Package 'ProteinGymR'

December 3, 2024

Title Programmatic access to ProteinGym datasets in R/Bioconductor

Version 1.1.0

Description The ProteinGymR package provides analysis-ready data resources from ProteinGym, generated by Notin et al., 2023. ProteinGym comprises a collection of benchmarks for evaluating the performance of models predicting the effect of point mutations. This package provides access to 1. Deep mutational scanning (DMS) scores from 217 assays measuring the impact of all possible amino acid substitutions across 186 proteins, 2. AlphaMissense pathogenicity scores for ~1.6 M substitutions in the ProteinGym DMS data, and 3. five performance metrics for 62 variant prediction models in a zero-shot setting.

License Artistic-2.0

URL https://github.com/ccb-hms/ProteinGymR

BugReports https://github.com/ccb-hms/ProteinGymR/issues

Depends R (>= 4.4.0)

Imports ExperimentHub, dplyr, forcats, ggdist, gghalves, ggplot2, purrr, queryup, spdl, tidyr, tidyselect

Suggests ComplexHeatmap, AnnotationHub, tibble, stringr, BiocStyle, knitr, testthat (>= 3.0.0)

VignetteBuilder knitr

Encoding UTF-8

biocViews ExperimentData, ExperimentHub, PackageTypeData, Homo_sapiens_Data, ReproducibleResearch, CellCulture, SequencingData, Proteome

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Config/testthat/edition 3

git_url https://git.bioconductor.org/packages/ProteinGymR

git_branch devel

git_last_commit 35282b0

git_last_commit_date 2024-10-29

Repository Bioconductor 3.21

Date/Publication 2024-12-03

Author Tram Nguyen [aut, cre] (ORCID: <https://orcid.org/0000-0003-4809-6227>),
Pascal Notin [aut],
Aaron Kollasch [aut],
Debora Marks [aut],
Ludwig Geistlinger [aut]

Maintainer Tram Nguyen <Tram_Nguyen@hms.harvard.edu>

Contents

am_scores 2
available_models 3
dms_corr_plot 4
dms_substitutions 6
zeroshot_DMS_metrics 7

Index **9**

`am_scores` *AlphaMissense scores for ProteinGym variants*

Description

AlphaMissense scores for ProteinGym variants

Usage

```
am_scores(metadata = FALSE)
```

Arguments

`metadata`	Logical, whether only experiment metadata should be returned. Default behavior is to return processed data with metadata included.

Details

`am_scores()` loads in the AlphaMissense pathogencity scores for substitutions matching those in the ProteinGym DMS assays. The table is taken from the AlphaMissense Supplementary Data by Cheng et al. 2023. See reference for details.

The columns contain:

`DMS_id`**:** Character, ProteinGym assay identifier.

`Uniprot_ID`**:** Character, UniProt accession identifier.

`variant_id`: Character, variant identifier string matching ProteinGym. Protein position in the middle, and the reference and mutant amino acid residues to the left and right of the position, respectively.

`AlphaMissense`: Numeric, AlphaMissense pathogenicity score.

Value

Returns a `data.frame()`.

References

Cheng et al. (2023) Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 391, eadg7492. DOI:10.1126/science.adg7492.

Examples

```
data <- am_scores()
data_meta <- am_scores(metadata = TRUE)
```

`available_models` *Benchmark Variant Effect Prediction Models*

Description

`benchmark_models()` plots one of the five model performance metrics ("AUC", "MCC", "NDCG", "Spearman", "Top_recall") for up to 5 user-specified variant effect prediction tools listed in `available_models()`. See reference for more details about the metrics and models.

Usage

```
available_models()

benchmark_models(
  metric = c("AUC", "MCC", "NDCG", "Spearman", "Top_recall"),
  models = available_models()
)
```

Arguments

`metric`	`character()` a model performance metric to benchmark ("AUC", "MCC", "NDCG", "Spearman", "Top_recall").
`models`	`character()` a character vector of up to five variant effect prediction models to compare. Valid models can be seen with `available_models()`.

Value

`benchmark_models()` returns a `ggplot` object visualizing a chosen model performance metric across several variant effect prediction models, ordered by highest to lowest mean performance score.

References

Notin, P., Kollasch, A., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y., & Marks, D. (2023). ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 64331-64379). Curran Associates, Inc.

Examples

```
# Currently support models
available_models()

benchmark_models(metric = "Spearman", models = c("Site_Independent",
"DeepSequence_single", "ESM2_15B", "GEMME", "CARP_640M"))

benchmark_models(models = "GEMME")
```

---

dms_corr_plot *Integrate ProteinGym DMS and AlphaMissense Pathogenicity Scores*

Description

`dms_corr_plot()` runs a Spearman correlation between ProteinGym deep mutational scanning (DMS) assay scores and AlphaMissense predicted pathogenicity scores. Returns a ggplot object for visualization.

