

# Annotation of Complex Proteomic Data Obtained Using Linear Ion Trap LC-MS/MS With ETD: Analysis Of Human Cerebrospinal Fluid Digests

R. G. Biringer<sup>1</sup>, Z. Hao<sup>1</sup>, H. Tran<sup>1</sup>, M. G. Harrington<sup>2</sup>, A. F. R. Hühmer<sup>1</sup>

<sup>1</sup>Thermo Fisher Scientific, San Jose, CA, United States, <sup>2</sup>Huntington Medical Research Institutes, Pasadena, CA, United States.

## Overview

**Purpose:** Automate annotation of LC-MS/MS results as a first step in a targeted analysis workflow and to determine the optimal protease or combination of proteases to maximize the number of unique protein identifications for alternating ETD-CID analysis of complex proteomes.

**Methods:** An LTQ XL™ with ETD equipped with a Surveyor™ chromatography system to examine proteolytic digests of human cerebrospinal fluid proteins. Results were annotated with GO annotations and locations of known and proposed posttranslational modifications using automated KDE-based workflows.

**Results:** LysC or a combination of LysC and ArgC provide superior proteome coverage. Annotations provide a global analysis of sample content as well as function, process, component, and posttranslational modification (PTM) information necessary for designing a targeted re-interrogation of existing analyses or for designing additional experiments.

## Introduction

LC/MS analysis of complex protein mixtures such as whole cell digests or digests of biological fluids generally produce abundant protein identifications and some understanding of the relative amounts of each present. Insight into biological meaning or simply which experiment to do next requires extensive and specific information about each identified protein. Fortunately, public databases such as NCBI protein, Swiss-Prot-TrEMBL, and Uniprot provide comprehensive descriptions of proteins, often providing clues for subsequent, more targeted experiments. However, without additional tools, the researcher has no alternative but to query these databases one protein at a time, a laborious and time consuming process. The main goal of this study was to establish annotation workflows for LC-MS/MS data of human cerebrospinal fluid (CSF) using a new set of tools that automatically retrieve pertinent information about each identified protein from public databases. The tools provide annotation including, but not limited to, GO classifications, sites of post-translational modifications, PubMed references and genomic information. Digests were prepared with several different proteases, and individually analyzed on a LTQ XL with ETD. Peptide sequences were annotated using KDE (knowledge discovery environment)-based workflow tools. Preliminary results indicate that the annotation capabilities presented provide specific information about proteins of the CSF and a convenient method to design iterative and targeted follow-up experiments.

## Methods

**Sample Preparation:** 360 mg solid urea was dissolved in 500 µL of human cerebrospinal fluid (CSF, 1.4 mg/mL protein) and then diluted to 1 mL with 100 mM ammonium bicarbonate at pH 8.0. The protein solution was reduced with dithiothreitol (3 mM, 1 hr.) and alkylated with iodoacetamide (12 mM, 1 hr.). Following the alkylation, the excess alkylating reagent was sequestered with an equi-molar amount of DTT (1 hr.). Samples were split into fractions containing 50 µg total protein and individual fractions were solvent exchanged for 100 mM ammonium bicarbonate, pH 8.0 using 5 kDa cutoff spin filters. Samples to be digested with ArgC were made 10 mM in CaCl<sub>2</sub> prior to addition of the protease. The alkylated CSF proteins were proteolytically digested overnight (~16 hrs) with either LysC, ArgC, GluC, trypsin or chymotrypsin at a 100:1 protein:protease ratio. Digestions were performed at 37°C for all but GluC digestion which was performed at 25°C. Proteolysis then was sequestered by reducing the pH to 3 with glacial formic acid.

**LC-MS Analysis of Samples:** 5 µL of each CSF digest (3.1 µg total peptide) was directly injected onto a peptide trap (CapTrap, Michrom). Peptides were eluted onto and through a C18 column, 25 cm X 100 µm with a 4 hr, 0-85% pseudo-exponential gradient (A:0.1% formic acid, B: 100% acetonitrile/0.1% formic acid) at a flow rate of 350 nL/min using a Surveyor HPLC equipped with a Micro AS and nanospray source (Thermo Fisher Scientific, San Jose). The eluted peptides were analyzed by a LTQ XL with ETD (Thermo Fisher Scientific, San Jose) using alternating CID/ETD fragmentation with supplemental activation and data-dependent MS/MS detection. Three replicates of each experiment were performed.

