A Quantitative Proteomics Design for Systematic Identification of Protease Cleavage Events*

Francis Impens‡§¶, Niklaas Colaert‡§, Kenny Helsens‡§‖, Bart Ghesquière‡§, Evy Timmerman‡§, Pieter-Jan De Bock‡§, Benjamin M. Chain**‡‡, Joël Vandekerckhove‡§, and Kris Gevaert‡§ §§

We present here a novel proteomics design for systematic identification of protease cleavage events by quantitative N-terminal proteomics, circumventing the need for time-consuming manual validation. We bypass the singleton detection problem of protease-generated neo-N-terminal peptides by introducing differential isotopic proteome labeling such that these substrate reporter peptides are readily distinguished from all other N-terminal peptides. Our approach was validated using the canonical human caspase-3 protease and further applied to mouse cathepsin D and E substrate processing in a mouse dendritic cell proteome, identifying the largest set of protein protease substrates ever reported and gaining novel insight into substrate specificity differences of these cathepsins. ***Molecular & Cellular Proteomics 9:2327–2333, 2010.***

Several protocols for proteome-wide identification of protease processing events were recently published. They all follow strategies in which N-terminal peptides, including neo-N-terminal peptides generated by protease action, are enriched from whole proteome digests before identification (*e.g.* Refs. 1–4). LC-MS/MS analyses of these peptides often yield hundreds of processing events identified in a single experiment (*e.g.* Refs. 3–5). The N-terminal COFRADIC[1] technology developed in our laboratory (6) has been successful in identifying cleavage events of both canonical (*e.g.* caspases-3 and -7 (7)) and non-canonical proteases (*e.g.* HtrA2/Omi (8)). Differential stable isotopic labeling in particular, necessary to univocally distinguish genuine neo-N-terminal peptides, allows analyzing control and protease-treated proteomes in a single run. However, this also introduces the most important bottleneck of the technology: verifying whether the peptide envelope of a neo-N-terminal peptide only carries the isotopic label of the protease-treated sample (see Fig. 1*A*) often had to be done manually for each identified peptide. This "singleton detection problem" can to some extent be automated by software routines such as ProteinProspector (http://prospector.ucsf.edu/prospector/mshome.htm), the MASCOT Distiller Quantitation Toolbox (www.matrixscience.com/distiller.html), and ICPLQuant (9), although these often need specific or proprietary data formats or can only handle MALDI-MS data (9), and researchers still need to individually check correct calling of a neo-N-terminal peptide (10).

To fully overcome this singleton detection problem, here we present and validate a method for highly automated, software-based quantification and annotation of protein processing events on a proteomics scale based on stable isotopic labeling and positional proteomics. We illustrate its strength by generating the largest set of cathepsin D and E substrates hitherto reported. Furthermore, differences in the specificity profiles of these non-canonical proteases are illustrated by the validation of a cleavage event specific for cathepsin E in filamin-A.

RESULTS

Rationale of Approach—In a typical proteomics hunt for protease substrates, two differently labeled proteomes are used; one is incubated with a protease of interest, whereas the second serves as a control (5, 7). Following mixing equal amounts of both samples and enrichment of N-terminal peptides, neo-N-terminal peptides reporting protease processing appear as singletons in MS spectra, whereas N-terminal peptides not affected by the added protease present themselves as doublets (Fig. 1*A*). Although the latter are readily quantified, the former are typically missed by quantification algorithms. In fact, such singleton peptides are best considered as extremely regulated peptides, and an accurate calling of their ratios appears cumbersome (see below and supplemental Fig. 1), explaining why such peptides often go undetected.

To circumvent this quantification problem, here we introduce a simple solution independent of the protocol used and its associated chemistries. We used SILAC (11) to label proteomes with light (L;$^{12}C_6$) or heavy (H; $^{13}C_6$) isotopic variants of arginine, although the principle of the method is broadly applicable and thus compatible with other isotopic labeling strategies. Arginine is used here because during COFRADIC

From the ‡Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium, §Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium, and **Division of Infection and Immunity, University College London, London WC1E 6BT, United Kingdom

Received, May 27, 2010

Published, MCP Papers in Press, July 13, 2010, DOI 10.1074/mcp.M110.001271

[1] The abbreviations used are: COFRADIC, combined fractional diagonal chromatography; SILAC, stable isotope labeling by amino acids in cell culture; L, light; H, heavy.

