Software for predicting missing protein values from RNA values

Short description

Software consists of two Python (Python 3.x, version3.7 or higher is recommended) modules:

1. "DLNetworkForProteinAbundancePreparation.py" and
2. "DLNetworkForProteinAbundancePrediction.py".

The first one is designed for preparation of gene and protein data from a raw data to establish one-to-one correspondence between identifiers and tissue names in these two files. In general it should be used to synchronize two csv-style files with headers. The prediction module uses prepared data for predicting missing protein values from the RNA values and optional gene context file. Technical details are described below.

1. Module "DLNetworkForProteinAbundancePreparation.py"

Input

File with gene information and file with protein information. Each file should be "csv or tsv style" text file with header row, titles and values should be delimited by some delimiter. Delimiters for both files may be different. In both files name of the column containing identifiers should be the same. If output files should be with the same delimiter different from original files, it should be specified.

Output

Two files – file corresponding to gene file and file corresponding to protein file. Only columns appearing in both input files will be present in output files. Order of columns will be the same in both files. First column will be the identifier column. Other columns are in the alphabetical order. Only rows having identifiers present in both input files, will be written in output files. Rows are sorted in ascending order by identifier column values.

Parameters

Name	Description	Mandatory
`-i, --i1`	The name of the first input file	YES
`-I, --i2`	The name of the second input file	YES
`-o, --o1`	The name of the first output file	YES
`-O, --o2`	The name of the second output file	YES
`-d, --d1`	Delimiter in the first input file (default – TAB)	NO
`-D, --d2`	Delimiter in the second input file (default – TAB)	NO
`-e, --id`	Name of the identifier column	YES
`-f, --dout`	Delimiter for the output files	NO
`-c, --cs`	Case sensitive	NO

Parameters are specified by the corresponding keyword, order of parameters may be arbitrary. Delimiters should be one character (tabulation character is denoted as "TAB").

Usage

```
Python DLNetworkForProteinAbundancePreparation.py <list of
parameters>
```

Error messages and warnings

Errors

```
Identifier <identifier_name> is not present in the file
<input_file_name> header.
```

The identifier specified in parameters is not specified in the input file header.

```
Data row length exceeds length of the header row in
<input_file_name>.
```

Number of fields in header row is less than in data row.

Ordinary system errors (like “File not found”) are also possible.

Warnings

```
Column <column_name> is not present in the file
<input_file_name> header.
```

The column name specified in one of input files is not present in the other - the mentioned column will not be present in the output file produced from this input file.

```
Identifier <identifier_value> is not present in
<input_file_name>.
```

Identifier field value specified in one of input files is not present in the other - row with the mentioned identifier value will not be present in the output file produced from this input file. Only five warnings of this kind will be reported. In case of more errors, warnings are tailed with “`... more warnings ...`”.

2. Module “DLNetworkForProteinAbundancePrediction.py”

Input

File with RNA information and file with protein information. Each file should be tab-delimited text file with header row. In both files name of the column containing identifiers should be the same. Optional gene context file may be specified. If context file is not specified, obtained results are close to obtained by linear regression. In the file with protein information missing values should be specified as 0 or NA.

Output

Output file is corresponding to the protein file where missing values are replaced by calculated ones.

Usage

```
Python DLNetworkForProteinAbundancePrediction.py <RNA_file>
<protein_file> <out_file> [<context_file>]
```