

# Intelligent Data Annotation Workflows Applied to the Characterization of Human Cerebrospinal Fluid



Andreas F. R. Hühmer<sup>1</sup>, Zhiqi. Hao<sup>1</sup>, Sivio Wiedrich<sup>2</sup>, Ryan Bowen<sup>2</sup>, William Newell<sup>2</sup>, Andreas Matern<sup>2</sup>, Roger G. Biringer<sup>1</sup>

<sup>1</sup>Thermo Fisher Scientific, San Jose, CA, United States, <sup>2</sup>Inforsense, London, United Kingdom.

## Overview

**Purpose:** Establish automated annotation workflows for LC-MS/MS data.

**Methods:** InforSense-based workflow tools were integrated into a new MS data analysis package (Proteome Discoverer) and applied to the evaluation of LC-MS/MS data of human cerebrospinal fluid (CSF).

**Results:** 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.

## 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 and UniprotKB/Swiss-Prot 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 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.

Digests were prepared with several different proteases, and individually analyzed on an LTQ XL™ mass spectrometer equipped with ETD, either with CID alone or with a combination of CID and ETD to achieve superior sequence coverage and additional protein identifications. Protein sequences were annotated using InforSense-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

**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 an 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 Proteome Discoverer 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. CID results were processed with the SEQUEST® search engine and ETD data processed with the Z.Core search engine. Data for three successive identical experiments were combined and evaluated.

**Workflow Development:** The annotation tools were developed using the InforSense® Platform ([www.inforsense.com](http://www.inforsense.com)) and were integrated into a new MS data analysis package, Proteome Discoverer 1.0. The workflows automatically retrieve pertinent information from public databases about each protein identified by Proteome Discoverer. Information retrieved by the workflows provide annotations including, but not limited to, GO (Gene Ontology, <http://www.geneontology.org>) classifications, sites of post-translational modifications, PubMed references, and genomic information.

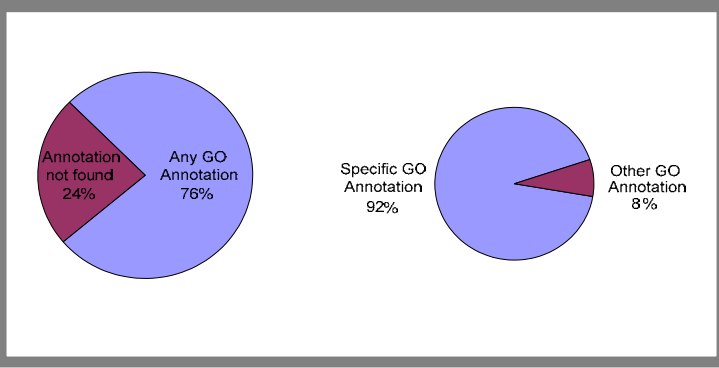
**Results Annotation:** InforSense-based workflows were executed through Proteome Discoverer 1.0 by the InforSense virtual machine to extract GO annotations and descriptive information from the public Swiss-Prot-TrEMBL (<http://www.expasy.org>) and NCBI (protein, <http://www.ncbi.nlm.nih.gov>) servers for each protein identified. Only results retrieved from the former server are depicted here. GO annotations for each of the three GO categories (component, function, and process) are individually sorted into biologically meaningful groups, including a catch-all "other GO group" for those that do not quite fit the others. Additional descriptive information is parsed into several categories, including locations of known and predicted posttranslational modifications.

## Results

### GO Annotations:

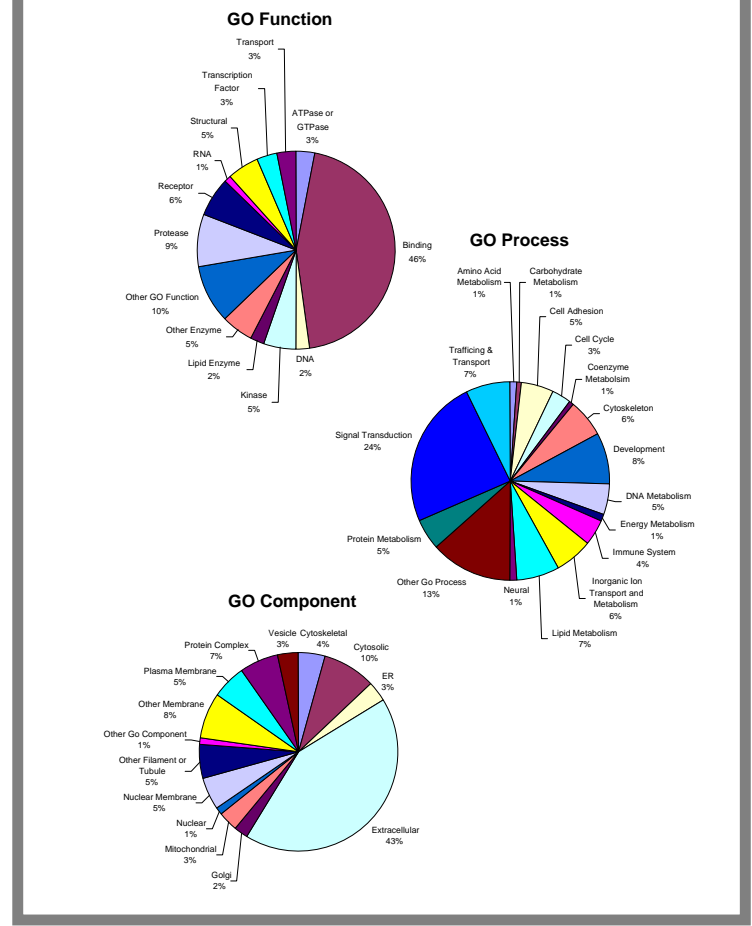
- On the average, 76% of the proteins identified had at least one GO annotation in the Swiss-Prot-TrEMBL database.
- On the average, 92% of the GO Annotated proteins could be sorted into a biologically meaningful group.

