opm software: an R package for analyzing Phenotype MicroArray data

Markus Göker, DSMZ, Germany
Johannes Sikorski, DSMZ, Germany
Lea Vaas, CBS, The Netherlands
Benjamin Hofner, FAU, Germany
Motivation

- robust statistical analysis of PM data
- flexible metadata management
- flexible production of high-quality graphics
- no restrictions regarding user-defined analyses
- reproducible research
- easy interaction with other software
- easily extendable by the user
- interactive or fully automated usage possible
Why R?

- *de facto* standard for free-software statistical computing
- all operating systems
- flexible and clean coding
- non-interactive and interactive use
- powerful GUIs/IDEs (e.g. RStudio™)
Input formats

- CSV from OmniLog® PM File Management/Kinetic Analysis software
- CSV from MicroStation® software
- YAML (produced by opm itself)

Facilities

- Single command reads entire directory structures
- All plate types can be input
- Automated split into plate types
- Batch conversion of large numbers of files
Leibniz-Institut DSMZ – German Collection of Microorganisms and Cell Cultures

import CSV into 'opm' package

aggregate (bootstrapped) curve parameters A, AUC, lambda, mu
discretize curve parameters
add metadata
manage metadata
Import/export in YAML format

OPMS object: contains
- raw kinetic values
- metadata
- aggregated curve parameters
- discretized curve parameters

data import and management

full data

if needed
query and subset for:
- metadata entries
- specific plates/wells
- specific time points

export discretized results in:
html or plain text in publication quality
NEXUS, PHYLIP or Hennig86 format

raw kinetic values

levelplot

xy plot

confidence-interval plot
heatmap

graphical and statistical analysis

any other statistical or graphical analysis within R

within R

manuscript for publication
third-party software
exploitation of phylogeny software such as RAxML, PAUP* or TNT
import/export into databases, exchange among labs

OmniLog
OmniLog PM respiration or growth experiments
export raw kinetic values as CSV
Metadata addition & manipulation

Goal: self-describing objects that carry all relevant information (data and metadata)

- if well-structured information from CSV files is present, it can be converted to metadata in 1 line of code
- unique plate identifiers assist in adding other meta-information from tabular input
- no limits regarding structure and amount of metadata
Plotting raw data: x/y plots
Plotting raw data: x/y plots (after subsetting)
Stored substrate information

- mapping to substrate names for almost all plates (PM21-25 to be added)
- KEGG IDs (=> Bioconductor/KEGGGraph)
- Metacyc IDs
- MeSH IDs
- CAS IDs
- error-tolerant search for names, plates, positions
- user-defined modifications (abbreviations etc.) for plots
Plotting raw data: level plots

E. coli vs. P. aeruginosa

Escherichia coli DSM18039  Escherichia coli DSM0083T  Pseudomonas aeruginosa DSM1707  Pseudomonas aeruginosa 429SC1
Aggregating: estimating curve parameters

- **Parametric models** (Gompertz, logistic etc.) work only well for rather regular curve shapes
- **Splines** outperform them in other cases

Aggregating: estimating curve parameters

- **parametric models** (Gompertz, logistic etc.) work only well for rather regular curve shapes
- **splines** outperform them in other cases

Aggregating: estimating curve parameters ...from splines
Aggregating: estimating curve parameters from splines

...but splines must be tuned
Plotting curve parameters: heat map
Plotting curve parameters: radial plot
Plotting curve parameters: principal-component analysis & biplot

Tools available in
- stats
- vegan
- biodiversityR
- ...
Multiple comparison of means

- user-defined comparisons
- inherent multiplicity adjustment
- significance of difference
- and effect size visible
- on original scale

Subsetting and exploration of details

Example: A (maximum)
Subsetting and exploration of details

Example: lambda (lag phase)
Discretization

- **multi-state characters**

```nexus
begin data;
dimensions ntax = 4 nchar = 96;
format datatype = standard missing = ?;
format symbols = "0123456789ABCDEFGHIJKLMNOPQRSTUVWXYZ";
charlabels 'A01 (Negative Control)' 'A02 (Dextrin)' 'A03 (D-Maltose)' 'A04 (D-Trehalose)' 'A05 (D-Cellobiose)' 'A06 (b-Gentiobiose)' 'A07 (Sucrose)' 'A08 (D-Turanose)' 'A09 (Stachyose)';
matrix
  'Escherichia coli DSM18039' 4B2311111TSS131104106RPN23220059DQP9H6007MM2LQQ3NL3M332MRRS3NLNN8N2NRVV2E201NN07T
  'Escherichia coli DSM30083T' ALPNFPGDSTSHN00GQOPQPPONNOBBP00QPN05F0PPDQPT80QIQI89QSSS6PPPQA444SVV4J4R3PQQM
  'Pseudomonas aeruginosa DSM1707' 450421111RRS11012C001R0K30B30O33IQO2G40C2523PPS3PKL0PQERST232N36203RV062QNP7061
  'Pseudomonas aeruginosa 429SC1' 453333334RRT4323203345Q04N53754NRPS4NN302525QPS6Q0QQRRRLRST332P372Q2VVQ82SPR2P81
end;
```