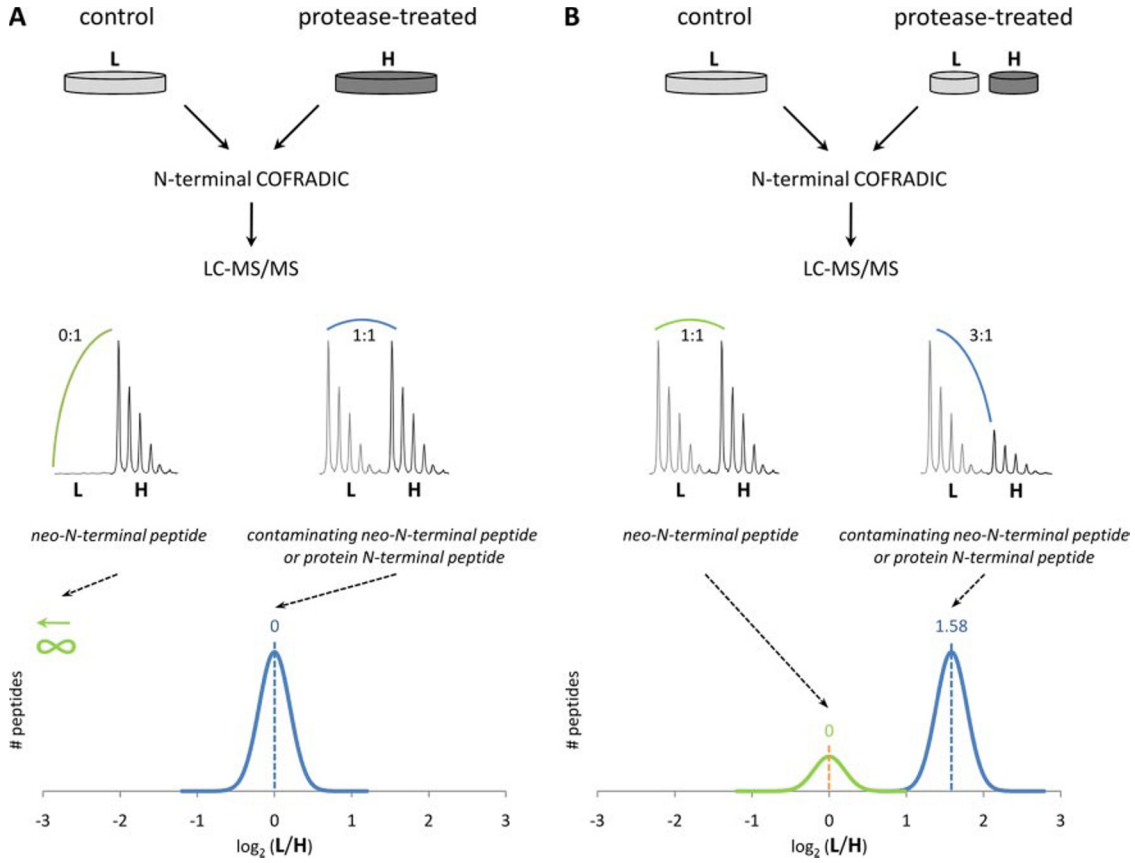
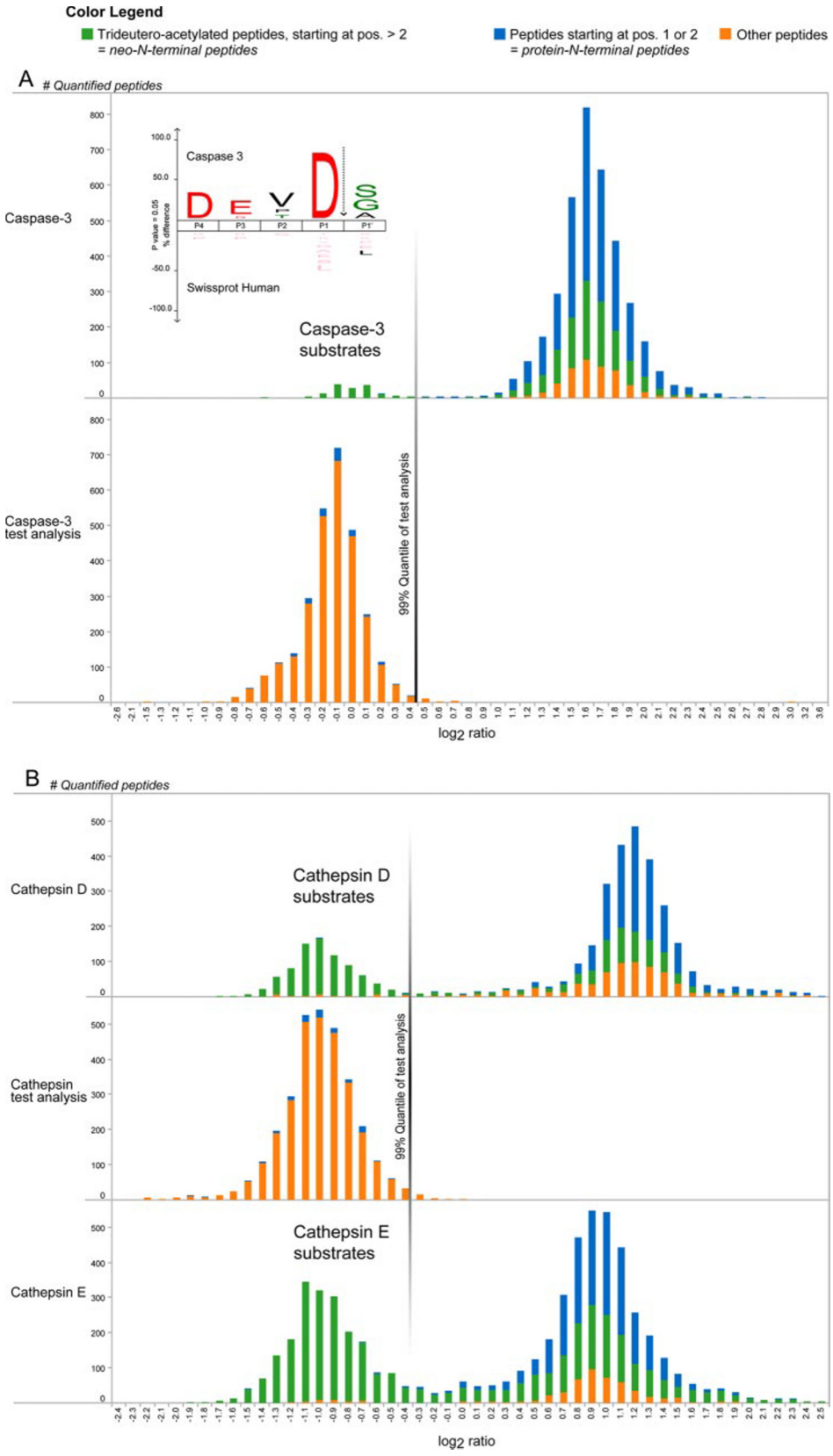


