pwOmics - Pathway-based Integration of time-series Omics Data using public database knowledge

Astrid Wachter
Medical Statistics, University Medical Center Göttingen, Germany

October 17, 2016

Contents

1 Introduction

Characterization of biological processes can be performed in great detail with the increased generation of omics data on different functional levels

of the cell. Especially interpretation of time-series omics data measured in parallel with different platforms is a complex but promising task, needing consideration of time-independent combination of omics data and additionally time-dependent signaling analysis. As each measurement technique shows a certain bias and has natural limitations in identifying full signaling responses (Yeger-Lotem et al. 2009), such cross-platform analysis is an up-to-date approach in order to connect biological implications on different signaling levels. Using diverse data types is expected to provide a deeper understanding of global biological functions and the underlying complex processes (al. 2012).
This is why computational data analysis tools for interpretation of data from proteomics and transcriptomics measurements in parallel are needed.
pwOmics is a tool for pathway-based level-specific data comparison and analysis of single time point or time-series omics data measured in parallel. It provides individual analysis workflows for the different omics data sets (see Figure 1) and in addition enables consensus analysis of omics data as shown in the workflow overview in Figure 2.

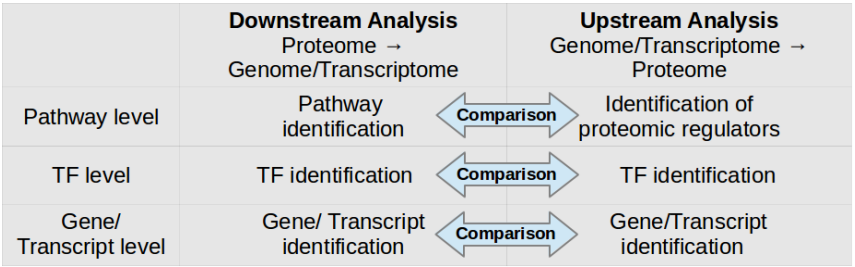

Figure 1: *pwOmics* downstream and upstream analysis.

Up to this point analysis is restricted to human species. In future an expansion of the package is possible dependent on available online open access database information.

2 Databases

As *pwOmics* is a package for data integration based on prior pathway and transcription knowledge data, it is necessary to define the databases to work with. Three different kinds of databases are necessary to do all analyses steps:

1. Pathway databases:
   The user can choose from Biocarta (Nishimura 2001), Reactome (Milacic et al. 2012; Croft et al. 2014), PID (Schaefer et al. 2009) from

