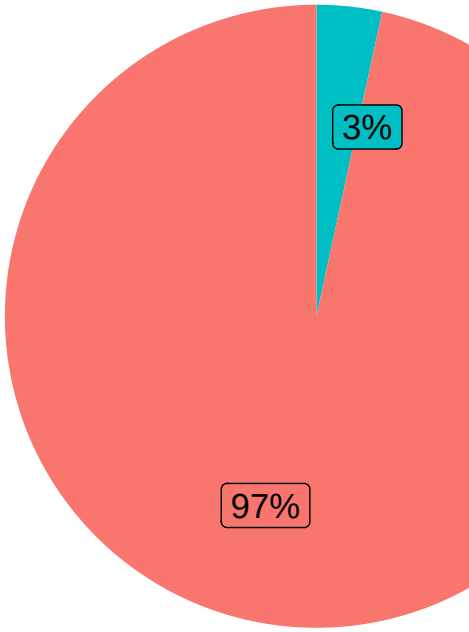
Percentage of GLU residue sides in GK regions for Extracellular space



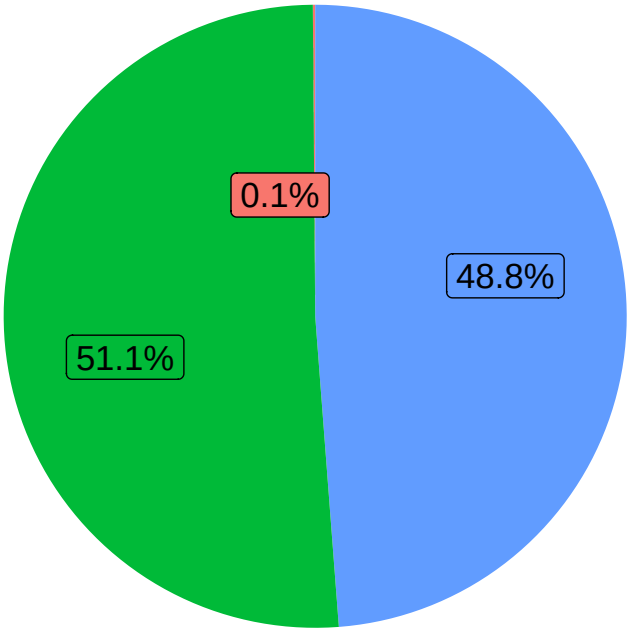
Percentage of SER residue sides in GK regions for Extracellular space



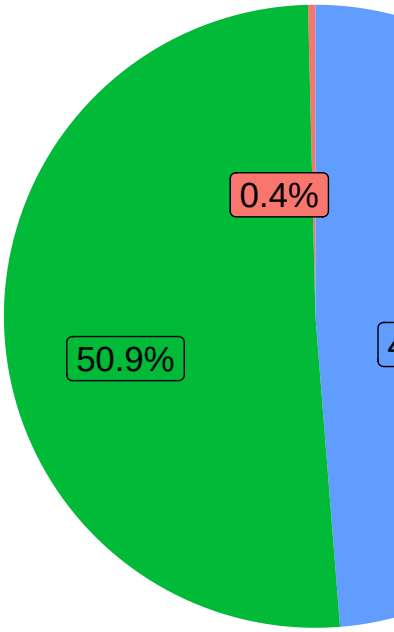
Percentage of PRO residue sides in for Extracellular space



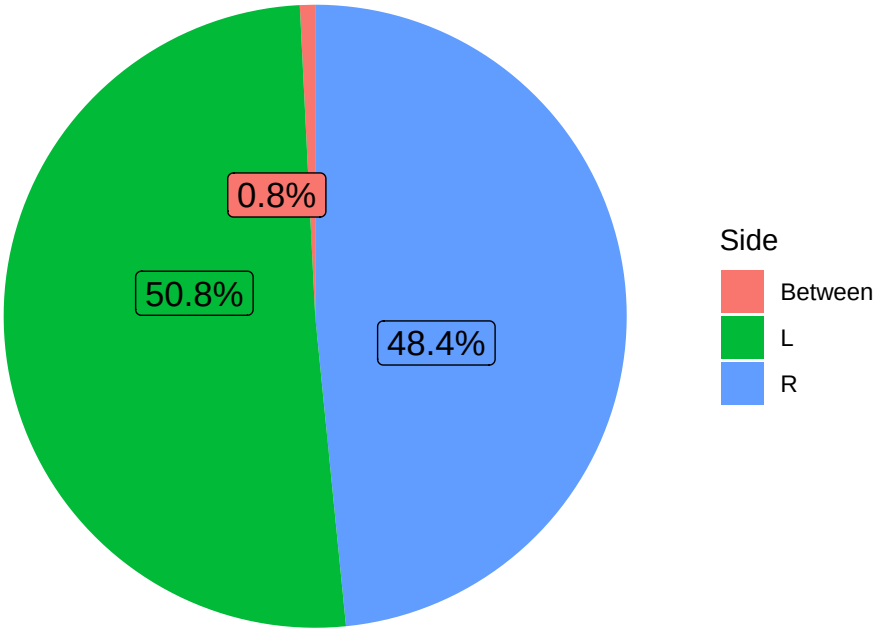
Percentage of LYS residue sides in GK + FL regions for Extracellular space



