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

Mitra, Vikram

Chapter 6. Summary and future outlook

Bottom-up proteomics experiments using liquid chromatography coupled to mass spectrometry (LC-MS(/MS)) has become the most widely used proteomics and metabolomics molecular profiling platform. Large-scale label-free LC-MS(/MS) studies performed with a large sample size that might need acquisition across several weeks, in multiple analysis sites or instruments. Quantitative data pre-processing workflows and quality assessment approaches has to be adopted in order to accurately process and integrate the large amount of data acquired in such types of studies (**chapter 1**).

This thesis contributed to the goal of investigating for the first time the presence of orthogonality, and its effect on the accuracy of retention time alignment of LC-MS(/MS) chromatograms in **chapter 3**. The assessment of orthogonality also allows determination of the error in identification transfer and MS1 based peak matching using retention time and m/z coordinates of peaks and features as described in **chapter 4**. The distinction between the correctable monotonic shift and non-correctable orthogonality for the two separation coordinates and the quantitative readout of MS1 data allowed the generalization of this concept in **chapter 2**. This generalisation allows to design new algorithms for monotonic shift correction and orthogonality assessment not restricted to retention time domain, where the orthogonality should be assessed after correction for monotonic shift. Assessment of orthogonality is also important in identifying parameters and conditions, which influence orthogonality. These parameters may then be used to minimise orthogonality leading to development of robust quantitative molecular profiling with LC-MS(/MS) approach. Further improvement of retention time alignment algorithms can be performed by developing quality control approaches, which does not provide one metric for the whole chromatographic pairs, but allows local assessment of the monotonic shift correction accuracy and assessment the genuine orthogonality. It is also important to develop new retention time alignment algorithms that correctly deals with retention time domains, which contain low number of common peaks between chromatograms.

There are many pre-analytical factors, which may influence a molecular profile acquired with LC-MS(/MS) and contains quantitative information on hundreds of thousands of compounds. To this end, **chapter 5** presented the application of multivariate ANOVA – the ANOVA simultaneous component analysis (ASCA) – approach to identify significant pre-analytical factors and peaks affected by factors in the most economical way using minimal number of analysed samples following fractional experimental design principle. In our approach, only

main effects were evaluated due to the fractional design model applied in our study. However, ASCA[1] allows assessment of interactions between factors and allows the inclusion of other types of experimental design models[2]. This can be explored not only for analysis of human serum samples depleted from the 16 most abundant proteins, but also for other proteomics and metabolomics sample types with different molecular complexity.

The quality control approach for assessing orthogonality after correction for monotonic shift is the continuation of the algorithm development work for LC-MS(/MS) pre-processing started in the group of Analytical Biochemistry. This work included the development of new time alignment algorithms[3–5], development and integration of comprehensive MS1 data processing pipelines such as threshold avoiding proteomics pipeline (TAPP)[6] and msCompare[7], and assessment of MS1 data pre-processing pipelines[7] and feature selection approaches[8] using differentially spiked sample sets. Large-scale studies require inclusion of more stringent quality control approaches, which considers that data pre-processing algorithms introduce errors and thus contribute to the overall technical variability. In the previous study using differentially spiked samples and msCompare workflow[7], my colleagues have assessed the overall MS1 quantification error using one score based on the ranks of spiked-in compounds amongst the list of the most differential ones. Stepwise optimisation should be performed both in terms of used parameters and in terms of the choice of applied algorithmic models (e.g. different approaches for peak/feature detection or different type of monotonic shift correction algorithms for retention time and m/z dimensions) to identify and further optimise the underperforming algorithm parts, and parameters.

