Evaluation of spectral libraries and sample preparation for DIA-LC-MS analysis of host cell proteins

A case study of a bacterially expressed recombinant biopharmaceutical protein

Heissel, Søren; Bunkenborg, Jakob; Kristiansen, Max Per; Holmbjerg, Anne Fich; Grimstrup, Marie; Mørtz, Ejvind; Kofoed, Thomas; Højrup, Peter

Published in:
Protein Expression and Purification

DOI:
10.1016/j.pep.2018.03.002

Publication date:
2018

Document version:
Accepted manuscript

Title

SWATH-MS based absolute quantification of host cell proteins associated with a biopharmaceutical protein.

Authors

Søren Heissel 1), Jakob Bunkenborg 2), Max Per Kristiansen 3), Anne Fich Holmbjerg 3), Marie Grimstrup 2), Ejvind Mørtz 2), Thomas Kofoed 2), Peter Højrup* 1)

*Corresponding author: php@bmb.sdu.dk

1) Department of Biochemistry and Molecular Biology, University of Southern Denmark, 5230 Odense M, Denmark

2) Alphalyse A/S, 5220 Odense SØ, Denmark

3) Statens Seruminstitut, 2300 Copenhagen S, Denmark

Abstract

Recombinantly expressed biopharmaceutical proteins often undergo a series of purification steps with the aim of removing contaminating material. Depending on the application of the protein, there are various requirements for the degree of purity, but host cell proteins (HCP) will in general remain in small amounts. We describe a method using microflow LC-MS/MS with extensive, multidimensional spectral libraries followed by 1D SWATH-acquisition for fast and robust absolute quantification across several orders of magnitude of the HCPs associated with a biopharmaceutical protein. The HCP content of this BP has never been studied by LC-MS, and the calculated purity was found to be in agreement with ELISA data. A set of evaluated reference proteins was spiked into the sample before and after digestion and evaluated for absolute label-free quantification through a Top3 quantification strategy. The label-free quantification was further evaluated using stable isotope standard (SIS) peptides from selected HCPs. The methodology provides a general, fast and cost-efficient identification and label-free absolute quantification strategy, which can easily be transferred to different samples for HCP characterization and purity assessment. The dynamic range of the HCPs was established, and it was shown that 50 % of the molar HCP amount originated from less than 10 proteins. The developed strategy can be deployed during development processes or quality control for other biopharmaceutical proteins to evaluate the amounts of distinct HCPs and to identify possible risk factors and antigenic protein contaminants. The speed of the analysis (a few days) along with the easy setup allow for evaluation of the purification process in parallel to process development.

Introduction

Recombinant biopharmaceutical proteins are often obtained as a crude mixture containing culture medium, DNA, lipids, proteins and other biological components[1] and are therefore subjected to a series of purification steps, ultimately yielding the purified biopharmaceutical protein. The protein impurities which almost always remain, called host cell proteins (HCPs), are of special interest, since these may cause unwanted immunogenic responses or possess proteolytic activity[2,3]. Manufacturers are required to monitor and document the HCP content throughout production; however, there are no regulations on the analytical methods applied. HCP content is usually studied using immunogenic methods such as ELISA assays[4–6] where polyclonal antibodies are raised through inoculation using proteins produced by the null cell line, i.e. the host cell-line which does not produce the biopharmaceutical protein.

In recent years, mass spectrometry (MS) has increasingly been applied for HCP-based workflows[3,5,7–9], offering identification of individual proteins which may aid in identifying possible risk factors and evaluate the purification process. The main hurdles when analyzing HCPs in biopharmaceutical protein samples by MS are the complexity and the high dynamic range, as the drug protein is often >100 times more abundant than the most abundant HCP. Due to the many challenges, a robust strategy for quantification of HCPs is needed. The biopharmaceutical protein used in this study (henceforth referred to as BP) is produced by Statens Serum Institut (DK) for diagnostic testing in the field of tuberculosis (TB) infections and was chosen, as it has been routinely studied by ELISA assays, providing a benchmark for the results obtained from the LC-MS/MS analyses. The BP is purified by tangential flow filtration (TFF) which removes contaminants below ~10 kDa. Due to remaining non-proteinous buffer components, which may interfere with the analyses, extensive sample clean-up is required before introducing the sample to the mass spectrometer. The most common workflow consists of protein purification, digestion in a urea-containing buffer and clean-up by solid phase extraction. Another strategy, which has shown great potential is the suspension trapping strategy, published by Zougman and coworkers[10], where proteins are precipitated onto a quartz filter trap and desalted before digestion directly on-trap. For absolute quantification of many proteins in a solution, several label-free strategies exist, which are based on spiking the sample with reference proteins in known amounts, providing cost-efficient and sufficiently accurate strategies for absolute protein quantification[11–14]. A number of studies by Doneanu and coworkers show how multidimensional LC-MS workflows can provide accurate absolute quantitative results of HCPs down to very low amounts[15,16], however, by applying multidimensional workflows, analysis time is significantly increased. Data dependent acquisition (DDA) was used to identify HCP peptides with high confidence and to select which transitions to monitor, thus functioning as spectral libraries. Quantitative data were recorded using the data independent

acquisition (DIA) strategy *Sequential Window Acquisition of all Theoretical* Mass *Spectra* (SWATH)[17], where ions are fragmented untargeted based on large, predefined selection windows. DIA has been reported to allow for accurate MS2 quantification of several thousands of peptides across a large dynamic range[18].

In this study we characterized the HCP content of a BP using DDA and SWATH-MS. Tryptic digests of the raw harvest media (henceforth referred to as P10), and the final commercial product (henceforth referred to as P15) were evaluated for generating the most comprehensive spectral library. P10 and P15 were analyzed both with and without pre-fractionation of the tryptic peptides by high-pH RP-HPLC, which has previously proven a feasible option for increasing the chromatographic power in proteomic workflows[19,20]. The generated spectral libraries can be used for analyses several months apart, making this method highly suitable for monitoring BPs during production, requiring only quantitative measurements. All quantitative analyses were recorded on a 1-dimensional separation of P15. Absolute HCP quantification was performed by spiking with a set of reference proteins in known amounts and evaluated using stable isotope standard (SIS) peptides originating from a set of selected HCPs. We evaluated acetone precipitation along with MCX purification and suspension trapping for establishing a sample preparation strategy which produces accurate quantitative results. Non-specific cleavage searches of the tryptic digested samples were performed in order to identify low-abundant non-specific products of the biopharmaceutical protein, which may be misinterpreted as HCP peptides and thereby introduce false positives. The HCPs associated with this BP have not previously been characterized by LC-MS, and this study shows that a catalog of potential immunogenic targets can be created fast and cost-efficient. The general setup developed in this study proved to be sensitive, quantitative and robust. Compared to ELISA assays, quantitative data for individual proteins were obtained. This additional information can speed up the development process and improve/simplify the down-stream process. The unique information regarding individual HCPs can improve the regulatory approval process, and in the end increase the possibility for bringing new safe drugs and vaccines to the market.