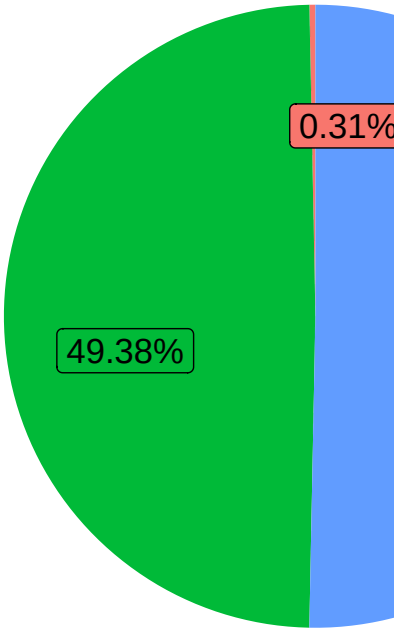
Percentage of ARG residue sid for Extracellular space



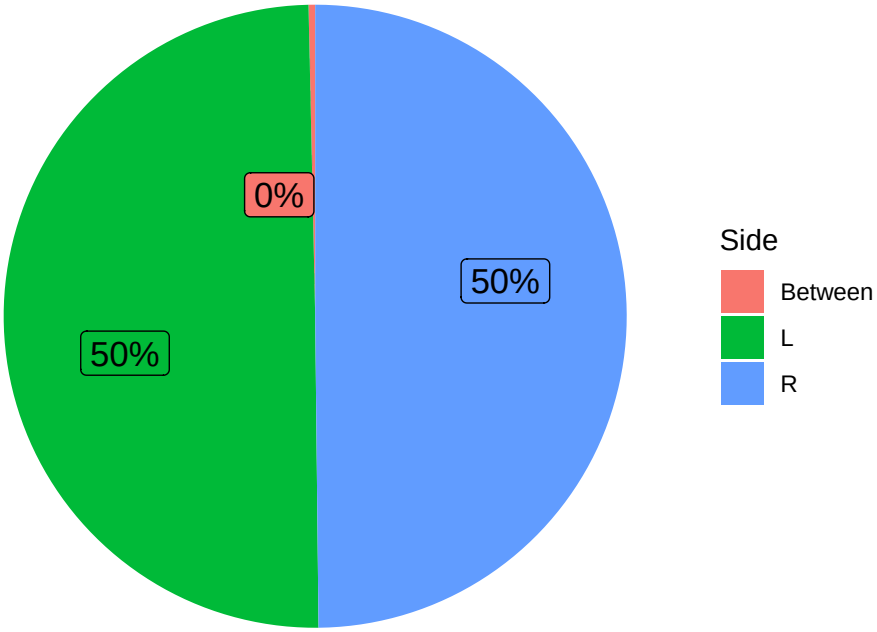
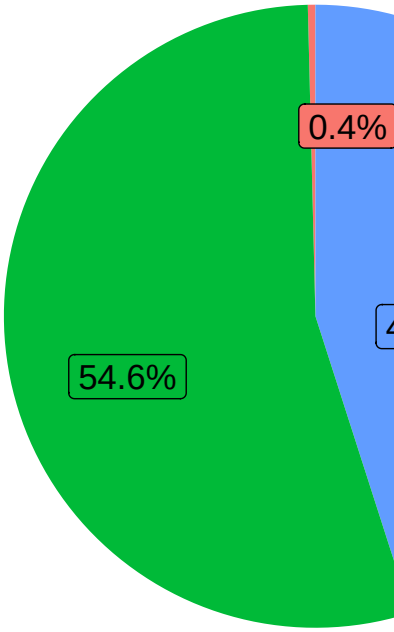
Percentage of ASP residue sides in GK + FL regions for Extracellular space



Percentage of GLU residue sid for Extracellular space



Percentage of SER residue sides in GK + FL regions for Extracellular space



Percentage of PRO residue sid for Extracellular space