Finally, one of the most important challenges in LC-MS(/MS) pre-processing is to build flexible workflows. msCompare[7], Taverna[9], Molgenis[10,11] and Galaxy[12,13] offer environments for creation of flexible workflows. These workflow managers take care and efficiently log the uploaded data, processing tools, workflows parameters, manage the processing activities such as job launching, job processes monitoring, results retrieval and therefore contribute to a FAIR (Findable, Accessible, Interoperable and Re-usable)[14] data management principle, which principle was recently fostered by the Dutch Techcenter for Life science (DTL). Galaxy was developed to deal with genomics data[13], but tools and workflows emerge rapidly to process LC-MS(/MS) based proteomics[12,15] and metabolomics[16] data and to integrate multi-omics data layers[12,15]. Additionally Galaxy now has a tool repository called ToolShed[17], which allows sharing of algorithms, modules and complete workflows for MS1 quantification, peptide/protein identification and for quality assessment of the raw and

processed data with other users facilitating dissemination of developed tools and workflows. These workflow managers not only allow integration of proteomics data but also allows successful integration of other omics datasets and support biological/clinical data interpretation. For example it is now possible to develop complex workflows for processing data obtained from proteogenomics clinical studies by integrating genomics and proteomics data, which includes pre-processing modules for genomics and LC-MS/MS data, multi-omics data integration layer and modules for downstream statistical analysis and biological functional data interpretation such as molecular pathway analysis.

Currently, I'm employed as a senior scientist at Proteome Sciences plc. My role within the company primarily deals with data processing and analysis of large-scale proteomics and phospho proteomics datasets. I'm contributing further to the development of novel algorithms for processing LC-MS(/MS) data for both label-free and TMT-labelled samples in large-scale clinical studies to identify molecular mechanisms in Alzheimer's disease, Hepatocellular cholangio carcinoma and pancreatic cancer[18–24].

References

(1) van der Greef, J.; Smilde, A. K.; Smilde, A. K.; Greef, J. van der; der, J. G. van; Smilde, A.; Stahle, L.; Wold, S.; Harrington, P.; Vieira, N.; Espinoza, J.; Nien, J.; Romero, R.; Yergey, A.; den, P. B. van; ter, C. B.; Smilde, A.; Jansen, J.; Hoefsloot, H.; Lamers, R.; Greef, J.; Timmerman, M.; Chang, W.; Thissen, U.; Ehlert, K.; Koek, M.; Jellema, R.; Hankemeier, T.; der, J. G. van; Wang, M.; Jansen, J.; Hoefsloot, H.; der, J. G. van; Timmerman, M.; Smilde, A.; Keppel, G.; Saufley, W.; Tokunaga, H.; Jansen, J.; Hoefsloot, H.; der, J. G. van; Timmerman, M.; West-erhuis, J.; Smilde, A.; Efron, B.; Tibshirani, R.; Sokal, R.; Rohlf, F.; Fisher, R.; Tusher, V.; Tibshirani, R.; Gilbert, C.; Anderson, M.; ter, C. B.; Anderson, M.; Keun, H.; Ebbels, T.; Bollard, M.; Beckonert, O.; Antti, H.; Holmes, E.; Lindon, J.; Nicholson, J.; Heijne, W.; Lamers, R.; van, P. B.; Groten, J.; van, J. N.; van, B. O. *J. Chemom.* **2005**, *19* (5–7), 376–386.

(2) Timmerman, M. E.; Hoefsloot, H. C. J.; Smilde, A. K.; Ceulemans, E. *Metabolomics* **2015**, *11* (5), 1265–1276.

(3) Christin, C.; Smilde, A. K.; Hoefsloot, H. C. J.; Suits, F.; Bischoff, R.; Horvatovich, P. L. *Anal. Chem.* **2008**, *80* (18), 7012–7021.

(4) Christin, C.; Hoefsloot, H. C. J.; Smilde, A. K.; Suits, F.; Bischoff, R.; Horvatovich, P. L. *J. Proteome Res.* **2010**, *9* (3), 1483–1495.

(5) Suits, F.; Lepre, J.; Du, P.; Bischoff, R.; Horvatovich, P. *Anal. Chem.* **2008**, *80*, 3095–3104.

