TopFIND, a knowledgebase linking protein termini with function

To the Editor: We present 'termini-oriented protein function inferred database' (TopFIND; http://clipserve.clip.ubc.ca/topfind/), a knowledgebase providing integrated information on translated protein N and C termini, their formation by proteolytic processing and their amino acid modifications. TopFIND is open to data contribution from users.

Among the fundamental characteristics of a protein are its N and C termini. Protein-termini isoforms are genetically encoded and also often generated during translation. After translation, protein termini are highly dynamic, being frequently trimmed by exopeptidases. Neo-termini can also be generated by endopeptidases after precise and limited proteolysis, termed processing. Necessary for the maturation of many proteins, processing can also occur after synthesis and maturation, often resulting in dramatic functional consequences. Aberrant proteolysis is also pathognomonic. Hence, proteolytic generation of pleiotropic stable forms of proteins, the universal susceptibility of proteins to proteolysis and its irreversibility distinguish proteolysis from many other highly studied post-translational modifications.

With recent advances of N terminomics[1,2] and C terminomics[3,4] in the emerging field of degradomics (**Supplementary Discussion**) and the start of the Human Proteome Project, *in vivo* information about the actual protein N and C termini, their proteolytic generation and post-translational modifications is rapidly accumulating. Nonetheless, this information has remained largely inaccessible.

TopFIND integrates information from the UniProt knowledgebase (UniProtKB), MEROPS peptidase database[5] and experimental terminomics studies (**Supplementary Methods**) of four organisms (*Homo sapiens*, *Mus musculus*, *Escherichia coli* and *Saccharomyces cerevisiae*) resulting in 53,849 protein entries, 69,036 N termini and 61,314 C termini (**Supplementary Table 1** and **Supplementary Fig. 1**). The terminomics studies included in TopFIND to date provide experimental evidence for just 6,226 N termini and 1,188 C termini, reflecting the unmet need for their continued identification and