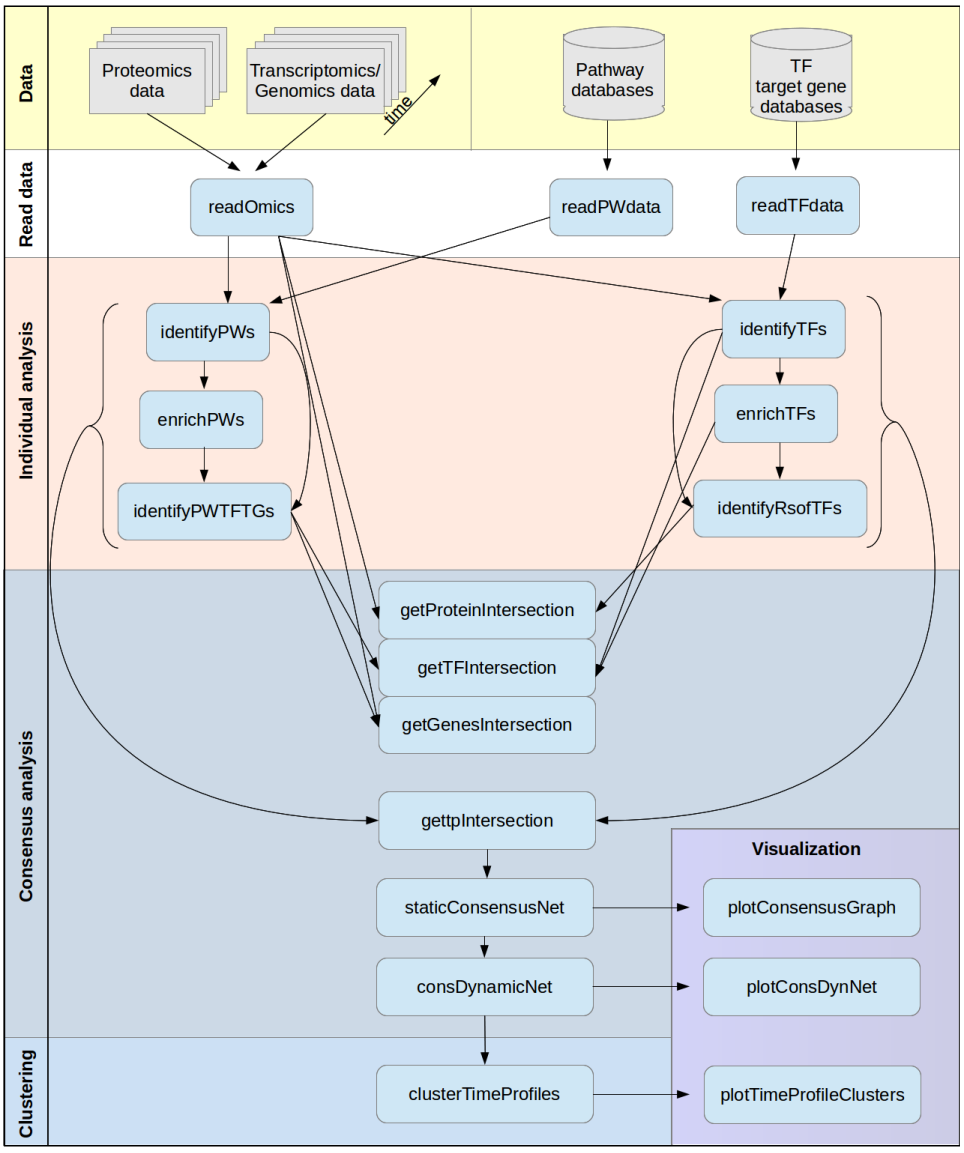

Figure 2: *pwOmics* workflow overview.

the National Cancer Institute (NCI) and KEGG (Kanehisa et al. 2014; Kanehisa and Goto 2000).

2. Protein-protein interaction (PPI) database:
   STRING (Franceschini et al. 2013).

3. Transcription factor (TF) - target gene databases:
   The user can choose from ChEA (Lachmann et al. 2010), Pazar (Portales-Casamar et al. 2009; Portales-Casamar et al. 2007) and/or decide to specify an own file e.g. based on a commercial database.

The pathway database information is used to identify the pathways of the differentially abundant proteins in the downstream analysis as well as upstream protein regulators of TFs in the upstream analysis. The PPI database STRING (Franceschini et al. 2013) was chosen to define the protein net for the consensus analysis. The TF - target gene databases information is necessary for the TF identification in pathways in the downstream analysis. Additionally the upstream TFs of differentially expressed genes/transcripts are identified in the upstream analysis based on this information.
In downstream analysis the pathway gene set information is used, whereas in the upstream analysis also the pathway topology information is exploited.

The database information is downloaded internally via *STRINGdb* and *AnnotationHub* ("AnnotationHub: Client to access AnnotationHub resources") package. In case the author is interested also in the metadata of the pathway database and TF - target database it can be received by

```
> library(pwOmics)
> library(AnnotationHub)
> ah = AnnotationHub()
> #pathway databases
> pw = query(ah, "NIH Pathway Interaction Database")
> pw[1]
```

```
AnnotationHub with 1 record
# snapshotDate(): 2016-10-11
# names(): AH22329
# $dataprovider: NIH Pathway Interaction Database
# $species: Homo sapiens
# $rdataclass: biopax
# $title: BioCarta.owl.gz
# $description: BioCarta BioPax file from NCI Pathway Interaction Database
```

```
# $taxonomyid: 9606
# $genome: hg19
# $sourcetype: BioPax
# $sourceurl: ftp://ftp1.nci.nih.gov/pub/PID/BioPAX/BioCarta.owl.gz
# $sourcelastmodifieddate: 2009-09-09
# $sourcesize: 338343
# $tags: c("BioCarta", "BioPax", "Pathway Interaction Database")
# retrieve record with 'object[["AH22329"]]'

> #TF-target databases
> chea = query(ah, "ChEA")
> chea[1]

AnnotationHub with 1 record
# snapshotDate(): 2016-10-11
# names(): AH22237
# $dataprovider: ChEA
# $species: NA
# $rdataclass: data.frame
# $title: chea-background.zip
# $description: ChEA background file, containing transcription factor data t...
# $taxonomyid: NA
# $genome: NA
# $sourcetype: Zip
# $sourceurl: http://amp.pharm.mssm.edu/result/kea/chea-background.zip
# $sourcelastmodifieddate: 2015-03-09
# $sourcesize: 3655103
# $tags: c("ChEA", "Transcription Factors")
# retrieve record with 'object[["AH22237"]]'

> pazar = query(ah, "Pazar")
> pazar[1]

AnnotationHub with 1 record
# snapshotDate(): 2016-10-11
# names(): AH22238
# $dataprovider: Pazar
# $species: NA
# $rdataclass: GRanges
# $title: pazar_ABS_20120522.csv
# $description: TF - Target Gene file from pazar_ABS_20120522
# $taxonomyid: NA
# $genome: NA
```

```
# $sourcetype: CSV
# $sourceurl: http://www.pazar.info/tftargets/pazar_ABS_20120522.csv
# $sourcelastmodifieddate: 2012-06-04
# $sourcesize: 120202
# $tags: c("Pazar", "Transcription Factors")
# retrieve record with 'object[["AH22238"]]'
```

In case you want to use TF - target gene information which is not part of the mentioned databases but e.g. part of a commercial database, a user-specified file can be used for the analysis. This file should be a '.txt' file with first column transcription factors and second column target gene symbols without a header, e.g.:

GATA-4 HAMP
c-Jun IL18
NF-kappaB
TLR2
MYB LTB
FOXO1A
TGFBR1
...

The STRING PPI-information is downloaded automatically while processing and analyzing the data: The *STRINGdb* package (Franceschini et al. 2013) is used here.

