A Comprehensive Analysis Pipeline for Phosphoproteomic Biomarker Discovery

David N. Perkins, Pedro Moreno-Cardoso, Nazrath Nawaz, Weronika E. Borek, Josie A. Christopher, Amy E. Campbell, Andrew Williamson, Arran D. Dokal

Kinomica Ltd, Alderley Park, Macclesfield, United Kingdom; COI: Authors affiliated with Kinomica Ltd are employees, co-founders and/or shareholders of the company.

Introduction

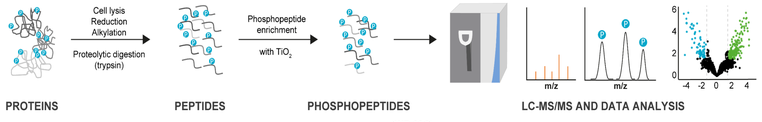
- Phospho-proteomic analysis by mass-spectrometry is a powerful and widely used approach for biomarker discovery where different acquisition methods are typically employed eg DDA, DIA
- There are few, if any, tools which allow the interactive discovery and quantification of potential peptides from multiple techniques and vendor formats
- Here, we describe a unified analytical phospho-proteomics pipeline that allows the identification and quantification of phosphorylated peptides using DDA, DIA and combined methods
- We demonstrate how these outputs, along with interactive exploration of the data, support downstream analyses (including machine learning) as part of a biomarker discovery pipeline

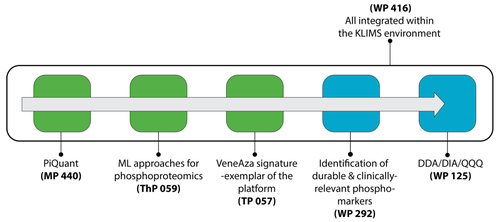
Methods

- For the analysis of our clinical and technical datasets we have developed a web based system
  - Supports multiple proteomics search engines eg. DIANN, Mascot and vendor data formats
  - All file conversions/manipulations are carried out automatically as appropriate
- Analytical parameters and search results are stored within a SQL database
  - Facilitates data tracking, quality control and further processing
  - When new data is acquired, an analysis can be automatically re-run
  - Multiple searches and downstream processing with different parameters are easy to perform with newly and previously acquired data
- DIA, DDA and Triple-Quadrupole data is quantified using in-house software (Piquant and au2select)
- These modules form part of our pipeline system ***(Fig 1),*** again all outputs are stored within our database
  - Peptide sequence, modifications and proteins are easily queried
  - The database contains over 50 million peptides, 7000 datafiles and 7500 separate analyses
- Pipeline tools are defined in the KLIMS database
  - Adding new analysis tools is as simple as filling in the database definitions
  - The web user interface for the tool may also be defined in a similar way
- Refinement of peptide identification and quantification is facilitated using a novel data viewer which integrates multiple data types

Fig 1– KLIMS system architecture

KLIMS is the basis of the KScan® workflow





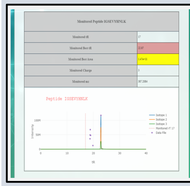
Other aspects of the full KScan® workflow are covered in **several other ASMS 2024 posters:**

Pipeline Analysis and Workflow

a) KLIMS General Workflow

Data Input

- Upload data from local files, instruments or PRIDE
- Files converted automatically for later analysis



Quality Control

- Constant measure of parameters eg number of spectra, TIC etc.
- Specified peptides are monitored to ensure consistency of retention times etc.
- Data files checked for presence of over 500 contaminants

Search Submission

- Multiple datasets and search engines can be selected simultaneously
- Data set changes (eg addition of extra files) are automatically re-searched

Pipeline Analysis

- Multiple analysis may be applied to datasets and search results
- Results stored in relational database
- Additional analysis steps are added automatically if needed eg file conversion

b) Acute Myeloid Leukemia (AML) KLIMS Biomarker Discovery Workflow

- AML is a haematological malignancy with poor prognosis
- Several treatments are approved for AML, but approaches to determine appropriate drugs for patients are problematic
- 182 bone marrow and peripheral blood samples were collected from 138 patients who were grouped into good and bad responders
- Our system was used to produce machine learning models that accurately predict patient response to AML treatments using phospho-proteomic data ***(Fig 2)***

