gene2pathway

March 24, 2012

`classificationModel`

Hierarchical Classification Model

Description

This file contains the hierarchical classification model to predict KEGG pathway branches for genes. The model uses a pruned KEGG hierarchy, where metabolic pathways are not distinguished further, and the KEGG hierarchy for "cellular processes" and "genetic information processing" is pruned at the 2nd level. By default the model uses bagging to improve prediction accuracy. Important: There exists one separate model file for each organism.

Format

List of class "model", where each model has the following entries:

W learned decision hyperplane normal vector

C dictionary of label vectors, which can be predicted individually or which can be used to predict combinations of them

detectors SVM models trained to separate one specific pathway branch from the rest of the hierarchy

used_domains InterPro domains used by the classifier to separate the specific branch from the rest of the hierarchy

alldomains all InterPro domains used to build feature vectors

allpathways hierarchy branches, which can be predicted

treesizes relative size of hierarchy below the corresponding branch

kegg_hierarchy a nested list with information (parent branches, pathway names, pathway IDs, hierarchy level) on all higher hierarchy branches for each pathway

Author(s)

Holger Froehlich

See Also

`classificationModelSignalTrans`

`classificationModelSignalTrans` *Hierarchical Classification Model for Signaling Transduction Pathways and Pathway Components*

Description

This file contains the hierarchical classification model to predict KEGG signaling pathways and pathway components for genes. The model contains only pathway components, to which a specified minimum number of genes could be mapped in the training phase (see `retrain.signaltrans`). Important: There exists one separate model file for each organism.

Format

List of class "model", where each model has the following entries:

W learned decision hyperplane normal vector

C dictionary of label vectors, which can be predicted individually or which can be used to predict combinations of them

detectors SVM models trained to separate one specific pathway branch from the rest of the hierarchy

used_domains InterPro domains used by the classifier to separate the specific branch from the rest of the hierarchy

alldomains all InterPro domains used to build feature vectors

allpathways hierarchy branches, which can be predicted

treesizes" relative size of hierarchy below the corresponding branch

kegg_hierarchy a nested list with information (parent branches, pathway names, pathway IDs, hierarchy level) on all higher hierarchy branches for each pathway

elemIDs a list of KEGG element IDs mapping to each pathway component - may be used to highlight pathway components with `color.pathway.by.elements`.

Author(s)

Holger Froehlich

See Also

`classificationModel`

gene2pathway *Pathway membership prediction*

Description

Predicts a gene's membership to a branch in the KEGG hierarchy via the contained InterPro domains.

Usage

```
gene2pathway(geneIDs=NULL, flyBase=FALSE, gene2Domains=NULL, organism="hsa", use
```

Arguments

geneIDs	a character vector of Entrez gene IDs or FlyBase identifiers (not necessary, if the argument gene2Domains is provided)
flyBase	Are FlyBase identifiers provided? Default: No
gene2Domains	By default associations between genes and InterPro domains are retrieved via biomaRt from Ensembl. Alternatively, the user can provide its own mapping of genes to InterPro domains in form of a list here (see details).
organism	KEGG letter code describing an organism. Please refer to <URL:http://www.genome.jp/kegg-bin/create_kegg_menu> for a complete list of organisms (and their letter codes) supported by KEGG.
useKEGG	Should KEGG information instead of a prediction be used when possible?
mc.cores	number of cores to use for parallelization; requires package 'doMC' to be loaded

Details

A hierarchical classification model based on SVMs and a ranking perceptron is used. This model is usually additionally bagged to improve prediction quality. The model is stored in the package data directory and is recommended to be retrained from time to time.

The KEGG hierarchy is taken from the package keggorthology. By default associations between genes and InterPro domains are retrieved automatically via biomaRt from Ensembl. Please refer to <URL:http://www.ebi.ac.uk/ensembl/> for a list of organisms supported by Ensembl. Alternatively to using Ensembl and biomaRt, the user can provide its own mapping of genes to InterPro domains in form of a list. This especially allows for using organisms, which are supported by KEGG, but not by Ensembl so far. The list has the form genes -> InterPro domains, and each list entry is named by a gene identifier of the corresponding gene. Entrez gene IDs or FlyBase identifiers have to be used.