3 Example dataset

The example dataset used here for demonstration purposes is the one presented in (Waters 2012), which comprises the mitogenic response of human mammary epithelial cells to epidermal growth factor (EGF). This dataset includes whole genome time course microarray data measured with NimbleGen whole genome 60-mer oligonucleotide arrays (Design Version 2003_10_27) at time points 0, 1, 4, 8, 13, 18 and 24 hr after EGF stimulation. The complementary proteomics data was measured with LC-FTICR (Fourier-transform ion cyclotron resonance-mass spectrometry coupled with advanced capillary liquid chromatography) at time points 0.25, 1, 4, 8, 13, 18 and 24 hours after EGF stimulation. Preprocessing of data was done as described in (Waters 2012) resulting in lists of significant genes and proteins for each time point as log10 expression ratios relative to the time 0 hr controls.

4 Data pre-processing

pwOmics is a package for secondary data analysis, i.e. it needs already pre-processed data as input for the analysis. The input required is

1. a list of all protein IDs measured,
2. a list of all gene/transcript IDs measured,
3. a list of differentially abundant proteins + log fold changes,
4. a list of differentially expressed genes/transcripts + log fold changes.

The IDs need to be gene symbols, both for protein and gene/transcript data. In case time-series data is analyzed inputs 3. and 4. needs to be specified for each time point. It is absolutely necessary, that all proteins and genes/transcript in inputs 3. and 4. are part of the lists of all protein IDs and all gene/transcript IDs, respectively.

The OmicsData object is the format used for data analysis in *pwOmics* package. It contains a list of four main elements:

1. OmicsD - here the omics data set, its description and the results are stored
2. PathwayD - here the chosen pathway databases and the generated Biopax model is stored
3. TFtargetsD - here the chosen TF-target gene databases and the combined TF-target gene information is stored
4. Status - The status variable equals '1' in case not all information needed for the analysis is read in yet and '2' after identification of the first upstream/downstream signaling levels. As the enrichment step is not necessarily part of the analysis and dependent on the pathway database and the TF-target gene database the identification of signaling molecules in further levels might not be successful, the status variable is not used in the further analysis.

Thus *pwOmics* reads in the omics data set provided by the user to the first element of the OmicsData object and further on stores all the results in this

part as well.

This is why the user has to provide the omics data set in a special format: A list needs to be generated with a protein list named 'P' as first element and a gene/transcript list named 'G' as second element. These lists contain as first element a data frame with all (unique) protein IDs and gene/transcript IDs in the first column, respectively, and as second element a list with data frames for each time point of measurement. The data frames have two columns with the first one containing the differentially abundant/expressed proteins or genes/transcripts as gene symbols and the second column containing the corresponding log fold changes, e.g.:

```
> data(OmicsExampleData)
> OmicsExampleData
```

Generated as in the following example:

```
OmicsExampleData = list(P = list(allPIDs,
                   list(PIDstp0.25, PIDstp1, PIDstp4, PIDstp8,
                        PIDstp13, PIDstp18, PIDstp24)),
                        G = list(allGIDs,
                   list(GIDstp1, GIDstp4, GIDstp8, GIDstp13,
                        GIDstp18, GIDstp24)))
```

```
> head(OmicsExampleData$P[[2]][[1]])
```

```
  GeneSymbol     X15min
1     MRPS17  0.6976049
2      RPS12 -1.0297977
3     SLC3A2 -1.2623327
4       RPL8  0.8304820
5       ACTB -2.4914461
6      ALDOA  0.8637013
```

In case the user only wants to analyze omics data from a single time point just one data frame has to be specified.
The time points do not have to be the same for protein and gene/transcript data and need to be specified when reading in the omics data set separately via the 'tp_prots' and 'tp_genes' parameters of the 'readOmics' function.

```
> data_omics = readOmics(tp_prots = c(0.25, 1, 4, 8, 13, 18, 24),
+                        tp_genes = c(1, 4, 8, 13, 18, 24),
```

```
+                         OmicsExampleData,
+                         PWdatabase = c("biocarta", "kegg", "nci",
+                                        "reactome"),
+                         TFtargetdatabase = c("chea", "pazar"))
```

If data from a single timepoint measurement should be analyzed the user simply assigns the experiment number '1' for these parameters:

```
#for single time point data set:
omics = list(P = list(allPIDs, list(PIDs_1)),
             G = list(allGIDs, list(GIDs_1)))
data_omics = readOmics(tp_prots = c(1),
                   tp_genes = c(1),
                   OmicsExampleData,
                   PWdatabase = c("biocarta", "kegg", "nci",
                                  "reactome"),
                   TFtargetdatabase = c("chea", "pazar"))
```

Additionally the selected databases have to be specified.

The stored information can be easily accessed via the following functions:

```
> getOmicsTimepoints(data_omics)
> head(getOmicsallProteinIDs(data_omics))
> head(getOmicsallGeneIDs(data_omics))
> head(getOmicsDataset(data_omics, writeData = FALSE)[[1]])
```

5 Individual analysis

As shown in Figure 1 the analysis is based on an individual analysis of the proteomic and the genomic/transcriptomic data. The downstream analysis and upstream analysis are described in the following subsections.

Prior to that the database information has to be read in. In a first step the TF- target information can be made accessible to the OmicsData object by:

```
data_omics = readTFdata(data_omics)
```

Via the 'TF_target_path' parameter the path of the user-specified file can be given. This information can be used additionally to the selected database content.

