Nota applicativa

Effect of Sample Loading on Protein Identifications

Craig Dorschel

Waters Corporation

For research use only. Not for use in diagnostic procedures.

Abstract

In this study, we present data demonstrating that the maximum number of protein identifications are obtained at an optimal sample load, below the level that results in overloading, and that few additional identifications result from the subsequent overloading of the analytical column. The data presented here were obtained using a 150 µm diameter column, packed with a hybrid reverse phase material. The results can be scaled for other column diameters.

Introduction

The goal of many proteomics experiments is the identification of proteins from complex biological mixtures. This has been dominated by LC-MS/MS solutions. In these experiments, the challenge is often to confidently assign peptide and protein sequences to the tandem MS data. Identification of proteins from single peptides using this approach has a high error rate, which has resulted in the erroneous reporting of large numbers of proteins with low sequence coverage in the literature.

The Waters Identity^{E} High Definition Proteomics System is designed to obtain the largest number of rigorous protein identifications possible for a given sample of a complex mixture of proteins, coupled with accurate quantitation. A crucial factor in obtaining good results is applying the optimum amount of protein digest sample to the UPLC Column. Too small a quantity will limit the number of peptides, and thus proteins, detected and identified. This adversely affects the amount of information about a sample that may be gained and further limits quantitative comparisons of samples.

In many proteomic experiments chromatography systems have been intentionally overloaded in the belief that a significant increase in the number of protein identifications will be obtained. This results in poor chromatographic performance, distorting both peak shapes and elution order, and challenges the detection system of the mass spectrometer.

Here, we present data demonstrating the maximum number of protein identifications obtained at an optimal sample load, below the level that results in overloading. Few additional identifications result from the subsequent overloading of the analytical column. The data presented here were obtained using a 150 µm diameter column, packed with a hybrid reverse phase material. The results can be scaled for other column diameters.

Since there is an optimum sample quantity for a given column, it is important for the analyst to have an easily obtained and accurate estimate of the total quantity of protein digest present in the sample. While there are a number of traditional colorimetric methods for estimating protein concentration prior to digestion, these methods can be skewed by factors such as sample turbidity. The Identity^E System provides a convenient alternative by estimating the absolute amount of each identified protein. By addition of a known amount of a standard digest to a sample digest, a survey injection will provide an estimate of the actual quantity injected, after which any necessary adjustments to the injection volume can be made to assure that the optimal quantity of sample has been analyzed.

Experimental

A dilution series of the MassPREP *E. coli* Digestion Standard (p/n: 186003196) was prepared in 3.0% Acetonitrile in 0.1% formic acid, with final concentrations of 50, 100, 250, 500, 1000, and 2500 ng/μL, respectively. Each sample also contained 250 fmol/μL of MassPREP Alcohol Dehydrogenase Digestion Standard (p/n: 186002328). Samples were analyzed in triplicate using 2 μL injections directly onto the analytical column. The data were processed in ProteinLynx Global SERVER version 2.3, and proteins identified using the Identity^E search algorithm, querying the EU *E. coli* databank (with the ADH sequence appended). The estimated absolute quantitation of every identified protein (containing at least three peptide identifications) was also provided by the Identity^E algorithm, using the added ADH digest as the quantitation standard.

LC Conditions

LC System:	nanoACQUITY UPLC System
Column:	ACQUITY UPLC BEH C_{18} Column 0.15 x 100 mm, 1.7 μm
Column Temp.:	35 ˚C
Flow Rate:	1.0 μL/min.
Mobile Phase A:	0.1% Formic Acid (Aq)

Mobile Phase B:	0.1% Formic Acid in Acetonitrile
Gradient:	35-95% B/10 min.

