Probleme und Lösungsansätze im Zusammenhang mit der Auswertung labelfreier LC-MS Daten

Jörg Kuharev

September 3, 2012

Table of contents

1 Einordnung der Pipeline

Analyse quantitativer lebelfreier LC-MS Daten erfolgt in mehreren Schritten und unter Einsatz verschiedener Software.

1.1 Software: Waters MassLynx

Waters MassLynx nimmt die Raw-Daten vom Instrument auf. Der erforderliche Speicherbedard hängt direkt von der Signal-Komplexität ab. Für Q-TOF Premier sind es etwa 1-5GB/h, oder im Durchschnitt 4GB für eine 2h LC-MS Messung. Für Synapt G2S hängt der Speicherbedarf noch zusätzlich davon ab, ob die Ion Mobility Fähigkeit genutzt wird, und beträgt etwa 5-20GB/h, im Durchschnitt werden 20GB für eine 2h LC-MS Messung ohne Ion Mobility belegt.

1.2 Software: Waters PLGS

Waters ProteinLynx Global Server (PLGS) wird für den Umgang mit der Raw-Daten eingesetzt. Wir überlassen PLGS

1. Signalverarbeitung: Peak Detektion
2. Datenbanksuche: Peptid- und Protein-Identification

1.3 Software: `ISOQuant`

`ISOQuant` ist eine Eigenentwicklung für die Aufbereitung der mit PLGS prozessierten Daten. Die Pipeline bildet u.a. folgende Verarbeitungsschritte ab:

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

2 Projektstruktur

Ein label-freies quantitatives LC-MS Experiment untersucht vergleichend mehrere Proteome. Jedes Proteom wird als eine unabhängige biologische Probe behandelt. Jede Probe wird zur statistischen Absicherung in technischen Replikaten mehrfach untersucht. Die Struktur eines solchen Experimentes kann hierarchisch beschrieben werden. PLGS schreibt eine feste Projektstruktur vor.

3 Projekt- und Datenhierarchie in PLGS

1. Project: allgemeine Projekt-bezogene Informationen
2. Expression Analysis: Parameter und Beschreibung der Analyse
3. Group: fasst mehrere Proben zusammen
4. Sample: fasst mehrere Messungen zusammen
5. Workflow: eine Messung, d.h. Peak-Liste sowie Peptid- und Proteinlisten

zusätzlich gilt:

- ohne Expression-Modul gibt es keine Gruppen, bzw. eine `Default`-Group
- Messungen werden standardmäßig zuerst der `Default`-Probe zugeordnet

4 Project Designer

- Projektumorganisation - neue Projektstruktur
- gleiche Hierarchieebenen wie PLGS Expression
- Groups und Samples frei definierbar
- Neuordnung vorhandener Workflows

5 Data Transfer

- reorganisierte Workflows durchsuchen
- Workflow-Daten in `MySQL` importieren
- je Projekt ein Datenbankschema
  - Reduktion der Datenmenge
  - logische Abgrenzung
- einfacher Zugriff
- neue Workflow-übergreifende Möglichkeiten