Secondly, the 'readPWdata' function takes the OmicsData object with the provided information about the omics data set and the path of the prepared '.RData' files from the pathway databases (see section 2) and automatically generates the corresponding genelists of the pathway data if 'loadgenelists = FALSE'. In this step the automatic definition of internal differing IDs for different pathway databases is necessary, which are stored in a new biopax model in the OmicsData object.

```
data_omicsPW = readPWdata(data_omics,
               loadgenelists = FALSE)
```

As the process of generating genelists with these IDs can take some time - especially for rather big databases such as Reactome Milacic et al. 2012; Croft et al. 2014 - the genelists for the different databases are automatically stored in the working directory and can be reused in another analysis when the corresponding path containing these files is given to the 'readPWdata' function as loadgenelists parameter.

```
data_omicsPW = readPWdata(data_omics,
             loadgenelists = "Genelist_reactome.RData")
```

Automatically the information of the selected databases and/or the corresponding user-specified file are merged. The file format (if this option is used) should be exactly as specified in section 2.

5.1 Downstream analysis

The downstream analysis is starting with the provided proteomic data (either single time point data or time-series data). The first step is the identification of the pathways in which the differentially abundant proteins play a role. *pwOmics* performs this searching step on the basis of the provided proteomic data set and the selected pathway database(s).

After reading in these information the user can follow the workflow for downstream analysis and identify the pathways in which the differentially abundant proteins are present:

```
data_omics = identifyPWs(data_omicsPW)
```

In a next step pathway enrichment can be conducted. The user can specify the multiple testing correction method as well as the significance level for this step. In case of few identified pathways this might result in too few

pathways for further analysis. In this case the enrichment step should be skipped.

```
data_omics = enrichPWs(data_omics, "BH", alpha = 0.05)
```

Following the workflow the next step is the identification of the transcription factors in these (enriched) pathways, which is done with the information provided by the chosen TF-target gene database. The user can choose if only the enriched pathways or all pathways should be considered for further analysis:

```
data_omics = identifyPWTFTGs(data_omics, only_enriched = FALSE)
```

For use of this function the working directory should contain the previously generated genelists.

The results of the downstream analysis can be easily accessed by the following functions:

```
getDS_PWs(data_omics)
getDS_TFs(data_omics)
getDS_TGs(data_omics)

#Access biopax model generated newly on basis of selected
#pathway databases:
getBiopaxModel(data_omics)
```

5.2 Upstream analysis

The upstream analysis is starting with the provided gene/transcript data (either single time point data or time-series data). It first of all identifies the upstream TFs of the differentially expressed genes/transcripts. This step is done with the provided/selected information of TF-target gene pairs.

Given this information, the identification of upstream TFs can be done:

```
data_omics = identifyTFs(data_omics)
```

Similarly as in the downstream analysis also in the upstream analysis an optional enrichment step can be conducted, but here on the TF level.

```
data_omics = enrichTFs(data_omics, "BH", alpha = 0.05)
```

Upstream of the (enriched) TFs the regulator proteins can be identified with the following function:

```
data_omics = identifyRsofTFs(data_omics, only_enriched = FALSE,
             noTFs_inPW = 1, order_neighbors = 10)
```

Again, the user can specify if only the enriched TFs or all TFs should be considered for further analysis. The identification of upstream regulators is done in the following way:

1. Identification of the pathways the previously identified TFs are part of.

2. Selection of pathways according to the user-specified parameter 'noTFs_inPW': Only those pathways are considered in further analysis with at least this number of TFs in the pathway. Minimum number of TFs in the pathway is 2.

3. Upstream regulators are identified for these TFs. This is done by finding first for each TF the pathway neighborhood according to the user-specified parameter 'order_neighbors'. This parameter specifies the order of the identified pathway neighborhood. Then the intersection of all identified neighborhoods for all TFs in a pathway is determined. The resulting pathway node set is defined here as the set of regulator proteins.

In case the pathways under consideration do not have pathway components in the downloaded biopax model, this will be indicated with a warning. This warning can be ignored by the user in regard to the following analysis steps.

The results of the upstream analysis can be accessed with the following functions:

```
getUS_TFs(data_omics)
getUS_PWs(data_omics)
getUS_regulators(data_omics)
```

6 Consensus analysis

The consensus analysis combines the results from upstream and downstream analysis by constituting in particular the comparative analysis of the results of the two different data sets. The intersection analysis simply compares the

results of the separate upstream and downstream analysis. The static consensus analysis enables setting up static consensus graphs for each time point measured in parallel. Finally, the consensus-based dynamic analysis provides the user with one final dynamic network obtained from the data changes over time based on dynamic bayesian network inference. The consensus-based dynamic analysis is self-evidently only conductable with time-series data sets measured for proteome and genome/transcriptome data in parallel.

6.1 Intersection analysis

The intersection analysis of *pwOmics* is a simple comparative analysis of the results of upstream and downstream analysis. Thus, it enables a comparison of single time point data and time-series data, the latter also for non-corresponding time points in the different data sets. The comparison is possible on the three different functional levels considered in this package: On the proteome level, the transcription factor level and gene/transcript level.

```
getProteinIntersection(data_omics,
                       tp_prot = 4,
                       tp_genes = 4)
getTFIntersection(data_omics,
                  tp_prot = 4,
                  tp_genes = 1)
getGenesIntersection(data_omics,
                     tp_prot = 4,
                     tp_genes = 13)
```

These functions not only enable a comparison of the same timepoints on the distinct levels, but for time-series data sets also for non-matching time points:
With the time resolution of measuring omics data in most cases being predefined by expected signaling changes and financial limitations the potential in the interpretation of the results is strongly confined to the experimental design decisions. Thus, measured signaling changes, which naturally consist of a superposition of diverse time-scales of transcriptional and translational processes and comprehend diverse frequency patterns (Yosef and Regev 2011), are dependent on the sampling. This means for some of the signaling axes it might be the case, that

- changes are not detected at all as their rate is too high,
- hopefully most are represented in the data and

- some might be so slow that their change is not considered significant and thus are excluded from analysis.