**Data-MS Analysis:** All MS data were analyzed with BioWorks™ 3.3.1 software using a human Swiss-Prot-TrEMBL database and the results filtered to a 5% maximal false positive rate using a reverse database search approach. Data for three successive identical experiments were combined and evaluated.

**Results Annotation:** KDE-based workflows (Inforsense) were written to extract GO annotations and descriptive information from the public Swiss-Prot-TrEMBL server ([www.expasy.org](http://www.expasy.org)) for each protein identified with BioWorks. GO annotations for each of the three GO categories (component, function, and process) were individually sorted into biologically meaningful groups, including a catch-all "other GO group" for those that did not quite fit the others. Additional descriptive information was parsed into several categories, including locations of known and predicted posttranslational modifications.

## Results

### Optimal Protease(s):

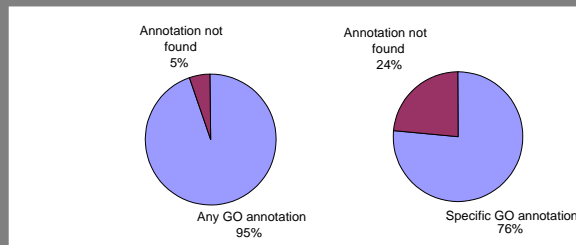
TABLE 1. Number of unique protein identifications as a function of protease and combined data from different proteases. Percent improvement is calculated in reference to the number of unique identifications observed for LysC (127 unique identifications).

Protease	Unique Protein Identifications	Data Combinations	Unique Protein Identifications	% improvement
LysC	127	LysC + ArgC	220	73
ArgC	101	LysC + GluC	152	20
GluC	59	LysC + chymo	137	8
Chymo	32	LysC + trypsin	189	49
Trypsin	113	trypsin + ArgC	179	41

- Lys C digests provide the largest number of unique protein identifications for data obtained with a single protease.
- Combining results from LysC and Arg C digests provides the greatest proteome coverage, a 73% improvement over LysC alone.

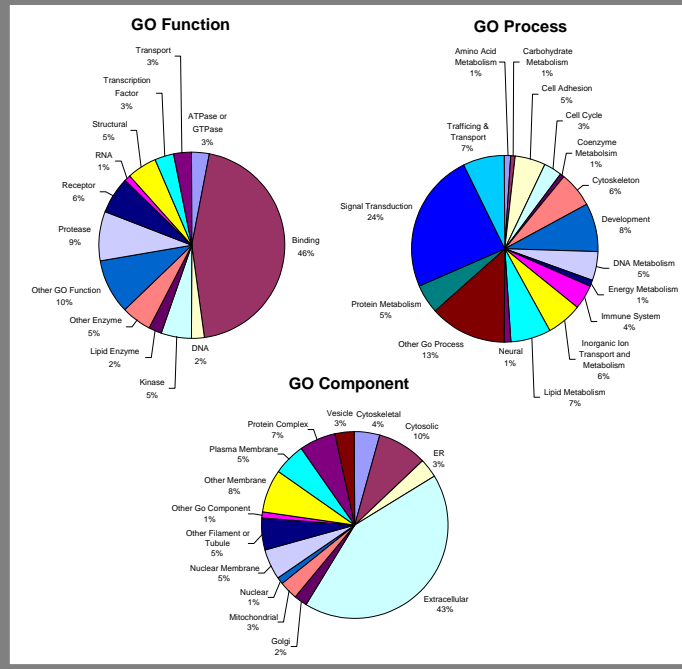
## GO Annotations:

FIGURE 1. Percentage of protein identifications with GO annotations on the Swiss-Prot-TrEMBL server. Overall average of all GO annotations presented.



- On the average, 95% of the proteins identified had at least one GO annotation in the Swiss-Prot-TrEMBL database.
- On the average, 76% of the proteins identified had a GO annotation from a particular category (i.e. component, function or process).

FIGURE 2. Grouped GO annotations for unique protein identifications from LysC digests. Group names and percentage of total identified proteins having GO annotations are given.



The distribution of GO annotations provides a distinct pattern that is characteristic of the sample type and preparation method.

- The distributions obtained using different proteases are quite similar, indicating that there is no significant protein-type selectivity advantage for any one protease.
- Individual GO annotations for each protein (not shown) provide insight for planning subsequent experimentation.

