Inferring the perturbation time from biological time course data –DEtime package

Jing Yang

09/03/2015

Abstract

Time course data is often used to study the dynamics in a biological process after perturation at certain time. Inferring the perturbation time under different scenarios in a biological process allows us to identify these critical moments and focus on any following activities in the process, which is of critical importance in understanding likely caucal relationships. In DEtime package, we propose a Bayesian method to infer the perturbation time from a control and perturbed system. A non-parametric Gaussian Process regression is applied in deriving the posterior distribution of the perturbation point. This vignette explains how to use the package. For further exposition of the algorithm, please refer to out paper(Jing Yang and Rattray, n.d.)

Description

This package implements the Gaussian regression framework for perturbation time point inferrence in a two sample case. The package contains two main functions: **DEtime_infer**, which is used to find out perturbation point of genes, and **DEtime_rank**, which is used to filter these silent genes before carrying out perturbation point inference by **DEtime_infer** function.

The package works on the time course data from a wild-type and a perturbed system. Acting upon pre-defined testing perturbation time, the package goes over these perturbation time candidates and derives their likelihoods. From Bayes' theory, under a uniform prior assumption, the posterior distribution of the tested perturbation time is derived from their corresponding likeliooods. *Maximum a posterior (MAP)*, *mean* or *median* of the posterior distribution can be taken as the solution to the estimated perturbation time point.

Details

```
Package:  DEtime
Type:     Package
Version:  1.0
Date:     2015-03-09
License:  GPL-3
```

Author(s)

Jing Yang

Maintainer

Jing Yang ynnjing@gmail.com

Functionss

DEtime_infer - Perturbation time inference

Description

DEtime_infer is the main function in DEtime Package, which applies a mixedGP kernel to time course data under control and perturbed conditions. It returns the posterior distribution of these predefined perturbation time candidates and relevant statistical estimations of the inferred perturbation time point.

Usage

DEtime_infer(times, ControlData, PerturbedData, times_test=NULL, gene_ID=NULL)

Arguments

- **times**: experimental time points at which the control and perturbed time course data are measured, pls note that if you have multiple replicates, times will have to be repeated for each replicated measurement;
- **ControlData**: The measured time course data under control condtion. The data is a matrix where each row represents the time course data for one particular gene and each column is the measured data at corresponding time point.
- **PerturbedData**: The measured time course data under perturbed condtion. The data is a matrix where each row represents the time course data for one particular gene. The columns for both **ControlData** and **PerturbedData** are ordered by the time sequencing followed by replicates, as shown in the Table below (the geneIDs are used for illustration purpose only, they are not included in the mearument data):

geneIDs	replicate 1	replicate 2
gene 1	t_1 t_2 ... t_n	t_1 t_2 ... t_n
gene 2	t_1 t_2 ... t_n	t_1 t_2 ... t_n

- **times_test**: perturbation time points which will be evalued by **DEtime_infer** function. **times_test** has to be in the range of times and evenly spaced. If this data is missing, **times_test** will be created by an evenly spcaed sequence between the min(**times**) and max(**times**) with 50 bins ;
- **gene_ID**: The ID of each gene investigated in the algorithm. If this value is missing, 1, 2, 3, ... will be used instead.

Returns

The function will return a **DEtimeOutput** object which contains:

- **result**: statistical estimations for the inferred perturbation time, which includes:
  - $**MAP**: *maximum a posterior* solution to the inferred perturbation time
  - $**mean**: mean of the posterior distribution of the inferred perturbation time
  - $**median**: median of the posterior distribution of the inferred perturbation time
  - $**ptl5**: 5 percentile of the posterior distribution of the inferred perturbation time
  - $**ptl95**: 95 percentile of the posterior distribution of the inferred perturbation time
- $**posterior**: posterior distribution of the tested perturbation time points
- $**model**: optimized GP model which will be used for later GP regression work

- $**best_param** : optimized hyperparameter for the optimized GP model
- $**originaltimes**: original experimental time points which will be used for future print or plot functions
- $**originaldata**: original measured time course data which will be used for future print or plot functions
- $**times_test**: tested perturbation time points
- $**gene_ID**: the ID of genes for the data

Details

Both control and perturbed data have to be measured at the same time points with the same number of replicates. Replicates are required to be obtained across all time points.