Experimental

All chemicals, proteins and reagents were purchased from Sigma-Aldrich, St. Louis, MO, USA, unless otherwise specified. The BP was provided by Statens Serum Institut (Copenhagen, Denmark).

Amino acid analysis

For determination of protein amounts, Amino acid analysis (AAA) was applied. Aliquots of protein- or peptide-containing samples were lyophilized by vacuum centrifugation and hydrolyzed using 6M HCl under vacuum at 110°C for 20 hours[21]. The samples were then analyzed by a Biochrom 30+ analyzer (Biochrom

Ltd., Cambridge, UK), using post-column derivatization by ninhydrin. The protein concentration in the BP sample was determined to be 0.336 µg/µL upon delivery.

Acetone precipitation strategy

Proteins from 300 µL (100.8 µg) of BP were precipitated by the addition of 4*volumes of ice-cold acetone followed by incubation overnight at -18°C. After incubation, the sample was centrifugated at 14000*g using a Minispin table-top centrifuge (Eppendorf) followed by removal of the supernatant. The pellet was washed with an ice-cold mixture of acetone/ethanol/water (2:2:1) followed by another centrifugation at 14000*g. The washing step was performed twice. The sample was lyophilized and redissolved in 7M urea, 2M thiourea in 50 mM ammonium bicarbonate (AMBIC). Cysteines were reduced using 10 mM dithiothreitol (DTT) at RT for 45 minutes and subsequently alkylated using 23 mM iodoacetamide (IAA) for 45 minutes at RT in the dark. An additional aliquot of DTT was added to quench excess IAA. The sample was diluted to a final concentration of 4.5 M urea, and proteins were digested by the addition of 44 mAU Lysyl Endopeptidase (Wako Chemicals, Japan) followed by incubation for 3 hours at RT. The sample was diluted to 9 times the starting volume using 50 mM AMBIC and modified trypsin (Promega) was added at 4 % w/w and incubated at 37°C for 20 hours. Digestion was stopped by the addition of trifluoroacetic acid (TFA) to a final concentration of 1 %. Peptides were purified by MCX mixed-bed cartridges (Waters, MA, USA) according to manufacturer specifications.

Suspension trap purification

50 µg of the BP sample was lyophilized and resolubilized in 2% SDS, 50 mM AMBIC. Cysteines were reduced and alkylated as previously described, with the exception of being reduced at 57°C. Proteins were purified as described by Zougman, Selby and Banks[10], with minor modifications. Briefly; the quartz traps were constructed using micro-quartz fiber filter paper (Ahlstrom Munktell, Sweden) stomped into a D200 pipette tip (Gilson) using a syringe. No C18 trap was included. A solution of 90 % methanol, 50mM AMBIC was used to precipitate the proteins. Tryptic digestion was done first on the filter at 47°C for 60 minutes using 1 % w/w modified trypsin (Promega), followed by elution to a low-binding Eppendorf tube using 50 mM AMBIC, and further digestion overnight with 2% w/w non-modified trypsin (Sigma) at 37°C.

High-pH fractionation

30 µg of tryptic digested P10 and P15 were dissolved in 50 µL 0.1 % triethylamine (TEA) and separated using a 1260 Infinity HPLC (Agilent Technologies, Germany) equipped with a 2.1*150mm Kinetex Evo C18 column (Phenomenex, USA) operating at 45°C and 200 µL/min. Solvent A consisted of 0.1 % TEA in water and solvent B consisted of 90 % ACN. 0.1 % TEA in water. Solvent B levels were kept at 2 % for 2 minutes,

and then increased to 40 % over 17 minutes, followed by a sharp increase to 90 % over 6 minutes, where it was kept for 5 minutes. 10 fractions were collected in 2-minute intervals and concatenated to yield 5 fractions.

Preparation of protein quantification standards

The reference-protein mix for quantification was constructed as a ladder, consisting of several proteins with varying MW in various amounts. All proteins were purchased from Sigma Aldrich. The ladder consisted of horse myoglobin, bovine carbonic anhydrase II, human serotransferrin and bovine alcohol dehydrogenase. Proteins were dissolved individually in 6 M urea, 50 mM TEAB (Fluka Analytical, St. Louis, MO, USA). Reduction and alkylation was carried out as described above. The protein ladders were constructed using two different strategies; one where the appropriate amounts of reference proteins (spanning several orders of magnitudes) were added to the BP samples, before any other sample preparation followed by protein purification, digestion and clean-up. This strategy was used for calculating the absolute amounts of the HCPs. The other strategy was to digest the reference proteins and purify the peptides individually, determine the amounts by AAA, mix the ladder and spike it into the sample immediately before injection into the MS. This strategy was used to evaluate the accuracy of the quantification strategy. The individual reference proteins were digested as described in the "acetone-MCX" section. All protein digests were purified using Oasis HLB cartridges (Waters, MA, USA) according to the manufacturer's specifications. Another mixture was created containing peptides from all reference proteins in near-equal amounts, and used for spectral library generation. The Hi3 quantification standard was purchased from Waters (Product number 186006011) and consisted of the 6 best ionizing peptides from the tryptically digested rabbit Phosphorylase B[22]. The lyophilized peptides were redissolved in UHQ H_2O, aliquoted and frozen for subsequent experiments.

Preparation of stable isotope standard peptides

A total of 21 tryptic peptides from 6 selected HCPs (A2RHZ5, A2RKI9, A2RMQ7, P0A3J1, A2RMV9 and A2RL56) were synthesized with a heavy-isotope lysine or arginine (JPT peptides, Berlin, Germany) and delivered non-purified and non-quantified. The 6 HCPs were selected based on their individual estimated quantity in order to span the entire dynamic range of the HCPs. All peptides were purified by RP-HPLC and quantified by AAA. See Supporting Information for procedure and AAA results.