FIG. 1. **Manual *versus* automated annotation of protease cleavage events.** *A*, in a typical setup, a heavy (*H*) labeled proteome is used for protease treatment, and the light (*L*) labeled proteome serves as a control. Following mixing and N-terminal COFRADIC sorting, neo-N-terminal peptides generated by the added protease are present as singletons, whereas all other N-terminal peptides are present as couples with (light/heavy) ratios around 1 (0 in $\log_2$ scale). *B*, a mixture of light and heavy labeled proteins (mixed in a 1:1 ratio) is treated with a protease, and as a result, neo-N-terminal peptides generated by the action of the added protease are now present in light/heavy ratios distributed around 1 (0 in $\log_2$ scale) and are clearly distinct from all other N-terminal peptides that come in ratios around 3 (1.58 in $\log_2$ scale). Both types of peptides are readily quantified, circumventing the need for manual validation.

isolation trideuteroacetylation is performed to block all primary amino groups, leaving only arginine as the trypsin-sensitive site (6). The proteome that will be incubated with the protease of interest is then made by mixing equal parts of L and H proteome preparations, whereas two parts of L proteins serve as control (Fig. 1*B*). Following protease incubation, both samples are mixed, and N-terminal peptides are isolated by N-terminal COFRADIC. As a result, genuine neo-N termini derived from the protease-treated sample now show up as doublets of peptide ions with equal intensities (L/H ratio = 1), whereas all other N-terminal peptides present in both samples (protein N-terminal peptides and neo-N-terminal peptides due to possible contaminating protease activity) also appear as doublets but with highly different ratio values ((L + L + L)/H = 3). Both types of peptides are readily quantified, resulting in clearly distinct ratio distributions of genuine neo-N-terminal *versus* all other N-terminal peptides (Fig. 1*B* and supplemental Fig. 1).