Examples

```
## read simulated example data
library("DEtime")
data(SimulatedData)
res <- DEtime_infer(times = times, ControlData = ControlData, PerturbedData=PerturbedData)
```

Functionss

DEtime_rank - Rank time course data by log-likelihood ratio

Description

DEtime_rank is the function used for filtering silent genes in DEtime Package. In this function, an independent GP and an integrated GP are applied to model the time course data under control and perturbed conditions, respectively. The log-likelihood ratio of the GP modeling result is used as an indication of the differential expression of the studied gene. A higher rank generally indicates better differential expression.

Usage

DEtime_rank(times, ControlData, PerturbedData, gene_ID=NULL, savefile=TRUE)

Arguments

- **times**: experimental time points at which the control and perturbed time course data are measured, pls note that the time points will have to be repeated if there are replicated measurements;
- **ControlData**: The measured time course data under control condtion. The data is a matrix where each row represents the time course data for one particular gene and each column represents the measurement at a specific time point;
- **PerturbedData**: The measured time course data under perturbed condtion. The data is a matrix where each row represents the time course data for one particular gene. The columns for both **ControlData** and **PerturbedData** are ordered by the time sequencing followed by replicates, as shown in the Table below (the geneIDs are used for illustration purpose only, they are not included in the mearument data):

geneIDs	replicate 1	replicate 2
gene 1	$t_1\ t_2\ \ldots\ t_n$	$t_1\ t_2\ \ldots\ t_n$
gene 2	$t_1\ t_2\ \ldots\ t_n$	$t_1\ t_2\ \ldots\ t_n$

- **gene_ID**: The ID of each gene investigated in the algorithm. If this value is missing, 1, 2, 3, ... will be used instead.
- **savefile**: A BOOLEAN parameter used to indicate if the ranking list will be saved in a file or not. If set to TRUE, the result will be saved in DEtime_rank.txt

Returns

The function will return a table which contains the gene_IDs as the first column and the associated loglikelihood ratio as the second column.

Details

Both control and perturbed data have to be measured at the same time points with the same number of replicates. Replicates are required to be obtained across all time points.

Examples

```
## read simulated example data
library("DEtime")
data(SimulatedData)
res <- DEtime_rank(times = times, ControlData = ControlData, PerturbedData=PerturbedData,
              gene_ID=gene_ID, savefile=TRUE)
```

print_DEtime - print the results from DEtime function

Description

The function prints the results returned from **DEtime_infer** function, which will show the **gene_ID** associated with **MAP**, **mean**, **median**, **ptl5** (lower 5 percentile) and **ptl95** (upper 5 percentile) of the posterior distribution of inferred perturbation time points.

Usage

print_DEtime(DEtimeOutput)

Argument

- **DEtimeOutput**: the returned value from **DEtime_infer** function

Example

```
library("DEtime")
## read simulated example data
data(SimulatedData)
res <- DEtime_infer(times = times, ControlData = ControlData, PerturbedData=PerturbedData)
print_DEtime(res)
```

```
## Perturbation point inference results from DEtime package:
## =========================================
```

```
##   gene_ID   MAP  mean median   ptl5 ptl95
## 1       1 4.490 4.595  4.490 4.4898 4.898
## 2       2 4.898 3.638  4.082 0.8163 4.898
## 3       3 5.306 4.789  4.898 2.4490 6.122
## ======================================
```

plot_DEtime - plot the results of DEtime function

Description

plot_DEtime plots the results returned from **DEtime_infer** function. The produced figures show the the posterior distribution of inferred perturbation time points on the upper panel and Gaussian Regression of the original data on the lower panel. Please note that the MAP solution of the perturbation point is taken as the optimized estimate to the perturbation point and Gaussian Regression is derived based upon this estimated perturabtion point.

Usage

plot_DEtime(DEtimeOutput, plot_gene_ID=NULL)