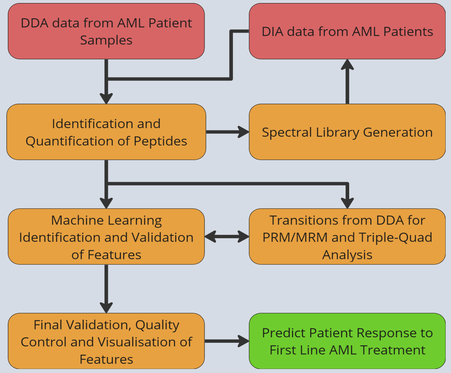
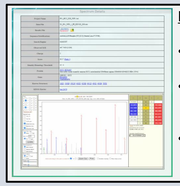


Fig 2– KLIMS workflow used to define predictive model for AML drug response



Peptide and Protein Search

- Fast interactive queries of whole database
- Search by sequence, modification, charge and retention time etc
- Comprehensive links to external databases eg BioGrid

c) Spectral Library Generation

- With appropriate queries it is possible for KLIMS to create targeted spectral libraries or transitions based on sample preparation, type, patient response, disease and drug treatment etc.
- Spectral libraries are produced from a simple user interface that allow various parameters to be set ***(Fig 3)***
- Retention time alignment is carried out using a reference dataset or standard iRT peptides ***(Fig 4)***

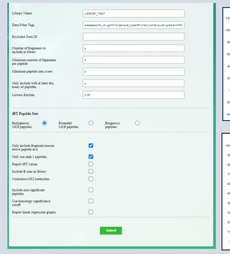


Fig 3– KLIMS interface for spectral library generation

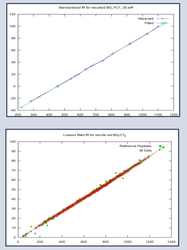


Fig 4 a, b– Retention time alignment

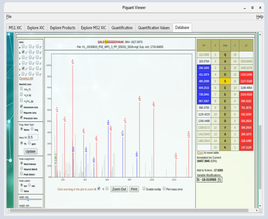


Fig 5 – Interactive viewer allows the user to see the DDA MS/MS spectra used to generate the spectral library

d) Data Viewer

- The viewer allows interactive visualisation of DIA, DDA and Triple-Quad data ***(Fig 6-8)***
- This can be accessed by clicking a link on the web interface

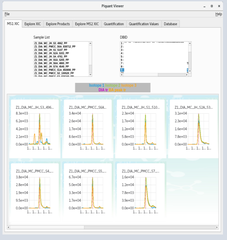


Fig 6– MS1 XIC from a DIA dataset

Fig 7– MS2 fragment XICs from a DIA dataset

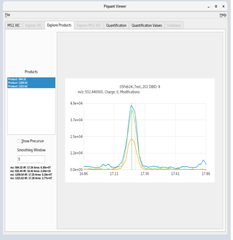


Fig 8– Transition XICs from a Triple-Quadrupole dataset

Conclusions and Comparisons

- KLIMS is a powerful tool which integrates seamlessly the processing of mass spectrometry data from a variety of diverse sources, eg DIA, DDA etc
- Other software such as Proteome Discoverer and Skyline do not have the range of visualisation and analysis tools that are integrated within KLIMS
- A unique, powerful feature of the system is the underlying relational database which allows for data tracking, searching and re-analysis
- Unlike other similar packages, KLIMS is virtually infinitely expandable and allows the simple addition of search engines and analysis software etc
- A simple web interface is used to add new analysis tools and web interface components, these are the stored in the relational database

References

Perkins DN et al.Electrophoresis. 20(18) pp3551-67 (1999).

Demichev, V et al. Nat Methods 17, pp41–44 (2020)

MacLean B et al. Bioinformatics. 26(7) pp966-8 (2010).