LC-MS/MS analyses

The peptides were separated using an Eksigent 425 microflow HPLC system (Sciex, Warrington, UK) equipped with a reversed phase CSH C18 column (0.3mm, 100mm, 1.9µm, Waters) and analyzed using a

TripleTOF 6600 Q-TOF mass spectrometer (Sciex, Warrington, UK) run in high-sensitivity mode. The flow rate was 5 µL/min and the column temperature was kept at 55°C. Solvent A consisted of 0.1% formic acid (FA) in water. Solvent B consisted of 0.1% FA in acetonitrile. During the 80 minute gradient the solvent B level went from 2 to 29 % in 60 minutes, from 29 to 55 % in 5 minutes, from 55 to 75 % in 2 minutes and remained at 75 % for 5 minutes and then returned to 2 % over 2 minutes and stayed at 2 % for the remaining 6 minutes. DDA analyses were carried out by a full MS1 spectrum with an accumulation time of 250 ms, followed by fragmentation of the top 25 most intense multiply charged ions. MS2 spectra were acquired with an accumulation time of 50 ms. The SWATH analyses consisted of a parent spectrum, followed by 52 MS/MS acquisitions with variable window sizes from m/z 349.5-1700. The window distribution consisted of a single window of 21 m/z from 349.5-370.5, followed by a single window of 15 m/z from 369.5-384.5, followed by 33 windows of 13 m/z stretching from 383.5-780.5 m/z. This was followed by 9 windows of 26 m/z from 779.5-1005.5 and 6 windows with a size of 51 m/z from 1004.5-1305.5, two windows with a size of 101 m/z from 1304.5-1505.5 and a single window of 195.5 m/z from 1504.5-1700. An accumulation time of 33 ms was utilized and the entire cycle time was approximately 1.9 seconds. The default charge state was set to 2+. All samples were analyzed in triplicates.

Database searching

The DDA files were recalibrated and converted to mgf-format using Protein Pilot version 5.0.1.0, 4895 (Sciex, Warrington, UK) and searched using the search engine MASCOT (Version 2.2.04, Matrix Science Ltd., London, UK)[23]. A mass tolerance of 10 ppm on the MS1 level and 0.1 Da on the MS2 level was applied. Oxidation of methionine was included as variable modification. Carbamidomethylation of cysteine was applied as a static modification. For database searching, the complete proteome of *Lactococcus lactis* strain MG1363 was retrieved from www.uniprot.org (2383 sequences), and the drug protein sequence was added manually. For experiments including protein spikes, the necessary proteins were added manually. All search results had a threshold 1 % FDR on the PSM level and were exported as MASCOT .DAT files for further analysis. For non-specific searches, Andromeda was utilized through MaxQuant[24] (v. 1.5.3.30). Peptide lengths of 7-25 amino acids were allowed. The data was searched with a 1 % FDR on PSM and protein level.

Label-free absolute quantification

The MASCOT search results were imported into Skyline version 3.5.0.9319[25] and used as spectral libraries. Only unique peptides were processed for quantification. The same mass deviations and parameters as the MASCOT search were applied. Quantification was performed on the P15 sample, spiked with reference proteins, recorded using DIA. The peak areas for the top 5 transitions (y- and b-ions) per peptide were exported. Subsequent data analysis was performed using R version 3.3.1[26] and Microsoft Excel version 14.0.7173.5000. Entries with a library dot product (dotp) < 0.7 and an isotope dot product (idotp) < 0.85 were discarded in order to filter out low-quality entries. Peak areas for all transitions were accumulated for each individual peptide to yield a peptide response. All proteins identified with less than 3 peptides were not included in the quantitative analyses. A mean intensity (summed peptide intensity / number of peptides, MeanInt) and Top3 intensity (summed intensity for the 3 most intense peptides) were exported. The signal and amount of the reference proteins were used for deriving absolute HCP amounts. For the comparison with heavy peptide-based quantification, a digest of P15 was purified by acetone-MCX and spiked with purified peptides.

Absolute quantification using heavy peptides

The heavy peptide standards were mixed in equimolar amounts and spiked into 2 µg of BP digest (acetone precipitated and MCX purified) in amounts ranging from 5 fmol to 1 pmol. All dilution steps were analyzed in triplicates using SWATH and quantified using Skyline. The intensity of the three most intense y-ions were summed and used for quantification. The observed responses from the native peptides were normalized against the heavy responses. Peptides which produced poor linearity throughout the dilution steps were discarded.

Results and discussion

Identified HCPs

In order to achieve a complete HCP characterization, it is important to have extensive identification of the peptides belonging to each HCP. Tryptic digests of P10 and P15 (1D and 2D) were analyzed by LC-MS. Using the 1-dimensional strategy, 608 HCPs were identified from P10, while 256 HCPs were identified from P15. P10, being the raw harvest media, had not been subjected to any purification and it is therefore expected that many HCPs are present at higher amounts compared to P15. The high-pH fractionations lead to a significant increase in the number of identified HCPs (1175 HCPs in P10 and 441 HCPs in P15). As seen from Figure 1A, a large subset of proteins were only identified in P10, representing HCPs which are removed to such a high degree during purification that they remain undetectable in the final product.

Utilizing different spectral libraries

The search results were used for generation of spectral libraries and applied to a 1D separated P15 digest, recorded using DIA in order to study the effects on quantification. The number of quantifiable proteins was significantly higher when using the 2D-separated samples as libraries. By combining all four strategies in the spectral library, additional peptides were identified, thereby allowing quantification of additional HCPs, which were not quantifiable using the individual libraries. The abundances of these additional proteins were at least two orders of magnitude below the most abundant HCP (as determined by their Top3 values) and therefore these HCPs did not contribute significantly to the combined HCP amount. These proteins do, however, still represent possible immunogenic targets, which are essential to monitor. Figure 1B shows the distribution of quantifiable proteins. The benefits of an extensive spectral library have previously been described[27,28] and the presented results suggest that a combination of the various libraries provides the most extensive library, which may be used for comprehensive quantification of HCPs in future batches of the BP, requiring only a 1-D separation of a tryptic digest, thereby saving analysis time.