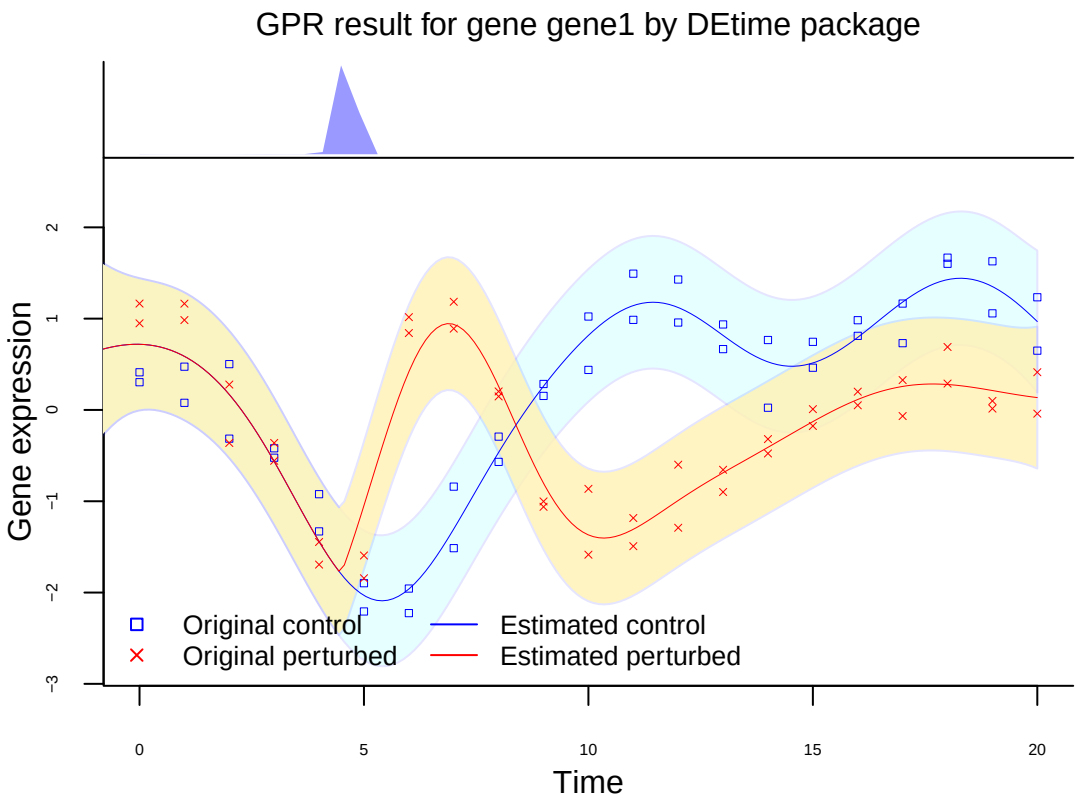
Argument

- **DEtimeOutput**: the returned value from **DEtime_infer** function
- **plot_gene_ID**: the gene_IDs of those genes whose GP regression and posterior distribution of the perturbation time points will be plotted. If not supplied, all the genes will be plotted.

Example

```
library("DEtime")
## read simulated example data
data(SimulatedData)
res <- DEtime_infer(times = times, ControlData = ControlData, PerturbedData=PerturbedData, gene_ID = ge
plot_DEtime(res,plot_gene_ID='gene1')
```

```
## gene1 is plotted
```



Run the package on the real data used in our paper and plot the one with top loglikelihood ratio

Descriptions

In this experiment, the aim is to study the transcriptional change occuring in Arabidopsis following inoculation with P. syringae pv. tomato DC3000 (PtoDC3000) versus the disarmed strain Pto DC3000hrpA

The data contain two different time series:

- infection of Arabidopsis with virulent Pseudomonas syringage pv. tomato DC3000, which leads to disease development (perturbed condition 1), referred as ControlData in the dataset
- infection of Arabidopsis with the disarmed strain DC3000hrpA (perturbed condition 2), referred as PerturbedData in the dataset

In this example, the perturbation time between perturbed condition 1 and perturbed condition 2 is inferred.

```
library("DEtime")
## import data
data(RealData)
## calculate the loglikelihood ratio for each gene
res_rank <- DEtime_rank(times = times, ControlData = ControlData, PerturbedData=PerturbedData, gene_ID=g
```

