

Automated Annotation of Complex Proteomic Data Obtained Using Linear Ion Trap LC-MS/MS with ETD: Analysis of Human Cerebrospinal Fluid

R. G. Biringer¹, Z. Hao¹, H. Tran¹, M. G. Harrington², A. F. R. Hühmer¹

¹Thermo Fisher Scientific, San Jose, CA, United States, ²Huntington Medical Research Institutes, Pasadena, CA, United States

Abstract

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

Methods: KDE (knowledge discovery environment)-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 protein 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. The annotation tool is integrated into a new MS data analysis package (Proteome Discoverer) that automatically retrieves pertinent information about each identified protein from public databases. Information retrieved by the workflows provide annotation including, but not limited to, GO (Gene Ontology, <http://www.geneontology.org>) classifications, sites of post-translational modifications (PTMs), PubMed[®] references, and genomic information. 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. Peptide sequences were annotated using KDE-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.

Materials & Methods

Sample Preparation: 360 mg solid urea was dissolved in 500 μ L 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 equimolar amount of DTT (1 hr). Samples were split into fractions containing 50 μ g 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 CaCl₂ 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 μ L of each CSF digest (3.1 μ g total peptide) was directly injected onto a peptide trap (CapTrap, Michrom). Peptides were eluted onto and through a C18 column, 25 cm X 100 μ m (Micro-Tech Scientific, San Diego) 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.

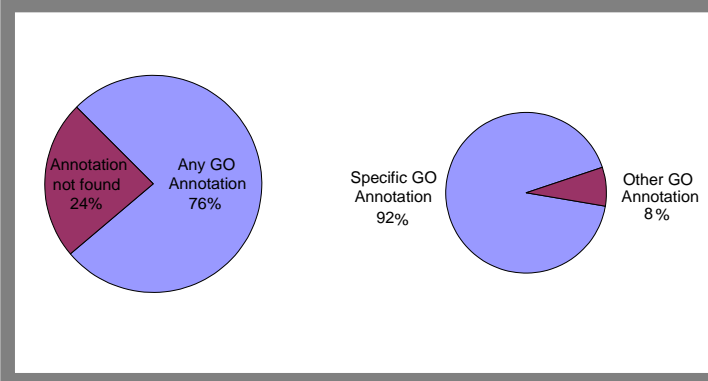
Data-MS Analysis: All MS data were analyzed with Proteome Discoverer using a human Swiss-Prot-TrEMBL database and the results filtered to a 5% maximal false positive rate using a reverse database search approach. CID results were processed with the Sequest search engine and ETD data processed with the ZCore search engine (see also poster V81-T). 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 (<http://www.expasy.org>) or NCBI (protein, <http://www.ncbi.nlm.nih.gov>) servers for each protein identified with Proteome Discoverer. Only results 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 one of the classifications. Additional descriptive information is parsed into several categories, including locations of known and predicted posttranslational modifications.

Results

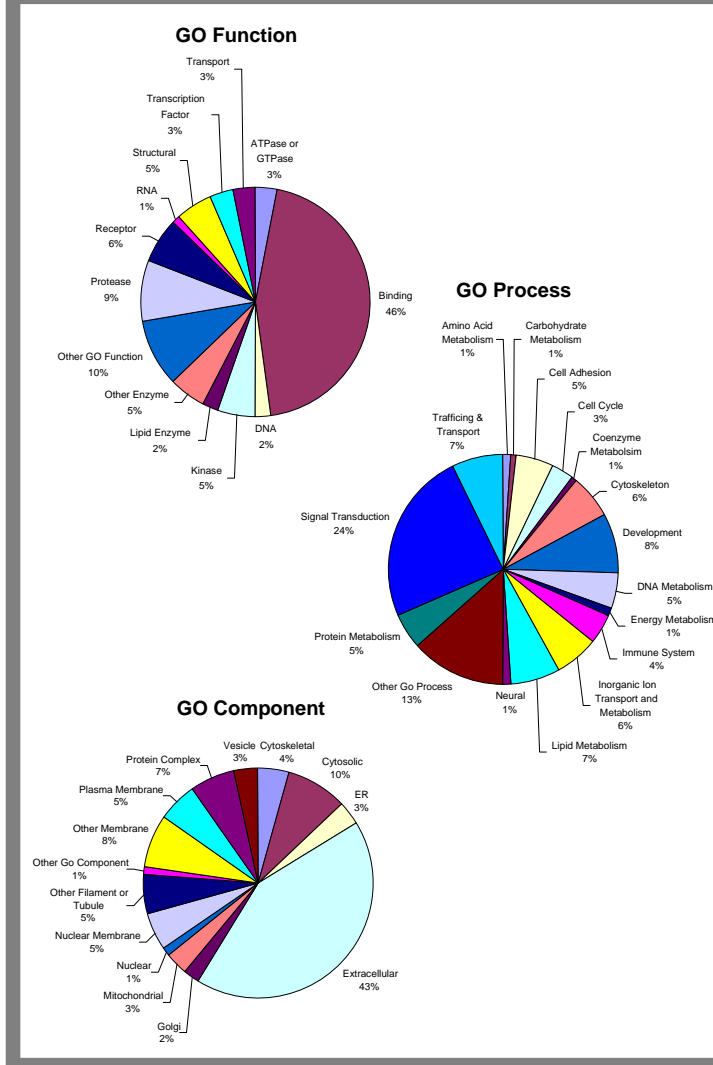
GO Annotations:

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.