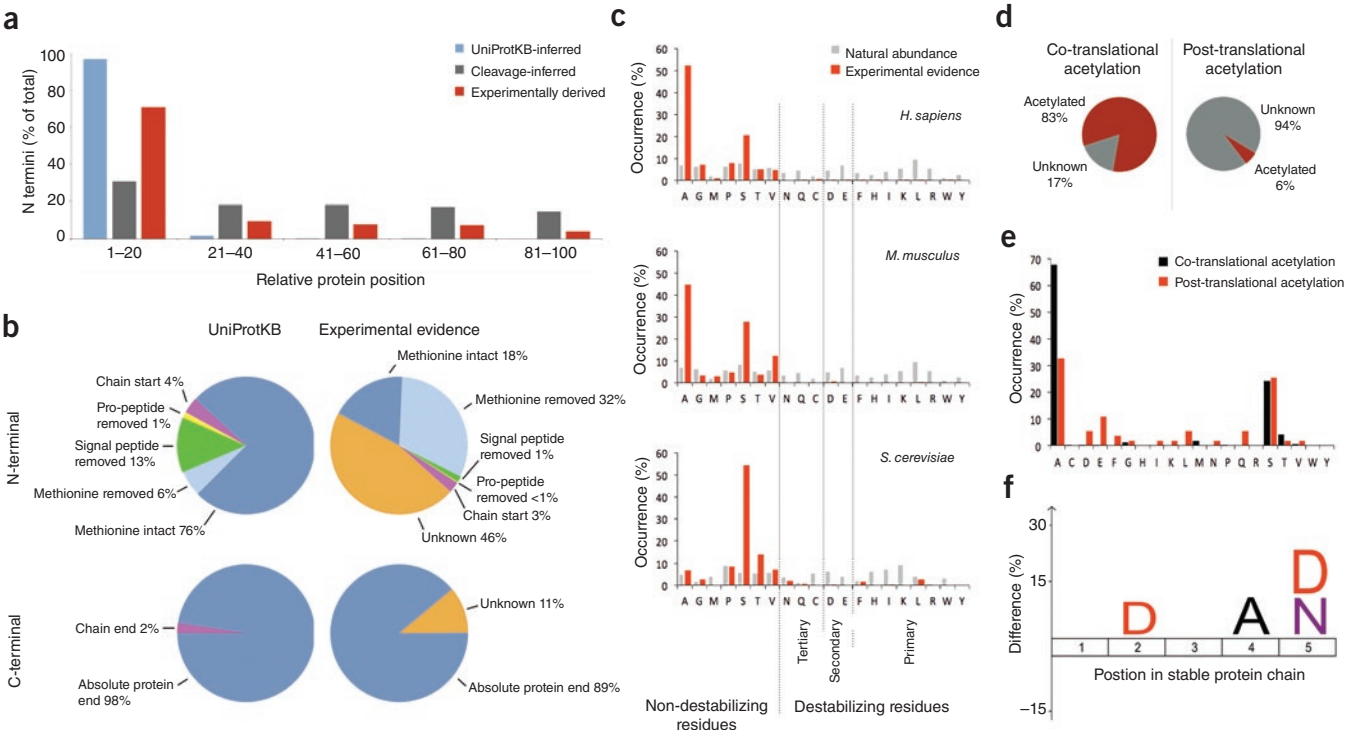


Figure 1 | TopFIND knowledgebase content, terminal acetylation and protein stability. (**a**) Distribution of N termini binned by their relative position in the sequences of all proteins. (**b**) UniProtKB-annotated and experimentally derived N and C termini in TopFIND. (**c**) N-terminal amino acid occurrence after methionine removal in human, mouse and yeast experimental data (2,281, 581 and 271 N-terminal identifications analyzed, respectively). Amino acids are categorized into primary, secondary and tertiary destabilizing residues according to the N-end rule (delineated by vertical lines). (**d**) Experimental evidence in TopFIND for the extent of co-translational (551 N termini) and post-translational (855 N termini) acetylation after protease cleavage in the mouse proteome. (**e**) Acetylated N-terminal amino acid occurrence after removal of initiating methionine (co-translational acetylation) or on neo–N termini (post-translational acetylation). (**f**) Sequence preference for post-translational N-terminal acetylation on stable protein chains in mouse (53 N termini analyzed). All internal (start ≥ third position in polypeptide) N-terminal sequences serve as the reference set.

annotation. By including experimental data, TopFIND reveals previously missing information regarding protein start and end sites, and the origin of termini. Indeed, >15% of all N termini and >10% of all C termini reported by TopFIND are not present in other knowledgebases. Cross-species, proteome-wide analysis of protein termini and terminal modifications by TopFIND revealed that N and C termini inferred based on genomic data did not agree with experimentally determined termini in >50% of cases. If we disregard the N termini originating from methionyl aminopeptidase-1 and -2 processing in the first two residues of the full-length sequence and C termini occurring from cleavages in the first 10 amino acids, then TopFIND provides evidence for 5,681 neo–N termini and 5,534 neo–C termini from 9,608 cleavage sites reported in the MEROPS database (several proteases might cleave at the same site) (**Supplementary Table 2**).

We derived the relative positional distribution of N termini along all protein sequences across species using TopFIND. In contrast to UniProtKB, TopFIND annotates 29% of the stable chains as starting distal to the expected protein maturation sites such as initiator methionine, signal peptide and pro-peptide removal points (**Fig. 1a**). The experimental evidence showed that 46% and 11% of all stable N and C termini, respectively, did not match any of these classic processing categories (**Fig. 1b** and **Supplementary Discussion**). The N-end rule links the N-terminal residue of an intracellular protein to its half-life *in vivo*. In a cross-species comparison of stable protein chains after methionine removal, we found that destabilizing N-terminal residues were virtually absent in mouse, human and to a lesser extent yeast (**Fig. 1c**), indicating that any such proteins are indeed degraded after translation and co-translational protein maturation, providing proteomic validation of the N-end rule.

We also applied TopFIND to study the extent and type of N-terminal amino acid modifications across species. There was direct information about modification, or lack thereof, for 44% of N-terminal peptides derived from UniProtKB (**Supplementary Fig. 2**), and when this information was available, it focused virtually exclusively on the presence or absence of N-terminal acetylation. As co-translational acetylation after initiating methionine removal is well-defined, we used this as a benchmark to assess the quality of TopFIND data. We found full agreement with recent reports on the relative amino-acid distribution for acetylation and the sequence preference for acetylation upon methionine removal (**Supplementary Fig. 3** and **Supplementary Discussion**).

Post-translational N-terminal acetylation is a recently predicted[6] and experimentally confirmed modification[7] that is expected to play an important role in determining protein stability *in vivo*. TopFIND analysis of all experimentally identified mouse N termini showed that 6% of the stable cleavage products had post-translational acetylation of the neo–N terminus (**Fig. 1d**). Comparison of the N-terminal amino acids susceptible to co-translational versus post-translational acetylation in mouse showed clear differences (**Fig. 1e**), but preferentially occurred at a terminal amino acid that is followed by a negatively charged amino acid at position 2 (**Fig. 1f** and **Supplementary Fig. 3b**).

To highlight the need to investigate other forms of terminal modifications we investigated pyroglutamate (pGlu) formation[7], a process that recently has been recognized to be enzymatically driven[8], and derived the sequence-specificity for its formation (**Supplementary Fig. 4a** and **Supplementary Discussion**).

An important use of TopFIND is to provide new information so as to formulate hypotheses springing from the mature and neo-termini of proteins. For example, we are interested in the widespread somatic mutations found in the important cancer suppressor, p53. Positional cross-correlation analysis suggests the loss of caspase 3 cleavage at Asp186-Gly187 and consequent reduced apoptosis upon somatic mutation in the sequence encoding Asp186, thereby diminishing the cancer protective activity of p53 (**Supplementary Fig. 5** and **Supplementary Discussion**). Additionally, a TopFIND-assisted analysis of Bap31 revealed a previously unknown stable C-terminal fragment having a potential function in transcriptional initiation of apoptosis (**Supplementary Fig. 6** and **Supplementary Discussion**).