FIGURE 1. Percentage of protein identifications with GO annotations on the Swiss-Prot-TrEMBL server. Overall average of all GO annotations (<http://www.geneontology.org>) presented.



- 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 (other data not shown) are quite similar, indicating that there is no significant protein-type selectivity advantage for any one protease.
- Individual GO annotations for each protein (Table 1) provide insight for planning subsequent experimentation.

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.



## Posttranslational Modifications:

TABLE 1. Selection of previously reported posttranslational modification data mined from the public Swiss-Prot-TrEMBL server (<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	Pyroglutamate	Hydroxylation	Hydroxyproline	Hydroxylysine	Hydroxyproline by PMA	Hydroxyproline by SBC	Hydroxyproline by SBC in vitro
Q00203	AFB1_HUMAN	AF-B1 complex subunit beta 1	278									
O00506	STK25_HUMAN	Serine/threonine-protein kinase 25	342	168								
O14578	CTRO_HUMAN	Citron Rho-interacting kinase	1971									
O70809	TMCC2_HUMAN	Cerebral protein 11										
P00352	AL1A1_HUMAN	Retinal dehydrogenase 1							2			
P00445	FAB_HUMAN	Coagulation factor VIII							365			
P01024	CO3_HUMAN	Complement C3							1			
P01860	IGHG3_HUMAN	Ig gamma-3 chain C region							24			
P00852	AFPA2_HUMAN	Fibrinogen alpha-2 chain							31			
P02675	FIBB_HUMAN	Fibrinogen beta chain							32			
P02751	FN1_HUMAN	Fibronectin precursor (FN)	2384									
P02763	AT1A1_HUMAN	Alpha-1-acid glycoprotein 1							19			
P02765	FETUA_HUMAN	Alpha-2-HS-glycoprotein	138									
P04264	KCC1_HUMAN	Kainin, type II cytoskeletal 1	21									
P05060	SCG1_HUMAN	Secretogranin-1	149						341			
P05090	APOD_HUMAN	Apolipoprotein D							21			
P06396	CELS_HUMAN	Celastrolin									465	651
POC0L4	CO4A_HUMAN	Complement C4-A							1422			
POC0L5	CO4B_HUMAN	Complement C4-B precursor							1417			
P10451	CSTP_HUMAN	Citosteronin	263	185								
P10809	CH60_HUMAN	60 kDa heat shock protein	70						227			
P11137	MAP2_HUMAN	Microtubule-associated protein 2 (MAP 2)	1799									
P13591	NCA11_HUMAN	Neural cell adhesion molecule 1	774									
P17856	KIF1A_HUMAN	Kinesin, type I cytoskeletal 1							633			
P18652	AT1A2_HUMAN	Alpha-1-acid glycoprotein 2							19			
P35663	CYLC1_HUMAN	Cyclin-1	541	543								
P36955	PECF_HUMAN	Pigment epithelium-derived factor							20			
P49454	CENPF_HUMAN	Centromere protein F	274									
P49792	RBP2_HUMAN	E3 SUMO-protein ligase RanBP2	2290	2450								
P51925	AF1_HUMAN	AF-1/MEF2 family member 1	588	768								
P68871	HB_HUMAN	Hemoglobin subunit beta									145	
Q01814	AT2B2_HUMAN	Plasma membrane calcium-transporting ATPase 2										
Q06966	AHMK_HUMAN	Neuroblast differentiation-associated protein	2911	243								
Q12934	BFSP1_HUMAN	Filensin									5	
Q13043	STK4_HUMAN	Serine/threonine-protein kinase 4										
Q13061	TRDN_HUMAN	Triadin	409									
Q13127	REST_HUMAN	RE-1-silencing transcription factor	864									
Q13315	ATM_HUMAN	Serine-protein kinase ATM	367	1985								
Q14684	RRP1B_HUMAN	RRP1-like protein B	513									
Q15057	CENBF_HUMAN	Centaurin-beta 2 (Cyt-b2)	775									
Q99996	AKAP9_HUMAN	A-kinase anchor protein 9	3869									
Q9P286	PAK7_HUMAN	Serine/threonine-protein kinase PAK 7										
Q9Y4E5	UBP15_HUMAN	Ubiquitin carboxy-terminal hydrolase 15	229									

- Although phosphorylation and acetylation are highlighted in Table 1, all posttranslational modifications and metal binding site information described on the Swiss-Prot-TrEMBL server were harvested.
- Sites of post translational modifications provide information necessary for a targeted re-analysis of the data or for making changes experimental in design.

## Conclusions

- Comprehensive annotation of database search results using GO terminology can be accomplished with InforSense-based workflow tools.
- Automated GO annotation capabilities provide specific biological context about complex protein mixture.
- Annotation results can provide critical information for additional data mining steps and targeted follow-up wet-lab experiments.

SEQUEST is a registered trademark of the University of Washington. InforSense is a registered trademark of InforSense Ltd. All other trademarks are the property of Thermo Fisher Scientific Inc. and its subsidiaries.