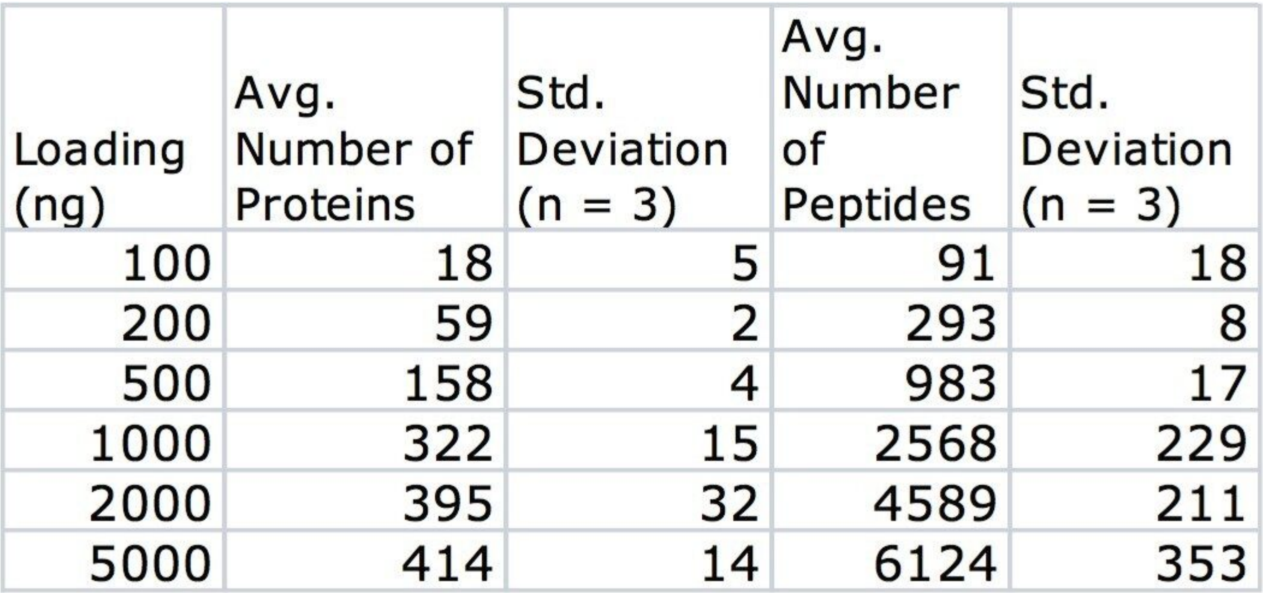
MS Conditions

MS System:	Waters Q-Tof Premier Mass Spectrometer
Ionization Mode:	ESI Positive
Capillary Voltage:	4000 V
Cone Voltage:	35 V
Nanoflow Gas:	2.3 Bar
Source Temp:	100 ˚C
Acquisition Range:	50-1990 *m/z*
Collision Energies:	MS 4 V, MS^E 15-40 V ramp

Results and Discussion

Protein and Peptide Identifications

The numbers of proteins and peptides identified for each injection of the sample loading study were obtained from the ProteinLynx Global SERVER data browser, and averaged for each loading level. The results are tabulated in Table 1.



Loading (ng)	Avg. Number of Proteins	Std. Deviation (n = 3)	Avg. Number of Peptides	Std. Deviation (n = 3)
100	18	5	91	18
200	59	2	293	8
500	158	4	983	17
1000	322	15	2568	229
2000	395	32	4589	211
5000	414	14	6124	353

Table 1. Results of Loading Study.

The relationship of the number of identified proteins to the amount of sample loaded is plotted in Figure 1. It can be seen that the number of identifications rises sharply as the loading increases from 100 to 500 ng and doubles again as the loading increases from 500 ng to 1 µg. From that point, the number of identifications does not increase dramatically and there is little gain in the number of proteins identified going from 2 to 5 µ g.

The number of peptides identified also rises sharply at first, proportional to the column loading, before becoming more gradual. In this instance there is still a substantial increase in peptide identifications going from 2 to 5 µg, but the bulk of these are lower intensity peptides related to the already identified proteins. While this increases the average sequence coverage by approximately three peptides, it does not enable the identification of a significant number of new proteins.