Reference proteins

The PhosB standard was spiked into a BP P15 digest in amounts ranging from 500 amol to 50 pmol per injection, and the Top3 signal of PhosB was normalized against the Top3 signal of a constant HCP in order to establish the linear range. The multi-point calibration curve was compared to a single-point calibration, based on extrapolation of the signal obtained with 500 fmol PhosB. The multi-point calibration curve was identical to the single-point calibration curve between 5 fmol and 5 pmol on-

column, which was determined as the linear range. The calibration plots are presented in Supporting Information Figure S-1.

A number of diverse protein candidates were evaluated in order to comprise a representative spike-in ladder. Linearity of the protein ladders was investigated using the Top3 method[14] and the MeanInt method[13]. The final ladder consisted of human serotransferrin, bovine carbonic anhydrase 2, horse alcohol dehydrogenase and horse myoglobin. These proteins were selected based on their varying MW, which in turn resulted in a varying number of peptide candidates for quantification. Other proteins were discarded either due to too few peptide candidates or low purity (e.g. β-casein). The quantification strategies were evaluated using the purified peptides spiked into the sample immediately before injection. The spiked proteins spanned 20 – 600 fmol on-column per injection. Calibration curves were constructed using both Top3 and MeanInt. The Top3 method showed the highest degree of linearity across the entire range (r^2=0.999) and was therefore chosen as the method of quantification. When using MeanInt, the amount of large proteins tended to be underestimated while overestimating the amounts of small proteins. The Top3 strategy may involve a "one-point calibration" strategy using the signal and amount of a single protein, followed by extrapolation. However, if this protein is not sufficiently digested, or some of the peptides are partially lost during processing, the signal and thereby the calibration curve is greatly affected. By including several proteins in the calibration curve, variations between proteins and in the sample preparation process may be compensated for. Excellent linearity was observed both when spiking with purified peptides and intact proteins. Spiking with intact proteins produced an r^2 of 0.996 for the acetone-MCX experiment and 0.976 for the suspension trap experiment with amounts ranging from 40 fmol to 2.4 pmol per 4 µg of sample. Supporting Figure S-1 shows the linearity of the different strategies.

Signal-to-amount ratios (Top3 signal/fmol) were calculated for all individual reference proteins in order to investigate possible variations between the reference proteins. When spiking with purified peptides, the ratios of all four proteins were consistent throughout the entire range, not varying greatly from each other nor from the ratio obtained using all proteins. When spiking with the intact reference proteins prior to any sample preparation, the 3 highest abundant proteins were found to produce similar ratios in the acetone-MCX experiment, but serotransferrin, which was spiked at 10 fmol per µg of sample, produced a much higher ratio. The deviation of the ratios suggests a difference in recovery rates as a result of inconsistencies during either the precipitation or digestion, but the remaining proteins allowed for compensation of this. The suspension trap experiment showed a higher degree of variance between the ratios of all reference proteins throughout the

entire range. It is possible that these variations are a result of the smaller precipitated amount, the precipitation time or the temperature, compared to the acetone-MCX strategy. The overall ratio produced from all reference proteins is quite consistent with the other experiments, however, inconsistent recovery rates will complicate the quantification. The acetone-MCX and suspension trap Top3 values were normalized against an internal reference protein and based on the correlation of the results, the recovery rates of the majority of HCPs was found to be consistent. Figure 2 shows the individual and collected ratios of the reference proteins for the different spiking strategies.

Evaluation of calculated amounts by SIS-peptides

Absolute protein quantification by spiking with heavy-isotope peptides is a very accurate strategy and is often performed using MRM on a triple-quadrupole instrument[29]. Recent studies show that SWATH-acquisition is, to a high degree, comparable to MRM in obtaining accurate quantitative results[17,30,31]. Of the 6 target-HCPs, A2RHZ5, A2RKI9, A2RMQ7 and P0A3J1 were repeatedly quantified with 3 or more peptides in the label-free experiment. The protein A2RL56 could not be identified in the analyses. A2RMV9 was only identified with two peptides, and did therefore not meet the requirements for quantification. A2RMV9 was however still included exclusively in the comparative analysis presented in Table 1, for evaluation purposes (Instead of a Top3 signal, the intensity of the two identified peptides were used to construct a "Top2 signal"). When calculating the absolute amounts of the SIS peptides, all peptides belonging to the same protein were found to be present in the same range, differing by a factor of less than 3.5. This variation may be due to inconsistencies in the sample preparation protocol, incomplete digestion[32] or loss of the SIS-peptide, either by contamination or lack of solubility. All the calculated peptide amounts along with the protein amount calculated by Top3 are listed in Table 1 together with the CV between the methods. A high level of agreement between the Top3 strategy and most of the SIS-peptides was observed. Only the protein A2RMQ7 produced results which were not in agreement with the Top3 strategy. This protein was, however, only identified in a single replicate and only by 3 peptides. A higher CV was generally observed for low-abundant proteins and proteins identified with a low number of peptides, which is in accordance with previous findings[33]. This suggests that increasing the number of peptide candidates also increases the accuracy and reliability of the Top3 quantification strategy. The deduced amount was found to be present in the same range as the SIS-peptides.

Absolute quantification of HCPs

In order to compensate for losses during sample preparation, the reference proteins were spiked into the sample before purification. In this case all losses of the reference proteins would reflect losses of HCPs. The quantitative SWATH data was combined with the previously obtained spectral

libraries. The general workflow for absolute quantification of the HCPs is shown in Figure 3. 100.8 µg of sample was precipitated with acetone and subjected to MCX purification (see Materials and Methods). This sample was compared to 50 µg of sample, which was purified by the suspension trap strategy. Both samples were spiked with the intact protein ladder before any initial sample preparation. 4 µg of each sample was analyzed (assuming no loss). Using the acetone-MCX strategy, 150 proteins could be quantified, while 115 proteins were quantified using the suspension trap strategy. The determined amounts were found to correlate well between the two methods (r^2=0.936). However, proteins were generally determined to be in higher abundance from the acetone-MCX data, which may be a result of the precipitation as described above. Figure 4A shows the correlation between the two sample preparation strategies. Table 2 shows the 20 most abundant HCPs along with the calculated amount/µg of starting material and the coefficient of variation (CV) between the two analyses. An average CV of 5.7 % was found between replicates for the acetone-MCX samples, and 5.9 % for the suspension trap experiments. When comparing the quantities between the two methods, an average CV of 23.1 % was observed.