As the corresponding signaling changes are not expected to be seen at the same time point in proteome data and gene/transcript data it is necessary to enable also the comparison of non-corresponding time points.
The possibility to compare such time points naturally cannot account for the changes not captured during measurement, however, it gives the possibility to consider also the time needed for regulatory control mechanisms in the interpretation of the measurement results - even if this shows considerable variations as well.
In case the user wants to compare the corresponding time points on the three levels simultaneously he can do so by using the following function:

```
gettpIntersection(data_omics)
```

6.2 Static consensus analysis

The static consensus analysis goes one step ahead and integrates the results gained from the comparative analysis of the corresponding time points to a consensus net for each time point. The change of this consensus net over time gives a first insight into the changes seen statically at the different time points. However, the static consensus nets do not yet include information gathered over time - as it is the case in the consensus-based dynamic analysis (see section 6.3). This is why the static consensus analysis is also applicable for single time point measurements.

The static consensus analysis is conducted by generation of a Steiner tree (Kleinberg and Tardos 2006) on basis of matching proteins and TFs identified in downstream and upstream analysis for each corresponding time point. The underlying graph used is the protein-protein-interaction (PPI) graph from the STRING database reduced to the connected nodes. The matching proteins and TFs are considered as terminal nodes and are connected via a shortest path-based approximation of the Steiner tree algorithm (Takahashi and Matsuyama 1980; Sadeghi and Fröhlich 2013) across the reduced PPI-STRING-graph. Subsequently knowledge of TF-target gene pairs from the chosen database is used to expand the graph with matching genes/transcripts from both upstream and downstream analysis. In case the consensus graph contains corresponding proteins and genes/transcripts, feedback loops are added automically.

```
consensusGraphs = staticConsensusNet(data_omics)
```

6.3 Consensus-based dynamic analysis

Unlike the static consensus analysis, the consensus-based dynamic analysis takes into consideration also the signaling changes over time by applying dynamic bayesian network inference. The packages used for the consensus-based dynamic analysis are *longitudinal* (Opgen-Rhein and Strimmer 2006; Rainer Opgen-Rhein 2006) to adjust the format of the data and the actual network inference part is done via the *ebdbNet* (Rau et al. 2010) package. This package includes an iterative empirical Bayesian procedure with a Kalman filter estimating the posterior distributions of the network parameters. The defined prior structure of the network is used for a straightforward estimation of hyperparameters via an expectation maximization (EM)-like algorithm and the dimension of the hidden states are determined via the singular value decomposition (SVD) of a block-Hankel matrix.

The nodes included into the network inference step are nodes which are part of the static consensus graphs from corresponding time points of the two different measurement types, i.e.

1. proteins identified in upstream and downstream analysis at the same time points,

2. Steiner nodes identified via static consensus analysis,

3. TFs identified in upstream and downstream analysis at the same time points and

4. genes/transcripts identified in upstream and downstream analysis at the same time points.

To apply dynamic network inference a reasonable number of measurements needs to be available. As in most cases of parallel protein and gene/transcript measurements only very few corresponding time steps are available it is necessary to artificially introduce additional time steps. This is done by generating smoothing splines applied on the log fold changes provided by the user under the simplifying assumption of a gradual change of signaling between the different time points.

This assumption, however, has to be applied consciously and carefully, as there might be higher frequency signaling components superimposed (see for a comprehensive analysis of temporal dynamics of gene expression (Yosef

and Regev 2011)). In theory a signal has to be sampled 2 times its maximal frequency in order to be able to reconstruct it exactly from time discrete measurements (Nyquist-Shannon sampling theorem (Shannon and Weaver 1949; Nyquist 1928)). This means only exact interpretation of those signaling axes are possible that have a frequency which is smaller than half of the sampling frequency. However, under certain preconditions on signal structure and the sampling operator reconstruction of the original signal can be done with a lower sampling rate (Blumensath and Davies 2009). This is an interesting starting point for a more comprehensive dynamic analysis of the expected signals and the sampling needed for an extensive data mining of omics data sets measured in parallel, but exceeds the scope of this package.

The number of time points generated additionally via smoothing splines is based on simulation results of *ebdbNet* analysis for median area under the curve (AUC) values of receiver operating characteristic (ROC) curves: In their results it was shown that a plateau at around 50 to 75 time points was reached. Thus in *pwOmics* 50 time points are predicted with smoothing splines in order to apply dynamic bayesian network inference on omics data sets measured in parallel.