Automated Identification of Caspase-3 Cleavage Sites—We validated our approach by screening for cleavage sites of the canonical protease caspase-3. As key proteases during apoptosis, caspases show an almost absolute requirement for an aspartic acid residue preceding the cleavage site (12), and this specificity was already broadly used to assign cleavage sites to caspase activity (13). The experiment was designed such that the L/H ratios of caspase-3-generated neo-N-terminal peptides were expected to be distributed close to 1, whereas the L/H ratios of all other N-terminal peptides were expected to be around 3. Here, human Jurkat T-cells were arginine SILAC-labeled as described before (5), and a mixed (L/H, 1:1) cell lysate was incubated with 150 nM recombinant human caspase-3; the same total amount of the light labeled cell lysate served as control. Furthermore, before adding recombinant caspase-3, endogenous caspase activity in both lysates was inhibited by cysteine alkylation (7). Following incubation for 1 h at 37 °C, the protease-treated and control samples were mixed and subjected to N-terminal COFRADIC sorting (14). The sorted peptides were then analyzed by LC-MS/MS on a linear trap quadrupole Orbitrap XL mass spectrometer. Spectra were searched with the MASCOT algo-


Color Legend
Trideutero-acetylated peptides, starting at pos. > 2
= neo-N-terminal peptides
Peptides starting at pos. 1 or 2
= protein-N-terminal peptides
Other peptides
A
Quantified peptides
Caspase-3
Caspase 3
P4
P3
P2
P1
P1'
P value = 0.05
% difference
Swissprot Human
Caspase-3 substrates
Caspase-3 test analysis
99% Quantile of test analysis
log2 ratio
B
Quantified peptides
Cathepsin D
Cathepsin D substrates
Cathepsin test analysis
99% Quantile of test analysis
Cathepsin E
Cathepsin E substrates
log2 ratio

rithm, and identified peptides were quantified by the MASCOT Distiller software (an evaluation of the MASCOT Distiller parameters is shown in supplemental Fig. 2). In addition, Peptizer (15) was used to reduce the number of potential false positive identifications. Note that full experimental details and spectra from neo-N-terminal peptides are provided as supplemental experimental procedures and supplemental spectra respectively.

As expected, two types of peptides are observed based on their L/H ratios that were distributed around values 1 and 3 (0 and 1.58 in $\log_2$ scale) (Fig. 2*A*, *upper panel*). Shotgun analysis on an aliquot of the protease-treated sample without prior isolation of N-terminal peptides confirmed that peptides with ratio values around 1 were from the protease-treated proteome (Fig. 2*A*, *lower panel*). Furthermore, the so-derived ratio distribution from this test analysis was used to determine a ratio cutoff value based on the boundary of a one-sided 99% quantile that served to assign peptides to one of both distributions in the first experiment (Fig. 2*A*, *lower panel*). A list of caspase-3 cleavage sites was then compiled from the neo-N-terminal peptides with an L/H ratio value centered around 1 by considering only peptides with start positions beyond the second amino acid of the protein that were further trideuteroacetylated at their N-terminal α-amino group. In total, 141 peptides fulfilled these criteria and pointed to 76 caspase-3-regulated cleavage events in 72 proteins (supplemental Table 1). Without exception, these peptides were generated by cleavage after aspartic acid, consistent with the known specificity of caspase-3 (Fig. 2*A*) (12). The fact that no other neo-N-terminal peptides were found as the result of caspase-3 cleavage validates our approach and indicates its sensitivity to identify highly confident protease cleavage sites without manual validation.

