Package 'JohnsonKinaseData'

July 22, 2025

Title Kinase PWMs based on data published by Johnson et al. 2023 and Yaron-Barir et al. 2024

Version 1.4.0

Date 2023-09-24

Description The packages provides position specific weight matrices (PWMs) for 303 human serine/threonine and 93 tyrosine kinases originally published in Johnson et al. 2023 (doi:10.1038/s41586-022-05575-3) and Yaron-Barir et al. 2024 (doi:10.1038/s41586-024-07407-y). The package includes basic functionality to score user provided phosphosites. It also includes pre-computed PWM scores (``background scores'') for a large collection of curated human phosphosites which can be used to rank PWM scores relative to the background scores (``percentile rank'').

License MIT + file LICENSE

URL https://github.com/fgeier/JohnsonKinaseData/

BugReports https://support.bioconductor.org/t/JohnsonKinaseData

Imports ExperimentHub, BiocParallel, checkmate, dplyr, stats, stringr, tidyr, purrr, utils

Suggests knitr, BiocStyle, ExperimentHubData, testthat (>= 3.0.0), rmarkdown

biocViews ExperimentHub, Homo_sapiens_Data, Proteome

VignetteBuilder knitr

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Config/testthat/edition 3

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

git_branch RELEASE_3_21

git_last_commit 1073ddd

git_last_commit_date 2025-04-15

Repository Bioconductor 3.21

Date/Publication 2025-07-22

Author Florian Geier [aut, cre] (ORCID:
<https://orcid.org/0000-0002-9076-9264>)

Maintainer Florian Geier <florian.geier@unibas.ch>

Contents

`JohnsonKinaseData-package`

JohnsonKinaseData: Kinase PWMs based on data published by Johnson et al. 2023 and Yaron-Barir et al. 2024

Description

The packages provides position specific weight matrices (PWMs) for 303 human serine/threonine and 93 tyrosine kinases originally published in Johnson et al. 2023 (doi:10.1038/s41586-022-05575-3) and Yaron-Barir et al. 2024 (doi:10.1038/s41586-024-07407-y). The package includes basic functionality to score user provided phosphosites. It also includes pre-computed PWM scores ("background scores") for a large collection of curated human phosphosites which can be used to rank PWM scores relative to the background scores ("percentile rank").

Author(s)

Maintainer: Florian Geier <florian.geier@unibas.ch> (ORCID)

See Also

Useful links:

- https://github.com/fgeier/JohnsonKinaseData/
- Report bugs at https://support.bioconductor.org/t/JohnsonKinaseData

`getBackgroundScores` *Get precomputed PWM scores for two sets of curated human phosphosites*

Description

Two sets of background scores are provided:

Usage

```
getBackgroundScores(phosphoAcceptor = c("Ser/Thr", "Tyr"))
```

Arguments

`phosphoAcceptor`
: Return background scores for either Ser/Thr or Tyr phosphosites

Details

1. Ser/Thr phosphosites published in Johnson et al. 2023 2. Tyr phosphosites published in Yaron-Barir et al. 2024

The background scores are derived from matching the corresponding PWMs to each of the sets. The resulting data frames contain the log2-odds score per phosphosite and PWM.

Value

A data frame with log2-odds scores per phosphosite (rows) and PWMs (columns)

References

Johnson, J.L., Yaron, T.M., Huntsman, E.M. et al. An atlas of substrate specificities for the human serine/threonine kinome. Nature 613, 759–766 (2023). https://doi.org/10.1038/s41586-022-05575-3

Yaron-Barir, T.M., Joughin, B.A., Huntsman, E.M. et al. The intrinsic substrate specificity of the human tyrosine kinome. Nature 629, 1174–1181 (2024). https://doi.org/10.1038/s41586-024-07407-y

Examples

```
bg <- getBackgroundScores(phosphoAcceptor='Tyr')
```

getKinaseAnnotation *Get annnotation data for all kinase PWMs*

Description

The annotation data records for each kinase PWM, the PWM matrix name, gene symbol and description, Uniprot ID, Entrez ID, acceptor specificity, kinase sub-type, as well as the kinase family.

Usage

```
getKinaseAnnotation()
```

Details

Kinase PWMs are either serine/threonine or tyrosine specific in their central phospho-acceptor. For dual-specific kinases, the tyrosine-specific PWM is indicated by the '_TYR' suffix in the PWM name.

Tyrosine kinases are further distinguished by sub-type and include receptor tyrosine kinases (RTK), non-receptor tyrosine kinases (nRTK) and non-canonical tyrosine kinases (ncTK) with dual-specificity.

Value

A data frame with columns MatrixName, GeneName, UniprotID, EntrezID, Description, AcceptorSpecificity, KinaseSubType and KinaseFamily

Examples

```
anno <- getKinaseAnnotation()
```

getKinasePWM *Get a list of position specific weight matrices (PWMs) for human kinases*

Description

The function returns a named list of 396 kinase PWMs. Among these are 303 serine/threonine kinases, 78 canonical tyrosine kinases and 15 non-canonical tyrosine kinases i.e. dual-specific kinases, indicated by the '_TYR' suffix.

Usage

```
getKinasePWM(includeSTfavorability = TRUE, matchAcceptorSpecificity = FALSE)
```

Arguments

includeSTfavorability
: Include serine vs. threonine favorability for the central phospho-acceptor?

matchAcceptorSpecificity
: Only score sites with a matching central phospho-acceptor?

Details

Each PWM stores the log2-odds score per amino acid (23 rows) and position (10 or 11 columns) in matrix format. Beside the 20 standard amino acids also phosphorylated serine, threonine and tyrosine residues are included.