After generation of these time points a block-Hankel matrix of autocovariances is constructed based on the time series abundance/expression data. For this the user needs to provide the *laghankel* parameter, specifying the maximum relevant time lag to be used in constructing the block-Hankel matrix. With a singular value decomposition (see function 'hankel' of *ebdbNet* package) the number of hidden states can be determined. Here, the user can specify the *cutoffhankel* parameter to choose the cutoff to determine the desired percent of total variance explained by the singular values. Additional parameters on convergence criteria and iterations performed can be specified. For further details the user is referred to (Rau et al. 2010).

```
library(ebdbNet)
library(longitudinal)
dynInferredNet = consDynamicNet(data_omics, consensusGraphs,
                 laghankel = 3,
                 cutoffhankel = 0.9)
```

7 Time profile clustering

An additional analysis option is clustering of co-regulation patterns over time. It provides information about the signaling molecules with common

dynamic behaviour and thus allows to draw conclusions in terms of signaling chronology. Time profile clustering is performed as soft clustering based on the *Mfuzz* package (Futschik 2012). The advantage of this clustering method is that a protein, TF or gene/transcript can be assigned to several clusters, thus reducing the sensitivity to noise and the information loss hard clustering exhibits. It is implemented as fuzzy c-means algorithm (Hathaway and Bezdek 1986) and iteratively optimizes the objective function to minimize the variation of objects within the clusters. The user needs to provide a 'min.std' threshold parameter if proteins or genes/transcripts with a low standard deviation should be excluded. In addition the maximum number of cluster centers which should be tested in the 'minimum distance between cluster centroid test' has to by given. This number is used as initial number to determine the data-specific maximal cluster number based on the number of distinct data points. For more details see (Futschik 2012) and (Schwämmle and Jensen 2010).

```
library(Mfuzz)
fuzzyClusters = clusterTimeProfiles(dynInferredNet,
                                    min.std = 0,
                                    ncenters = 12)
```

8 Visualization

To complement the results from the different comparisons and analyses (accessible via the 'get...' functions) the *pwOmics* package provides visualization functions for the different analyses. The consensus graphs of the static analysis for one or more corresponding time points can be plotted with the following function (see Figures 3 and 4):

```
plotConsensusGraph(consensusGraphs, data_omics)
```

The consensus-based dynamic analysis result can be visualized as follows (see Figure 5):

```
plotConsDynNet(dynInferredNet, sig.level = 0.65)
```

Here, the parameter 'sig.level' is the significance level used as cutoff for plotting edges in the network and has to be specified in the range between 0 and 1. Furthermore the user can indicate if unconnected nodes should be removed and provide additional *igraph* (Csardi and Nepusz 2006) layout parameters.

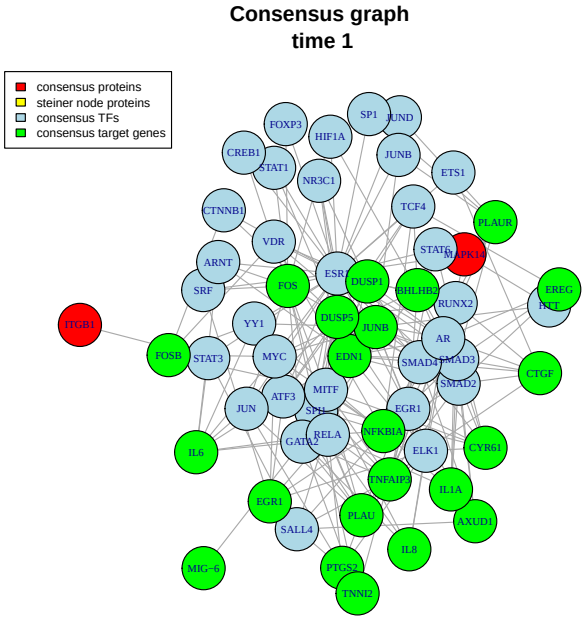

Figure 3: *pwOmics* static consensus graph: Time point 1 hr.

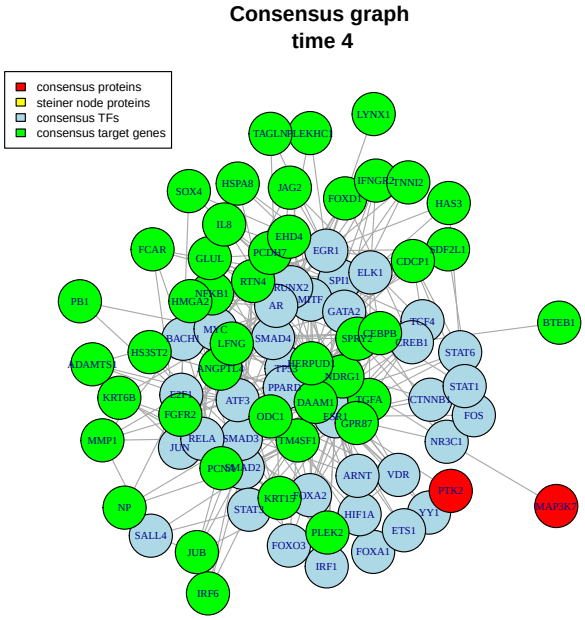

Figure 4: *pwOmics* static consensus graph: Time point 4 hrs.