The quantified HCPs were plotted against their theoretical pI and MW, which was calculated using the *Compute pI/MW tool*, available on the ExPASy server[34]. Figure 5 shows a bubble plot of the quantified HCPs. It is seen that intense HCPs are present at MWs and pIs both below and above that of BP, which is in accordance with the TFF purification strategy, which removes components below a certain size. This information may be used for evaluation of the purification strategy, if certain HCPs can be identified as risk factors.

From the quantitative data, the internal dynamic ranges were investigated. It was found that the majority of the overall HCP signal arose from less than 10 proteins. 80 % of the cumulative amount was determined to consist of only 31 proteins for the acetone-MCX data and 28 proteins for the suspension trap data. The remaining proteins may not contribute much to the overall HCP amount, however, these may still induce unwanted immune reactions. It is therefore essential to know the identity and amounts of these proteins. Figure 4B shows the cumulative percentages of the quantified HCPs. The HCP measured amounts are presented in Supporting Table S-3. The presented data suggest that the acetone-MCX strategy is most suitable for HCP-quantitation, as reference proteins showed a better linearity and allowed for quantification of more HCPs than the suspension trapping strategy. The higher yield of HCPs and better linearity of HCPs is believed to be a result of the powerful precipitation conditions along with the high capacity of the MCX cartridges. However, it is possible that suspension trapping may be equally efficient with minor method optimization. This

method may be of interest, as it offers protein purification and digestion with minimal costs in chemicals and time.

Unspecific cleavage products

Even though trypsin is described as a highly specific protease, non-specific cleavage tends to occur, leading to semi-tryptic and non-specific peptides, which further complicates the process of database searching[35]. These non-specific cleavages occur at a much lower rate, leading to much lower intensities of these peptides, compared to their fully tryptic counterparts[36,37]. If the HCPs are subjected to non-specific cleavage it may not affect the results, but if the BP is subjected to non-specific cleavage, these cleavage products may be in the intensity range of the HCPs and be misinterpreted as tryptic HCP peptides. This can lead to severe overestimation of contaminants and false identifications.

It was determined that 13 % of the identified HCP peptides were semi-tryptic and 1 % were non-specific. For the BP, 73 % of the identified peptides by count were found to be semi-tryptic and 23 % were found to be non-specific. The relatively low number of fully tryptic BP peptides is most likely a result of the high amount of substrate for unspecific cleavages. The non-specific BP cleavage products were found in such high intensity that they accounted for most of the previously unexplained peaks of the base peak chromatogram. Supporting Figure S-2 shows the elution profiles for tryptic and non-tryptic BP peptides along with the cleavage specificity for BP and the HCPs. No significant chymotryptic activity was observed.

The semi-tryptic and non-specific BP peptides were cross-referenced with the fully tryptic HCP peptides identified by MASCOT in order to identify BP-peptides misinterpreted as HCP-peptides, and the peptides were evaluated manually using their elution times and fragment spectra. The non-specific BP peptides *SLLDEGK* and *WDATATELNNALQ* were assigned as *SIEDLGK* (A2RMJ0) and WDGAENTQVEAVK (A2RN08), respectively. The peak areas of these two peptides were among the largest observed in the sample, even though both proteins were only identified by a single peptide, which is quite uncommon as the number of identified peptides commonly reflects the amount of a given protein[38]. This illustrates the importance of identifying non-specific cleavage products, especially when dealing with high-dynamic range samples such as biopharmaceutical proteins. For future analyses, searches should be performed in multiple passes, first excluding all peptide-spectrum matches deriving from BP peptides and then searching the unassigned spectra for HCP information. Aside from reassigning BP-peptides, this will in general increase the number of identifications[39].

Assessing purity of the biopharmaceutical protein

Proteins represented by less than 3 peptides were excluded, as these were deemed not to be reliably quantifiable. It was found that proteins represented by either one or two peptides comprised 8 % of the total HCP signal (measured against Top3 signals of HCPs represented with ≥ 3 peptides). It was assessed that these proteins did not contribute significantly to the overall purity and could therefore be excluded from the quantitative analysis, but may still act as immunogenic targets, and should therefore still be reported as identified.

The purity of the sample was calculated based on the sum of the quantified HCPs. The MW of the HCPs was used to calculate the measure of *total HCP mass / total mass* (which was determined by AAA).

Since the proteins may exist as fragments, this could lead to possible overestimation of the HCP amounts. However, by requiring at least 3 peptides from each protein, proteins with broad sequence coverages are selected. The quantitative data showed a total HCP amount of 113 ng HCP/µg of sample (11.3 %) calculated using the acetone-MCX strategy and 66 ng HCP/µg of sample (6.6 %) calculated using the suspension trap strategy. The HCP content has previously been determined by ELISA assays by SSI to be between 7.3 and 11.4 %, which is in agreement with the obtained results.

Conclusions

This work presents a reproducible strategy for LC-MS/MS-based absolute quantification of host cell proteins associated with a biopharmaceutical protein. The entire workflow may be performed within a few days. By combining samples from different steps in the purification process, extended spectral libraries were created and used to guide the SWATH-based quantification. When combining several various libraries, additional proteins were quantified in the low-abundant range in the final product. Non-tryptic cleavage products were evaluated in order to achieve higher certainty in the identified HCPs, which was determined to be easily influenced by minor side-cleavage products of the BP. This issue, although rarely addressed, is applicable to all proteomics experiments with high dynamic-range samples.

Label-free absolute quantification was performed by spiking with reference proteins and quantifying with Top3. The quantification strategy was verified by SIS-peptides.

The results were used to study the identity and quantity of >100 HCPs and determine the internal dynamic range. It was determined that proteins represented by less than 3 peptides did not add

significantly to the total intensity, and could therefore be excluded from the quantitative analyses. The HCP content of the sample was thus assessed to be 6.6 to 11.3 % of the total protein amount. The acetone-MCX strategy was found to yield more identifications than the suspension trapping strategy, resulting in a better linearity of the reference proteins. In conclusion, the strategy may be used to study the HCP content of new biopharmaceutical proteins in the pipeline as an orthogonal method to ELISA assays. Due to the speed and depth of information, the present method is particularly suitable for obtaining information of in-process samples and thus guide the purification process.

Associated content

Supporting information available:

- Purification and quantification of crude stable isotope labeled standard peptides.
- List of identified HCPs.
- Calibration plots for the different spiking strategies.
- List of quantified HCPs.
- Description of cleavage specificity for the BP and associated HCPs.