Cathepsin D and E Cleavage Events—In a second experiment, a similar approach was used to map cleavage events of the non-canonical mouse cathepsins D and E. Both cathepsins are intracellular aspartic proteases with little specificity differences thus far reported (16). Although cathepsin D is ubiquitously expressed in lysosomes, cathepsin E is a non-lysosomal protease mainly present in immune cells (17). Because it was shown that cathepsin E functions in antigen processing in dendritic cells (18), lysates of arginine SILAC-labeled primary mouse dendritic cells were used to screen for cleavage events. Such non-immortalized, primary cells pose difficulties for SILAC labeling as such cell populations are not able to fully incorporate the SILAC label into proteins, resulting in an incompletely labeled proteome. Given the experimental setup of the caspase-3 experiment, we decided to turn incomplete labeling of this primary cell population into an opportunity for automated quantification and annotation of cathepsin cleavage sites. Therefore a freeze-thaw lysate of such an incompletely (68% H, 32% L) arginine SILAC-labeled cell population was incubated with 20 nM recombinant mouse cathepsin D or E; a lysate of non-labeled (L) dendritic cells to which pepstatin was added to inhibit endogenous cathepsin D and E activity served as control. Following incubation for 15 min at 37 °C, control and protease-treated samples were mixed in equal amounts and further analyzed as described above.

Peptide L/H ratios were now distributed around 0.5 (-1 in $\log_2$ scale) and 2 (1 in $\log_2$ scale; Fig. 2*B*). A test analysis was again performed to determine the ratio cutoff value (Fig. 2*B*, *middle panel*), and genuine neo-N-terminal peptides were considered based on the same criteria as described above. In this way, 790 neo-N-terminal peptides (584 cleavage sites in 340 proteins) were quantified in the cathepsin D study, and 1,967 neo-N-terminal peptides (1,231 cleavage sites in 639 proteins) were quantified in the cathepsin E study (supplemental Tables 2 and 3). 314 cleavage sites in 202 proteins were found in both analyses in line with the similar subsite specificity reported previously (16). Similar specificities are also supported by the sequence logos shown in Fig. 3 (iceLogo (19)). In accordance with previous studies, these subsite specificity profiles indicate that cleavage mainly occurs between hydrophobic residues with a high preference for leucine or phenylalanine residues preceding the cleavage site. However, subtle differences between both proteases appear (Fig. 3), and to validate these, we incorporated the cathepsin E cleavage site identified in filamin-A in a peptide substrate and monitored its cleavage by cathepsins D and E by reversed phase HPLC. This site was chosen particularly because it contains several features of cathepsin E-specific sites as revealed by the iceLogo analyses and has no obvious counterpart in the identified cathepsin D substrates. As expected, only cleavage

FIG. 2. **Identified human caspase-3 and mouse cathepsin D and E cleavage sites.** *A*, given the experimental design described in the main text, caspase-3-generated neo-N-terminal peptides show light/heavy ratios of about 1 (0 in $\log_2$ scale), whereas all other N-terminal peptides hold ratio values distributed around 3 (1.58 in $\log_2$ scale) (*upper panel*). An aliquot of the protease-treated sample analyzed by shotgun proteomics yielded quantified peptides shown in the *lower panel*, and the border of a one-sided 99% quantile of these peptide ratios was used as ratio cutoff value (this cutoff value was 0.45 and is indicated with a *vertical line*). The *inset* shows an iceLogo of all 141 (76 unique) neo-N-terminal peptides generated by caspase-3, which revealed the canonical caspase-3 recognition site DEVD. Further note that many neo-N-terminal peptides with ratios distributed close to 3 are the result of signal peptide removal (as the case for mitochondrial proteins) rather than being generated by contaminating protease activity during caspase-3 incubation. *B*, freeze-thaw lysates of partially (68% [$^{13}C_6$]Arg) SILAC-labeled primary mouse dendritic cells were incubated with cathepsin D or E. Upon mixing with a control lysate, cathepsin-generated neo-N-terminal peptides have light/heavy ratios of about 0.5 (-1 in $\log_2$ scale), whereas all other N-terminal peptides hold ratio values around 2 (1 in $\log_2$ scale) (*upper* and *lower panels*). A test analysis to determine a ratio cutoff value was performed similarly to the caspase-3 experiment (*middle panel*; the cutoff value was now -0.31).