6 Datenbankstruktur

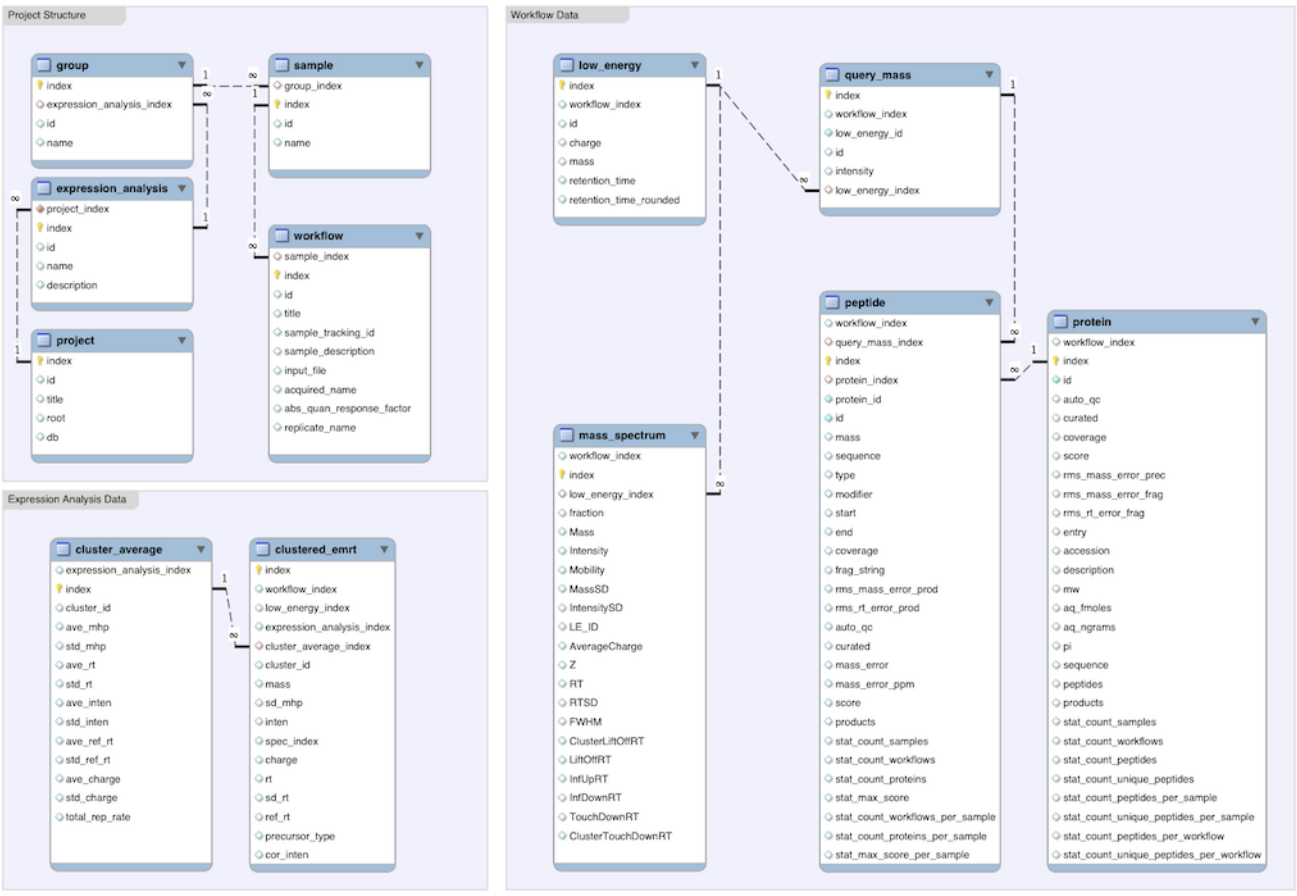

Figure 1: ERM-Diagram of a project database

7 Collecting Statistical Information - Peptide

- 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

8 Collecting Statistical Information - 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

9 Fraction Time Shifting - Problem

- zusammengeführte LC-Fraktionen (zu einem Workflow)
- unabhängig erfasste Signale
- gemeinsame Peak-Liste
- Reihenfolge korrespondierender Signale ist inkonsistent
- RT-Alignment geht nicht!

10 Fraction Time Shifting - Lösung

- Peaks einzelner LC-Fraktion in getrennte Zeiträume verschieben
- Signalreihenfolge konsistent
- RT-Alignment geht!

11 Retention Time Alignment

11.1 Problem

- LC instabil
- korrespondierende Signale zu unterschiedlichen Zeiten
- Zeitversatz ist nicht linear

11.2 Lösung

- chronologisch konsistente Sequenz von übereinstimmenden Signalen suchen
- Zeitverschiebungen zwischen Fundstellen linear interpolieren
- Projektion der Retentionszeiten auf gemeinsame Referenz

Die jeweiligen Sequenzen der Übereinstimmungen werden durch paarweises `Time Warping` der Messungen gegen die Referenzmessung gefunden. Der verwendete Algorithmus ist eine Abwandlung des `Dynamic Time Warping` (DTW) mit linearisierter Speicherkomplexität und gesteigerter Performance durch parallel ausführbare Rekursion.

11.3 Ergebnis

Referenz-Retentionszeiten der korrespondierenden Peaks ähneln einander.