- **binary characters (positive, negative, weak/ambiguous)**

- **two partitioning algorithms**

- optionally applied group-wise

- **two concepts of “ambiguity”**
Plotting discretized curve parameters: heat map
13-10

Positive for γ-amino-n-butyric acid, δ-amino-valeric acid, butyric acid, capric acid, caproic acid, 4-hydroxy-benzoic acid, β-hydroxy-butyric acid, malonic acid, quinic acid, L-alaninamide, L-arginine, L-histidine, L-homoserine, 4-hydroxy-L-proline (trans), L-isoleucine, L-leucine, L-lysine, L-ornithine, L-pyroglutamic acid, L-valine, D,L-carnitine, putrescine and dihydroxy-acetone.

Negative for negative control, chondroitin sulfate C, α-cyclodextrin, β-cyclodextrin, γ-cyclodextrin, dextrin, gelatin, glycerogen, inulin, laminarin, mannan, pectin, N-acetylgalactosamine, N-acetyl-neuraminic acid, β-D-allose, amygdalin, D-arabinose, D-arabitol, L-arabitol, arbutin, 2-deoxy-D-ribose, m-erythritol, D-fucose, 3-O-β-D-galactopyranosyl-D-arabinose, β-gentiobiose, L-glucose, D-lactitol, D-melezitose, maltitol, α-methyl-D-glucoside, β-methyl-D-glucoside, 3-O-methyl-D-glucose, β-methyl-D-glucuronic acid, α-methyl-D-mannoside, β-methyl-D-xylpyranoside, palatinose, D-raffinose, D-salicin, sedoheptulose, L-sorbose, stachyose, D-tagatose, turanose, xylitol, N-acetyl-D-glucosaminotol, citraconic acid, D-glucosamine, 2-hydroxy-benzoic acid, γ-hydroxy-butyric acid, α-keto-valeric acid, itaconic acid, 5-keto-D-glucuronic acid, D-lactic acid methyl ester, melibionic acid, oxalic acid, oxalomallic acid, D-ribo-1,4-lactone, sebacic acid, succinic acid, D-tartaric acid, L-tartaric acid, acetamide, N-acetyl-l-glutamic acid, glycine, L-methionine, L-phenylalanine, butylamine (sec), D,L-octopamine, 2,3-butanediol, 2,3-butanediol and 3-hydroxy-2-butanone.

Ambiguous for sorbic acid.

13-9

Positive for γ-amino-n-butyric acid, δ-amino-valeric acid, caproic acid, 4-hydroxy-benzoic acid, β-hydroxy-butyric acid, malonic acid, quinic acid, L-arginine, 4-hydroxy-L-proline (trans), L-isoleucine, L-pyroglutamic acid, L-valine, D,L-carnitine and putrescine.

Negative for negative control, chondroitin sulfate C, α-cyclodextrin, β-cyclodextrin, γ-cyclodextrin, dextrin, gelatin, glycerogen, inulin, laminarin, mannan, pectin, N-acetylgalactosamine, N-acetyl-neuraminic acid, β-D-allose, amygdalin, D-arabinose, D-arabitol, L-arabitol, arbutin, 2-deoxy-D-ribose, m-erythritol, D-fucose, 3-O-β-D-galactopyranosyl-D-arabinose, β-gentiobiose, L-glucose, D-lactitol, D-melezitose, maltitol, α-methyl-D-glucoside, β-methyl-D-glucoside, 3-O-methyl-D-glucose, β-methyl-D-glucuronic acid, α-methyl-D-mannoside, β-methyl-D-xylpyranoside, palatinose, D-raffinose, D-salicin, sedoheptulose, L-sorbose, stachyose, D-tagatose, turanose, xylitol, N-acetyl-D-glucosaminotol, citraconic acid, D-glucosamine, 2-hydroxy-benzoic acid, γ-hydroxy-butyric acid, α-keto-valeric acid, itaconic acid, 5-keto-D-glucuronic acid, D-lactic acid methyl ester, melibionic acid, oxalic acid, oxalomallic acid, D-ribo-1,4-lactone, sebacic acid, succinic acid, D-tartaric acid, L-tartaric acid, acetamide, N-acetyl-l-glutamic acid, glycine, L-methionine, L-phenylalanine, butylamine (sec), D,L-octopamine, 2,3-butanediol, 2,3-butanediol and 3-hydroxy-2-butanone.