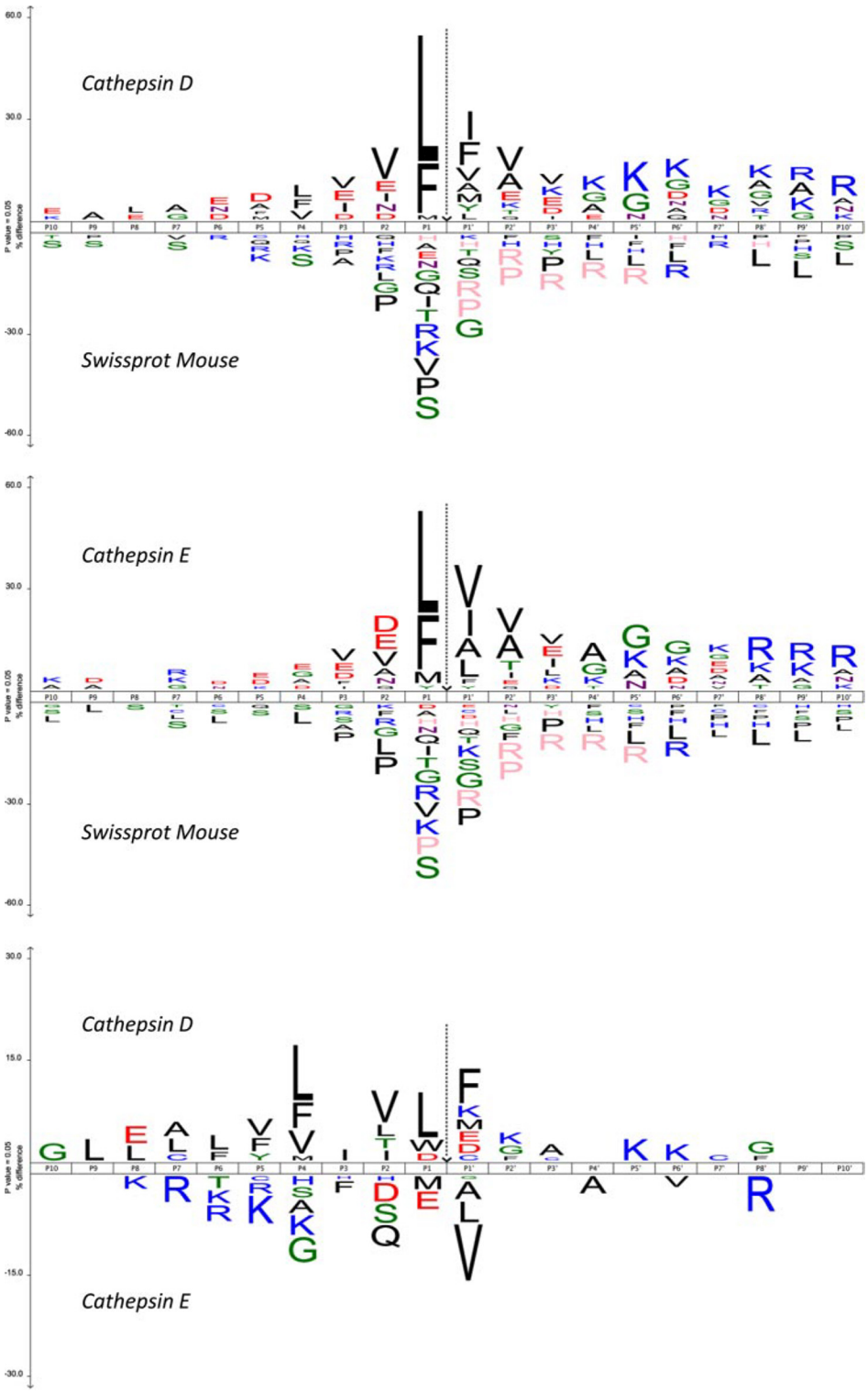


FIG. 3. **Sequence patterns at cathepsin D and E cleavage sites.** In the *upper* and *middle panels*, the amino acid frequency at every subsite is compared with sampled frequencies in the mouse proteins stored in the Swiss-Prot database (negative control). Only residues that are statistically over-represented (*upper part* of the iceLogo) or under-represented (*lower part* of the iceLogo) at a 95% confidence level are depicted. Residues that were never observed at specific positions are shown in a *pink* color. In the *lower panel*, the cleavage site environments of the substrates of cathepsin D are compared with those of cathepsin E. Highly similar substrate specificity profiles were obtained for both cathepsins (*upper* and *middle panels*), although subtle specificity differences between both proteases appeared, especially at the unprimed site region (*lower panel*).

by cathepsin E was detected, supporting our observed differences in the specificity profiles of both cathepsins (supplemental Fig. 3) and pointing to the fact that cathepsins D and E, although strongly related, might hold different specificities and cleave different substrates.

