

The Application of New Software Tools to Quantitative Protein Profiling Via Isotope-coded Affinity Tag (ICAT) and Tandem Mass Spectrometry

I. STATISTICALLY ANNOTATED DATASETS FOR PEPTIDE SEQUENCES AND PROTEINS IDENTIFIED VIA THE APPLICATION OF ICAT AND TANDEM MASS SPECTROMETRY TO PROTEINS COPURIFYING WITH T CELL LIPID RAFTS*[§]

Priska D. von Haller†§, Eugene Yi‡, Samuel Donohoe, Kelly Vaughn, Andrew Keller, Alexey I. Nesvizhskii, Jimmy Eng, Xiao-jun Li, David R. Goodlett, Ruedi Aebersold, and Julian D. Watts¶

Lipid rafts were prepared according to standard protocols from Jurkat T cells stimulated via T cell receptor/CD28 cross-linking and from control (unstimulated) cells. Co-isolating proteins from the control and stimulated cell preparations were labeled with isotopically normal (d0) and heavy (d8) versions of the same isotope-coded affinity tag (ICAT) reagent, respectively. Samples were combined, proteolyzed, and resultant peptides fractionated via cation exchange chromatography. Cysteine-containing (ICAT-labeled) peptides were recovered via the biotin tag component of the ICAT reagents by avidin-affinity chromatography. On-line micro-capillary liquid chromatography tandem mass spectrometry was performed on both avidin-affinity (ICAT-labeled) and flow-through (unlabeled) fractions. Initial peptide sequence identification was by searching recorded tandem mass spectrometry spectra against a human sequence data base using SEQUEST™ software. New statistical data modeling algorithms were then applied to the SEQUEST™ search results. These allowed for discrimination between likely “correct” and “incorrect” peptide assignments, and from these the inferred proteins that they collectively represented, by calculating estimated probabilities that each peptide assignment and subsequent protein identification was a member of the “correct” population. For convenience, the resultant lists of peptide sequences assigned and the proteins to which they corresponded were filtered at an arbitrarily set cut-off of 0.5 (i.e. 50% likely to be “correct”) and above and compiled into two separate datasets. In total, these data sets contained 7667 individual peptide identifications, which represented 2669 unique peptide sequences, corresponding to 685 proteins and related protein groups. *Molecular & Cellular Proteomics* 2:426–427, 2003.

DATASET DESCRIPTION

Individual lipid raft preparations were made from 2.5×10^8 Jurkat T cells (control or stimulated) as described elsewhere.¹ Stimulation was via cross-linking of the T cell receptor and CD28 coreceptor with monoclonal antibodies (clones OKT3 and 9.3, respectively) for 2 min at 37 °C. Isotope-coded affinity tag (ICAT)² labeling of proteins co-isolating with the lipid rafts was according to standard protocols (1–3),¹ labeling the control and stimulated samples with the d0- and d8-ICAT reagents, respectively. Following sample pooling and tryptic proteolysis, peptide fractionation was via cation exchange, followed by avidin affinity chromatography, again according to standard procedures (1, 3).¹ This protocol was performed twice, with different cell preparations, under identical conditions.

ICAT-labeled peptide fractions pools from the two iterations of the experiment, as well as the avidin-affinity flow-through fractions (unlabeled peptides) from one iteration of the experiment, were analyzed by microcapillary-liquid chromatography tandem mass spectrometry. This was done in an automated fashion, according to standard in-house protocols (1, 4), using an LCQ-DECA ion-trap mass spectrometer (ThermoFinnigan, San Jose, CA) equipped with an in-house built micro-spray device. For reasons of convenience and to facilitate subsequent comparisons, the tandem mass spectrometry data were separated into three, smaller, data subsets: 1) ICAT-labeled avidin affinity-purified fractions from experiment 1 (ICAT 1); 2) ICAT-labeled avidin affinity-purified fractions from experiment 2 (ICAT 2); and 3) the unlabeled avidin affinity flow-through fractions from experiment 1 (Flow-through 1). These three data subsets represented a combined total of

From the Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103

Received, May 8, 2003, and in revised form, June 24, 2003

Published, MCP Papers in Press, June 25, 2003, DOI 10.1074/mcp.D300002-MCP200

¹ von Haller, P. D., Yi, E., Donohoe, S., Vaughn, K., Keller, A., Nesvizhskii, A. I., Eng, J., Li, X., Wollscheid, B., Goodlett, D. R., Aebersold, R., and Watts, J. D., manuscript in preparation.