We believe that TopFIND will be a useful information repository for protein original and neo-N- and C-termini, terminal modifications and proteolytic processes. TopFIND also provides a powerful new means of hypothesis generation that can inspire new projects, leading to insights into protein processing, function and cell physiology.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank P.F. Huesgen for fruitful discussions. P.F.L. is supported through a Feodor Lynen Research Fellowship of the Alexander von Humboldt Foundation and the German Federal Ministry of Education and Research, and a Michael Smith Foundation for Health Research/Breast Cancer Society of Canada Research Trainee Award. C.M.O. is supported by a Canada Research Chair in Metalloproteinase Proteomics and Systems Biology. This work was supported by a grant from the Canadian Institutes of Health Research and from a program project grant in Breast Cancer Metastases from the Canadian Breast Cancer Research Alliance with funds from the Canadian Breast Cancer Foundation and the Cancer Research Society and with an Infrastructure Grant from the Michael Smith Foundation for Health Research.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Philipp F Lange[1–3] & Christopher M Overall[1–3]

[1]Centre for Blood Research, University of British Columbia, Vancouver, British Columbia, Canada. [2]Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, British Columbia, Canada. [3]Department of Oral Biological and Medical Sciences, University of British Columbia, Vancouver, British Columbia, Canada.
e-mail: chris.overall@ubc.ca

PUBLISHED ONLINE 7 AUGUST 2011; CORRECTED ONLINE 12 AUGUST 2011 (DETAILS ONLINE); DOI:10.1038/NMETH.1669

1. Doucet, A. *et al. Mol. Cell. Proteomics* **7**, 1925–1951 (2008).
2. Impens, F. *et al. Proteomics* **10**, 1284–1296 (2010).
3. Schilling, O. *et al. Nat. Methods* **7**, 508–511 (2010).
4. Van Damme, P. *et al. Nat. Methods* **7**, 512–515 (2010).
5. Rawlings, N.D., Barrett, A.J. & Bateman, A. *Nucleic Acids Res.* **38**, D227–D233 (2010).
6. Hwang, C.-S., Shemorry, A. & Varshavsky, A. *Science* **327**, 973–977 (2010).
7. Kleifeld, O. *et al. Nat. Biotechnol.* **28**, 281–288 (2010).
8. Jawhar, S. *et al. J. Biol. Chem.* **286**, 4454–4460 (2010).

Nature Methods

TopFIND, a knowledgebase linking protein termini with function

Philipp F Lange & Christopher M Overall

Supplementary Figure 1	Protein information elements as visualized by TopFIND
Supplementary Figure 2	Cross-species comparison of terminal modifications
Supplementary Figure 3	Co-translational acetylation
Supplementary Figure 4	Terminal pyroglutamate formation and acetylation
Supplementary Figure 5	p53 cleavage by caspase 3 and site specificity as provided by TopFIND
Supplementary Figure 6	Information on Bap31 provided by TopFIND
Supplementary Table 1	Database content
Supplementary Discussion	
Supplementary Methods	

Note: Supplementary Table 2 is available on the Nature Methods website.

SUPPLEMENTARY DISCUSSION

TopFIND has been developed to harness the functional information that can be gained from protein termini and protein processing and linking this with existing knowledge to make it accessible to the scientific community. Four components are central to this. (i) The underlying database schema which reflects functional relationships between the molecules and processes of interest (**Supplementary Methods**). (ii) Comprehensive coverage of existing and new information, which is achieved in TopFIND by periodic mining of relevant external databases, literature curating (**Supplementary Table 1** and **Supplementary Methods**) and direct contribution by the scientific community. (iii) Data analysis to derive statistical information (**Fig. 1** and **Supplementary Figs. 2, 3** and **4**), to identify new elements (**Supplementary Table 2**), and new links between elements (**Supplementary Fig. 5** and **6**). (iv) A powerful interface to retrieve information (**Supplementary Fig. 1**), put this information into perspective and generate new knowledge and hypotheses (**Supplementary Figs. 5** and **6**).

Data generation

Recently new methodologies for the high throughput gel-free mass spectrometry-based identification and characterization of protein N termini have been developed by a number of laboratories[1-3] resulting in an explosion of N terminomics data. Overall, C-termini are intrinsically more difficult to determine due to the lower reactivity of carboxyl groups. But still, huge advances have been made with recent chromatographic separation approaches[4,5] and a new quantitative C-terminal peptide enrichment technique termed carboxyl-terminal assisted isotope labeling of substrates (C-TAILS)[6] by our laboratory. Together, these render both N and C termini of proteins in complex *in vivo* samples accessible.

These novel techniques have led to a number of N-terminome analyses including the blood N terminome[7], the mitochondrial N-terminome[8] and the N terminome during apoptosis[2,9]. The number of tissues, cell types and subcellular compartments investigated as well as the depth of coverage will increase dramatically with the recently announced Human Proteome Project (HPP)[10]. The aim of this international project is to discover at least one protein for each human gene and provide tools for their thorough characterization. This extends to the identification and cataloging of all stable fragments (chains), termini and their modifications, such as acetylation, cyclization and citrullination in the healthy state.

Equally important as the identification of protein termini is the characterization of the underlying proteolytic processes that form the stable protein chains and their defining start and end termini. Since proteolytic processing often alters the functional state of a protein the precise location of the exact N and C termini of a protein is critical to know. Until recently, characterization of proteases and their substrates was mostly achieved by thorough biochemical characterization *in vitro*. Now the ability to rapidly enrich and identify N and C terminal peptides from complex proteomes has led to a dramatic increase in identification of protease substrates as well as their cleavage sites *in vivo*[11].

	UniProtKB inferred	cleavage inferred	experimental evidence	total	new
proteins	53,849	-	-	53,849	-
cleavages	9,608	-	-	9,608	-
N termini	58,655	5,681	6,226	69,507	10,548
C termini	54,381	5,534	1,188	61,314	6,482

Supplementary Table 1 | Database content Number of proteins, termini and cleavages accessible through TopFIND. Unique termini data derived from cleavages or experiments submitted by users not reported by any other knowledgebase (new) is provided in addition to UniProtKB inferred data.