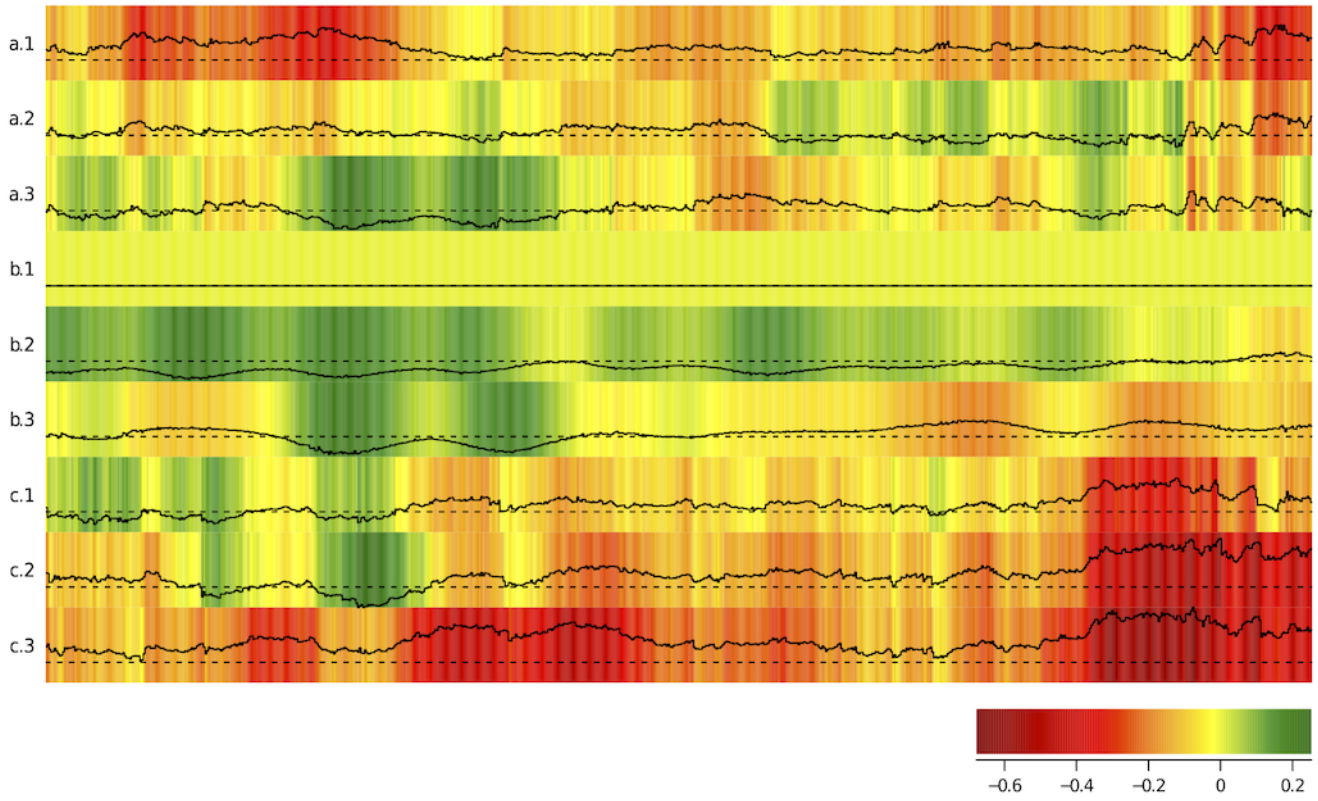

Figure 2: Nicht-lineare Retentionszeitverschiebungen

11.4 Evaluierung

DTW gewährleistet per Definition die mathematisch optimale Lösung, bzw. eine der möglichen optimalen Lösungen, falls mehrere vorhanden. Eine Evaluierung der Methode entfällt. Die Übertragbarkeit von DTW auf dieses Problem wurde beschrieben.

12 Peak Clustering

Zur vergleichender Analyse müssen Signale, die von gleichen Peptid-Ionen stammen, identifiziert werden.