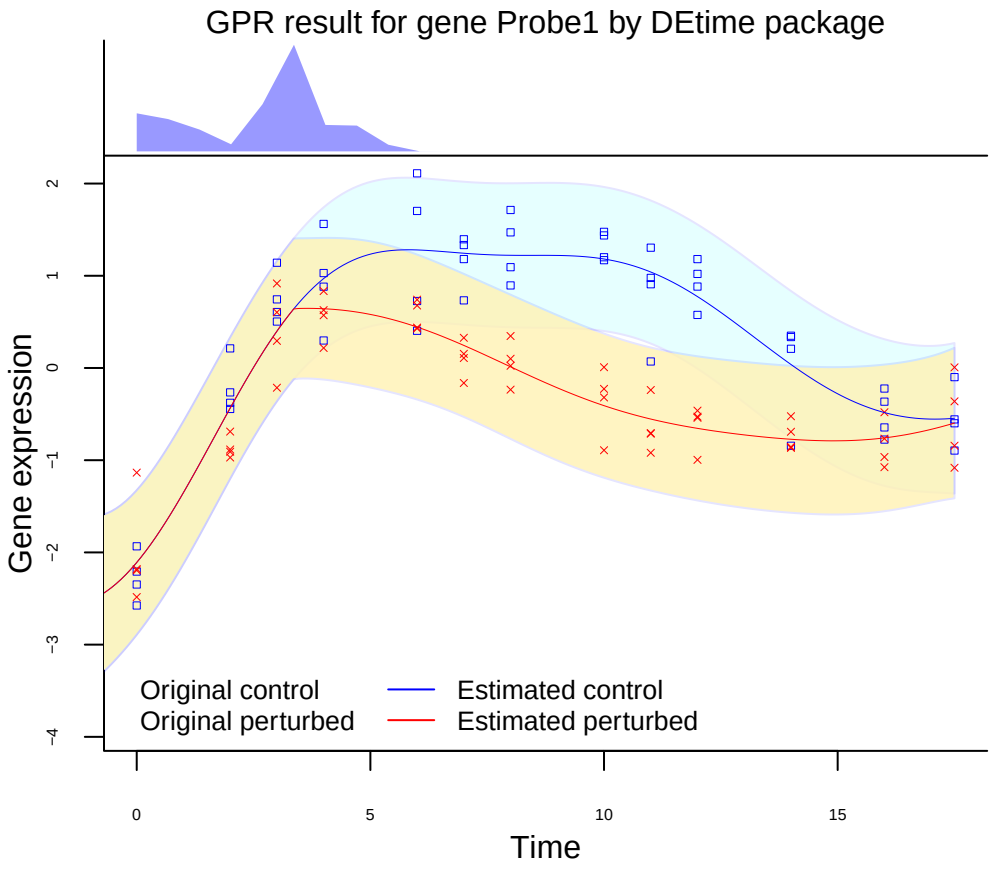
```
## rank list saved in DEtime_rank.txt
```

```
## inferring the perturbation point by DEtime_infer
res <- DEtime_infer(times = times, ControlData = ControlData, PerturbedData=PerturbedData, times_test=t
## Print a summary of the results
print_DEtime(res)
```

```
## Perturbation point inference results from DEtime package:
## ======================================
##    gene_ID     MAP   mean  median   ptl5  ptl95
## 1   Probe1  3.3654  2.615  3.3654 0.0000  4.712
## 2   Probe2  1.3462  5.499  4.0385 0.0000 16.154
## 3   Probe3  6.0577  4.116  4.0385 0.6731  7.404
## 4   Probe4  8.7500  7.479  8.0769 1.3462 14.135
## 5   Probe5  8.7500  7.938  8.0769 3.3654 13.462
## 6   Probe6  8.7500  7.387  8.0769 3.3654  9.423
## 7   Probe7 17.5000 11.674 14.1346 3.3654 17.500
## 8   Probe8  8.7500  8.122  8.7500 4.0385 10.769
## 9   Probe9 16.1538  9.104  9.4231 0.6731 16.827
## 10 Probe10  2.0192  3.319  2.0192 0.0000  9.423
## 11 Probe11  1.3462  6.978  2.6923 0.0000 17.500
## 12 Probe12 10.7692  9.500 10.0962 0.6731 16.827
## 13 Probe13  3.3654  9.367  8.0769 2.0192 17.500
## 14 Probe14  0.6731  1.592  0.6731 0.0000  6.731
## 15 Probe15  9.4231 10.262 10.7692 1.3462 17.500
## 16 Probe16  2.0192  1.876  2.0192 0.0000  4.038
## 17 Probe17 13.4615 12.089 12.7885 6.7308 14.808
## 18 Probe18  2.6923  2.510  2.6923 1.3462  3.365
## 19 Probe19 17.5000 13.853 14.8077 6.0577 17.500
## 20 Probe20  0.0000  9.359 10.0962 0.0000 16.827
## ======================================
```

```
## plot the gene with top loglikelihood ratio
plot_DEtime(res, plot_gene_ID=gene_ID[which.max(res_rank[,2])])
```

```
## Probe1 is plotted
```



References

Jing Yang, Murray R. Grant, Christopher A. Penfold, and Magnus Rattray. n.d. "Inferring the Perturbation Time from Biological Time Course Data." *Bioinformatics*.