The central phospho-acceptor position of each PWM is at position `0` (column 6). For serine/threonine specific kinases this position quantifies the favorability of serine over threonine, but can be omitted when setting 'includeSTfavorability=FALSE'.

The specificity of a kinase PWM is controlled by parameter 'matchAcceptorSpecificity'. If set to 'TRUE', sites without a matching acceptor are scored with '-Inf'.

Value

A named list of numeric matrices (PWMs).

References

Johnson, J.L., Yaron, T.M., Huntsman, E.M. et al. An atlas of substrate specificities for the human serine/threonine kinome. Nature 613, 759–766 (2023). https://doi.org/10.1038/s41586-022-05575-3

Yaron-Barir, T.M., Joughin, B.A., Huntsman, E.M. et al. The intrinsic substrate specificity of the human tyrosine kinome. Nature 629, 1174–1181 (2024). https://doi.org/10.1038/s41586-024-07407-y

Examples

```
pwms <- getKinasePWM()
```

---

getScoreMaps *Map log2-odds score to percentile rank*

Description

For each kinase PWM, get a function that maps its log2-odds score to the percentile rank in the background score distribution. The percentile rank of a given score is the percentage of scores in corresponding background score distribution that are less than or equal to that score. The background score distribution per PWM is derived from matching each PWM to either the 85'603 unique phosphosites published in Johnson et al. 2023 (serine/threonine PWMs) or the 6659 unique phosphosites published in Yaron-Barir et al. 2024 (tyrosine PWMs).

Usage

```
getScoreMaps()
```

Details

Note: since the background sites don't contain non-central phosphorylated residues (phospho-priming), the percentile rank of an input site which includes phospho-priming will be capped to 100, if its PWM score exceeds the largest observed background score for that PWM.

Internally, stats::approxfun is used to linearly interpolate between the PWM score and its 0.1% - quantile in the distribution over background scores. This approximation allows for a lower memory footprint compared with the full set of background scores.

Value

A named list of functions, one for each kinase PWM. Each function is taking a vector of log2-odds scores as input and returns the corresponding percentile ranks.

Examples

```
maps <- getScoreMaps()
```

---

processPhosphopeptides *processPhosphopetides*

Description

Process phospho-peptides to a common format used for PWM matching

Usage

```
processPhosphopeptides(
  sites,
  onlyCentralAcceptor = TRUE,
  allowPhosphoPriming = TRUE,
  upstream = 5L,
  downstream = 5L
)
```

Arguments

`sites`	Character vector with phospho-peptides
`onlyCentralAcceptor`	Process only the central phospho-acceptor residue?
`allowPhosphoPriming`	Allow phospho-acceptors at non-central positions? These should be indicated by the lower case letters s, t or y.

upstream	Number of bases upstream of central phospho-acceptor in the processed output
downstream	Number of bases downstream of central phospho-acceptor in the processed output

Details

Phosphorylated residues are recognized either by lower case letters (s, t or y) or the phosphorylated residue is followed by an asterisk (S*, T* or Y*).

If a peptide reports several phosphorylated residues, parameter `onlyCentralAcceptor` allows for two processing options: (1) By default, only the central phospho-acceptor of each phospho-peptide is considered. Here central is defined as the left-closest position to `floor(nchar(site)/2)+1`. (2) All phospho-acceptors are considered as central in which case the phospho-peptide is replicated and aligned for each of its phosphorylated residues. In this case the output sites are not in parallel to the input peptides.

In both cases, non-central phospho-acceptors are indicated by lower case letters (s, t, or y). These residues enable phospho-priming of the site. If phospho-priming is disabled (parameter `allowPhosphoPriming`) these residues are converted to upper case letters.

If a site does not follow the phosphorylation patterns described above, the central residue defined by position `floor(nchar(site)/2)+1` is considered the default phospho-acceptor site.

The input sites are truncated and/or padded such that the processed sites are of width `upstream+downstream+1`. By default the central phospho-acceptor is surrounded by 5 upstream and 5 downstream residues.

A warning is raised if the central phospho-acceptor is not serine, threonine or tyrosine.

Value

A tibble with columns: `sites`, `processed`, `acceptor`

Examples

```
procSites <- processPhosphopeptides(c("AGLLS*DEDC", "RtEKGS*N", "ETGKDN"))
```

---

scorePhosphosites *Match kinase PWMs to processed phosphosites*

Description

`scorePhosphosites` takes a list of kinase PWMs and a vector of processed phosphosites as input and returns a matrix of match scores per PWM and site.

Usage

```
scorePhosphosites(
  pwms,
  sites,
  scoreType = c("lod", "percentile"),
  BPPARAM = BiocParallel::SerialParam()
)
```

Arguments

pwms	List with kinase PWMs as returned by getKinasePWM.
sites	A character vector with phosphosites. Check processPhosphopeptides for the correct phosphosite format.
scoreType	Log2-odds score or percentile rank.
BPPARAM	A BiocParallelParam object specifying how parallelization should be performed.

Details

The match score is either the log2-odds score (`lod`) or the percentile rank (`percentile`) in the background score distribution.

Value

A numeric matrix of size `length(sites)` times `length(pwms)`.

See Also

getKinasePWM for getting a list of kinase PWMs, processPhosphopeptides for the correct phosphosite format, and getScoreMaps for mapping PWM scores to percentile ranks

Examples

```
score <- scorePhosphosites(getKinasePWM(), c("TGRRHTLAEV", "LISAVSPEIR"))
```

Index