Value

gene2Path	mapping of gene IDs to corresponding KEGG pathway names
byKEGG	indicates by TRUE/FALSE for each gene whether the mapping information was obtained directly from KEGG or whether it was predicted
scores	confidence scores for the prediction (0, if no prediction was performed): see notes for details
votes	fraction of votes for individual pathway predictions

Note

By default a bagged model prediction is used, i.e. each of the individual sub-models is giving a vote for a specific output. The final output is determined by the majority of the votes for each hierarchy branch separately. The corresponding fraction voting for a specific branch may be interpreted as its probability. In the ideal case all individual branch probabilites should always be close to 1, if the gene maps to that part of the KEGG hierarchy, and close to 0 otherwise. A cumulative measure of confindence is thus the average over all probabilities > 0.5 and one minus the average over all probabilities < 0.5. We combine both measure by taking the average of both and report it as a reliability score.

If the user decides to retrain a model WITHOUT using bagging, then the reliability score is simply the margin between the highest and the second highest ranked solution. This margin should be larger 2 for good confidence.

Author(s)

Holger Froehlich

See Also

`retrain`, `classificationModel`

Examples

```
## Not run:
gene2pathway("FBgn0030327", flyBase=TRUE, organism="dme")

## End(Not run)
```

---

gene2pathway.signaltrans

Pathway membership prediction

Description

Predicts a gene's membership to a KEGG signaling pathway and/or pathway component via the contained InterPro domains.

Usage

```
gene2pathway.signaltrans(geneIDs=NULL, flyBase=FALSE, gene2Domains=NULL, organis
```

Arguments

`geneIDs`	a character vector of Entrez gene IDs or FlyBase identifiers (not necessary, if the argument gene2Domains is provided)
`flyBase`	Are FlyBase identifiers provided? Default: No
`gene2Domains`	By default associations between genes and InterPro domains are retrieved via biomaRt from Ensembl. Alternatively, the user can provide its own mapping of genes to InterPro domains in form of a list here (see details).

organism	KEGG letter code describing an organism. Please refer to <URL:http://www.genome.jp/kegg-bin/create_kegg_menu> for a complete list of organisms (and their letter codes) supported by KEGG.
useKEGG	Should KEGG information instead of a prediction be used when possible?
mc.cores	number of cores to use for parallelization; requires package 'doMC' to be loaded

Details

A hierarchical classification model based on SVMs and a ranking perceptron is used. This model is usually additionally bagged to improve prediction quality. The model is stored in the package data directory and is recommended to be retrained from time to time.

The KEGG hierarchy is taken from the package keggorthology. By default associations between genes and InterPro domains are retrieved automatically via biomaRt from Ensembl. Please refer to <URL:http://www.ebi.ac.uk/ensembl/> for a list of organisms supported by Ensembl. Alternatively to using Ensembl and biomaRt, the user can provide its own mapping of genes to InterPro domains in form of a list. This especially allows for using organisms, which are supported by KEGG, but not by Ensembl so far. The list has the form genes -> InterPro domains, and each list entry is named by a gene identifier of the corresponding gene. Entrez gene IDs or FlyBase identifiers have to be used.

Value

gene2Path	mapping of gene IDs to corresponding KEGG pathway names
byKEGG	indicates by TRUE/FALSE for each gene whether the mapping information was obtained directly from KEGG or whether it was predicted
scores	confidence scores for the prediction (0, if no prediction was performed): see notes for details
elemIDs	KEGG elements mapping to the corresponding predicted pathway components, if there are any, otherwise NULL. May be used to highlight pathway components with color.pathway.by.elements.
votes	fraction of votes for individual pathway component predictions

Note

By default a bagged model prediction is used, i.e. each of the individual sub-models is giving a vote for a specific output. The final output is determined by the majority of the votes for each hierarchy branch separately. The corresponding fraction voting for a specific branch may be interpreted as its probability. In the ideal case all individual branch probabilites should always be close to 1, if the gene maps to that part of the KEGG hierarchy, and close to 0 otherwise. A cumulative measure of confindence is thus the average over all probabilities > 0.5 and one minus the average over all probabilities < 0.5. We combine both measure by taking the average of both and report it as a reliability score.

If the user decides to retrain a model WITHOUT using bagging, then the reliability score is simply the margin between the highest and the second highest ranked solution. This margin should be larger 2 for good confidence.

Author(s)

Holger Froehlich

See Also

retrain.signaltrans, classificationModelSignalTrans

Examples

```
## Not run:
 gene2pathway.signaltrans("1443")

## End(Not run)
```

`getComponents` *KEGG pathway information*

Description

1. get connected pathway components; 2. get all elements of a given pathway; 3. color certain elements in a pathway.

