Description of ISOQuant methods

Jörg Kuharev

2012-05-09

Table of contents

1 analysis pipeline

The quantitative analysis of label free MS proteomics experiments is done by pipelined usage of Waters Protein Lynx Global Server (PLGS) followed by the automated analysis using our in-house developed analysis pipeline `ISOQuant`.

In the analysis workflow PLGS is used for peak detection as well as for peptide and protein identification. After that the experiment data is automatically imported into the analysis pipeline and processed in multiple stages.

Steps of the analysis are:

- PLGS Analysis:
    1. PLGS peak detection
    2. PLGS peptide and protein identification
- `ISOQuant` Analysis:
    1. project design
    2. data transfer
    3. collecting statistical information about peptides and proteins
    4. retention time alignment
    5. peak clustering
    6. intensity normalization
    7. cluster annotation
    8. protein homology filtering
    9. peptide intensity redistribution
    10. protein quantification
    11. report creation

For working with merged LC fractions, two additional methods are provided:

- fraction time shifting (done before retention time alignment)
- fraction time unshifting (done after peak clustering)

2 PLGS Analysis

In PLGS können Messdaten innerhalb eines Projektes hierarchisch organisiert werden. So können zusammenhängende Messungen (Workflows), z.B. technische Replikate einer biologischen Probe zu einer logischen Probe (Sample) zusammengefasst werden. Unter Verwendung des Expression^{R}-Moduls werden die analysierten Messungen und ihre Daten zwangsweise hierarchisch geordnet. Diese Hierarchie setzt sich (von oben nach unten) wie folgt zusammen:

1. Project
   enthält allgemeine Projekt-bezogene Informationen sowie eins oder mehrere `Expression Analysis`-Elemente (wir betrachten jede einzelne `Expression Analysis` als ein eigenständiges Projekt!)
2. Expression Analysis
   enthält Parameter und Beschreibung der Analyse sowie eins oder mehrere `Group`-Elemente
3. Group
   fasst mehrere Proben, z.B. biologische Replikate, zu einer Gruppe zusammen und enthält somit eins oder mehrere `Sample`-Elemente
4. Sample
   fasst mehrere Messungen, i.d.R. technische Replikate einer Probe zusammen und enthält somit eins oder mehrere `Workflow`-Elemente
5. Workflow
   enthält detaillierte Informationen über eine einzelne Messung und fasst Ergebnisse der Peak-Detektion sowie der Peptid- und Proteinidentifikation:
   - Peak-Liste
   - Peptid- und Proteinlisten

In Abwesenheit des Expression-Moduls können einzelne Messungen lediglich in Proben zusammengefasst werden, d.h. die Zuordnung zu den Gruppen und Expressionsanalysen entfällt. Ungeordnete und neue Messungen werden von PLGS automatisch der Probe `Default` zugeordnet.

3 ISOQuant Analysis Methods

In this chapter we describe our analysis methods implemented in `ISOQuant`.

3.1 project design

`Project Designer`-Modul ermöglicht dem Benutzer (unabhängig der von PLGS oder PLGS-Expression^R vorgegebenen Projektstruktur) Messungen innerhalb eines Projektes hierarchisch neuzuorganisieren. Aus Kompatibilitätsgründen werden die gleichen Hierarchieebenen verwendet, wie in Kapitel 2 beschrieben. Dabei können neue Groups und Samples frei definiert sowie vorhandene Workflows neugeordnet werden.

3.2 Data transfer

Die mit dem Projektdesigner ausgewählten, reorganisierten Workflows werden durchsucht und die mit ihnen verknüpften Daten in das Datenbankmanagementsystem `MySQL` importiert. Um die verwaltete Datenmenge zu reduzieren und die Daten einzelner Projekte innerhalb der relationalen Datenbank voneinander logisch abzugrenzen, wird für jedes zu importierende Projekt ein gesondertes Datenbankschema verwendet. Eine solche Projektdatenbank hat einen festen Strukturellen Aufbau, der mit dem in der Abbildung 1 gezeigten ERM-Diagramm beschrieben wird.