(6) Suits, F.; Hoekman, B.; Rosenling, T.; Bischoff, R.; Horvatovich, P. *Anal. Chem.* **2011**, *83* (20), 7786–7794.

(7) Hoekman, B.; Breitling, R.; Suits, F.; Bischoff, R.; Horvatovich, P. *Mol. Cell. Proteomics* **2012**, *11* (6), M111.015974.

(8) Christin, C.; Hoefsloot, H. C. J.; Smilde, A. K.; Hoekman, B.; Suits, F.; Bischoff, R.; Horvatovich, P. *Mol. Cell. Proteomics* **2013**, *12* (1), 263–276.

(9) Wolstencroft, K.; Haines, R.; Fellows, D.; Williams, A.; Withers, D.; Owen, S.; Soiland-Reyes, S.; Dunlop, I.; Nenadic, A.; Fisher, P.; Bhagat, J.; Belhajjame, K.; Bacall, F.; Hardisty, A.; Nieva de la Hidalga, A.; Balcazar Vargas, M. P.; Sufi, S.; Goble, C. *Nucleic Acids Res.* **2013**, *41* (Web Server issue), W557-61.

(10) Kanterakis, A.; Deelen, P.; van Dijk, F.; Byelas, H.; Dijkstra, M.; Swertz, M. A. *BMC Res. Notes* **2015**, *8* (1), 359.

(11) Swertz, M. A.; Dijkstra, M.; Adamusiak, T.; van der Velde, J. K.; Kanterakis, A.; Roos, E. T.; Lops, J.; Thorisson, G. A.; Arends, D.; Byelas, G.; Muilu, J.; Brookes, A. J.; de Brock, E. O.; Jansen, R. C.; Parkinson, H. *BMC Bioinformatics* **2010**, No. Suppl 12, S12.

(12) Boekel, J.; Chilton, J. M.; Cooke, I. R.; Horvatovich, P. L.; Jagtap, P. D.; Käll, L.; Lehtiö, J.; Lukasse, P.; Moerland, P. D.; Griffin, T. J. *Nat. Biotechnol.* **2015**, *33* (2), 137–139.

(13) Koboldt, D. C.; Ding, L.; Mardis, E. R.; Wilson, R. K.; Ding, L.; Mardis, E.; Wilson, R.; Voelkerding, K.; Dames, S.; Durtschi, J.; Sana, M.; Iascone, M.; Marchetti, D.; Palatini, J.; Galasso, M.; Volinia, S.; Wooley, J.; Godzik, A.; Friedberg, I.; Chistoserdova, L.; Kunin, V.; Copeland, A.; Lapidus, A.; Mavromatis, K.; Gilbert, J.; Dupont, C.; Oinn, T.; Addis, M.; Ferris, J.; Marvin, D.; Hull, D.; Wolstencroft, K.; Stevens, R.; Goble, C.; Ludäscher, B.; Altintas, I.; Berkley, C.; Taylor, I.; Shields, M.; Wang, I.; Harrison, A.; Taylor, I.; Shields, M.; Wang, I.; Harrison, A.; Giardine, B.; Riemer, C.; Hardison, R.; Linke, B.; Giegerich, R.; Goesmann, A.; Deelman, E.; Singh, G.; Su, M.; Blythe, J.; Gil, Y.; Kesselman, C.; Mehta, G.; Vahi, K.; Berriman, G.; Good, J.; Laity, A.; Jacob, J.; Katz, D.; Shah, S.; He, D.; Sawkins, J.; Druce, J.; Quon, G.;