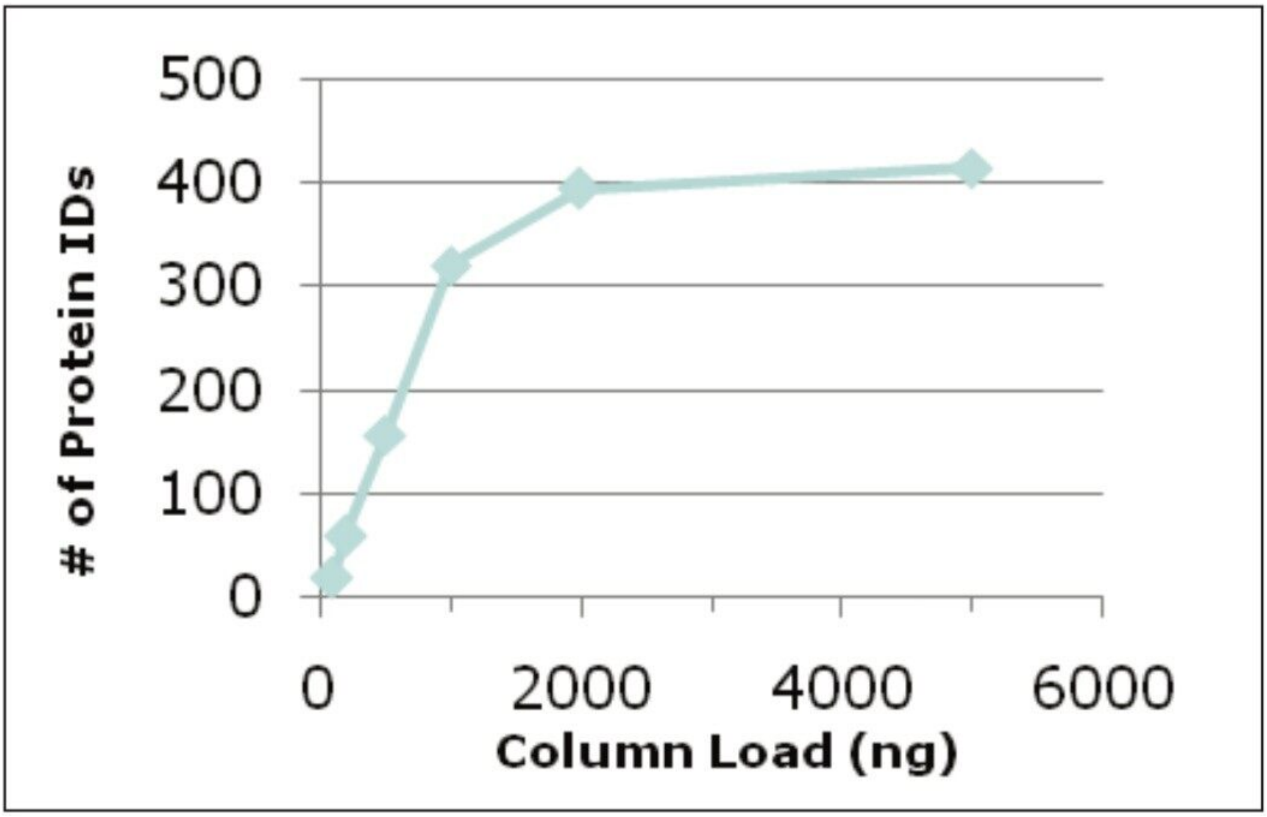


Figure 1. Relationship of the number of protein identifications to the amount of sample loaded on the UPLC column.

Effect on Chromatography

As sample quantities are increased on a given liquid chromatography column, non-ideal chromatographic behavior will eventually begin to manifest itself. When the sample quantity exceeds the capacity of the stationary phase to establish equilibrium, excess sample will remain in the mobile phase, resulting in peak broadening and decreased retention times. In gradient elution, especially when no separate trapping column is used as in the present case, a portion of the stationary phase is occupied by the non-mobile sample. With very large sample loads this can result in a shortening of the effective column length, especially early in the run, again resulting in shorter retention times.

Finally, high sample concentrations can result in non-linear behavior of both the mass analyzer and eventually the electrospray process itself. Non-linearity of any part of the MS system will have a significant effect on the chromatographic peak heights and the calculated peak areas.

To examine the current data for evidence of overloading with increasing sample load, the behavior of a single peptide was examined in detail. The chosen peptide – one of modest ionization efficiency – is the T25 tryptic peptide of elongation factor TU, having the sequence STCTGVEMFR, where the cysteine has been carbamidomethylated. Extracted ion chromatograms based on the monoisotopic doubly charged ion of the peptide were studied.

Figure 2 shows a trend of increasing peak width at half intensity, where the peak width has increased by 50% from the lowest sample loads up to 2 µg and has doubled for a 5 µg load.

Retention times and peak heights were also affected. Retention times were constant over three injections at any sample load level, but compared to the 100 ng load, elution was 90 seconds earlier at 1 µg, 2 minutes 30 seconds earlier at 2 µg, and 4 minutes 30 seconds earlier at 5 µg. Such retention time shifts might be encountered between sets of samples exhibiting a significant fold change of an abundant protein. Expression ^{E} High Definition Proteomics System is fully capable of tracking retention time shifts of this magnitude.

Peak broadening is also accompanied by a reduction in peak height. In this dataset, peak heights are linear up to only 1 µg oncolumn, thus peak height would be a poor choice on which to base quantitation. Identity^{E} and Expression^{E} Systems base quantitative comparisons on the deconvoluted peak areas for all isotopes and charge states. This method gives good linearity for the EF TU T25 peptide throughout the range of the experiment.