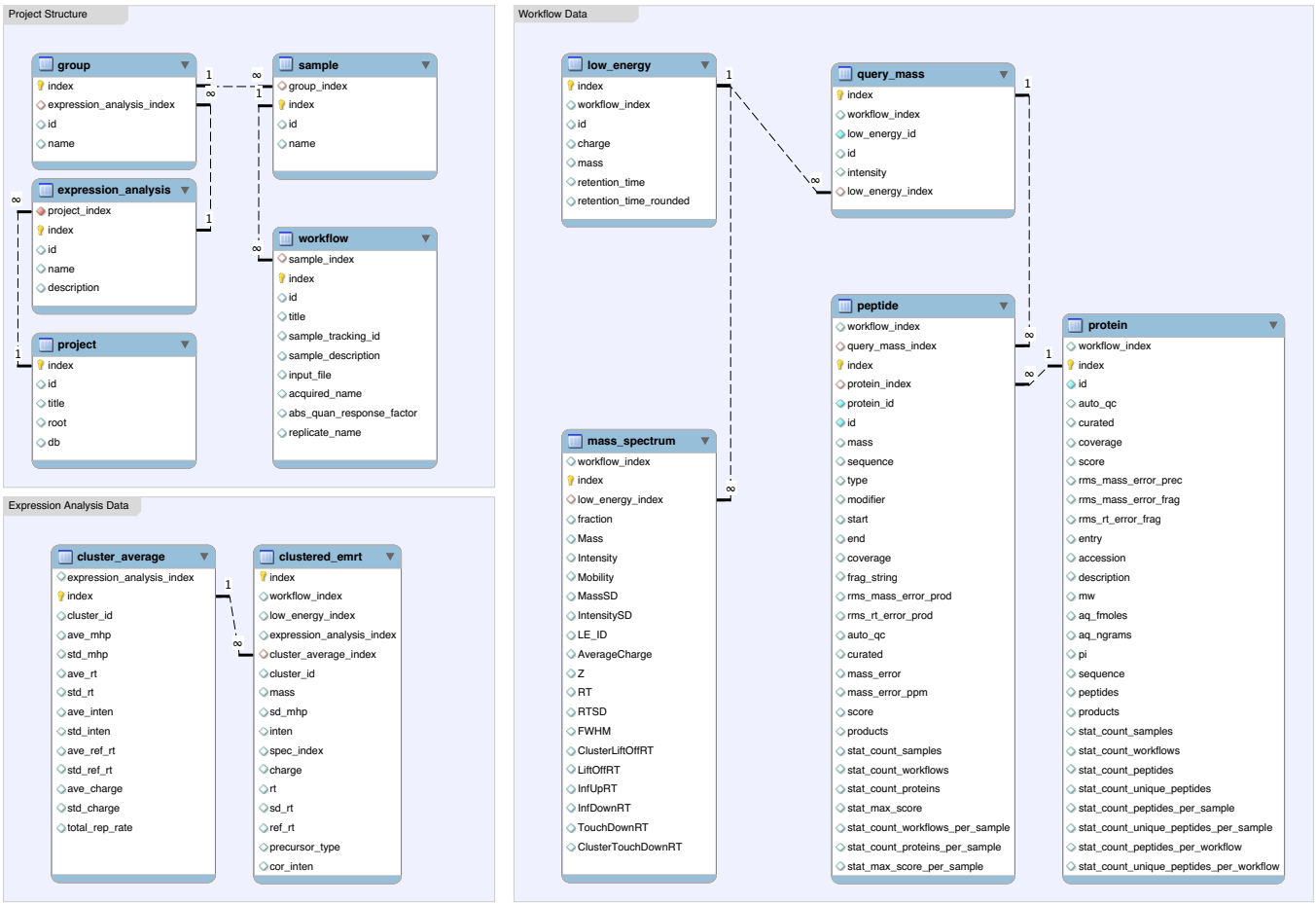

Figure 1: ERM-Diagram of a project database

3.3 Collecting statistical information

In diesem Schritt werden für die gespeicherten Peptide und Proteine Messung-übergreifend innerhalb eines Projektes nützliche statistische Informationen gesammelt und gespeichert, über die PLGS nicht oder nur unzureichend verfügt.

