Mastering data pre-processing for accurate quantitative molecular profiling with liquid chromatography coupled to mass spectrometry

Mitra, Vikram

Dutch Summary (Samenvatting)

Experimenten voor het ontdekken van eiwit biomarkers worden meestal uitgevoerd met behulp van een vloeistof chromatograaf gekoppeld aan een massaspectrometer (LC-MS). Voor de kwantificatie van de analieten, welke geïdentificeerd kunnen worden als biomarkers, worden meestal labelvrije methoden gebruikt. Bij labelvrije kwantificatie wordt er per monstermeting LC-MS(/MS) data verkregen en deze data wordt pas samengevoegd bij de datavoorbewerking. Een nadeel van het op grote schaal uitvoeren van labelvrije LC-MS(/MS) experimenten is dat de verschillende metingen op verschillende dagen, laboratoria en instrumenten plaatsvinden. Om datasets van hoge kwaliteit te produceren worden verschillen op technische niveaus verwijderd voorafgaand aan statistische analyse voor ontdekking van biomarkers. Verschillende workflows software en algoritmes zijn in de afgelopen jaren ontwikkeld om labelvrije LC-MS datasets te kunnen verwerken. Nog steeds is er onderzoek gaande naar innovatieve dataverwerkingsmethoden om de labelvrije LC-MS(/MS) studies naar biomarkers te ondersteunen. Dit proefschrift begint met een objectieve review over de gebruikte LC-MS(/MS) databeschrijving- en gegevensverwerkingsmethoden. In het tweede hoofdstuk worden de twee belangrijkste bronnen onderzocht van verschuiving van pieken die voorkomen in de m/z en retentietijd scheidingscoördinaen en het kwantitatief uitlezen (iontellingen) van LC-MS(/MS) gegevens, namelijk: monotone verschuiving en orthogonaliteit. In dit hoofdstuk wordendehuidige methoden en algorithmes voor correctie van monotone verschuiving en beoordeling van orthogonaliteit beschreven. Dit wordt gevolgd door het identificeren en het voorstellen van kritische overwegingen voor succesvolle uitlijning van retentietijd van een dataset in hoofdstuk drie. Daarnaast worden de succesvolle tijduitlijning criteria besproken. Eén van de voorstellen wordt besproken in hoofdstuk vier. Dit voorstel is een methode om de hoeveelheid identificaties van individuele MS datasets te verbeteren door het gebruik van “identificatie vervoer”. In het laatste hoofdstuk van dit proefschrift wordt de implementatie van de ANOVA – simultane component analyse (ASCA) methode beschreven voor het identificeren van essentiële experimentele factoren die invloed hebben op de kwaliteit van eiwit datasets.

Het ontdekken van biomarkersmet behulp van labelvrije LC-MS(/MS) datasets vereist robuuste statistische toetsen van data matrices. In het begin worden de standaard dataverwerkingsalgoritmes die gebruikt worden om labelvrije LC-MS data voor te bewerken samengevat. Hoofdstuk 2 introduceert de drie belangrijkste dimensies die uit LC-MS(/MS)

data gehaald kunnen worden, namelijk; massa/lading verhouding (m/z), retentietijd (RT) en ion intensiteit. Het verschil in experimentele condities beïnvloedt deze dimensies en heeft effect op de totale kwaliteit van de dataset. In dit hoofdstuk worden in detail de twee types veranderingen besproken, monotone verschuivingen en orthogonaliteit, die optreden in de drie dimensies van enkelvoudige LC-MS(/MS) data. We verwachten dat de monotone verschuivingen kunnen worden gecorrigeerd, terwijl de omvangrijke orthogonaliteit alleen kan worden vastgesteld, omdat de correctie nauwkeurige modellering van de toevallige fysisch-chemische effecten vereist.

De bepaling van de overeenstemming tussen verschillende metingen kan een triviaal probleem zijn als elk peptide dezelfde m/z waarde heeft en precies op dezelfde retentietijd elueert. Vanwege systematische en component-specifieke verschuivingen in de retentietijd dimensie, zullen retentietijd correcties voorkomen bij de uitdagende data voorbewerkingsstap van de labelvrije LC-MS(/MS) datasets. Een groot aantal tijduitlijnmethodes is ontwikkeld voor accurate voorbewerking van LC-MS(/MS) datasets. Deze methodes gaan er standaard van uit dat soortgelijke componenten elueren in dezelfde volgorde, maar ze testen niet of deze aanname klopt. Als deze veronderstelling niet geldig is, betekent het dat uitlijning gebaseerd op de monotone retentietijd functie accuraatheid verliest voor de pieken die een wijziging in de retentietijd hebben. Om dit probleem aan te pakken wordt in hoofdstuk 3 een voorstel gedaan voor een kwaliteitscontrole methode. Deze methode stelt vast of de LC-MS(/MS) datasets kunnen worden uitgelijnd met dezelfde uitlijningsmethode, gebaseerd op statistische tests voor het corrigeren van retentietijd verschuivingen. Het algoritme zal als eerste de aanwezigheid van een adequaat aantal gemeenschappelijke pieken (>~100 gemeenschappelijke piekparen) bevestigen. Daarna wordt bepaald of de waarschijnlijkheid voor het in stand houden van de elutievolgorde van deze gemeenschappelijke pieken voldoende hoog is (> 0.01). Als laatste wordt de retentietijd uitlijning uitgevoerd op twee LC-MS chromatogrammen. Deze procedure is toegepast op LC-MS en LC-MS/MS datasets van twee verschillende inter-laboratorium proteomics analyse studies. Daaruit blijkt dat bij een groot aantal van de pieken in de chromatogrammen verkregen van verschillende laboratoria, de elutievolgorde verandert met aanzienlijke retentietijd verschillen.

