Small HCPs in a 12 kDa Protein Drug Analyzed by GeLC-MS/MS

Rikke Raaen Lund¹, Marie Grimstrup¹, Maheen Sayeed², Karin Abarca Heidemann², Ejvind Mørtz¹ ¹Alphalyse A/S, Odense, Denmark, ²Rockland Immunochemicals Inc., Limerick, PA, USA

Introduction

The content of low molecular weight host cell proteins (HCPs) in purified protein drugs is often difficult to evaluate, due to their low immunogenicity and poor ability to be visualized in gel-based total protein stains. The proteome of commonly used expression organisms, such as E. coli and Chinese Hamster cells, contains 30-40% proteins with a molecular weight below 20kDa, and these are easily missed in both in gel separations, Western blots and ELISA quantitation of the total HCP-content. To provide unbiased analysis of small as well as larger HCPs, we introduce the use of a mass spectrometry-based orthogonal method, well known from proteomics, called GeLC-MS/MS.

Here, we analyze an in process protein drug, a 12 kDa protein produced in E. coli, as well as the corresponding null cell lysate, using 1D-PAGE and nano-flow LC-MS/MS (GeLC-MS/MS) to achieve high coverage of small HCPs.

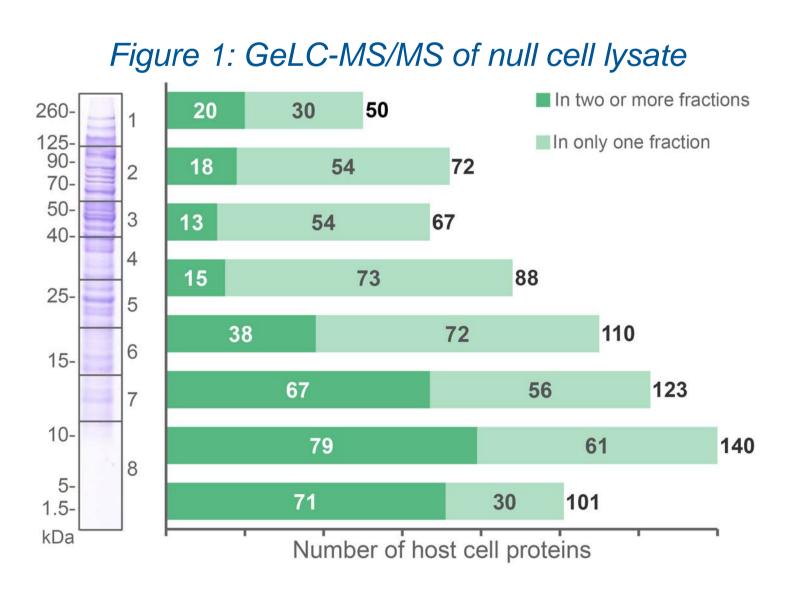
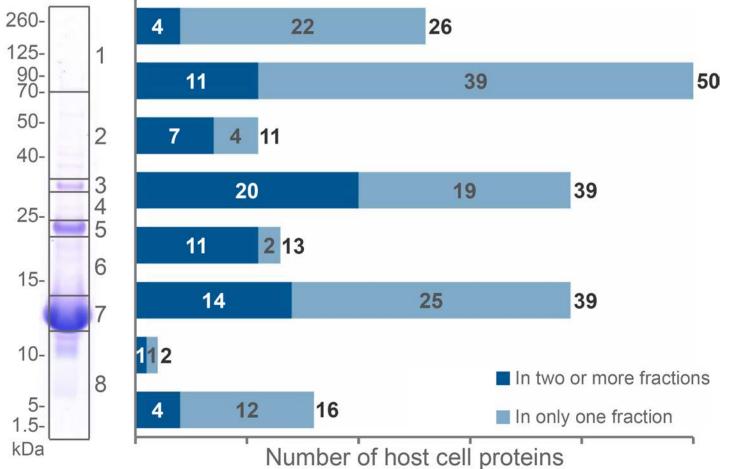


Figure 2: GeLC-MS/MS of the in process sample



Proteins were separated by 1D-PAGE and stained for isolation and visualization of the high concentration protein drug in separate gel fractions. Most proteins, 78% in the null cell lysate and 82% in the in process sample were only identified in a single gel fraction (light green and blue bars, fig. 1 and 2, respectively). Further, the fractionation leads to a high number of protein identifications in the fractions without the protein drug (fig 2). Protein separation by 1D-PAGE prior to LC-MS/MS leads to identification of a high percentage of small HCPs: 46% of the HCPs identified in the null cell lysate and 37% in the in process sample were

smaller than 20kDa (fig. 3, table 1 and 2). Further, the HCPs that were identified covered the entire pl range and 99% of the molecular weight range of the total *E. coli* proteome (fig. 3). Examples of small HCPs is given in table 3 and 4.