Usage

```
getComponents(pathway.id, organism="hsa")

get.elements.by.pathway(pathway.id)

color.pathway.by.elements(pathway.id, elements)
```

Arguments

`pathway.id`	KEGG pathway ID, e.g. "path:hsa04012"
`organism`	organism according to 3-letter KEGG abbreviation
`elements`	KEGG element IDs: character vector of numbers

Details

All functions use the KEGG SOAP service.

Value

getComponents: a list with the entries

`geneIDs`	Entrez gene IDs mapping to each pathway component
`elemIDs`	KEGG element IDs mapping to each pathway component

get.elements.by.pathway: list, see <URL http://www.genome.jp/kegg/soap/doc/keggapi_manual.html> for details

color.pathway.by.elements: an URL of a colored gif file, see <URL http://www.genome.jp/kegg/soap/doc/keggapi_manu for details

Author(s)

Holger Froehlich

Examples

```
## Not run:
  comp = getComponents("path:hsa04020") # get all connected components
  color.pathway.by.elements("path:hsa04020", comp$elemIDs[[2]]) # mark first component

## End(Not run)
```

---

internal *internal functions*

Description

internal functions: do not call these functions directly.

Usage

```
various
```

Arguments

various

Value

various

Author(s)

Holger Froehlich

---

retrain *Retrain classification model*

Description

Retrains the hierarchical classification model. This way new information from InterPro and KEGG databases can be incorporated to give better predictions. Retraining should be done on a regular basis from time to time.

Usage

```
retrain(minnmap=30, level1Only="Metabolism", level2Only="Genetic Information Pro
```

Arguments

minnmap	prune hierarchy branches with < minnmap mapping genes
level1Only	for these hierarchy branches only the first level is used
level2Only	for these hierarchy branches only the first and the second levels are used
organism	KEGG letter code describing an organism. Please refer to <URL:http://www.genome.jp/kegg-bin/create_kegg_menu> for a complete list of organisms (and their letter codes) supported by KEGG.
gene2Domains	By default associations between genes and InterPro domains are retrieved via biomaRt from Ensembl. Alternatively, the user can provide its own mapping of genes to InterPro domains in form of a list here (see details).
remove.duplicates	remove genes having the same InterPro domains prior training. Default: Don't do this
use.bagging	use bagging
nbag	number of models to average over
mc.cores	number of cores to use for parallelization; requires package 'doMC' to be loaded

Details

A hierarchical classification model based on SVMs and a ranking perceptron algorithm is trained. This model is usually additionally bagged to improve prediction quality. The method produces a "classificationModel_[organism].rda" (e.g. "classificationModel_hsa.rda") file, which should be stored in the package data directory. Once a new model has been trained, the complete package should be detached and reloaded.

The KEGG hierarchy is taken from the package keggorthology. By default associations between genes and InterPro domains are retrieved automatically via biomaRt from Ensembl. Please refer to <URL:http://www.ebi.ac.uk/ensembl/> for a list of organisms supported by Ensembl. Alternatively to using Ensembl and biomaRt, the user can provide its own mapping of genes to InterPro domains in form of a list. This especially allows for using organisms, which are supported by KEGG, but not by Ensembl so far. The list has the form genes -> InterPro domains, and each list entry is named by the Entrez gene ID of the corresponding gene.

Value

The model structure. See classificationModel for details.

Author(s)

Holger Froehlich

See Also

gene2pathway, classificationModel

Examples

```
## Not run:
retrain(organism="dme") # retrain classification model for drosophila

## End(Not run)
```

`retrain.signaltrans` *Retrain classification model for signaling pathways*

Description

Retrains the hierarchical classification model for signaling pathway components. This way new information from InterPro and KEGG databases can be incorporated to give better predictions. Retraining should be done on a regular basis from time to time.

Usage

```
retrain.signaltrans(minnmap=15, organism="hsa", gene2Domains=NULL, remove.duplic
```

Arguments

`minnmap`	prune hierarchy branches with < minnmap mapping genes
`organism`	KEGG letter code describing an organism. Please refer to <URL:http://www.genome.jp/kegg-bin/create_kegg_menu> for a complete list of organisms (and their letter codes) supported by KEGG.
`gene2Domains`	By default associations between genes and InterPro domains are retrieved via biomaRt from Ensembl. Alternatively, the user can provide its own mapping of genes to InterPro domains in form of a list here (see details).
`remove.duplicates`	remove genes having the same InterPro domains prior training
`use.bagging`	use bagging
`nbag`	number of models to average over
`mc.cores`	number of cores to use for parallelization; requires package 'doMC' to be loaded

Details

A hierarchical classification model based on SVMs and a ranking perceptron algorithm is trained. This model is usually additionally bagged to improve prediction quality. The method produces a "classificationModelSignalTrans_[organism].rda" (e.g. "classificationModelSignalTrans_hsa.rda") file, which should be stored in the package data directory. Once a new model has been trained, the complete package should be detached and reloaded.

