Package 'QSARdata'

January 20, 2025

Type Package

Title Quantitative Structure Activity Relationship (QSAR) Data Sets

Version 1.3

Date 2013-07-16

Author Max Kuhn

Maintainer Max Kuhn <mxkuhn@gmail.com>

Description Molecular descriptors and outcomes for several public domain data sets

License GPL

LazyLoad yes

Depends R (>= 2.10)

URL http://qsardata.r-forge.r-project.org/

NeedsCompilation no

Repository CRAN

Date/Publication 2013-07-16 18:30:26

Contents

AquaticTox *Fathead Minnow Acute Aquatic Toxicity*

Description

These data were compiled and described by He and Jurs (2005). The data set consists of 322 compounds that were experimentally assessed for toxicity. The outcome is the negative log of activity (but is labled as "activity"). The structures and outcomes were obtained from http://www.qsarworld.com/index.php.

The package contains none sets of molecular descriptors: atom pair distances, Daylight fingerprints (http://www.daylight.com/dayhtml/doc/theory/theory.finger.html), Dragon descriptors (http://www.talete.mi.it/products/dragon_plus.htm), MOE2D, MOE2D fingerprints, MOE3D, PipelinePilot fingerprints (http://accelrys.com/products/pipeline-pilot/) and QuickProp descriptors (http://www.schrodinger.com/products/14/17/).

For fingerprints, the 500 most variable bits were selected whenever possible.

Usage

```
data(AquaticTox)
```

Format

The data consist of several data frames. The first column of the descriptor data frames is called "Molecule" representing the compounds.

AquaticTox_AtomPair Atom pair descriptors

AquaticTox_Daylight_FP Daylight fingerprints (http://www.daylight.com/dayhtml/doc/theory/theory.finger.html)

AquaticTox_Dragon Dragon descriptors (http://www.talete.mi.it/products/dragon_plus.htm)

AquaticTox_Lcalc LCALC descriptors

AquaticTox_moe2D 2 dimensional MOE descriptors

AquaticTox_moe2D_FP 2 dimensional MOE fingerprints

AquaticTox_moe3D 3 dimensional MOE descriptors

AquaticTox_PipelinePilot_FP PipelinePilot fingerprints (http://accelrys.com/products/pipeline-pilot/)

AquaticTox_QuickProp QuickProp descriptors

AquaticTox_Outcome a data frame with columns for the molecule name and the outcome (for merging)

References

He and Jurs. Assessing the reliability of a QSAR model's predictions. Journal of Molecular Graphics and Modelling (2005) vol. 23 (6) pp. 503-523

Examples

```
data(AquaticTox)
head(AquaticTox_Outcome)
```

---

bbb2 *Blood-Brain Barrier Data*

Description

These data were compiled and described by Burns et al. (2004). The data set consists of 80 compounds that were designated as either crossing the blood-brain barrier or not crossing. The structures and outcomes were obtained from http://www.qsarworld.com/index.php.

The package contains none sets of molecular descriptors: atom pair distances, Daylight fingerprints (http://www.daylight.com/dayhtml/doc/theory/theory.finger.html), Dragon descriptors (http://www.talete.mi.it/products/dragon_plus.htm), MOE2D, MOE2D fingerprints, MOE3D, PipelinePilot fingerprints (http://accelrys.com/products/pipeline-pilot/) and QuickProp descriptors.

For fingerprints, the 500 most variable bits were selected whenever possible.

There are compounds with missing data for some descriptors.

The "2" in the name is due to another data set in the **caret** package for blood-brain barrier data (with numeric outcomes). These are a completely different set of compounds and have no connection.

Usage

```
data(bbb2)
```

Format

The data consist of several data frames. The first column of the descriptor data frames is called "Molecule" representing the compounds.

bbb2_AtomPair Atom pair descriptors

bbb2_Daylight_FP Daylight fingerprints (http://www.daylight.com/dayhtml/doc/theory/theory.finger.html)

bbb2_Dragon Dragon descriptors (http://www.talete.mi.it/products/dragon_plus.htm)

bbb2_Lcalc LCALC descriptors

bbb2_moe2D 2 dimensional MOE descriptors

bbb2_moe2D_FP 2 dimensional MOE fingerprints

bbb2_moe3D 3 dimensional MOE descriptors

bbb2_PipelinePilot_FP PipelinePilot fingerprints (http://accelrys.com/products/pipeline-pilot/)

bbb2_QuickProp QuickProp descriptors

bbb2_Class a factor with levels "Crosses" and "DoesNot"

bbb2_Outcome a data frame with columns for the molecule name and the outcome (for merging)

References

Burns et al. A mathematical model for prediction of drug molecule diffusion across the blood-brain barrier. The Canadian Journal of Neurological Sciences (2004) vol. 31 (4) pp. 520-527

Examples

```
data(bbb2)
head(bbb2_Outcome)
```

---

caco *Caco-2 Permeability Data*

Description

These data were compiled and described by Pham-The et al. (2013). The data set consists compounds that were designated as high, medium or low permeability. The structures and outcomes were obtained from the supporting information at http://doi.wiley.com/10.1002/minf.201200166. These data are from Table SI1 and Table SI4. Some compounds failed in descriptor calculations so the total sample size here is 3796 compounds.