Data contribution

TopFIND is designed for easy contribution of new datasets obtained by the described current and future high throughput approaches. The system is method independent, allowing data obtained by different methodologies, each with its advantages and shortcomings, to be integrated, combined and compared alongside. To facilitate the judgment process it encourages information on data quality and reliability to be stored alongside the main data. For example, a peptide assignment confidence in mass spectrometric approaches is included. Rather than imposing a slow approval and curating process, laboratories can contribute their data in a rapid and straightforward manner along with detailed information on how this data has been obtained and where it has been published. Upon data upload, the TopFIND knowledgebase curators are automatically notified who then perform a check for file integrity and completeness of the provided experimental metadata followed by the actual data integration. Users can decide at the time of data mining (see below) if the data satisfies their quality criteria and should be included in the representation or not. This guarantees access to new data in a timely manner without compromising data quality.

To contribute to the identification of cleavage sites, protease inhibitions, or N or C termini in TopFIND, one simply follows the step-by-step instructions laid out on the 'contribution' page. Briefly one needs to: (i) Create an account and log in; (ii) enter the experimental metadata and related publications. Save this as new evidence. Raw data can be referenced through a link to the appropriate repository file like PRIDE[12]. (iii) From here new cleavage sites, inhibitors, N or C termini can be added by the simultaneous import of many entries of one kind by uploading a .csv file that follows the simple format rules outlined on the contribution page. After successfully importing the data a confirmation message is sent to the contributors e-mail address.

Subsequently the data is cross checked by the database curator and made available to the public.

Data mining

The main entry point for information retrieval in TopFIND is through a protein entry. A specific protein can (i) either be selected from the list of proteins or (ii) searched for by its UniProt acc, common names and abbreviations or MEROPS code. If more than one matching protein is found, a search result page will be displayed. The protein or search results can be further narrowed down by species, function and chromosome location. Search and filter functionality also encompasses amino acid modifications of protein termini. Genomic location data has been incorporated so as to fully integrate TopFIND into the chromosome-centered approach of the Human Proteome Project (HPP). All proteins can therefore be searched for and filtered by the genomic location of their encoding genes. Additionally it is possible to search by protease classification (MEROPS family or clan), which retrieves the respective members as well as their combined substrates.

Result output

TopFIND output is organized into seven sections some of which are hidden when no information is available. Positional information is accompanied by a small graphical representation throughout the page. The full-length sequence is shown as a blue line, a region (such

as a protein domain) as a grey bar and a cleavage site or terminus visualized by a vertical red line.

i) Protein and protein isoform information: This section in TopFIND provides basic information retrieved from UniProtKB and MEROPS. Protein names and species and, if applicable, protease classification and isoforms are listed. UniProtKB curated annotation is presented alongside the amino acid sequence. For further background information, links to the appropriate entries at UniProtKB and MEROPS are provided (**Supplementary Fig. 1a**).

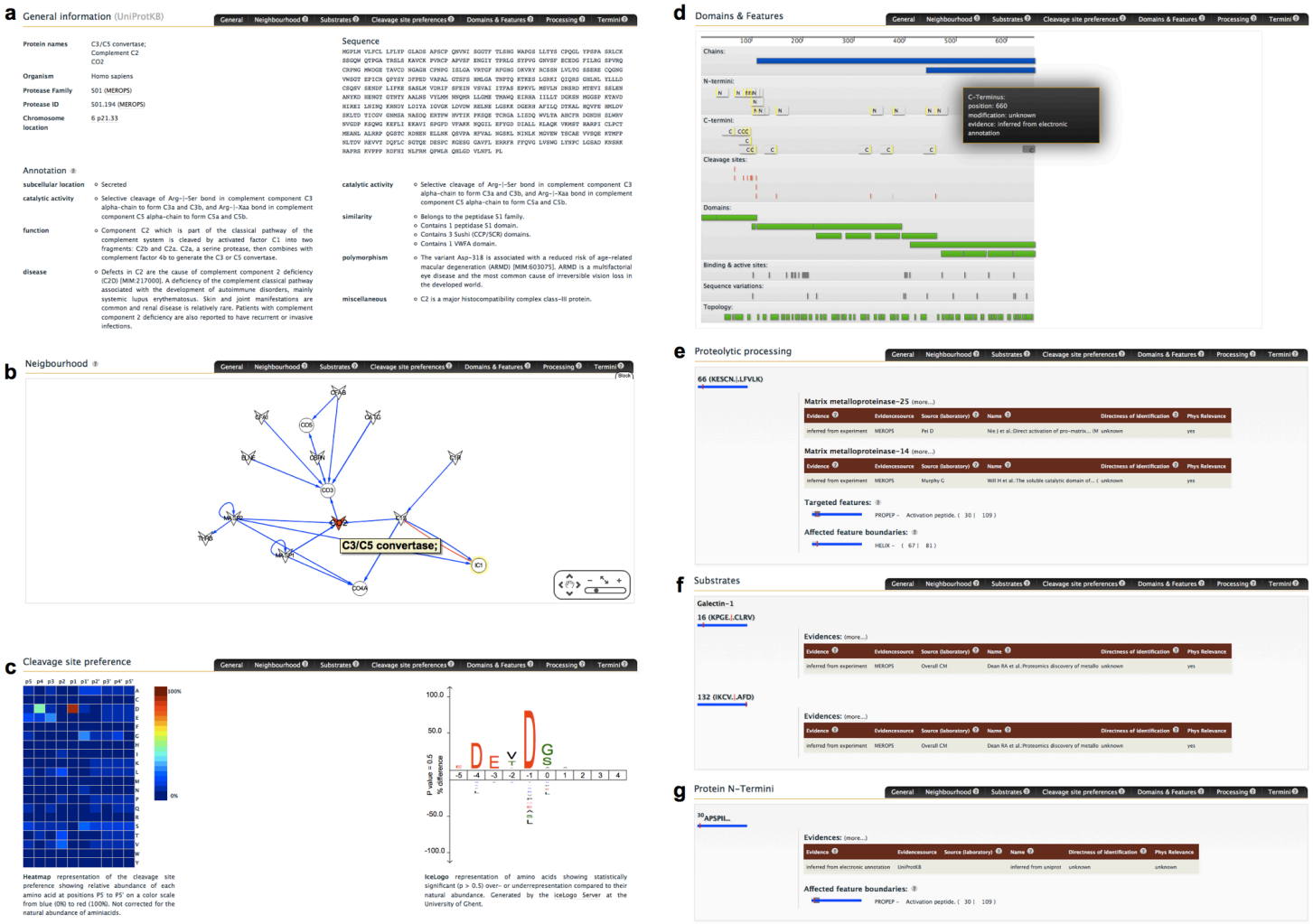
(ii) Network neighborhood: A network view of the interplay with other proteins is shown. The displayed protein is highlighted in red. Cleavages are visualized as blue arrows, inhibitory activity as red "T" bars and protein interactions as grey lines. Proteases are symbolized by a "V" shape and other proteins by a circle. Some proteins are combined for conciseness and clarity of the network. The user can zoom and pan as well as move around single or groups of proteins for enhanced clarity (**Supplementary Fig. 1b**).