Acknowledgements

This work was supported by Innovation Fund Denmark (grant number 34-214-1). The authors thank the employees at Sciex, Warrington, UK for introduction to instrumentation, especially Milla Neffling, Steve Taylor and Nick Morrice. The authors furthermore thank the Villum Center for Bioanalytical Sciences for providing the instrumentation on which preliminary analyses were performed and Ingrid Kromann from SSI for providing the material.

Conflict of interests

The authors declare no conflict of interest.

List of abbreviations

- BP: Biopharmaceutical protein
- DDA: Data dependent acquisition
- DIA: Data independent acquisition
- DTT: Dithiothreitol
- ELISA: Enzyme-linked immunosorbent assay
- HCP: Host cell protein

- IAA: Iodoacetamide
- LC-MS: Liquid chromatography-Mass spectrometry
- MW: Molecular weight
- P10: Nonpurified harvest media containing the biopharmaceutical protein
- P15: Biopharmaceutical protein after all purification steps.
- SIS: Stable isotope standard
- SWATH: Sequential Window Acquisition of all Theoretical Mass Spectra
- TEA: Triethylamine
- TFF: Tangential flow filtration

References

(1) Eaton, L. C. J. Chromatogr. A 1995, 705, 105–114.

(2) Gao, S. X.; Zhang, Y.; Stansberry-Perkins, K.; Buko, A.; Bai, S.; Nguyen, V.; Brader, M. L. Biotechnol. Bioeng. 2011, 108, 977–982.

(3) Hogwood, C. E. M.; Bracewell, D. G.; Smales, C. M. Curr. Opin. Biotechnol. 2014, 30, 153–160.

(4) Wang, X.; Hunter, A. K.; Mozier, N. M. Biotechnol. Bioeng. 2009, 103, 446–458.

(5) Zhu-Shimoni, J.; Yu, C.; Nishihara, J.; Wong, R. M.; Gunawan, F.; Lin, M.; Krawitz, D.; Liu, P.; Sandoval, W.; Vanderlaan, M. Biotechnol. Bioeng. 2014, 111, 2367–2379.

(6) Krawitz, D. C.; Forrest, W.; Moreno, G. T.; Kittleson, J.; Champion, K. M. Proteomics 2006, 6, 94–110.

(7) Xu, D.; Mane, S.; Sosic, Z. Electrophoresis 2015, 36, 363–370.

(8) Reisinger, V.; Toll, H.; Mayer, R. E.; Visser, J.; Wolschin, F. Anal. Biochem. 2014, 463, 1–6.

(9) Levy, N. E.; Valente, K. N.; Choe, L. H.; Lee, K. H.; Lenhoff, A. M. Biotechnol. Bioeng. 2014, 111, 904–912.

(10) Zougman, A.; Selby, P. J.; Banks, R. E. Proteomics 2014, 14, 1006–1000.

(11) Arike, L.; Valgepea, K.; Peil, L.; Nahku, R.; Adamberg, K.; Vilu, R. J. Proteomics 2012, 75, 5437–5448.

(12) Schwanhäusser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M. Nature 2011, 473, 337–342.

(13) Ahrné, E.; Molzahn, L.; Glatter, T.; Schmidt, A. Proteomics 2013, 13, 2567–2578.

(14) Silva, J. C.; Gorenstein, M. V.; Li, G.-Z.; Vissers, J. P. C.; Geromanos, S. J. Mol. Cell. Proteomics 2005, 5, 144–156.

(15) Doneanu, C. E.; Anderson, M.; Williams, B. J.; Lauber, M. A.; Chakraborty, A.; Chen, W. Anal. Chem. 2015, 87, 10283–10291.

(16) Doneanu, C. E.; Xenopoulos, A.; Fadgen, K.; Murphy, J.; Skilton, S. J.; Prentice, H.; Stapels, M.; Chen, W. MAbs 2012, 4, 24–44.

(17) Gillet, L. C.; Navarro, P.; Tate, S.; Rost, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. Mol. Cell. Proteomics 2012, 11, O111.016717-O111.016717.

(18) Egertson, J. D.; MacLean, B.; Johnson, R.; Xuan, Y.; MacCoss, M. J. Nat. Protoc. 2015, 10, 887–903.

(19) Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C. Anal. Chem. 2005, 77, 6426–6434.

(20) Batth, T. S.; Francavilla, C.; Olsen, J. V. J. Proteome Res. 2014, 13, 6176–6186.

(21) Højrup, P. Methods in Molecular Biology; Springer: New York, 2015

(22) Waters Corporation. *[CARE AND USE MANUAL] Hi 3 STANDARDS*; Milford, MA, USA, 2012.

(23) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Electrophoresis 1999, 20, 3551–3567.

(24) Cox, J.; Mann, M. Nat. Biotechnol. 2008, 26, 1367–1372.

(25) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. Bioinformatics 2010, 26, 966–968.

(26) R Core Team. R Core Team (2016): R Foundation for Statistical Computing, Vienna, Austria 2016.

(27) Zi, J.; Zhang, S.; Zhou, R.; Zhou, B.; Xu, S.; Hou, G.; Tan, F.; Wen, B.; Wang, Q.; Lin, L.; Liu, S. Anal. Chem. 2014, 86, 7242–7246.

(28) Schubert, O. T.; Gillet, L. C.; Collins, B. C.; Navarro, P.; Rosenberger, G.; Wolski, W. E.; Lam, H.; Amodei, D.; Mallick, P.; MacLean, B.; Aebersold, R. Nat. Protoc. 2015, 10, 426–441.

(29) Elliott, M. H.; Smith, D. S.; Parker, C. E.; Borchers, C. J. Mass Spectrom. 2009, 44, 1637–1660.

(30) Liu, Y.; Hüttenhain, R.; Surinova, S.; Gillet, L. C. J.; Mouritsen, J.; Brunner, R.; Navarro, P.; Aebersold, R. Proteomics 2013, 13, 1247–1256.

(31) Nakamura, K.; Hirayama-Kurogi, M.; Ito, S.; Kuno, T.; Yoneyama, T.; Obuchi, W.; Terasaki, T.; Ohtsuki, S. Proteomics 2016, 16, 2106–2117.

(32) Percy, A. J.; Yang, J.; Chambers, A. G.; Simon, R.; Hardie, D. B.; Borchers, C. H. J. Proteome Res. 2014, 13, 3733–3747.