Het succes van proteomics experimenten met LC-MS(/MS) zit in de mogelijkheid voor het identificeren en kwantificeren van een grote hoeveelheid peptiden. In de afgelopen jaren zijn verschillende analytische verbeteringen gemaakt voor het vergroten van het aantal peptiden dat kan worden geïdentificeerd in een enkele eiwitanalyse uitgevoerd met een LC-

MS(/MS). Door det hoge complexiteit van het biologische monster kan het identificeren met een hoge resolutie massaspectrometer alleen plaatsvinden in kleine fracties van de detecteerbare peptiden. De eerste reden is dat gecofragmenteerde peptide-ionen aanwezig zijn in hetzelfde geïsoleerde massabereik van de ionen van interesse. Verdere redenen zijn de te lage hoeveelheid van de peptiden, bemonstering van het peptide ionin een laag intesniteitsgebied, post-translationele modificatie van peptiden en peptidesequenties die is niet in de eiwitsequentie database aanwezig zijn. Fragmenten van meerdere peptide ionen tegelijk produceren complexe spectra van fragment ionen, die niet succesvol gekarakteriseerd kunnen worden door zoekalgoritmen doordat ze lage zoekscores opleveren. In Hoofdstuk 4 van dit proefschrift wordt de juistheid van de “identification transfer” methode geëvalueerd. Dit is een methode die PSM identificatiewaardes in een LC-MS(/MS) analyse versterkt. Dit hoofdstuk bevat voorbeeld MS/MS spectra die zijn geproduceerd van gemixte of chimerische spectra of van eptide ionen van lage intensiteit, en tevens wordt aangetoond dat zulke spectra lage spectrale kwaliteit hebben, waarbij pieken ongeïdentificeerd blijven door gebruikelijke database zoek algoritmen. Om spectrale kwaliteit te meten, hebben we een “Xrea” score in onze workflow geïmplementeerd om het percentage van spectra met lage kwaliteit te bepalen. De identification transfer workflow transformeert spectrale identificatie informatie van geïdentificeerde MS/MS spectra naar ongeïdentificeerde spectra gebaseerd op overeenkomsten van m/z en retentietijd waarden van pieken tussen LC-MS(/MS) metingen. Een verhoging van 34.37% in het identificeren van individuele MS datasets kan worden gevonden na gebruik van de “identification transfer” methode. De mogelijkheid om identificatie van ionen met een lage ion intensiteit te transformeren, staat toe dat meer omvangrijke datasets gegeneraliseerd kunnen worden. Dit kan betekenen dat de detectie van significante differentiële regulatie tussen monsters mogelijk wordt.

De totale concentratie van peptiden in de gegeven monsters wordt beïnvloed door een aantal experimentele factoren. Variaties in eiwitten komen voor in de bias van de uitkomst van de gebruikte statistische analyse en validatiestappen. Het is daarom belangrijk om de factoren die kritisch zijn voor een eiwit experiment te controleren en begrijpen. In hoofdstuk 5 worden humane serum monsters, die verkregen zijn in een experimentele design studie, geanalyseerd. Dit geeft weer dat we een screening konden doen naar experimentele factoren zoals de verblijftijd in de autosampler op 4°C, het stoppen van de digestiestap met een zuur of niet, het type bloedbuisje, verschil in hemolyse concentraties van de monsters, verschil in stollingstijd van het bloed in de autosampler van de LC-MS(/MS), het aantal vries-