Lett, D.; Zheng, G.; Xu, T.; Ouellette, B.; Reich, M.; Liefeld, T.; Gould, J.; Lerner, J.; Tamayo, P.; Mesirov, J.; Kuehn, H.; Liberzon, A.; Reich, M.; Mesirov, J.; Rowe, A.; Kalaitzopoulos, D.; Osmond, M.; Ghanem, M.; Guo, Y.; Ghanem, M.; Curcin, V.; Wendel, P.; Guo, Y.; Curcin, V.; Ghanem, M.; Goble, C.; Bhagat, J.; Aleksejevs, S.; Cruickshank, D.; Michaelides, D.; Newman, D.; Borkum, M.; Bechhofer, S.; Roos, M.; Li, P.; Roure, D. De; Abouelhoda, M.; Alaa, S.; Ghanem, M.; Elmroth, E.; Hernandez, F.; Tordsson, J.; Aalst, W. van der; Hofstede, A.; Kiepuszewski, B.; Barros, A.; Shields, M.; Ludäscher, B.; Weske, M.; Mcphillips, T.; Bowers, S.; McPhillips, T.; Bowers, S.; Ludäscher, B.; Kahn, G.; Macqueen, D.; Ashford, E.; David, L.; Rice, P.; Longden, I.; Bleasby, A.; Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Altschul, S.; Gish, W.; Miller, W.; Myers, E.; Lipman, D.; Altschul, S.; Madden, T.; Schäffer, A.; Zhang, Z.; Schwartz, S.; Wagner, L.; Miller, W.; Thompson, J.; Higgins, D.; Gibson, T.; Notredame, C.; Higgins, D.; Heringa, J.; Edgar, R.; Kerk, D.; Templeton, G.; Moorhead, G.; Pond, S. K.; Wadhawan, S.; Chiaromonte, F.; Ananda, G.; Chung, W.; Taylor, J.; Nekrutenko, A.; Team, T.; Huson, D. D.; AF, A.; Qi, J.; Schuster, S.; Venter, J.; Remington, K.; Heidelberg, J.; Halpern, A.; Rusch, D.; Eisen, J.; Wu, D.; Paulsen, I.; Nelson, K. *Brief. Bioinform.* **2010**, *11* (5), 484–498.

(14) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. .; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B.; Roche, D. G.; Kruuk, L. E. B.; Lanfear, R.; Binning, S. A.; Bechhofer, S.; Benson, D. A.; Berman, H.; Henrick, K.; Nakamura, H.; Wenger, M.; Crosas, M.; White, H. C.; Carrier, S.; Thompson, A.; Greenberg, J.; Scherle, R.; Lecarpentier, D.; Martone, M. E.; White, E.; Sandve, G. K.; Nekrutenko, A.; Taylor, J.; Hovig, E.; Wolstencroft, K.; Bauch, A.; Sansone, S.-A.; González-Beltrán, A.; Maguire, E.; Sansone, S.-A.; Rocca-Serra, P.; González-Beltrán, A.; Harland, L.; Groth, P.; Berman, H. M.; Bourne, P. E.; Berman, H. M.; Watenpaugh, K.; Westbrook, J. D.; Fitzgerald, P. M. D.; Rose, P. W.; Kinjo, A. R.; Gutmanas, A.; Starr, J.; Musen, M. A. *Sci. Data* **2016**, *3*, 160018.

(15) Fan, J.; Saha, S.; Barker, G.; Heesom, K. J.; Ghali, F.; Jones, A. R.; Matthews, D. A.; Bessant, C. *Mol. Cell. Proteomics* **2015**, *14* (11), 3087–3093.