Die gesammelten Informationen umfassen für jedes Peptid:

- In wie vielen Samples wurde dieses Peptid identifiziert
- In wie vielen Workflows wurde dieses Peptid identifiziert
- Wie vielen Proteinen wurde dieses Peptid zugeordnet
- Maximaler Score der Identifikation über alle Workflows
- In wie vielen Workflows im aktuellen Sample wurde dieses Peptid identifiziert
- Wie vielen Proteinen wurde dieses Peptid zugeordnet im aktuellen Sample
- Maximaler Score der Identifikation über alle Workflows im aktuellen Sample

für jedes Protein:

- In wie vielen Samples wurde dieses Protein identifiziert
- In wie vielen Workflows wurde dieses Protein identifiziert
- Wie viele Peptide können diesem Protein zugeordnet werden
- Wie viele Peptide können ausschließlich diesem Protein zugeordnet werden
- In wie vielen Workflows wurde dieses Protein identifiziert im aktuellen Sample
- Wie viele Peptide können diesem Protein zugeordnet werden im aktuellen Sample
- Wie viele Peptide können ausschließlich diesem Protein zugeordnet werden im aktuellen Sample

3.4 fraction time shifting

Beim Zusammenführen zusammengehöriger LC-Fraktionen in PLGS zu einem Workflow tauchen die ursprünglich unabhängig erfassten Signale (Peaks) in einer gemeinsamen Peak-Liste auf. Die absolute zeitliche Reihenfolge der korrespondierenden Peaks in mehreren zusammengeführten Workflows verändert sich. Das Retentionszeitalignment kann zunächst auf diese Workflows nicht angewandt werden. Die Peaks jeder einzelnen LC-Fraktion werden in getrennte Zeiträume verschoben, indem die Retentionszeit jedes Peaks einer Fraktion um einen festen Wert vergrößert wird. Dieser Zeitversatz wird als Produkt der auf die nächsten 10 Minuten aufgerundeten Gradientendauer und der Fraktionsnummer errechnet. Mit dem `fraction time shifting` wird gewährleistet, dass Peaks einer einzelnen Fraktion in ihrer zetlichen Abfolge zusammenhängend vorliegen, d.h. mit den Peaks einer anderen Fraktion nicht durchmischt werden.

3.5 Retention time alignment

Das Retentionszeitalignment projiziert Retentionszeiten der Workflows auf eine gemeinsame Referenz. Als Referenz wird automatisch eines der Workflows angenommen. (welches ist gerade die Referenz?) Nach Retentionszeitalignment ähneln die Referenz-Retentionszeiten der korrespondierenden Peaks in unterschiedlichen Workflows einander.

3.6 Peak clustering

This step builds clusters of corresponding peaks from different runs ending up in a so called EMRT table (Algorithm: Hierarchical-Non-Hierarchical clustering)

3.6.1 Space Transformation

For an easy application of mathematical clustering methods we project LC-MS peaks into a homogeneous geometric space by converting peak's masses to relative values as ppm followed by a resampling using an instrument dependent ppm-resolution-value and resampling retention times by user defined time-resolution-value.

Sort peaks by mass in ascending order.

$n = numberofpeaks$

$mass_0 = 0$

for each peak (at its i_{th} position) we calculate spatial coordinates as follows:

- $x_i = \frac{log(mass_i) - log(mass_{i-1})}{log(1 + massResolution)}$
- $y_i = \frac{rt_i}{timeResolution}$
- $z_i = \frac{driftTime_i}{driftResolution}$

3.7 Intensity normalization

This step corrects systematic errors introduced by divergence of technical conditions in different LC-MS runs

3.8 Cluster annotation

At this stage peak cluster are annotated by peptides using different filtering criteria. Clustering annotation improves peptide quantification and reproducibility due to minimizing false or missing peptide identifications. In fact this step is the relative peptide quantification.

3.9 Protein homology filtering

There is no possibility to determine abundances of homologue proteins when they are identified by a set of equal peptides. We pick homologue proteins with the highest probability of identification and reassign peptides to them. So only one of homologue proteins family appears in the final quantification.

3.10 Peptide intensity redistribution

A lot of identified peptides map to different proteins. Abundances of such shared peptides do not provide a direct way to compute protein abundances but confirm protein identification. Based on abundances of uniquely assigned peptides we redistribute abundances of shared peptides.

3.11 Protein quantification

At this stage of analysis statistically ensured protein abundances are calculated by using TOP3 method applied to previously calculated peptide abundances.

3.12 Report creation

Finally, results of the performed analysis are exported as a set of uniform reports enabling different views at the analyzed data as well as simplifying data interpretation and further analysis.