- 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 2. Grouped GO annotations for unique protein identifications from LysC digests. Group names and percentages 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 (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.

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	Sulfhydryl	Hydroxylation	N-acetylmethionine	N-acetylcysteine	Phosphoserine by PKA	Phosphoserine by PKC	Phosphoserine by SRC	Phosphoserine by SRC in vitro
O00203	AP3B1_HUMAN	AP-3 complex subunit beta-1	276											
O00506	STK25_HUMAN	Serine/threonine-protein kinase 25	342	168										
O14578	CTRD_HUMAN	Cation Rho-interacting kinase	1971											
O75069	TMC22_HUMAN	Cembril protein 11												
P00352	ALIA1_HUMAN	Retinal dehydrogenase 1												
P00451	FAB_HUMAN	Coagulation factor VIII												
P01024	CD3_HUMAN	Complement C3												
P01860	IGHG3_HUMAN	Ig gamma-3 chain C region												
P02652	APOA2_HUMAN	Apolipoprotein A-II												
P02675	FIBB_HUMAN	Fibrinogen beta chain												
P02751	FINC_HUMAN	Fibrinectin precursor (FN)	2384											
P02763	A1AG1_HUMAN	Alpha-1-acid glycoprotein 1												
P02765	FETUA_HUMAN	Alpha-2-HS-glycoprotein	138											
P04264	K2C1_HUMAN	Keratin, type II cytoskeletal 1	21											
P05060	SCG1_HUMAN	Secretogranin-1	149											
P05090	APOD_HUMAN	Apolipoprotein D												
P06396	CELS_HUMAN	Celsin												
P0C0L4	CO4A_HUMAN	Complement C4-A												
P0C0L5	CO4B_HUMAN	Complement C4-B precursor												
P10451	CSTP_HUMAN	Catepsin B	263	185										
P10809	CH80_HUMAN	80 kDa heat shock protein	70											
P11137	MAP2_HUMAN	Microtubule-associated protein 2 (MAP 2)	1799											
P12591	NCA11_HUMAN	Neural cell adhesion molecule 1	774											
P17838	KIF8_HUMAN	B-phosphotubulin, liver type												
P18652	A1AG2_HUMAN	Alpha-1-acid glycoprotein 2	541	543										
P35683	CYLC1_HUMAN	Cyclin-1												
P36955	PEDE_HUMAN	Pigment epithelium-derived factor												
P49454	CENPF_HUMAN	Centromere protein F	274											
P49792	RBP2_HUMAN	E3 SUMO-protein ligase RanBP2	2290	2450										
P51825	AFF1_HUMAN	AFA1/AF2 family member 1	588	766										
P68871	HBB_HUMAN	Hemoglobin subunit beta												
Q01814	AT2B2_HUMAN	Plasma membrane calcium-transporting ATPase 2												
Q09695	AHMK_HUMAN	Neuroblast differentiation-associated protein	2911	243										
Q12934	BFSP1_HUMAN	Filensin												
Q13043	STK4_HUMAN	Serine/threonine-protein kinase 4												
Q13051	TRDN_HUMAN	Tradin	409											
Q13177	REST_HUMAN	RE1-silencing transcription factor	864											
Q13315	ATM_HUMAN	Serine-protein kinase ATM	367	1985										
Q14684	RRP1B_HUMAN	RRP1-like protein B	513											
Q15057	CENBE2_HUMAN	Centasin-beta 2 (Cen-b2)	773											
Q99996	AKAP9_HUMAN	A-kinase anchor protein 9	3869											
Q8P286	PAK7_HUMAN	Serine/threonine-protein kinase PAK 7												
Q9Y4E9	UBP15_HUMAN	Ubiquitin carboxyl-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 posttranslational modifications provide information necessary for a targeted re-analysis of the data or for making changes in experimental design.

Conclusions

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

© 2008 Thermo Fisher Scientific Inc. All rights reserved. InforSense is a registered trademark of InforSense Ltd. PubMed is a registered trademark of the National Library of Medicine. SEQUEST is a registered trademark of the University of Washington. All other trademarks are the property of Thermo Fisher Scientific Inc. and its subsidiaries.

Thermo
SCIENTIFIC