(iii) Cleavage site specificity: If the protein is a protease then information on its preference for specific amino acids in the region of the cleavage site is given (S5-S5'). The information is visualized both as a heat map and iceLogo (see **Supplementary Methods** for differences and reasoning behind both) (**Supplementary Fig. 1c**).

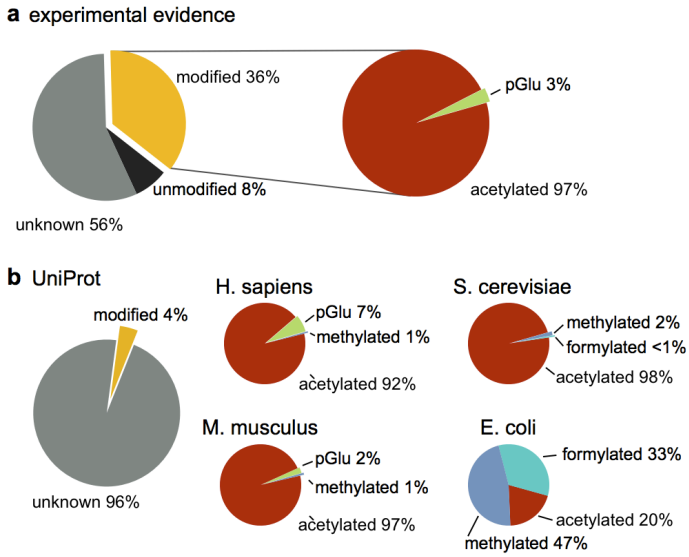
(iv) Domains and features: The linear organization of termini, cleavage sites and stable chains as well as features and domains along the primary protein sequence is graphically represented. Moving the mouse over an element will provide additional information such as the exact position, type and description (**Supplementary Fig. 1d**).

(v) Proteolytic processing: For every processing site within the protein, all proteases known to cleave at this site are listed alongside a brief table indicating the evidence that led to this observation. Detailed metadata is available through a link (**Supplementary Fig. 1e**).

If applicable, features (such as mutations) or feature boundaries (such as the end of a signal peptide) that overlap with the cleavage site are listed in TopFIND and depicted as grey bars on a blue protein backbone along with a vertical red line indicating the site of cleavage.



Supplementary Figure 1 | Protein information elements as visualized by TopFIND (**a**) General protein information displaying protein names and abbreviations, a link to the UniProtKB entry, the Merops ID (if applicable) linking to the respective entry on the Merops website, the chromosome location of the encoding gene, isoforms and the amino acid sequence. Protein annotation is summarized as derived from UniProtKB. (**b**) A fully interactive network view showing protein interactions through cleavages (blue arrows), protease inhibition (red T bars) and general UniProtKB derived protein-protein interactions (grey line) visualized using Cytoscape Web. (**c**) Cleavage site preferences visualized as heat map (left) and iceLogo (see Online Methods for details on differences between these). (**d**) Linear protein representation depicting the order of termini, cleavage sites, protein domains, sequence variations or topological features. (**e**) Proteolytic processing of the displayed protein is grouped by amino acid position and protease with short inline information on the supporting evidences with links to the details. Protein features that can be affected by the processing are indicated. (**f**) Proteolytic activity is grouped by substrate and position with brief reports on the supporting evidences displayed inline. (**g**) Protein Termini are displayed in the same way.



Supplementary Figure 2 | Cross-species comparison of terminal modifications (**a**) Relative extent of N terminal modification combined for all four species as derived from published experimental datasets imported into TopFIND (n_{total}=9,619). Only one study reported modifications other than acetylation[3] (n_{total}=3,464). (**b**) Relative extent of N terminal modifications combined for all four species as derived from UniProt (n_{total}=70,243). Modified termini are broken down by species and modification (*H. sapiens* n_{total}=1,421; *M. musculus* n_{total}=1,151; *S. cerevisae* n_{total}=192; *E. coli* n_{total}=30).