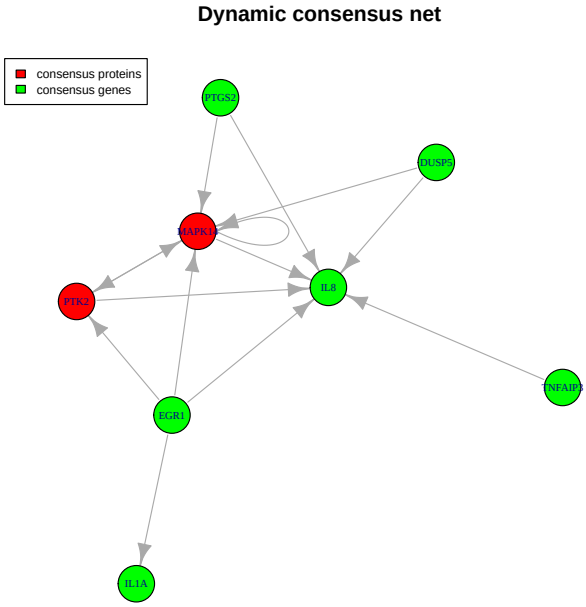

Figure 5: *pwOmics* dynamic network graph.

However, as the user can access the networks easily *tkplot* from the *igraph* R package is a nice interactive graph drawing alternative. In addition plot parameters can be easily changed as the result networks are of class 'igraph'.

In order to plot the results from time profile clustering (see Figure 6) the following function can be used:

```
plotTimeProfileClusters(fuzzyClusters)
```

The different colours represent the different clusters. The legend is only shown if the number of genes and proteins is not too large. Otherwise the user can easily access this information by having a look to the output of the 'clusterTimeProfiles' function which provides information about cluster centers, the number of data points in each cluster of the closest hard clustering, cluster indices, and additional parameters explained in detail in the 'mfuzz' documentation. In the legend the attachments '_g' and '_p', respectively, indicate, if the node originally derives from protein or gene/transcript measurements.

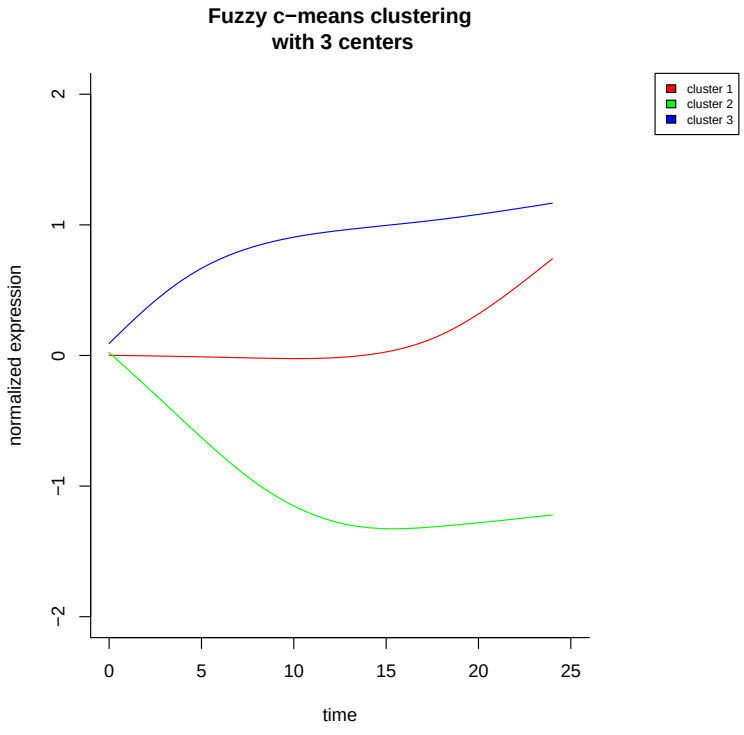

Figure 6: *pwOmics* time profile clusters.

9 Session Information

- R version 3.3.1 (2016-06-21), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=C`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, utils
- Other packages: AnnotationHub 2.6.0, BiocGenerics 0.20.0, pwOmics 1.6.0
- Loaded via a namespace (and not attached): AnnotationDbi 1.36.0, Biobase 2.34.0, BiocInstaller 1.24.0, DBI 0.5-1, GenomeInfoDb 1.10.0, GenomicRanges 1.26.0, IRanges 2.8.0, KernSmooth 2.23-15, R6 2.2.0, RColorBrewer 1.1-2, RCurl 1.95-4.8, RSQLite 1.0.0, Rcpp 0.12.7, S4Vectors 0.12.0, STRINGdb 1.14.0, XML 3.98-1.4, XVector 0.14.0, biomaRt 2.30.0, bitops 1.0-6, caTools 1.17.1, chron 2.3-47, curl 2.1, data.table 1.9.6, digest 0.6.10, gdata 2.17.0, gplots 3.0.1, gsubfn 0.6-6, gtools 3.5.0, hash 2.2.6, htmltools 0.3.5, httpuv 1.3.3, httr 1.2.1, igraph 1.0.1, interactiveDisplayBase 1.12.0, magrittr 1.5, mime 0.5, plotrix 3.6-3, plyr 1.8.4, png 0.1-7, proto 0.3-10, rBiopaxParser 2.14.0, shiny 0.14.1, sqldf 0.4-10, stats4 3.3.1, tools 3.3.1, xtable 1.8-2, zlibbioc 1.20.0

References

Yeger-Lotem, Esti et al. (2009). "Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity". eng. In: *Nature Genetics* 41.3, pp. 316–323. ISSN: 1546-1718. DOI: 10.1038/ng.337.

al., Boris Kholodenko et (2012). "Computational Approaches for Analyzing Information Flow in Biological Networks". In: *Science Signaling* 5, re1.

Nishimura, Darryl (2001). "BioCarta". In: *Biotech Software & Internet Report* 2.3, pp. 117–120. ISSN: 1527-9162. DOI: 10.1089/152791601750294344. URL: http://online.liebertpub.com/doi/abs/10.1089/152791601750294344 (visited on 01/02/2015).