(16) Bundy, J. G.; Davey, M. P.; Viant, M. R.; Sharma-Oates, A.; Viant, M. R.; Bundy, J.; Davey, M.; Viant, M.; Schadt, E.; Lamb, J.; Yang, X.; Zhu, J.; Edwards, S.; Guhathakurta, D.; Valdes, A.; Glass, D.; Spector, T.; Whitehead, A.; Nicholson, J.; Connelly, J.; Lindon, J.; Holmes, E.; Castillo, S.; Gopalacharyulu, P.; Yetukuri, L.; Orešič, M.; Dunn, W.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Haug, K.; Salek, R.; Conesa, P.; Hastings, J.; Matos, P.; Rijnbeek, M.; Southam, A.; Payne, T.; Cooper, H.; Arvanitis, T.; Viant, M.; Weber, R.; Viant, M.; Weber, R.; Southam, A.; Sommer, U.; Viant, M.; Beckonert, O.; Keun, H.; Ebbels, T.; Bundy, J.; Holmes, E.; Lindon, J.; Giacomini, F.; Corguillé, G.; Monsoor, M.; Landi, M.; Pericard, P.; Pétéra, M.; Blanc, A.; Brooke, J.; Fellows, D.; Soldati, M.; Pérez-Suárez, D.; Marassi, A.; Orvis, J.; Crabtree, J.; Galens, K.; Gussman, A.; Inman, J.; Lee, E.; Wolstencroft, K.; Haines, R.; Fellows, D.; Williams, A.; Withers, D.; Owen, S.; Goecks, J.; Nekrutenko, A.; Taylor, J.;

Sheynkman, G.; Johnson, J.; Jagtap, P.; Shortreed, M.; Onsongo, G.; Frey, B.; Kirwan, J.; Weber, R.; Broadhurst, D.; Viant, M.; Payne, T.; Southam, A.; Arvanitis, T.; Viant, M.; Smith, C.; Want, E.; O'Maille, G.; Abagyan, R.; Siuzdak, G.; Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H.; Hrydziuszko, O.; Viant, M.; Parsons, H.; Viant, M.; Luan, H.; Meng, N.; Liu, P.; Feng, Q.; Lin, S.; Fu, J.; Luan, H.; Meng, N.; Liu, P.; Feng, Q.; Lin, S.; Fu, J.; Blankenberg, D.; Kuster, G.; Bouvier, E.; Baker, D.; Afgan, E.; Stoler, N. *Metabolomics* **2009**, *5* (1), 3–21.

(17) Blankenberg, D.; Von Kuster, G.; Bouvier, E.; Baker, D.; Afgan, E.; Stoler, N.; Taylor, J.; Nekrutenko, A.; Nekrutenko, A. *Genome Biol.* **2014**, *15* (2), 403.

(18) Britton, D.; Zen, Y.; Quaglia, A.; Selzer, S.; Mitra, V.; Lößner, C.; Jung, S.; Böhm, G.; Schmid, P.; Prefot, P.; Hoehle, C.; Koncarevic, S.; Gee, J.; Nicholson, R.; Ward, M.; Castellano, L.; Stebbing, J.; Zucht, H. D.; Sarker, D.; Heaton, N.; Pike, I.. *PLoS One* **2014**, *9* (3), e90948.

(19) Russell, C. L.; Mitra, V.; Hansson, K.; Blennow, K.; Gobom, J.; Zetterberg, H.; Hiltunen, M.; Ward, M.; Pike, I. *J. Alzheimer's Dis.* **2016**, *55* (1), 303–313.

(20) Russell, C. L.; Heslegrave, A.; Mitra, V.; Zetterberg, H.; Pocock, J. M.; Ward, M. A.; Pike, I. *Rapid Commun. Mass Spectrom.* **2017**, *31* (2), 153–159.

(21) Zen, Y.; Britton, D.; Mitra, V.; Pike, I.; Sarker, D.; Itoh, T.; Heaton, N.; Quaglia, A. *Histopathology* **2014**, *65* (6), 784–792.

(22) Böhm, G.; Prefot, P.; Jung, S.; Selzer, S.; Mitra, V.; Britton, D.; Kuhn, K.; Pike, I.; Thompson, A. H. *J. Proteome Res.* **2015**, *14* (6), 2500–2510.

(23) Liang, H. C.; Russell, C.; Mitra, V.; Chung, R.; Hye, A.; Bazenet, C.; Lovestone, S.; Pike, I.; Ward, M. *J. Proteome Res.* **2015**, *14* (12), 5063–5076.

(24) Zen, Y.; Britton, D.; Mitra, V.; Pike, I.; Heaton, N.; Quaglia, A. *Histopathology* **2016**, *68* (6), 796–809.