(vi) Substrates: For each substrate a link to the respective entry as well as a list of cleavage sites and corresponding evidence in short tabular format is given (**Supplementary Fig. 1f**).
(vii) Termini: N and C termini are listed according to their position. Evidence information is given in brief tabular format and a link to full details, publications and data repository is provided. Termini are placed into perspective by listing features and domains such as propeptides whose start or endpoint overlaps with the terminus (**Supplementary Fig. 1g**).

Result filtering

By default, all known data are incorporated into the results page. This can be customized through a filter accessible at the top right of the page. After filtering, only data that is backed with evidence matching the given filter settings is incorporated. This will affect the termini, cleavage sites and substrates listed. Also the network view will only consider matching connections at the root level and the cleavage site preference is only built on data matching the criteria.

For example when studying a biological process in the brain, retrieving information that is relevant *in vivo* and has been reported specifically for the brain might provide the best insight whereas for development of protease inhibitors for drug discovery information obtained by a specific method *in vitro* might be more relevant.

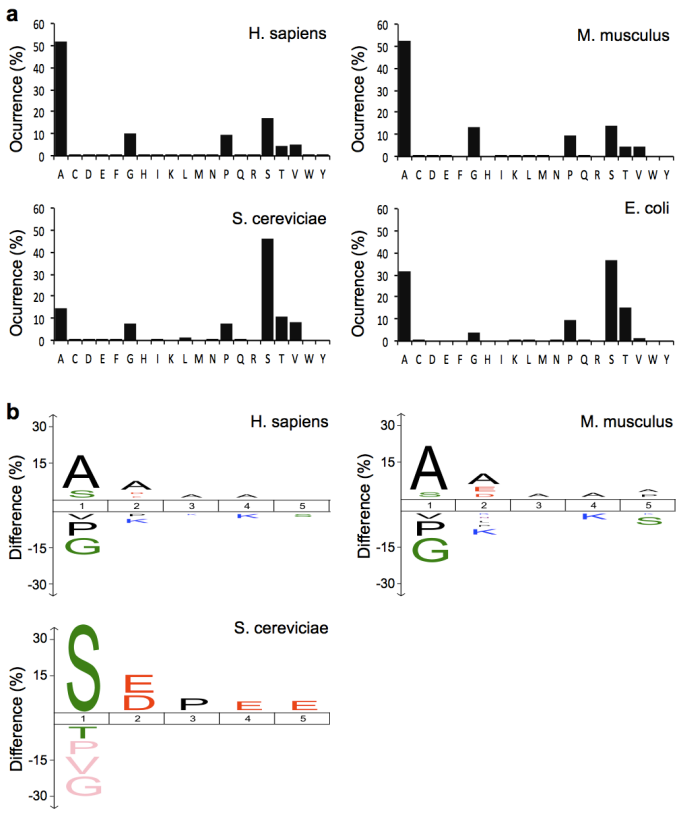
Data export

Displayed evidence metadata, termini and cleavage sites can be downloaded in a tab delimited .xls format for external use following the provided links. A concise RESTful query scheme provides access manual and computational access to customized subsets of evidence metadata, termini and cleavages. In addition the full underlying database can be downloaded as compressed .sql dump to be used for efficient external data mining.

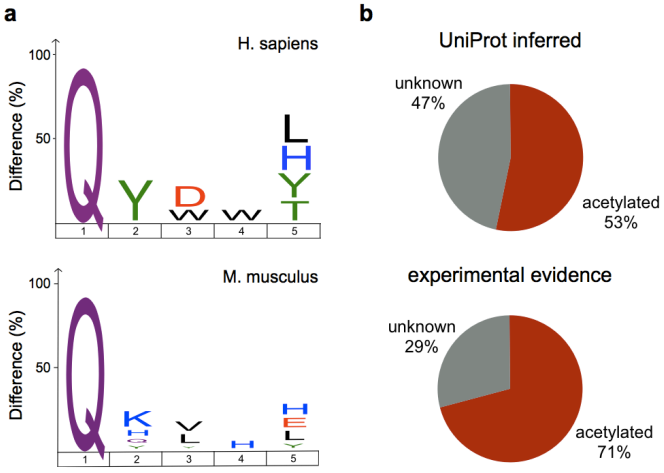
TopFIND based proteome-wide termini analyses

Quality control and co-translational acetylation. Co-translational acetylation after proteolytic removal of the initiation methionine is intensively studied and well understood[13,14]. Hence we evaluated the quality and comparability of TopFIND data by assessing the relative amino acid distribution (**Supplementary Fig. 3a**) and the sequence preference for acetylation of the N terminal amino acid after methionine removal (**Supplementary Fig. 3b**) and compared this with current literature. We found that the amino acid preference for methionine removal by the methionyl aminopeptidases to be alanine at position 2 for all species except S *cerevisiae*, followed by serine, proline, glycine and then threonine and valine for man, with some species variability. Both the sequence preference for acetylation and methionine removal are in full agreement with recent reports[13,14]. These data served as reference and so was not integrated into the TopFIND database prior to this analysis. All these data are now accessible through TopFIND.