The package contains none sets of molecular descriptors: atom pair distances, Dragon descriptors (http://www.talete.mi.it/products/dragon_plus.htm), PipelinePilot fingerprints (http://accelrys.com/products/pipeline-pilot/) and QuickProp descriptors.

For fingerprints, the 1000 most variable bits were selected whenever possible.

Usage

```
data(caco)
```

Format

The data consist of several data frames. The first column of the descriptor data frames is called "Molecule" representing the compounds. The original identifiers were chewed-up during the descriptor calculations and have been give unique but arbitrary values to merge across descriptor sets.

caco_AtomPair Atom pair descriptors

caco_Dragon Dragon descriptors (http://www.talete.mi.it/products/dragon_plus.htm)

caco_PipelinePilot_FP PipelinePilot fingerprints (http://accelrys.com/products/pipeline-pilot/)

caco_QuickProp QuickProp descriptors

caco_Outcome a data frame with columns for the molecule name and the outcome (for merging)

References

Pham-The, H., Gonzalez-Alvarez, I., Bermejo, M., Garrigues, T., Le-Thi-Thu, H., & Cabrera-Perez, M. A. (2013). The Use of Rule-Based and QSPR Approaches in ADME Profiling: A Case Study on Caco-2 Permeability. Molecular Informatics.

Examples

```
data(caco)
head(caco_Outcome)
```

`MeltingPoint` *Melting Point Data*

Description

Karthikeyan et al (2005) presented data where they used chemical descriptors to model the melting point of compounds (i.e. transition from solid to liquid state). They assembled 4401 compounds: 4126 for model training and 275 compounds as a final validation set. They calculated 2D and 3D MOE chemical descriptors.

Usage

```
data(MeltingPoint)
```

Format

The descriptors are contained in a data frame called `MP_Descriptors` and the melting points are in a numeric vector `MP_Outcome`. The original data set indicators are in a factor vector called `MP_Data` with levels "Test" and "Train"

References

Karthikeyan et al. General melting point prediction based on a diverse compound data set and artificial neural networks. Journal of chemical information and modeling (2005) vol. 45 (3) pp. 581-90

Examples

```
data(MeltingPoint)
head(MP_Descriptors)
```

`Mutagen` *Mutagenicity Data*

Description

Kazius et al (2005) investigated using chemical structure to predict mutagenicity (the increase of mutations due to the damage to genetic material). An Ames test was used to evaluate the mutagenicity potential of various chemicals. There were 4,337 compounds included in the data set with a mutagenicity rate of 55.3$%$. Using these compounds, the **DragonX** software (http://www.talete.mi.it/) was used to generate a baseline set of 1,579 predictors, including constitutional, topological and connectivity descriptors, among others. These variables consist of basic numeric variables (such as molecular weight) and counts variables (e.g., number of halogen atoms).

Usage

```
data(Mutagen)
```

Format

The descriptors are contained in a data frame called `Mutagen_Dragon` and the outcomes are in a factor vector `Mutagen_Outcomes` with levels "mutagen" and "nonmutagen"

References

Kazius et al. Derivation and validation of toxicophores for mutagenicity prediction. Journal of medicinal chemistry(Print) (2005) vol. 48 (1) pp. 312-320

Examples

```
data(Mutagen)
head(Mutagen_Dragon)
```

PLD *Drug-Induced Phospholipidosis Data*

Description

These data were compiled and described by Goracci et al. (2013). The data set consists compounds that were designated as phospholipidosis inducers or non-inducers. The structures and outcomes were obtained from the supporting information at http://pubs.acs.org/doi/abs/10.1021/ci400113t. These data are from their curated database although some compounds failed in descriptor calculations so the total sample size here is 324 compounds (instead of 331).

The package contains none sets of molecular descriptors: atom pair distances, Dragon descriptors (http://www.talete.mi.it/products/dragon_plus.htm), LCALC descriptors , PipelinePilot fingerprints (http://accelrys.com/products/pipeline-pilot/), QuickProp descriptors (http://www.chem.ac.ru/Chemistry/Soft/QIKPROP.en.html) and VolSurf descriptors (http://www.moldiscovery.com/soft_volsurf.php).

For fingerprints, the 500 most variable bits were selected whenever possible.

Usage

```
data(PLD)
```

Format

The data consist of several data frames. The first column of the descriptor data frames is called "Molecule" representing the compounds.

PLD_AtomPair Atom pair descriptors

PLD_Dragon Dragon descriptors (http://www.talete.mi.it/products/dragon_plus.htm)

PLD_PipelinePilot_FP PipelinePilot fingerprints (http://accelrys.com/products/pipeline-pilot/)

PLD_QuickProp QuickProp descriptors

PLD_VolSurfPlus VolSurf descriptors

PLD_LCALC LCALC descriptors

PLD_Outcome a data frame with columns for the molecule name and the outcome (for merging)

References

Goracci, L., Ceccarelli, M., Bonelli, D., & Cruciani, G. (2013). Modeling Phospholipidosis Induction: Reliability and Warnings. Journal of Chemical Information and Modeling, 53(6), 1436-1446. doi:10.1021/ci400113t

Examples

```
data(PLD)
head(PLD_Outcome)
```

Index