## Posttranslational Modifications:

TABLE 2. Selection of previously reported posttranslational modification data mined from the public Swiss-Prot-TrEMBL server ([www.expasy.org](http://www.expasy.org)) for each unique protein identification obtained from LysC digests. Numbers in grid refer to the sequence position.

ACCESSION	TREMBL	NAME	Phosphoserine	Phosphothreonine	Phosphotyrosine	Phosphoserine	Sulfotyrosine	N-acetylation	Hydroxylation	Hydroxylation by PKA	Potential	Phosphotyrosine by SBC	Phosphotyrosine by SBC in vitro
O02033	AFB1_HUMAN	AF-3 complex subunit beta-1	276										
O00506	STK25_HUMAN	Serine/threonine-protein kinase 25	342	168									
O14578	CTRO_HUMAN	Citron Rho-interacting kinase	1971										
O15089	TMCC2_HUMAN	Central protein 11											
P00352	AL1A1_HUMAN	Retinal dehydrogenase 1						2					
P00451	F8S_HUMAN	Coagulation factor VIII							365				
P01024	CO3L_HUMAN	Complement C3								1			
P01860	IGHG3_HUMAN	Ig gamma-3 chain C region								24			
P02852	APOA2_HUMAN	Apolipoprotein A-II								31			
P02675	FIBB_HUMAN	Fibrinogen beta chain								32			
P02751	FINC_HUMAN	Fibronectin precursor (FN)	2384							19			
P02763	A1AG1_HUMAN	Alpha-1-acid glycoprotein 1											
P02765	FETUA_HUMAN	Alpha-2-HS-glycoprotein								138			
P04264	KC1_HUMAN	Keratin, type II cytoskeletal 1								21			
P05060	SCG1_HUMAN	Secretogranin-1	149							341			
P05090	APOD_HUMAN	Apolipoprotein D									21		
P06366	GELS_HUMAN	Gelsolin										465	651
P0C0L4	CO4A_HUMAN	Complement C4-A									1422		
P0C0L5	CO4B_HUMAN	Complement C4-B precursor									1417		
P10451	O5TP_HUMAN	Osteopontin	263	185									
P10809	CH60_HUMAN	60 kDa heat shock protein	70		227								
P11137	MAP2_HUMAN	Microtubule-associated protein 2 (MAP 2)	1799										
P13591	NCAM1_HUMAN	Neural cell adhesion molecule 1	774										
P17558	KPL_HUMAN	B-phosphotransferase, liver type							633				
P19552	A1AG2_HUMAN	Alpha-1-acid glycoprotein 2								19			
P35663	CYLC1_HUMAN	Cyclin-1											
P36955	PEDF_HUMAN	Pigment epithelium-derived factor	541	543									
P49454	CENPF_HUMAN	Centromere protein F											
P49792	RBP2_HUMAN	E3 SUMO-protein ligase RanBP2	274										
P51825	AFF1_HUMAN	AF4/FMR2 family member 1	2290	2450									
P68871	HBB_HUMAN	Hemoglobin subunit beta	588	766									
Q01814	AT2B2_HUMAN	Plasma membrane calcium-transporting ATPase 2									145		
Q09666	AHNAK_HUMAN	Neuroblast differentiation-associated protein	2911	243									
Q12934	BFSP1_HUMAN	Flilensin										5	
Q13043	STK4_HUMAN	Serine/threonine-protein kinase 4								433			
Q13061	TRDN_HUMAN	Triadin											409
Q13127	REST_HUMAN	RE1-silencing transcription factor	864										864
Q13315	ATM_HUMAN	Atm-protein kinase ATM	367	1985									367
Q14684	RRP1B_HUMAN	RRP1-like protein B	515										515
Q15057	CENB2_HUMAN	Centaurin-beta 2 (Cn1-b2)	775										775
Q59996	AKAP9_HUMAN	A-kinase anchor protein 9	3869										3869
Q92926	PAK7_HUMAN	Serine/threonine-protein kinase PAK 7											
Q914E8	UBP15_HUMAN	Ubiquitin carboxyl-terminal hydrolase 15	225										

Although phosphorylation and acetylation are highlighted in Table 2, all posttranslational modifications and metal binding site information described on the Swiss-Prot-TrEMBL server were harvested.

Sites of posttranslational modifications provide information necessary for a targeted re-interrogation of the data or for making changes experimental in design.

## Conclusions

- Combining results from LysC and Arg C digests significantly enhances proteome coverage.
- Automation of GO and PTM annotation searching has been accomplished, providing an expedient first step in a targeted proteomics workflow.