



Universiteit
Leiden
The Netherlands

Intelligent workflows for automated analysis of mass spectrometry-based proteomics data

Güler, A.T.

Citation

Güler, A. T. (2022, April 7). *Intelligent workflows for automated analysis of mass spectrometry-based proteomics data*. Retrieved from <https://hdl.handle.net/1887/3281870>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3281870>

Note: To cite this publication please use the final published version (if applicable).

Nederlandse Samenvatting

Massaspectrometrie is een krachtige techniek voor *proteomics* omdat het de benodigde hoge gevoeligheid en snelheid levert voor de analyse van het complexe en dynamische proteoom. Echter, het gebruik van deze techniek levert massaspectrometrische data op die ook zeer complex is, waarvoor een geavanceerde meerstappen analyse nodig is. Het gebruik van wetenschappelijke *workflows* voor de analyse van massaspectrometrie data is niet geheel nieuw. Desalniettemin zijn vernieuwende aanpakken hard nodig in deze tijd van grote (en complexe) data teneinde de *workflows* autonoom en intelligent te maken door middel van metadata, en flexibel en adaptief door een modulair ontwerp. De focus van dit proefschrift ligt op het creëren van intelligente *workflows* voor de analyse en het contextualiseren van massaspectrometrische *proteomics* data. De ontwikkelde *workflows* en *tools* kunnen eenvoudig worden aangepast voor additionele functionaliteiten en hergebruik in verschillende data analyse *workflows*. Het onderzoek beschreven in dit proefschrift beperkt zich niet tot de analyse van massaspectrometrie data, maar behandelt ook methoden voor literatuur onderzoek, welke cruciaal zijn voor het ontwerpen van experimenten, en de interpretatie en contextualiseren van de experimentele resultaten. **Hoofdstuk 1** geeft een overzicht van de kernconcepten van massaspectrometrische *proteomics* en computationele methoden die in de volgende hoofdstukken worden toegepast. Tevens worden uitdagingen en hoe deze momenteel worden aangepakt kort besproken.

Hoofdstuk 2 introduceert wetenschappelijke *workflows* en hun voordelen in meerstappen analyses. Zelfs in hetzelfde subveld kunnen verschillende auteurs verschillende termen gebruiken om te refereren aan dezelfde domein entiteit. Wanneer het literatuuronderzoek handmatig met specifieke trefwoorden uitgevoerd wordt, kunnen deze overlappende studies over het hoofd gezien worden. De *workflows* die hier gepresenteerd worden maken geautomatiseerde literatuuronderzoeken eenvoudiger zonder te verdwalen in complexe bibliometrische methoden. Het snel verkrijgen van een overzicht van een veld kan handig zijn voor auteurs bij het uitvoeren van interdisciplinaire studies en meta-analyses. Verder kunnen onderzoekers hiermee ook expertiselabs en andere onderzoekers vinden voor samenwerkingen. **Hoofdstuk 3** presenteert meer geavanceerde *workflows* voor

bibliometrie met het vermogen om gebruik te maken van *web services* en statistische analyses uit te voeren op de verzamelde data. Na presentatie van hoe de *web services* geïntegreerd kunnen worden in het Taverna workflowsysteem, worden workflows voor citatie netwerken en biomoleculaire interacties getoond. De output van deze processen kan gevisualiseerd worden in het populaire VOSviewer of Cytoscape, daarbij gebruikmakend van krachtige bestaande *tools*. Literatuuranalyses zijn zeer nuttig voor de interpretatie van de uitkomsten van een studie, deze binnen de context te visualiseren, en een routekaart te verkrijgen voor ontwerp van toekomstige experimenten; dit hoofdstuk bespreekt hoe deze processen geautomatiseerd kunnen worden.

Hoofdstuk 4 presenteert een robuuste tool voor integratie en anatomische visualisatie van kwantitatieve *omics* data van model organismen. Eén vernieuwing van deze *tool* is dat het gebruik maakt van anatomische ontologieën om automatisch anatomische informatie te visualiseren ongeacht de resolutie van de input data. De *tool* beweegt door de anatomische ontologie hiërarchie om het geschikte niveau van orgaan en weefsel details te visualiseren. Vervolgens wordt van de gebruiker een algemeen en simpel data formaat gevraagd voor visualisatie van de data. Aangezien dit data formaat geen op *omics*-gebaseerde aannames maakt, kan data van verschillende *omics* experimenten worden geïntegreerd, en probleemloos worden gevisualiseerd.

Hoofdstuk 5 presenteert een herkalibratie tool die de accuraatheid van de massabepaling verbetert door middel van geautomatiseerde interne kalibratie. Aan de hand van opnieuw gekalibreerde data werden meer peptiden geïdentificeerd met hogere zekerheid. De gemeten massa's zijn gekalibreerd met de accurate peptide identificaties uit de originele data of uit andere metingen van hetzelfde of vergelijkbare samples. Deze techniek maakt gebruik van mzXML metadata om automatisch de meeste geschikte kalibratiefunctie te selecteren voor de gebruikte massaspectrometer. Het output format is hetzelfde als de input dus deze techniek kan in principe eenvoudig worden verbonden met elke *bottom-up proteomics* data analyse *workflow* die gebruik maakt van mzXML.

Tenslotte, in **Hoofdstuk 6**, worden de methoden en concepten zoals data beschikbaarheid en hergebruik, automatisering en data integratie besproken in de

context van het onderzoek dat gepresenteerd wordt in dit proefschrift. Huidige en toekomstige *proteomics* data analyse wordt besproken waarbij inzicht wordt gegeven in wat noodzakelijk is voor het uitvoeren van geautomatiseerde analyses die gebruikmaken van beschikbare data uit publieke archieven met als doel bestaande kennis uit de literatuur te versnellen. Net als veel andere software en ondersteunende *tools* is het workflowsysteem, Taverna, dat wordt gebruikt in dit proefschrift, onderhevig aan veroudering. Echter, de concepten en methoden die in dit proefschrift gepresenteerd worden zorgen ervoor dat deze *tools* hiertegen bestand zijn en bieden een sterke basis voor de toekomst van *proteomics* data analyse.