(33) Anderle, M.; Roy, S.; Lin, H.; Becker, C.; Joho, K. Bioinformatics 2004, 20, 3575–3582.

(34) Wilkins, M. R.; Gasteiger, E.; Bairoch, A.; Sanchez, J.; Williams, K. L.; Appel, R. D.; Hochstrasser, D. F. Methods Mol. Biol. 2005, 112, 531–552.

(35) Bunkenborg, J.; Espadas, G.; Molina, H. J. Proteome Res. 2013, 12, 3631–3641.

(36) Picotti, P.; Aebersold, R.; Domon, B. Mol. Cell. Proteomics 2007, 6, 1589–1598.

(37) Burkhart, J. M.; Schumbrutzki, C.; Wortelkamp, S.; Sickmann, A.; Zahedi, R. P. J. Proteomics 2012, 75, 1454–1462.

(38) Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M. Mol. Cell. Proteomics 2005, 4, 1265–1272.

(39) Kertesz-Farkas, A.; Keich, U.; Noble, W. S. J. Proteome Res. 2015, 14, 3027–3038.

Tables

Table 1 - Amounts of selected HCP peptides determined by spiking with SIS-peptides. The absolute amounts of the individual proteins as determined by Top3 quantification are also listed. A2RMV9

was only identified with 2 peptides in the label-free analyses, but was included in the table for evaluation purposes.

Protein Accession	Amount in fmol (Top3)	Amount in fmol calculated by SIS peptides [fmol] (CV in %)		
		Peptide 1	Peptide 2	Peptide 3
A2RHZ5	652.24	736.68 (6.08)	2215.51 (54.51)	643.83 (0.65)
A2RKI9	926.33	937.84 (0.62)	1920.28 (34.92)	
P0A3J1	125.92	96.04 (13.46)	41.71 (50.24)	145.92 (7.36)
A2RMQ7	22.62	132.51 (70.84)	304.80 (86.18)	93.87 (61.16)
A2RMV9	*26.21	67.08 (43.81)	43.19 (24.47)	

Table 2 – The 20 most abundant HCPs associated with the BP (Sorted according to abundance measured by the acetone-MCX strategy). The amounts were determined by Top3 quantification using a set of reference proteins spiked in before digestion for both the acetone-MCX (AM) and suspension trapping (ST) strategy.

Protein name	Accession	fmol /µg - AM	fmol/µg - ST	CV [%]
Uncharacterized protein	A2RM18	492	367	15
HU-like DNA-binding protein	A2RIK5	323	181	28
Non-heme iron-binding ferritin	A2RNH7	258	158	24
Secreted 45 kDa protein	P22865	250	90	47
50S ribosomal protein L7/L12	A2RKI9	181	93	32
Probable N-acetylmuramidase	A2RHZ5	133	110	10
Oligopeptide-binding protein oppA	A2RJ53	128	88	18
Basic membrane protein A	A2RK47	112	37	50
Serine protease Do-like HtrA	A2RNT9	111	61	29
Manganese ABC transporter substrate binding protein	A2RKC2	79	14	70
Mid-cell-anchored protein Z	A2RJC3	78	48	24
Phosphonate ABC transporter, phosphonate-binding protein	A2RI25	78	67	7
Cell surface antigen I/II	A2RL18	67	35	31
Putative secreted protein	A2RI00	64	41	22
Putative secreted protein	A2RJA8	53	N/A	N/A
50S ribosomal protein L11	A2RNF4	43	30	18
50S ribosomal protein L18	A2RNN7	42	27	22
Uncharacterized protein	A2RN73	42	28	21

Chaperone protein DnaK	P0A3J1	39	41	3
Foldase protein PrsA	A2RME7	37	9	61

511

512

Figure legends

513

514 Figure 1 – (A) Venn diagram showing the number of identified host cell proteins. (B) Venn diagram
515 showing quantifiable host cell proteins when using different DDA files as spectral libraries.
516 Identifications were performed using MASCOT and quantifiable proteins/peptides were selected
517 based on the criteria described in the text.

518 Figure 2 – Signal/fmol ratios of reference proteins spiked into the BP sample. The ratios are shown
519 for myoglobin (blue), carbonic anhydrase (red), alcohol dehydrogenase (green) and serotransferrin
520 (purple). The ratio used for quantification is based on all proteins and is shown in cyan.

521 Figure 3 – Flowchart describing the general workflow applied for absolute quantification of HCPs.

522 Figure 4 - A) The absolute HCP amounts quantified in the acetone-MCX strategy plotted against the
523 absolute HCP amounts quantified in the suspension trap strategy. B) The cumulative molar amounts
524 of the 100 most abundant HCPs plotted against the number of proteins. The quantities are based on
525 the Top3 signal of quantified HCPs, normalized against the internal reference protein calibration
526 curve. The figure shows amounts determined both for the acetone-MCX (black) and suspension
527 trapping (white) strategy.

528 Figure 5 – Bubble plots showing the pI (x-axis), MW (y-axis) and amounts (bubble size) of the HCPs
529 using the acetone-MCX strategy (A, blue) and the suspension trapping strategy (B, red). The BP is
530 represented by the purple dot. The amount of BP does not correspond to the size of the bubble