Cross-species comparison of terminal modifications. Recent reports indicate that 70-80% of all proteins in human and about 50% of all yeast proteins are co-translationally acetylated at their N terminus, often after methionine removal[13]. However, UniProtKB so far only reports about 50% of the proteins to be terminally acetylated after initiating methionine removal. Notably the analysis of N-termini derived from terminomics studies by TopFIND identifies about 70% of all methionine-removed proteins as being fully or partially acetylated, which matches the reported average between all four species (**Supplementary Fig. 4b**). As briefly mentioned in the paper,



Supplementary Figure 3 | Co-translational acetylation (**a**) Species comparison of the relative occurrence of amino acids at the N terminus after initiator methionine removal analyzed by TopFIND. (**b**) Over and under represented amino acids in acetylated N termini after methionine removal. Sequences of all N termini starting at position 2 of the translated amino acid sequence and being identified as acetylated by UniProtKB or experimental datasets are plotted against all N termini starting at position 2 by iceLogo. (*H. sapiens* n=1,533, *M. musculus* n=802, *S. cerevisiae* n=103, P value = 0.05).



Supplementary Figure 4 | Terminal pyroglutamate formation and acetylation (**a**) IceLogo representation of the amino acid composition of N-termini carrying a pGlu modification. (*H. sapiens* n=89, *M. musculus* n=56). All N terminal sequences served as reference set. (**b**) Extent of N-terminal acetylation after initiator methionine removal as reported by UniProt (n_{total}=3,432) or imported datasets (n_{total}=3,103).

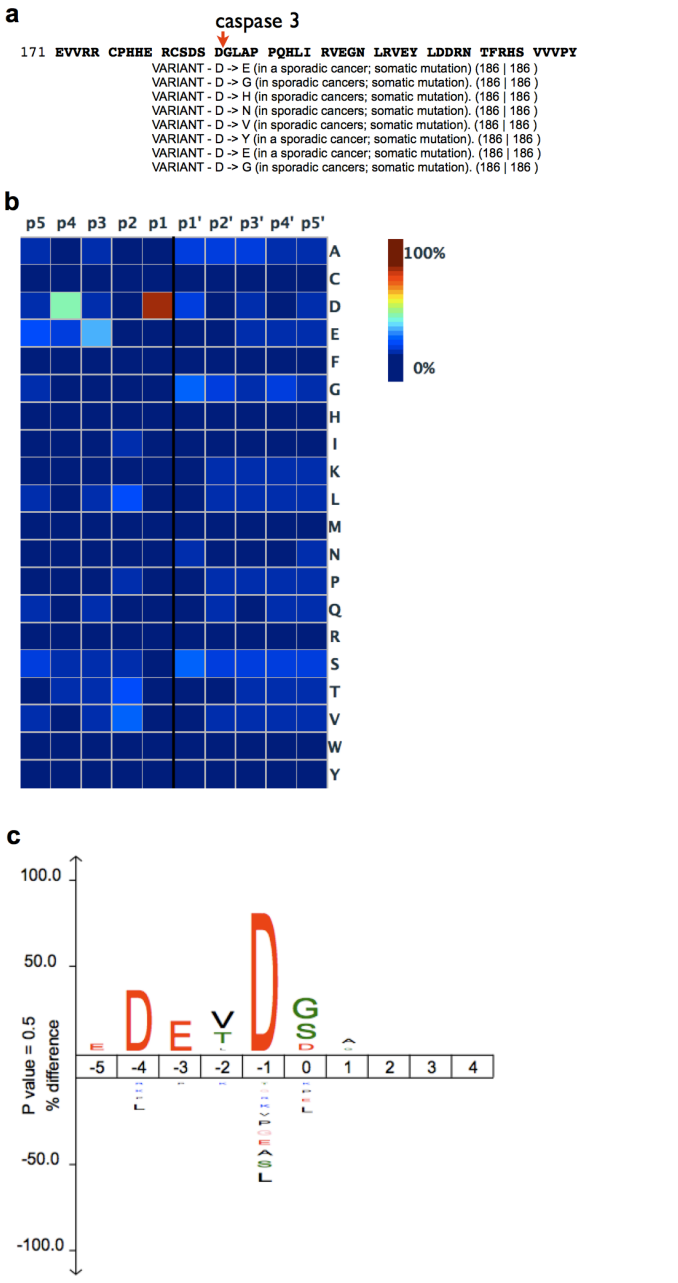
the great underrepresentation of N-terminal modifications other than acetylation in large-scale terminome approaches is worrisome and extremely prevalent as shown by TopFIND integration (**Supplementary Fig. 2a, b**). For example, while pyroglutamate formation is known to prevent processing by most aminopeptidases and may initiate Aβ aggregation through an increase in hydrophobicity[15] little is known about its proteome-wide extent and sequence specificity. This highlights the need to include terminal posttranslational modifications other than acetylation in N-terminome studies. TopFIND data integration allowed us to derive specificity profiles for pyroglutamate formation in mouse and man derived from 89 (human) and 56 (mouse) N-terminal sequences reported by UniProtKB and one experimental study[3] (**Supplementary Fig. 4a**). Notably, this profile correlates well with the specificity determined by enzyme kinetics[16].

Application

TopFIND is designed to provide the user with new information and hypotheses on a protein of interest by integrating all available knowledge on the protein, its termini and proteolytic processing. We present two examples to evaluate the potential of TopFIND for researchers not primarily interested in protein termini or proteolytic processing.