12.1 Problem

Nicht alle Signale werden als ein Peptid erkannt, Manche Signale werden fehlannotiert. Eigenschaften korrespondierender Signale, wie (Referenz-)Retentionszeiten, Massen, Ionen-Mobilität stimmen nicht exakt überein und überlappen sich mit den Eigenschaften anderer Signale. Korrespondierende Signale müssen in den Peak-Listen verschiedener Messungen gesucht und gruppiert werden.

12.2 Lösung

Korrespondenz der Signale wird unabhängig ihrer Annotation mit Hilfe geometrischer Clustering-Verfahren. Umgesetzte Algorithmen

- Hierarchical-Non-Hierarchical Clustering: hierarchisches Clustering mit Abbruch bei Erreichen einer vorgegebenen Entfernung.
- DBSCAN: dichte-basiertes Clustering

13 Intensity Normalization

Die Annahme, dass die Mehrheit der Proteine keine Unterschiede zwischen untersuchten Proteomen in der Expression aufweisen, können systematische Fehler der detektierten Mengen (Signalintensität) beobachtet werden.

13.1 Problem

Die beobachteten systematischen Fehler sind nicht-linear und haben i.d.R. mehrere ungeklärte Ursachen.

13.2 Lösung

Signale jeder einzelnen Messung werden auf multidimensionale systematische Fehler gegenüber dem Durchschnitt der jeweiligen Signal-Cluster in Abhängigkeit von der Intensität, Retentionszeit, Masse sowie Ionenmobilität untersucht und korrigiert.

14 Cluster Annotation

14.1 Problem

Weisen Signale innerhalb eines Clusters inkonsistente Annotationen auf, so wird die beste Annotation nach vorgegebenen Filterkriterien, wie Replikationsrate, Identifikationsscore, etc. bestimmt.

14.2 Lösung

Die beste Annotation wird zur Korrektur der Annotation aller Signale des betroffenen Clusters verwendet. Damit werden Lücken in der Signalannotation geschlossen und in vielen Fällen eine vergleichende Quantifizierung der Peptide zwischen den Proben erst ermöglicht.

15 Protein Inference

Ein Protein kann über mehrere Peptide identifiziert worden sein, gleichzeitig kann das gleiche Peptid mehreren Proteinen abstammen. Dies führt dazu, dass große Netzwerke der nicht-eindeutigen Identifikation entstehen können.

16 Protein Inference - Protein Homology Filtering

16.1 Problem

Homologe Proteine werden häufig durch nicht eindeutig zuzuordnenden (`shared`) Peptide identifiziert. In diesen Fällen gibt es keine Möglichkeit ein Protein aus einer solchen Familie homologer Proteine stabil zu quantifizieren.

17 Lösung

Wir identifizieren das Protein aus einer solchen Familie, das mit der höchsten Wahrscheinlichkeit identifiziert wurde und nehmen ferner an, dass alle `shared` Peptide aus diesem Protein stammen. Nur das jeweils wahrscheinlichste homologe Protein erscheint dann in der finalen Proteinquantifizierung.

18 Protein Inference - Peptide Intensity Redistribution

Proteininferenz verbindet nicht nur homologe Proteine. Wird ein Peptid mehreren Proteinen zugeordnet, so bestimmt PLGS (Algorithmus nicht nachvollziehbar) welche Menge dieses Peptides welchem Protein zuzuordnen ist. Die Kenntnis über die Mengenverteilung eindeutig zugeordneter Peptide erlaubt Rückschlüsse auf die wahrscheinliche Menge von `shared`-Peptiden und ihre umverteilung.

19 Protein Quantification

Liegen Peptidmengen und Zuordnungen zu Proteinen vor, können Rückschlüsse auf die ursprüngliche Proteinmengen gezogen werden. Wir setzen die TOP3-Methode für die Proteinquantifizierung ein[1]. Sie basiert auf der Beobachtung, dass die Proteinmenge mit der durchschnittlichen Menge der best-ioniesierenden Peptide dieses Proteins korreliert.

20 Report Creation

Die Ausgabe erfolgt in Form von mehreren standardisierten Berichten, die unterschiedliche Aspekte der Daten darstellen und dadurch eine einfache Interpretation der Ergebnisse ermöglichen.

[1] Silva, J. C., Gorenstein, M. V., Li, G.-Z., Vissers, J. P. C. & Geromanos, S. J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Mol. Cell Proteomics 5, 144–156 (2006).