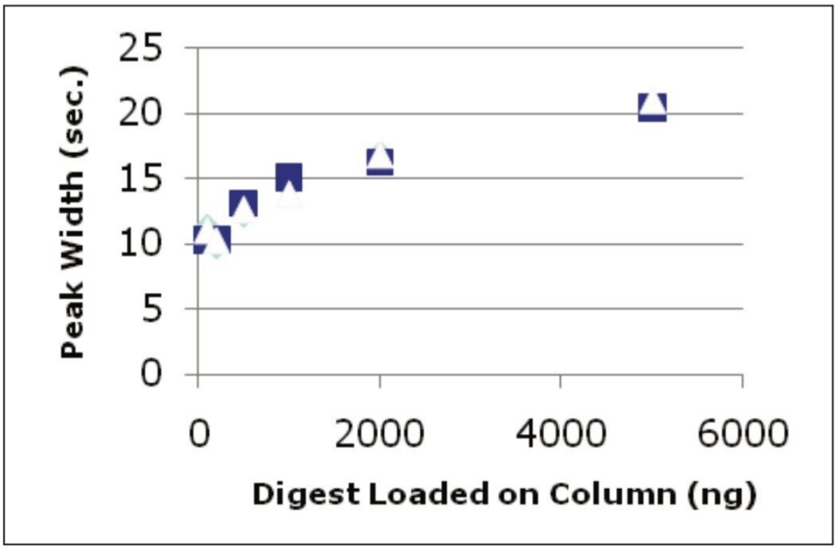
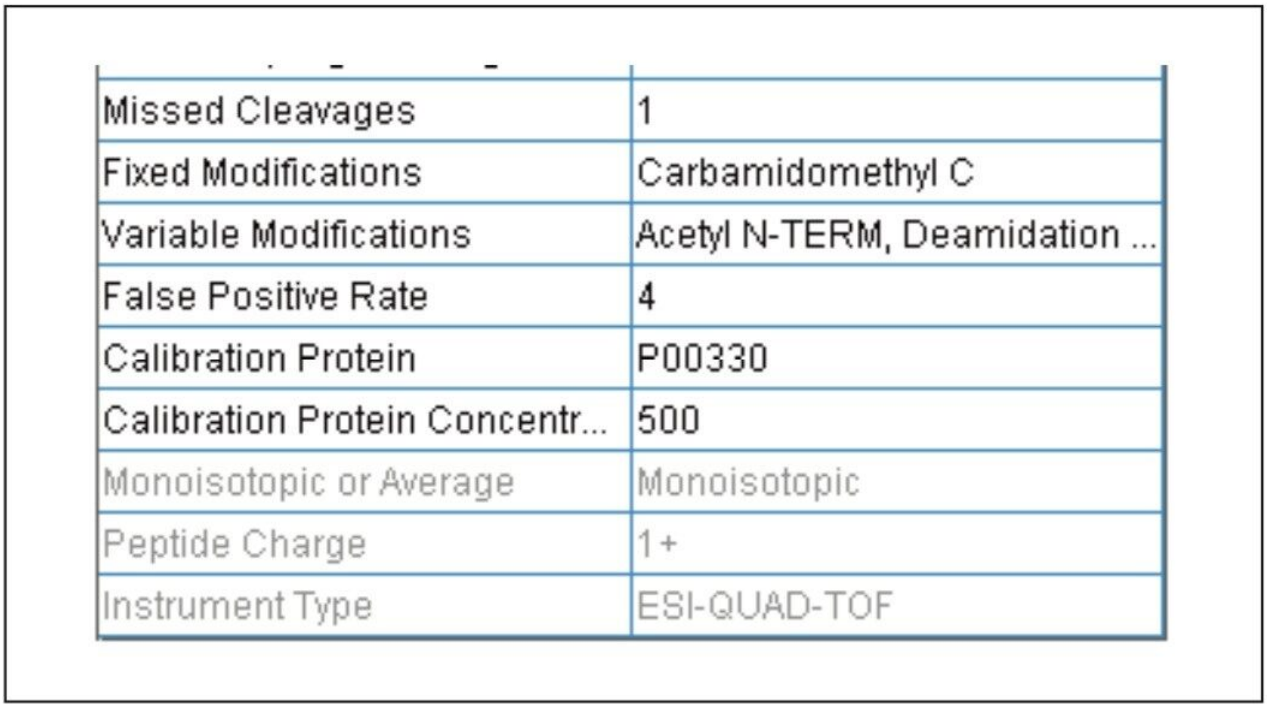


Figure 2. Effect of increasing sample load for the peak width of peptide T25 for EF TU.