DISCUSSION

Here we have presented a simple strategy, combining differential stable isotope labeling (SILAC) with differential sample mixing to avoid the appearance of singleton peptides. In this study, we provide proteome-wide analyses of high confidence protease cleavage sites, classification of which is performed completely automatically. The setup is arranged such that peptides reporting protease cleavage sites (neo-N-terminal peptides) are measured with L/H isotope ratios distributed distinctly from peptides that are common to both the protease-treated and control sample. As no manual interference is necessary, overall analysis time can be shortened considerably, opening opportunities for multiple large scale automated screenings of protease activities. It is further of note that the procedure appears generally suitable for proteome analyses where singleton detection problems appear as is also the case for interactomics or chemical proteomics studies. In this regard, it is worth mentioning that peptide ratio differences have been introduced before to study protein-protein interactions by quantitative proteomics (20). However, in that particular method, peptides from non-informative contaminants carried L/H ratios close to 1, optimal for quantification, whereas peptides from true binders were present as highly regulated peptides that are intrinsically difficult to quantify (supplemental Fig. 1). Clearly, our approach should also lead to a more straightforward detection of true binders in such experiments.

Recently, auf dem Keller and co-workers (4, 21) reported a statistics-based platform for quantitative N-terminome analysis using the TAILS (terminal amine isotopic labeling of substrates) technology. Although no advanced sample mixing was used, the authors also introduce a ratio cutoff value to distinguish between N-terminal and neo-N-terminal peptides following labeling by an iTRAQ (isobaric tag for relative and absolute quantification)-like reagent and quantification in MS2 mode (21). In their approach, neo-N-terminal peptides are present as highly regulated peptides. However, the ratio distributions of N-terminal and neo-N-terminal peptides are much broader than those presented here, and as a result, more advanced statistical analysis is required to distinguish between both types of peptides. In contrast, the higher accuracy of MS1 quantification, which we applied here, resulted in sharp and amply spaced ratio distributions, as can be judged from the caspase-3 data (Fig. 2*A*). In fact, the higher accuracy of MS1 quantification can be attributed to the fact that every peptide is quantified several times through multiple mass spectrometer scans, whereas in MS2 mode, quantification is often based on a single tandem mass spectrometry quantification event (22).

Acknowledgments—We thank Veronique Jonckheere for providing data for the evaluation of the MASCOT Distiller quantification software and Prof. Lennart Martens for discussions on this topic.

* This work was supported in part by research grants from the Fund for Scientific Research-Flanders (Belgium) (Project G.0042.07), the Concerted Research Actions (Project BOF07/GOA/012) from the Ghent University, and the Inter University Attraction Poles (Grant IUAP06).

[S] This article contains supplemental experimental procedures, Tables 1–3, Figs. 1–3, and spectra.

¶ A research assistant of the Research Foundation-Flanders (FWO-Vlaanderen).

‖ Supported by a Ph.D. grant from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

‡‡ Supported by Biotechnology and Biological Sciences Research Council Grant BB/D005469/1 under the Selective Chemical Intervention in Biological Systems initiative.

§§ To whom correspondence should be addressed: Dept. of Medical Protein Research and Biochemistry, VIB and Faculty of Medicine and Health Sciences, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium. Tel.: 32-92649274; Fax: 32-92649496; E-mail: kris.gevaert@vib-ugent.be.