The sensitivity of the method was estimated by parallel analysis of the purified protein drug with two spiked-in standards, these standards were both identified at 50ppm.



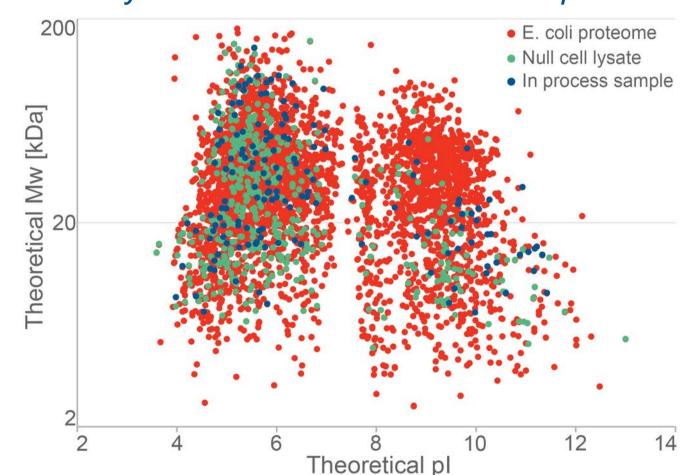


Table 2: Distribution of high and low molecular weight HCPs in the in process sample

1D-PAGE

Figure 3: Molecular weight and pl of the HCPs identified by GeLC-MS/MS and the E.coli proteome

Table 1: Distribution of high and low molecular weight HCPs in the null cell lysate

Protein size	HCPS Identified
HMW (≥20kDa)	297
LMW (<20kDa)	256
Total	553

Protein size	HCPs identified
HMW (≥20kDa)	96
LMW (<20kDa)	56
Total	152

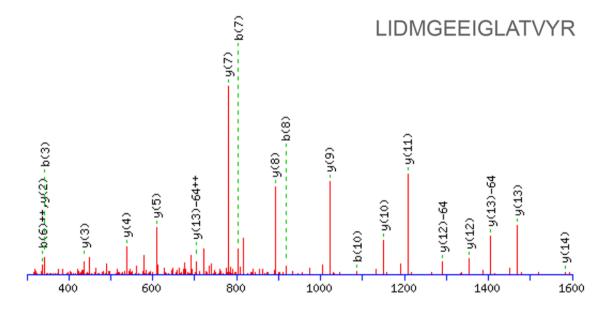
Nano Flow LC-MS/MS

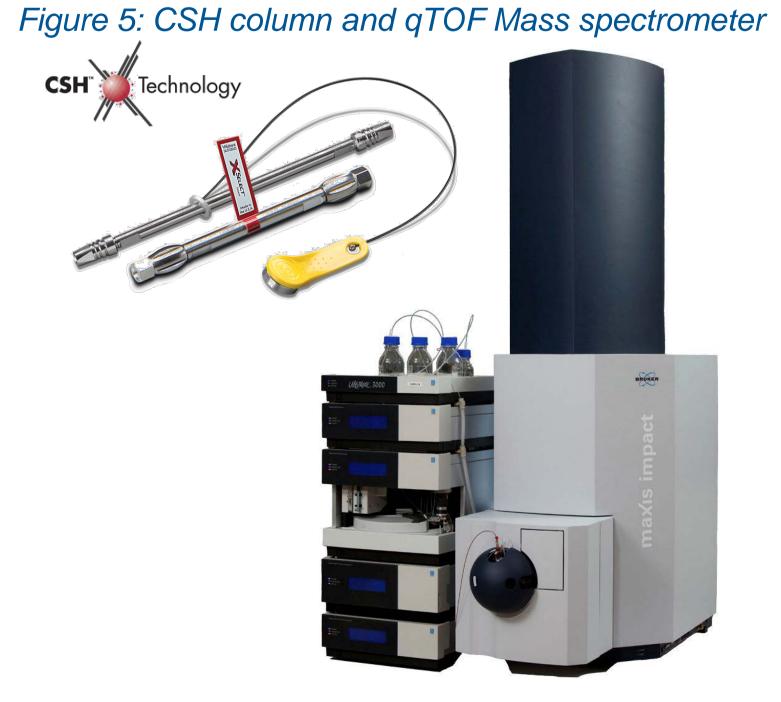
The proteins contained in the gel fractions were digested enzymatically into peptides using trypsin. Separation of the complex peptide mixture was achieved by nanolitre flow HPLC using a CSH (charged surface hybrid, Waters) column. This material has a high loading capacity and excellent peak shape in formic acid mobile phases. This enables highly sensitivity and accurate MS detection of low level HCPs using a qTOF mass spectrometer (Bruker Corp. fig. 4 and 5).