dooi cycli en verschillende trypsine-eiwit ratio's. Het experimentele design werd gevolgd door een twee niveaus van factorial design met een resolutie van vier. Het design heeft zestien monsters nodig waarbij de belangrijkste effecten geen bijdrage leveren aan een interactie van twee factoren effecten. Data die is voorbewerkt met de Treshold Avoiding Proteomics Pipeline produceert een data-matrix die kwantitatieve informatie bevat van de peptide pieken die waargenomen zijn in de zestien monsters. De voorbewerkte data is gefilterd op basis van een t-toets en een lijst wordt opgesteld met pieken die beïnvloed worden door experimentele factoren. Simulatie studies zijn uitgevoerd om de prestatie van de voorselectie op basis van de t-toetst vast te stellen. De intensiteit van de pieken in de gefilterde voorbewerkte data is geanalyseerd met een ANOVA-ASCA (Simultaneous Component Analysis) methode. ASCA is een multivariate statistische methode die ons toe staat omde belangrijkste experimentele factoren die effect hebben op de kwaliteit van de data te bepalen. Een significante waarde (P-waarde) voor het verschil tussen de instelniveaus van de factoren onthult de factoren die invloed hebben op de gemeten peptide intensiteit. Gevonden is dat de gebruikte trypsine-eiwit ratio, het stop van de trypsine reactie met zuur en verschillende hemolyse gehalten invloed hebben op de kwaliteit van de data. De meest beïnvloede peptide pieken behoren tot een specifieke groep die zijn geïdentificeerd met data verkregen uit het ASCA model. De resultaten geven aan dat de peptiden van het hemoglobine eiwit complex en hemopexin erg beïnvloed worden door het variëren van hemolyse gehalten, terwijl peptiden afkomstig uit Apolipoprotein A-I/A-II en van de zware keten van de Inter-Alpha-Trypsine Inhibitor voornamelijk beïnvloed worden door de trypsine-eiwit ratio. Peptiden die afkomstig zijn uit Alpha-1-antichymotrypsine worden significant beïnvloed door de stop procedure van de digestie met zuur. Experimentele designs van LC-MS(/MS) analyses van complexe eiwitmonsters gevolgd door ASCA analyse staan dus identificatie van de meest invloedrijke factoren toe, en brengen peptiden die het meest worden beïnvloed door één van de factoren aan het licht.

De laatste jaren is er een exponentiële groei aan beschikbare ruwe LC-MS(/MS) data in publieke platforms. Het voordeel van het herdefiniëren en kwantificeren van peptides van deze ruwe data kan de omvang van de identificatie van nieuwe biomarkers vergroten om mechanismen van ziekten te begrijpen. Voorafgaand van de herdefiniëring van zulke nieuwe QC metrics datasets, zal een routinematige voor-filteringsstap moeten worden uitgevoerd om LC-MS(/MS) datasets met een hoge kwaliteit te kunnen analyseren. Echter, succesvolle integratie van deze hoge kwaliteit LC-MS(/MS) datasets verkregen uit variërende platforms, aanwezig in een databank, kunnen lijden aan orthogonaliteit in het LC gebied. Accurate

modellering van orthogonaliteit, met een methode die fysische-chemische informatie gebruikt, zal de onzekerheid van het piek vergelijkingsproces verlagen door het gebruik van de m/z en retentietijd coördinaten van meerdere datasets. Na het minimaliseren van de orthogonaliteit en monotone verschuivingen gebruiken de retentietijd voorspellingsalgoritmes peptide sequentie informatie, zoals moleculaire massa, lading en hydrofobiciteit, om de nauwkeurige retentietijd van een peptide in de LC-MS(/MS) meting te kunnen voorspellen. Met de ontwikkeling van deze accurate retentietijd voorspellingsmodellen is het mogelijk om preciezer de locatie van de pieken te kunnen voorspellen tussen alle chromatogrammen. Dit zorgt voor accuraat corresponderende identificatiestussen datasets. Met minimale of geen retentietijd verschillen tussen de metingen zal de peptide identificatie overdracht methode toestaan dat overdracht tussen verkregen sets preciezer is en meer complete datasets produceert. Datasets met minimale missende datapunten kunnen worden gebruikt voor onpartijdige statistische analyse. Tot slot, analyse van multi-variate eiwit datasets hebben accurate beoordeling van pre-analytische factoren nodig die effect hebben op de monsterkwaliteit en op de peptideconcentraties. Dus door zorgvuldig de factoren die effect hebben op de kwaliteit van de monster te modelleren, kan worden toegestaan dat verschillen in niet gerelateerde technische niveaus worden verwijderd en tevens het vervangen van primaire biologische verschillen.

Momenteel ben ik senior wetenschapper bij Proteome Sciences plc. Mijn rol binnen het bedrijf betreft voornamelijk gegevensverwerking en analyse van grootschalige proteomics en fosfo proteomics datasets. Ik draag bij aan de ontwikkeling van nieuwe algoritmes voor de verwerking van LC-MS (/ MS) data voor zowel etiketvrije en TMT-gelabelde samples in grootschalige klinische studies om moleculaire mechanismen in de ziekte van Alzheimer, hepatocellulair cholangiocarcinoom en pancreaskanker te identificeren.