REFERENCES

1. Van Damme, P., Martens, L., Van Damme, J., Hugelier, K., Staes, A., Vandekerckhove, J., and Gevaert, K. (2005) Caspase-specific and non-specific in vivo protein processing during Fas-induced apoptosis. *Nat. Methods* **2,** 771–777
2. Schilling, O., and Overall, C. M. (2008) Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nat. Biotechnol.* **26,** 685–694
3. Mahrus, S., Trinidad, J. C., Barkan, D. T., Sali, A., Burlingame, A. L., and Wells, J. A. (2008) Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. *Cell* **134,** 866–876
4. Kleifeld, O., Doucet, A., auf dem Keller, U., Prudova, A., Schilling, O., Kainthan, R. K., Starr, A. E., Foster, L. J., Kizhakkedathu, J. N., and Overall, C. M. (2010) Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nat. Biotechnol.* **28,** 281–288
5. Van Damme, P., Maurer-Stroh, S., Plasman, K., Van Durme, J., Colaert, N., Timmerman, E., De Bock, P. J., Goethals, M., Rousseau, F., Schymkowitz, J., Vandekerckhove, J., and Gevaert, K. (2009) Analysis of protein processing by N-terminal proteomics reveals novel species-specific substrate determinants of granzyme B orthologs. *Mol. Cell. Proteomics* **8,** 258–272
6. Gevaert, K., Goethals, M., Martens, L., Van Damme, J., Staes, A., Thomas, G. R., and Vandekerckhove, J. (2003) Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat. Biotechnol.* **21,** 566–569
7. Demon, D., Van Damme, P., Vanden Berghe, T., Deceuninck, A., Van Durme, J., Verspurten, J., Helsens, K., Impens, F., Wejda, M., Schymkowitz, J., Rousseau, F., Madder, A., Vandekerckhove, J., Declercq, W., Gevaert, K., and Vandenabeele, P. (2009) Proteome-wide substrate analysis indicates substrate exclusion as a mechanism to generate caspase-7 versus caspase-3 specificity. *Mol. Cell. Proteomics* **8,** 2700–2714
8. Vande Walle, L., Van Damme, P., Lamkanfi, M., Saelens, X., Vandekerckhove, J., Gevaert, K., and Vandenabeele, P. (2007) Proteome-wide identification of HtrA2/Omi substrates. *J. Proteome Res.* **6,** 1006–1015
9. Brunner, A., Keidel, E. M., Dosch, D., Kellermann, J., and Lottspeich, F. (2010) ICPLQuant—a software for non-isobaric isotopic labeling proteomics. *Proteomics* **10,** 315–326
10. Colaert, N., Helsens, K., Impens, F., Vandekerckhove, J., and Gevaert, K. (2010) Rover: a tool to visualize and validate quantitative proteomics data from different sources. *Proteomics* **10,** 1226–1229
11. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids

in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics.* **1,** 376–386
12. Stennicke, H. R., Renatus, M., Meldal, M., and Salvesen, G. S. (2000) Internally quenched fluorescent peptide substrates disclose the subsite preferences of human caspases 1, 3, 6, 7 and 8. *Biochem. J.* **350,** 563–568
13. Demon, D., Van Damme, P., Berghe, T. V., Vandekerckhove, J., Declercq, W., Gevaert, K., and Vandenabeele, P. (2009) Caspase substrates: easily caught in deep waters? *Trends Biotechnol.* **27,** 680–688
14. Staes, A., Van Damme, P., Helsens, K., Demol, H., Vandekerckhove, J., and Gevaert, K. (2008) Improved recovery of proteome-informative, protein N-terminal peptides by combined fractional diagonal chromatography (COFRADIC). *Proteomics* **8,** 1362–1370
15. Helsens, K., Timmerman, E., Vandekerckhove, J., Gevaert, K., and Martens, L. (2008) Peptizer, a tool for assessing false positive peptide identifications and manually validating selected results. *Mol. Cell. Proteomics* **7,** 2364–2372
16. Yasuda, Y., Kageyama, T., Akamine, A., Shibata, M., Kominami, E., Uchiyama, Y., and Yamamoto, K. (1999) Characterization of new fluorogenic substrates for the rapid and sensitive assay of cathepsin E and cathepsin D. *J. Biochem.* **125,** 1137–1143
17. Zaidi, N., and Kalbacher, H. (2008) Cathepsin E: a mini review. *Biochem. Biophys. Res. Commun.* **367,** 517–522
18. Chain, B. M., Free, P., Medd, P., Swetman, C., Tabor, A. B., and Terrazzini, N. (2005) The expression and function of cathepsin E in dendritic cells. *J. Immunol.* **174,** 1791–1800
19. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J., and Gevaert, K. (2009) Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods* **6,** 786–787
20. Selbach, M., and Mann, M. (2006) Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK). *Nat. Methods* **3,** 981–983
21. auf dem Keller, U., Prudova, A., Gioia, M., Butler, G. S., and Overall, C. M. (2010) A statistics based platform for quantitative N-terminome analysis and identification of protease cleavage products. *Mol. Cell. Proteomics* **9,** 912–927
22. Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R., and Mann, M. (2010) Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* **7,** 383–385