GeLC-MS/MS advantages

- 1D-PAGE requires very little sample preparation and has a wide mass and pl range, leading to unbiased sample analysis
- Protein digest combined with nano flow LC-MS/MS provides sensitive and measurement of multiple peptides from each protein as well as sensitive and accurate measurement of peptide fragment masses

Figure 4: MS/MS spectrum of the a peptide from *Ferric Uptake Regulator (in process sample)*





accurate

HCP Identification Criteria

HCPs are identified by comparing the mass data to the theoretical masses of the host cell proteome. Peptides are assigned a score according to how well they match and these are summarized to a protein score. Since small proteins have fewer peptides they also have lower proteins scores even at the same molar level as larger proteins.

We have used conservative inclusion criteria that accounts for this: 1% false discovery rate, a minimum score of 25 for all peptides, and a two peptide minimum for proteins larger than 20kDa. These criteria allows highly confident HCP identification with a low false positive rate.

> Table 3: Top 30 small proteins in the null cell lysate

HCP no	Accession no	Protein	Name	l N	lass	pl	Score			
22	tr C6ECE6	Riboson	nal protein L9	15	5759	6.17	3350			
29	tr C6EGC3	Riboson	Ribosomal protein S7			10.3	2701			
31	tr C6ELG9	6,7-dim	6,7-dimethyl-8-ribityllumazine synthase			5.15	2656			
42	tr C6EL60	PTS syst	em, glucose su	ofamily, IIA subunit 18	3240	4.73	2336			
46	tr C6EFQ8	Redoxin	domain anatai		7005	1 75	2252			
48	tr C6EE34	50S ribo	osoma Ta	able 4: Top 30 s	sma	all I	HC	D _S		
50	tr C6ECC2	Inorgan								
55	tr C6EGN0	DNA-bir	nding	in the in proces	SS S	san	nple	ý		
58	tr C6EE18	Histone	famil							
59	tr C6EGG1	HCP no	Accession no	Protein Name				Mass	pl	Score
60	tr C6EE33	3	tr C6EJQ1	Ferric uptake regulator, Fur family				17012	5.68	3455
61	tr C6EGG3	6	tr C6EGI0	Peptide deformylase				19430	5.23	2765
62	tr C6EGE5	14	tr C6EC51	Methionine-R-sulfoxide reductase				15783	5.58	1881
70	tr C6EGF1	20	tr C6EE34	50S ribosomal protein L10				17757	9.04	1367
72	tr C6EG32	22	tr C6EGC3	Ribosomal protein S7				17593	10.3	1082
73	tr C6EG67	31	tr C6ECU9	Glutaredoxin				13042	4.75	906
75	tr C6ECE9	40	tr C6EGF4	Ribosomal protein L29				7269	9.98	625
78	tr C6ECU9	46	tr C6EGG3	30S ribosomal protein S5				17534	10.23	591
79	tr C6EIW8	63	tr C6EE36	50S ribosomal protein L11				14923	9.64	399
82	tr C6EAL7	65	tr C6EIX0	DNA protection during starvation p	orotein			18684	5.72	390
83	tr C6EIX0	70	tr C6EE23	Regulator of sigma D				18288	5.65	371
86	tr C6EDI0	71	tr C6EIZ8	Molybdopterin synthase sulfur car	rier sul	ounit		8734	4.38	370
88	tr C6EAL6	73	tr C6EBF4	Ferritin				19468	4.77	332
96	tr C6EAC8	75	tr C6EKH8	Iron-sulfur cluster assembly scaffol	ld prot	ein IscU		14011	4.82	301
99	tr C6EA12	80	tr C6EE73	50S ribosomal protein L31				8094	9.46	271
105	tr C6ECY3	81	tr C6ECE9	30S ribosomal protein S6				15163	5.25	262
107	tr C6EJY6	83	tr C6EG68	ATP synthase F1, delta subunit				19434	4.94	254
108	tr C6EC07	88	tr C6EF45	Uncharacterized protein GN=ECBD	_4172	-		10323	5.18	232
110	tr C6EG68	90	tr C6EJQ0	Flavodoxin				19896	4.21	224
112	tr C6EGF4	91	tr C6EGG2	Ribosomal protein L18				12762	10.41	222
		95	tr C6EGN0	DNA-binding protein				15587	5.43	202
		96	tr C6EGG1	50S ribosomal protein L6				18949	9.71	201
		97	tr C6EK87	UspA domain protein				15925	6.03	200
		102	tr C6EHC0	Acyl carrier protein				8634	3.98	190
		103	tr C6EAC8	Iron-sulfur cluster assembly access	ory pro	otein		12264	4.11	189
		105	tr C6EGF1	50S ribosomal protein L22				12219	10.23	185
		111	tr C6EGG9	Ribosomal protein S11				13950	11.33	158
		112	tr C6EL60	PTS system, glucose subfamily, IIA	subuni	t		18240	4.73	153
		113	tr C6EGK5	Acetyl-CoA carboxylase, biotin carb			rotein	16733	4.66	152
		115	tr C6EK04	Thioredoxin				15887	5	147