It is clearly important for the analyst to have a reasonable estimate of the quantity of protein digest injected in each case. The absolute quantitation feature (Reference 1) of the Identity^{E} System provides a convenient means to do so. A known quantity (in fmol) of a protein digest, such as MassPREP Alcohol Dehydrogenase Digest Standard, is added to each sample. The amount injected and the accession number for the standard are specified in the databank search workflow setup (Figure 3). The quantity of each identified protein (in fmol and ng), is returned with the search results, and can be viewed in the ProteinLynx Browser. It is convenient to make a survey injection to determine the concentration for each biological sample. Following the databank search the resulting protein table is copied into a spreadsheet, such as Microsoft Excel, the data filtered for duplicate molar quantities arising from homologs, and the nanogram amounts summed. If the quantity thus calculated differs from the expected value by more than 20%, the injection volume can be adjusted accordingly, or in extreme cases, a new digestion could be performed provided there is sufficient sample.



Missed Cleavages	1
Fixed Modifications	Carbamidomethyl C
Variable Modifications	Acetyl N-TERM, Deamidation ...
False Positive Rate	4
Calibration Protein	P00330
Calibration Protein Concentr...	500
Monoisotopic or Average	Monoisotopic
Peptide Charge	1+
Instrument Type	ESI-QUAD-TOF

Figure 3. Workflow Editor showing entries for absolute quantitation (Calibration Protein and Calibration Protein Concentration).

The absolute quantitation results for the proteins identified in at least 2 of 3 injections at the 2 µg sample load (344 proteins) are illustrated in Figure 4, where the dynamic range of the identified proteins spans 2.5 orders of magnitude.

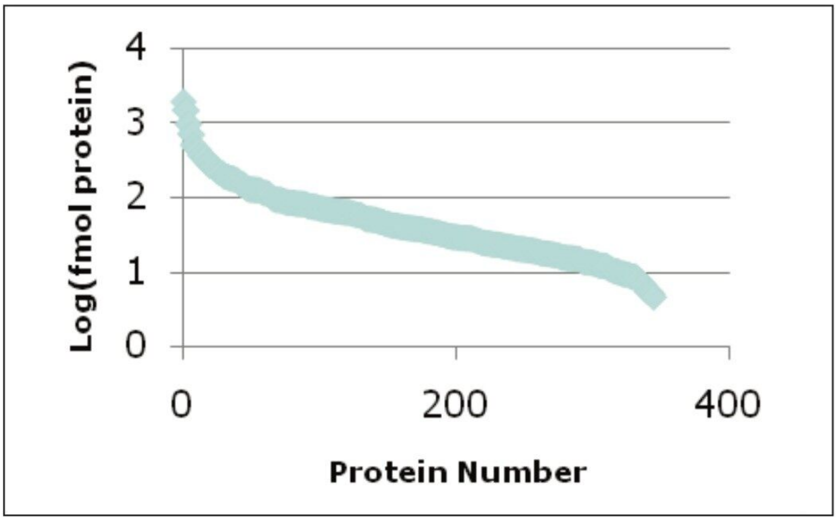


Figure 4. Range of concentrations for 344 E. coli proteins identified in at least 2 of 3 injections of protein digest at optimum sample loading.

Conclusion

Optimized chromatography is necessary to obtain the highest quality results possible with the Identity^E System. The optimum sample quantity for a 150 µm diameter column was determined to be 2 µg, based on the number of protein identifications, peptide coverage, and chromatographic behavior. If less than 1 µg is applied to this column, the number of protein identifications will be severely limited. If a larger quantity is injected, there is a risk that chromatography will be degraded affecting both qualitative and quantitative results.

Because chromatography scales by the ratio of the squares of column diameters, the optimum protein digest load for 75 µm diameter columns would be 500 ng.

Absolute quantitation provides a unique and simple means to assess the concentration of the protein digest sample to assure that the optimum amount has been analyzed, as well as providing valuable information regarding the relative amounts of each protein within the sample. This will provide the analyst the most rigorous protein identifications possible.

References

1. Silva, J.C.; Gorenstein, M.V.; Li, G.Z.; Vissers, J.P.C.; and Geromanos, S.J.; *Molecular and Cellular Proteomics* 5, 144–156 (2006).

Featured Products

ACQUITY UPLC M-Class System <https://www.waters.com/134776759>

ProteinLynx Global SERVER (PLGS) <https://www.waters.com/513821>

720002467, February 2008