Milacic, Marija et al. (2012). "Annotating cancer variants and anti-cancer therapeutics in reactome". eng. In: *Cancers* 4.4, pp. 1180–1211. ISSN: 2072-6694. DOI: 10.3390/cancers4041180.

Croft, David et al. (2014). "The Reactome pathway knowledgebase". eng. In: *Nucleic Acids Research* 42.Database issue, pp. D472–477. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1102.

Schaefer, Carl F. et al. (2009). "PID: the Pathway Interaction Database". eng. In: *Nucleic Acids Research* 37.Database issue, pp. D674–679. ISSN: 1362-4962. DOI: 10.1093/nar/gkn653.

Kanehisa, Minoru et al. (2014). "Data, information, knowledge and principle: back to metabolism in KEGG". eng. In: *Nucleic Acids Research* 42.Database issue, pp. D199–205. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1076.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes". eng. In: *Nucleic Acids Research* 28.1, pp. 27–30. ISSN: 0305-1048.

Franceschini, Andrea et al. (2013). "STRING v9.1: protein-protein interaction networks, with increased coverage and integration". In: *Nucleic Acids Research* 41.Database issue, pp. D808–D815. ISSN: 0305-1048. DOI: 10.1093/nar/gks1094. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531103/ (visited on 01/02/2015).

Lachmann, Alexander et al. (2010). "ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments". eng. In: *Bioinformatics (Oxford, England)* 26.19, pp. 2438–2444. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btq466.

Portales-Casamar, Elodie et al. (2009). "The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences". en. In: *Nucleic Acids Research* 37.suppl 1, pp. D54–D60. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkn783. URL: http://nar.oxfordjournals.org/content/37/suppl_1/D54 (visited on 01/02/2015).

Portales-Casamar, Elodie et al. (2007). "PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation". en. In: *Genome Biology* 8.10, R207. ISSN: 1465-6906. DOI: 10.1186/gb-2007-8-10-r207. URL: http://genomebiology.com/2007/8/10/R207/abstract (visited on 01/02/2015).

Morgan M Carlson M, Tenenbaum D and Arora S. "AnnotationHub: Client to access AnnotationHub resources". In: R package version 2.22.0.

Waters, K.M. et al. (2012). "Network Analysis of Epidermal Growth Factor Signaling Using Integrated Genomic, Proteomic and Phosphorylation Data". In: *PLoS ONE* 7, e34515.

Yosef, Nir and Aviv Regev (2011). "Impulse control: Temporal dynamics in gene transcription". In: *Cell* 144.6, pp. 886–896. ISSN: 0092-8674. DOI: 10.1016/j.cell.2011.02.015. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3148525/ (visited on 01/03/2015).

Kleinberg, J and E Tardos (2006). *Algorithm Design*. Pearson, Boston, MA.

Takahashi, H and A Matsuyama (1980). "An approximate solution for the Steiner problem in graphs". In: *Math. Jap.* 24, pp. 573–577.

Sadeghi, Afshin and Holger Fröhlich (2013). "Steiner tree methods for optimal sub-network identification: an empirical study". In: *BMC Bioinformatics* 14, p. 144.

Opgen-Rhein, R and K Strimmer (2006). "Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data". In: *Proceedings of the 4th International Workshop on Computational Systems Biology, WCSB 2006*, pp. 73–76.

Rainer Opgen-Rhein, Korbinian Strimmer (2006). "Inferring gene dependency networks from genomic longitudinal data: a functional data approach". In: *REVSTAT – Statistical Journal Volume* 4.1, pp. 53–65. ISSN: 1645-6726.

Rau, Andrea et al. (2010). "An empirical Bayesian method for estimating biological networks from temporal microarray data". In: *Statistical Applications in Genetics and Molecular Biology* 9.1.

Shannon, CA and W Weaver (1949). *The mathematical theory of communication*. University of Illinois Press.

Nyquist, H (1928). "Certain topics in telegraph transmission theory". In: *Transactions of the A. I. E. E.* Pp. 617–644.

Blumensath, Thomas and Michael E. Davies (2009). "Sampling Theorems for Signals From the Union of Finite-Dimensional Linear Subspaces". In: *IEEE Transactions on Information Theory* 55.4, pp. 1872–1882. DOI: 10.1109/TIT.2009.2013003. URL: http://dx.doi.org/10.1109/TIT.2009.2013003.

Futschik, Matthias (2012). *Mfuzz: Soft clustering of time series gene expression data*. R package version 2.22.0. URL: http://itb.biologie.hu-berlin.de/~futschik/software/R/Mfuzz/.

Hathaway, R and J Bezdek (1986). "Local convergence of the fuzzy c-means algorithm". In: *Pattern Recognition* 19, pp. 477–480.

Schwämmle, Veit and Ole Nørregaard Jensen (2010). "A simple and fast method to determine the parameters for fuzzy c–means cluster analysis". en. In: *Bioinformatics* 26.22, pp. 2841–2848. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btq534. URL: http://bioinformatics.oxfordjournals.org/content/26/22/2841 (visited on 01/03/2015).

Csardi, Gabor and Tamas Nepusz (2006). "The igraph software package for complex network research". In: *InterJournal* Complex Systems, p. 1695. URL: http://igraph.org.