Usage

```
dms_corr_plot(uniprotId, alphamissense_table, dms_table)
```

Arguments

`uniprotId`	`character()` a valid UniProt accession identifier.
`alphamissense_table`	a table containing AlphaMissense predictions for variants matching ProteinGym substitution mutants. The default is the supplemental table from the AlphaMissense paper. Alternatively, a user-defined `tibble::tbl_df` or `data.frame` can be supplied.
`dms_table`	a table containing deep mutational scanning (DMS) assay scores for mutations. The default table loads substitutions from ProteinGym. Alternatively, a user-defined `tibble::tbl_df` or `data.frame` can be supplied.

Details

For `dms_corr_plot()`, `alphamissense_table` columns must include:

- `UniProt_id`: UniProt accession identifier.
- `mutant`: Mutant identifier string matching the dms_table format. Protein position in the middle, and the reference and mutant amino acid residues to the left and right of the position, respectively.
- `AlphaMissense`: AlphaMissense pathogenicity score.

`dms_table` columns must include:

- `UniProt_id`: UniProt accession identifier.
- `mutant`: Mutant identifier string matching AlphaMissense variants. Specifically, the set of substitutions to apply on the reference sequence to obtain the mutated sequence (e.g., A1P:D2N implies the amino acid 'A' at position 1 should be replaced by 'P', and 'D' at position 2 should be replaced by 'N').
- `DMS_score`: Experimental measurement in the DMS assay. Higher values indicate higher fitness of the mutated protein.

Value

`dms_corr_plot()` returns a `ggplot` object visualizing the Spearman correlation between experimental DMS scores and AlphaMissense predicted scores and prints the r and p-value of the analysis to console. Generally, a stronger negative correlation corresponds to a tighter relationship between the two measures.

References

Cheng et al., Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 381, eadg7492. DOI:10.1126/science.adg7492.

Notin, P., Kollasch, A., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y., & Marks, D. (2023). ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 64331-64379). Curran Associates, Inc.

Examples

```
dms_corr_plot(uniprotId = "Q9NV35")
```

dms_substitutions *ProteinGym Deep Mutational Scanning (DMS) Scores for Substitutions*

Description

ProteinGym Deep Mutational Scanning (DMS) Scores for Substitutions

Usage

```
dms_substitutions(metadata = FALSE)
```

Arguments

metadata	Logical, whether only experiment metadata should be returned. Default behavior is to return processed data with metadata included.

Details

`dms_substitutions()` loads in ProteinGym deep mutational scanning assays (DMS) scores for substitutions in 217 studies. The data is provided by Notin et. al 2023. See reference for details.

Each assay includes 6 columns:

`UniProt_id`: Character, UniProt accession identifier.

`DMS_id`: Character, ProteinGym assay identifier.

`mutant`: Character, set of substitutions to apply on the reference sequence to obtain the mutated sequence (e.g., A1P:D2N implies the amino acid 'A' at position 1 should be replaced by 'P', and 'D' at position 2 should be replaced by 'N').

`mutated_sequence`: Character, full amino acid sequence for the mutated protein.

`DMS_score`: Numeric, experimental measurement in the DMS assay. Higher values indicate higher fitness of the mutated protein.

`DMS_score_bin`: Factor, indicates whether the DMS_score is above the fitness cutoff (1 is fit, 0 is not fit).

Value

Returns a `list()` object of 217 individual assays.

References

Notin, P., Kollasch, A., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y., & Marks, D. (2023). ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in Neural Information Processing Systems (Vol. 36, pp. 64331-64379). Curran Associates, Inc.

Examples

```
data <- dms_substitutions()
data_meta <- dms_substitutions(metadata = TRUE)
```

`zeroshot_DMS_metrics` *Model performance metrics for DMS substitutions in the zero-shot setting*

Description

Model performance metrics for DMS substitutions in the zero-shot setting

Usage

```
zeroshot_DMS_metrics(metadata = FALSE)
```

Arguments

`metadata`	Logical, whether only experiment metadata should be returned. Default behavior is to return processed data with metadata included.

Details

`zeroshot_DMS_metrics()` loads in the five model performance metrics for ("AUC", "MCC", "NDCG", "Spearman", "Top_recall") calculated on the DMS substitutions in the zero-shot setting.

Each data.frame columns contain:

- "DMS_ID": Showing the assay name for the 217 DMS studies.
- Columns 2:63: Corresponding to the average performance score of each of the 61 models tested.
- "Number_of_Mutants": Number of protein mutants evaluated.
- "Selection_Type": Protein function grouping.
- "UniProt_ID": UniProt protein entry name identifier
- "MSA_Neff_L_category": Multiple sequence alignment category.
- "Taxon": taxon group.

Value

Returns a `list()` object with five `data.frame()` corresponding to a model metric table.

References

Notin, P., Kollasch, A., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y., & Marks, D. (2023). ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in Neural Information Processing Systems (Vol. 36, pp. 64331-64379). Curran Associates, Inc.

Examples

```
data <- zeroshot_DMS_metrics()
data_meta <- zeroshot_DMS_metrics(metadata = TRUE)
```

Index

am_scores, 2
available_models, 3

benchmark_models *(*available_models*)*, 3

data.frame, *4*
data.frame(), *3*, *7*
dms_corr_plot, 4
dms_substitutions, 6

list(), *6*, *7*

tibble::tbl_df, *4*

zeroshot_DMS_metrics, 7