Conclusion

The HCPs that were identified covered 99% of the entire *E. coli* proteome in terms of molecular weight and pl. This shows that the developed GeLC-MS/MS method has no inherent limitations with respect to pl or molecular weight for HCP identification. The obtained protein identity enables an in-depth analysis of each individual HCP and a more detailed risk assessment.



HCPs Characteristics and Individual Risk Assessment

Identification of the protein names and database accession numbers enables an in-depth analysis of each individual HCP present in the drug sample. Important features for a drug risk assessment include:

- Immunological properties and presence of human B- and T-cell epitopes
- Homology to human proteins with important biological function
- Homology to the drug protein
- Enzymatic activity to modify or cleave the drug product constituents
- Hormone or hormone-like activity

Known information can be found at http://www.uniprot.org/ including:

- Molecular function, biological process, ligand binding and cellular component
- Post translational modifications and processing
- Expression, interaction and structure
- Protein family and domains
- Sequence variations
- Publications

Prediction of protein features can be investigated at http://www.cbs.dtu.dk/services, including (fig. 6):

- Immunological features
- Post translational modifications
- Protein structure and function

Figure 6: Examples of services available for prediction of protein features

immunological leatures						
ArrayPitope						
Residue-level epitope mapping of antig	ens based on peptide microarray data					
BepiPred »						
Linear B-cell epitopes						
DiscoTope » Post-translatio	nal modifications of proteins					
Discontinuous B-ce						
HLArestrictor DictyOGlyc						
	glycosylation sites (trained on <i>Dictyostelium discoideum</i> proteins)					
LYRA NetAcet	Protein function and structure					
MHCcluster NetCGlyc »	ArchaeaFun					
MHC class 1 clustri	Enzyme/non-enzyme and enzyme class (Archaea)					
Prediction of neo-er Coronavirus 3C-like	Identification of Rossmann folds and prediction of FAD, NAD and	I NADP specificity				
INECTION »	ICPHmodels					
Proteasomal cleave Glycation of a amin	Protein structure from sequence: distance constraints					
	Ldietaneo D					
Integrated class I a Notochia	Protein distance constraints					
NotCTI nan »	I Fosturo D					
Pan-specific integra O-GaiNAC (mucin t	Functional differences of protein variants					
Binding of peptides Linear moun atlas in	Co-evolving amino acids in proteins					
INetMHCcons »	INetDiseaseSNP					
Binding of peptides Generic phosphory	Predicts whether a single non-synonymous SNP causes a disease or is invariant					
NetMHCII »	INetSurfP »					
Binding of peptides Generic phosphory	Protein secondary structure and relative solvent accessibility					
NotMUCInan » INCLETIOSIC	INotTurnD					
Pan-specific binding Kinase specific pro	β-turns and β-turn types in proteins					
INotMHChan *	IProtEun					
Pan-specific binding Serine and threening	Protein functional category and enzyme class (Eukarya)					
Stability of peptide: Postfranslational c	Reduction of sequence similarity in a data set					
Netlepi	SigniSite					
r-cell epitopes resul	identification of residue-level genotype-prienotype correlations in	protein multiple seq	uence alignments			
	TMHMM »					
Identifying sequenc Arginine and lysine	Transmembrane helices in proteins					
PickPocket »	VarDom					
Dinaing of populator	Domaine in the malana antigen family i rEwr i					
VDJsolver »	wKinMut-2					
Analysis of human immunoglobulin VD	Prediction and annotation of the consequences of mutations in th	ne human kinome				