Characters exported by opm version 0.8.17

Organisms: 1, 13-10; 2, 13-9.

Symbols: -, negative reaction; w, weak reaction; +, positive reaction.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A01 (Negative Control)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A02 (L-Arabinose)</td>
<td>-/+</td>
<td>-</td>
</tr>
<tr>
<td>A03 (N-Acetyl-D-Glucosamine)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A04 (D-Saccharic Acid)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>A05 (Succinic Acid)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>A06 (D-Galactose)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A07 (L-Aspartic Acid)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>A08 (L-Proline)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>A09 (D-Alanine)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>A10 (D-Trehalose)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A11 (D-Mannose)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A12 (Dulcitol)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B01 (D-Serine)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B02 (D-Sorbitol)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B03 (Glycerol)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>B04 (L-Fucose)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B05 (D-Glucuronic Acid)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>B06 (D-Glucuronic Acid)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>B07 (D,L-α-Glycerol-Phosphate)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B08 (D-Xylose)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>B09 (L-Lactic Acid)</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
---
- metadata:
  Strain: 13-10
  Replicate: II
  csv_data:
- measurements:
  Hour:
  - 0.00
  - 0.25
- A01:
  - 30.0
  - 24.0
- aggregated:
  A01:
  mu: 4.910216
  lambda: -2.095738
  A: 52.18975
  AUC: 4422.683
- aggr_settings:
  method: grofit
  options:
  neg.nan.act: FALSE
  clean.bootstrap: TRUE
- discretized:
  A01: FALSE
  A02: FALSE
- disc_settings:
  method: kmeans
  options:
  cutoffs: 157.1985
  datasets: 4
  software: opm
  version: 0.8.17
---

[...]

- stored metadata
- raw measurements
- estimated curve parameters
- discretized values
- discretization settings

[...]

- parameter estimation settings

[...]

- software: opm
  version: 0.8.17
Code example

library(opm)

x <- read_opm(getwd(), convert = "grp", include = list("csv"))
x <- lapply(x, function(item) {
    md <- to_metadata(csv_data(item))
    md <- do.call(rbind, strsplit(md$`Strain Number`, " ", fixed = TRUE))
    colnames(md) <- c("Strain", "Replicate")
    metadata(item) <- to_metadata(md)
    item
})

x <- lapply(lapply(x, do_aggr, boot = 0L, cores = 8L), do_disc, cutoff = FALSE)

file.copy(grep("\.[^.]css$", opm_files("auxiliary"), value = TRUE), 
        "opm_styles.css", overwrite = TRUE)
opm_opt(css.file = "opm_styles.css")

for (plate in names(x)) {
    write(to_yaml(x[[plate]]), file = sprintf("Data_%s.yml", plate))
    text <- listing(x[[plate]], as.groups = "Strain", html = TRUE)
    write(phylo_data(text), sprintf("Description_%s.html", plate))
    text <- phylo_data(x[[plate]], format = "html", as.labels = "Strain")
    write(text, sprintf("Table_%s.html", plate))
    pkgutils::mypdf(sprintf("Plot_%s.pdf", plate))
    print(xy_plot(x[[plate]], include = list("Strain", "Replicate")))
dev.off()
}

- read files
- set up metadata
- aggregate & discretize
- for all plate types in input: create formatted text, table, and plots
Leibniz-Institut DSMZ – German Collection of Microorganisms and Cell Cultures

'opm'

import CSV into 'opm' package

OPMS object: contains
- raw kinetic values
- metadata
- aggregated curve parameters
- discretized curve parameters

aggregate (bootstrapped) curve parameters
- A, AUC, lambda, mu

discretize curve parameters

add metadata

manage metadata

import/export in YAML format

within R

data import and management

data analysis and export

full data

if needed

export discretized results in:
- html or plain text in publication quality
- NEXUS, PHYLIP or Hennig86 format

query and subset for:
- metadata entries
- specific plates/wells
- specific time points

graphical and statistical analysis

raw kinetic values

levelplot

xy plot

confidence-interval plot

heatmap

any other statistical or graphical analysis within R

manuscript for publication

third-party software

exploitation of phylogeny software such as RAxML, PAUP* or TNT

import/export into databases, exchange among labs
Embedding opm

SIMS lab management system
CeBiTec, Bielefeld, Germany
Summary: opm

- robust statistical analysis of PM data
- flexible metadata management
- flexible production of high-quality graphics
- no restrictions regarding user-defined analyses
- reproducible research
- easy interaction with other software
- easily extendable by the user
- interactive or fully automated usage possible
opm availability

- http://opm.dsmz.de/
- manual, tutorial, mailing list etc.
- open project, hosted at R-Forge
- programmers and advanced users are invited to join