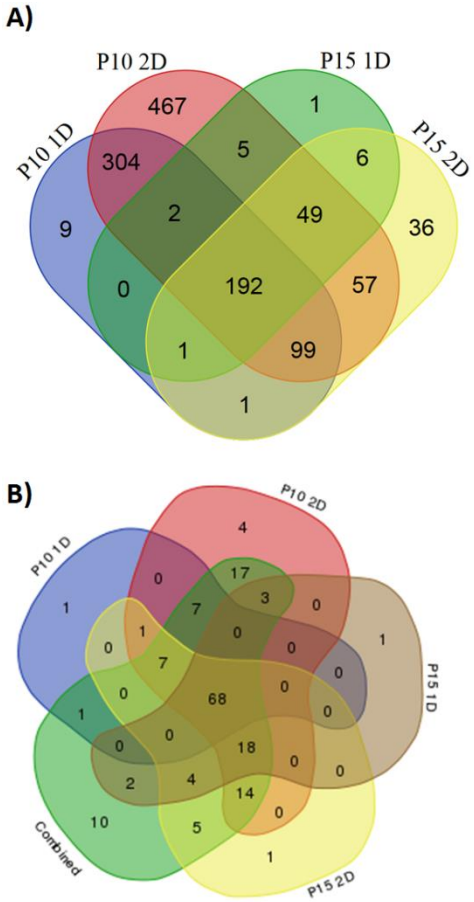
531

532

Figures

533

534 Figure 1



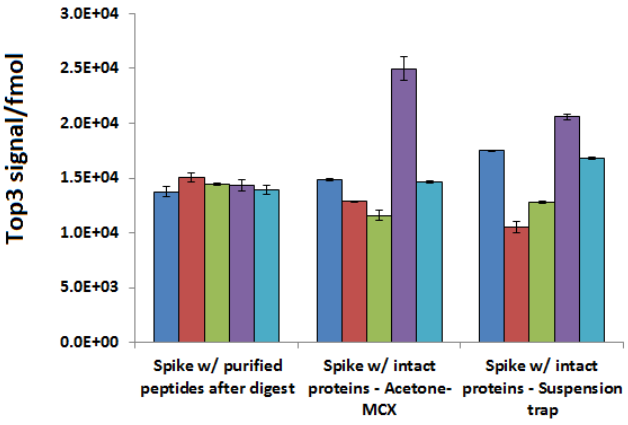
535

536

537 Figure 2

538

539



540

Figure 3

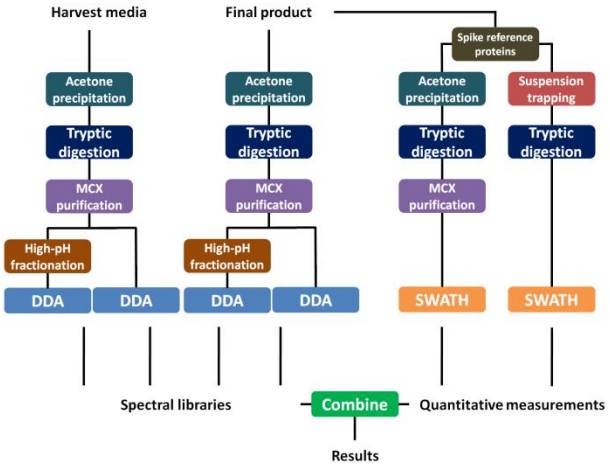
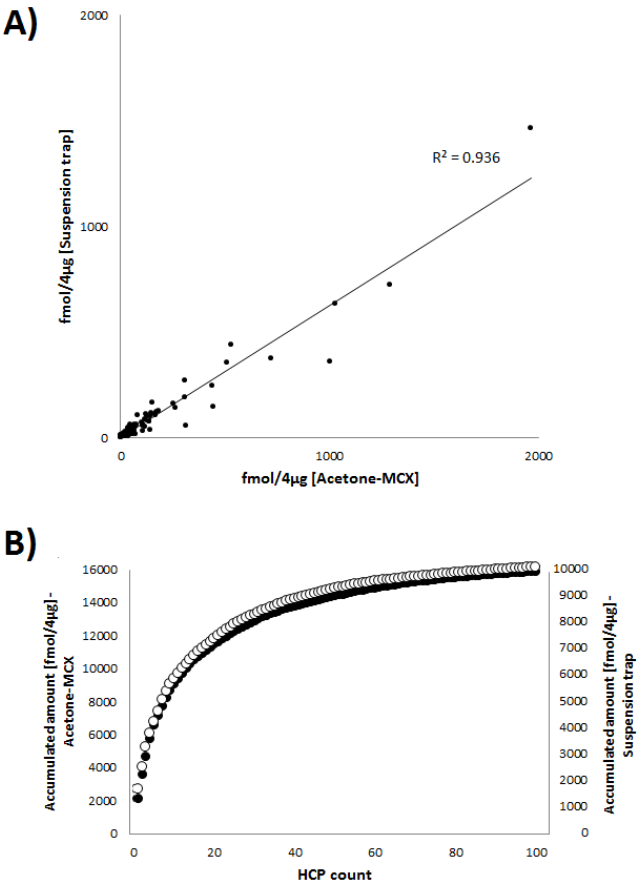
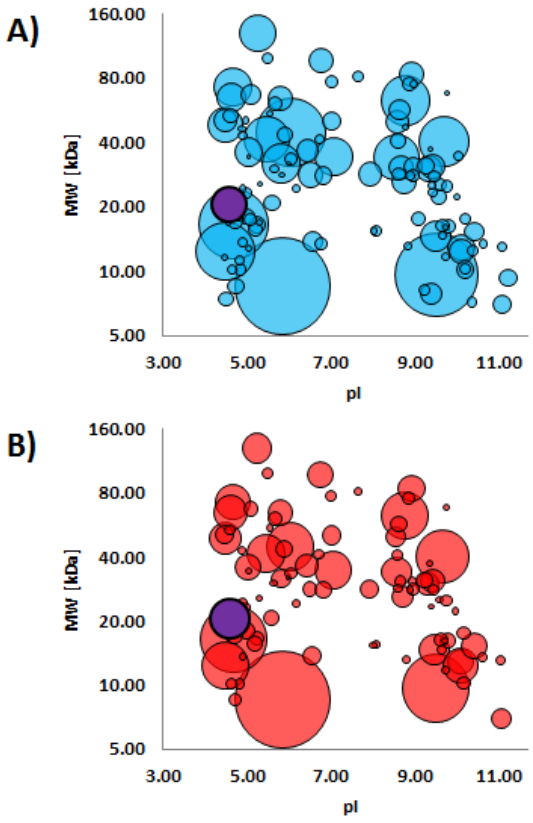


Figure 4



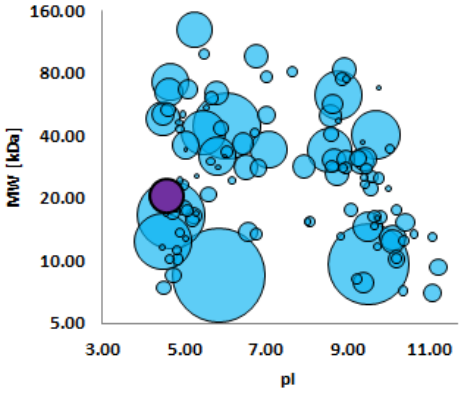
547 Figure 5



548

549

550 For TOC graphic only



551

552

553