The KEGG hierarchy is taken from the package keggorthology. Labels for the training set are obtained via the function `getComponents`, which uses the KEGG SOAP service. By default associations between genes and InterPro domains are retrieved automatically via biomaRt from Ensembl. Please refer to <URL:http://www.ebi.ac.uk/ensembl/> for a list of organisms supported by Ensembl. Alternatively to using Ensembl and biomaRt, the user can provide its own mapping of genes to InterPro domains in form of a list. This especially allows for using organisms, which are supported by KEGG, but not by Ensembl so far. The list has the form genes -> InterPro domains, and each list entry is named by the Entrez gene ID of the corresponding gene.

Value

The model structure. See `classificationModelSignalTrans` for details.

Author(s)

Holger Froehlich

See Also

gene2pathway.signaltrans, classificationModelSignalTrans

Examples

```
## Not run:
retrain.signaltrans() # retrain classification model for signal transduction pathways for

## End(Not run)
```

---

run.crossvalidation

Assessment of Prediction Performance via Cross-validation

Description

Evaluate the prediction performance of a gene2pathway model via a repeated cross-validation scheme.

Usage

```
run.crossvalidation(nfolds=10, repeats=10, stratified=TRUE, signaltrans.only=FAL
```

Arguments

nfolds	number of cross-validation folds
repeats	number of repeats of the cross-validation procedure
stratified	Ensure that during bagging each class is represented
signaltrans.only	do cross-validation for model predicting pathway components of signaling pathways
minnmap	prune hierarchy branches with < minnmap mapping genes
nbag	number of models to average over
level1Only	for these hierarchy branches only the first level is used
level2Only	for these hierarchy branches only the first and the second levels are used
organism	KEGG letter code describing an organism. Please refer to <URL:http://www.genome.jp/kegg-bin/create_kegg_menu> for a complete list of organisms (and their letter codes) supported by KEGG.
gene2Domains	By default associations between genes and InterPro domains are retrieved via biomaRt from Ensembl. Alternatively, the user can provide its own mapping of genes to InterPro domains in form of a list here (see details).
seed	seed value for random number generator: influences splitting of data into training and test
DIR	directory where to save diagnostic plots
mc.cores	number of cores to use for parallelization; requires package 'doMC' to be loaded

Details

A gene2pathway model is trained and tested within a repeated cross-validation scheme. The method produces boxplots (saved as PDFs in the directory passed in the DIR argument) of the accuracy (1 - loss), sensitivity, specificity and F1 values summarized over all pathways. Additionally it produces separate boxplots of F1-values for all pathways in the top KEGG hierarchy level, at the 2nd KEGG hierarchy level and for all pathways individually.

Value

`cv`	a matrix of nfolds*repeats rows and as many columns as labels with predictions of the model
`groups`	used groups in the cross-validation procedure
`used_domains`	used InterPro domains by the prediction model
`evaluation`	a list with average loss, sensitivity, specificity and F1-value for each pathway

Author(s)

Holger Froehlich

See Also

`retrain`, `gene2pathway`

Examples

```
## Not run:
 run.crossvalidation(signaltrans.only=T, repeats=1, nfolds=2)

## End(Not run)
```

---

test.overrepresentation

Test statistical overrepresentation of KEGG pathways in a list of genes

---

Description

Test the statistical overrepresentation of KEGG pathways in a group of genes using Fisher's exact test. The analysis can either be based on all KEGG pathways predicted by `gene2pathway`/`gene2pathway.signal` or on original KEGG annotation only.

Usage

```
test.overrepresentation(genesOfInterest, predpath, KEGGonly=FALSE, cutoff=0.1, m
```

Arguments

genesOfInterest	a character vector of gene identifiers (see gene2pathway, gene2pathway.signaltrans) for a gene list of interest
predpath	predictions of gene2pathway or gene2pathway.signaltrans
KEGGonly	use KEGG annotation only
cutoff	p-value significance cutoff
min.conf	filter predictions such that only those with a confidence score > min.conf are considered
adj.method	multiple testing correction method. Default: Benjamini-Yekutieli
mc.cores	number of cores to use for parallelization; requires package 'doMC' to be loaded

Value

Table with two columns: KEGG pathway and adjusted p-value (adjustment according to Benjamini-Yekutieli)

Index