² The abbreviation used is: ICAT, isotope-coded affinity tag.

101,799 tandem mass spectrometry spectra, which were in turn searched against a locally maintained human protein sequence data base using SEQUESTTM software (5). Search parameters used included provision for both unmodified and oxidized (+16 Da) methionine, as well as for d0-ICAT (+442.2 Da)- and d8-ICAT (+450.2 Da)-labeled cysteine, at a mass tolerance ± 3 Da with no proteolytic enzyme specified.

SEQUESTTM output files for each of the three separate data subsets were next separately submitted to PeptideProphetTM (6) for statistical data modeling and the generation of p_{comp} scores for each peptide assigned by SEQUESTTM. The output files generated by PeptideProphetTM for all three data subsets were finally combined and submitted to ProteinProphetTM (7), again for statistical data modeling and the generation of P_{comp} scores for each protein identified. p_{comp} is the computed probability, for each peptide sequence assignment made by data base searching, that it is a member of the population of "correct" assignments, on a scale of 0 (for "incorrect") to 1 (for "correct") (6). Likewise, P_{comp} is the computed probability, for each potential protein identification inferred from the observed peptide data, that it is a member of the population of "correct" identifications, again on a scale of 0 (for "incorrect") to 1 (for "correct") (7).

The final data sets of observed peptides and the proteins they represented were separately filtered for size at an arbitrarily chosen cut-off of $p_{\text{comp}} \geq 0.5$ and $P_{\text{comp}} \geq 0.5$, respectively (*i.e.* peptides and proteins that are 50% likely to be correct and higher). The final peptide list contained 7,667 separate peptide assignments, given in supplementary Table I, which represented 2,669 unique peptide sequences. The final protein list contained 685 separate protein and related protein group identifications and is given in supplementary Table II.

* This work was supported in part by grants from the National Institutes of Health (RO1-AI-41109-01 and RO1-AI-51344-01 to R.A.

and J.W., respectively), the National Heart, Lung, and Blood Institute Proteomics Center at the Institute for Systems Biology (N01-HV-28179), and a fellowship awarded by the Swiss National Science Foundation to P.D.H. We thank Oxford GlycoSciences (UK) for additional generous financial support. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ The on-line version of this article (available at <http://www.mcponline.org>) contains Supplemental Tables I and II.

‡ P.D.H. and E.Y. contributed equally to this work.

§ Current address: MacroGenics, 1441 North 34th Street, Seattle, WA 98103.

¶ To whom correspondence should be addressed. Tel.: 206-732-1283; Fax: 206-732-1299; E-mail: jwatts@systemsbiology.org.

REFERENCES

1. von Haller, P. D., Yi, E., Donohoe, S., Vaughn, K., Keller, A., Nesvizhskii, A. I., Eng, J., Li, X., Goodlett, D. R., Aebersold, R., and Watts, J. D. (2003) The application of new software tools to quantitative protein profiling via ICAT and tandem mass spectrometry: II. Evaluation of tandem mass spectrometry methodologies for large-scale protein analysis, and the application of statistical tools for data analysis and interpretation. *Mol. Cell. Proteomics* **2**, 428–442
2. Smolka, M. B., Zhou, H., Purkayastha, S., and Aebersold, R. (2001) Optimization of the isotope-coded affinity tag-labeling procedure for quantitative proteome analysis. *Anal. Biochem.* **297**, 25–31
3. Han, D. K., Eng, J., Zhou, H., and Aebersold, R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* **19**, 946–951
4. Yi, E. C., Marelli, M., Lee, H., Purvine, S. O., Aebersold, R., Aitchison, J. D., and Goodlett, D. R. (2002) Approaching complete peroxisome characterization by gas-phase fractionation. *Electrophoresis* **23**, 3205–3216
5. Eng, J., McCormack, A. L., and Yates, J. R. 3rd (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989
6. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392
7. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.*, in press.