Cancer causing somatic mutations in p53 prevent cleavage and cell death. The first example examines the intensely studied cellular tumor antigen p53 (p53) that displays a complex set of activities in health and disease with transcriptional activation being its most prevalent function in cancer[17]. Entering the full name 'cellular tumor antigen p53' or its UniProt accession into TopFIND retrieves the p53 protein page. A quick screen of the page immediately brings attention to an unusual number of amino acid variants listed in the proteolytic processing section. A closer look shows that p53 is known to be processed between Lys24↓Leu25 and Lys305↓Arg306 by granzyme K and at Asp21↓Leu22 and Asp186↓Gly187 by caspase 3. The detailed view of the evidence easily retrieves the original publications, which show a role of granzyme K cleavage of p53 in cytolysis[18] and p53 cleavage by caspase 3 in transcription-independent apoptosis[19].

TopFIND reports not only the cleavage along with the underlying evidence, but uniquely, TopFIND also shows the features and domains that might be affected by cleavage or, in turn, might affect the cleavage. Herein lies one of the key novel and extremely useful features of TopFIND. The cleavage by caspase 3 at position Asp186↓Gly187 for example, is shown to be in a region where several somatic mutations that lead to cancer have been reported (**Supplementary Fig. 5a**). To investigate the possibility that these



Supplementary Figure 5 | p53 cleavage by caspase 3 and site specificity as provided by TopFIND (**a**) Excerpt of p53 sequence and caspase 3 cleavage sites indicated by red arrow. Annotated sequence variants are listed below. (**b**) Heat map representation showing relative abundance of each amino acid at positions P5 to P5' on a color scale from blue (0%) to red (100%). (**c**) IceLogo representation of amino acids showing statistically significant (p< 0.5) over or under representation compared to their natural abundance.

mutations could affect the ability of the protease to cleave such a mutant of p53 the corresponding protease protein page is displayed by following the link on the protease name. In the cleavage site preference section the amino acid preference of caspase 3 is displayed as a heat map (**Supplementary Fig. 5b**) and as an iceLogo[20] (**Supplementary Fig. 5c**). Both images clearly indicate that all listed mutant forms (D186E, D186G, D186H, D186N, D186V, D186Y) will be almost certainly *not* cleaved here by caspase 3 due to the absolute requirement for an aspartate at P1. The same can be observed for granzyme K, which shows a strong preference for arginine or lysine at P1, which makes it highly unlikely that mutants of p53 (K24N, K305E, K305M, K305N, K305R) will be cleaved.

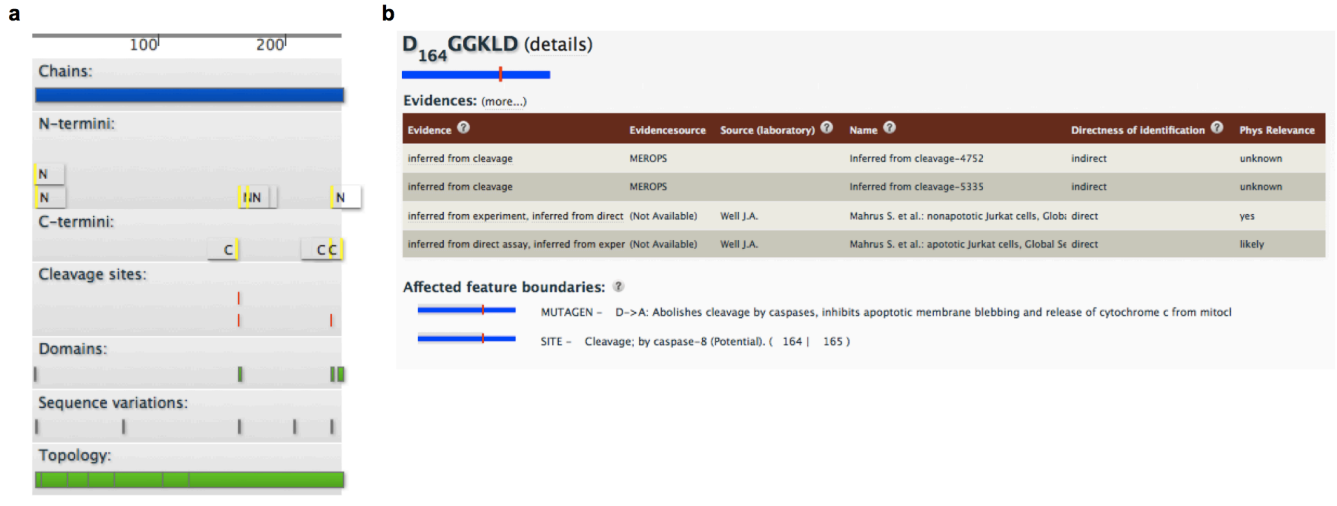
This correlation highlighted by TopFIND suggests that these somatic mutations might promote cancer by disabling cleavage of p53 by granzyme K or caspase 3, respectively thereby preventing the normal induction of transcription-independent cytolysis or apoptosis. However, mutation K305R replaces lysine with arginine, which is preferred by granzyme K and this ought not prevent cleavage. This renders the absence of granzyme K cleavage as the molecular mechanism underlying the effects of the somatic mutations at Lys24 and Lys305 either unlikely, or indicates a more complex mechanism involving several mutations, or several independent mechanisms for each mutation at these positions. In contrast, for mutations of Asp186 a lack of caspase 3 cleavage seems to be the prevalent mechanism. Indeed, Sayan and colleagues show a consistent reduction in transcription-independent apoptosis for non-cleavable naturally occurring p53 mutants compared to wild-type p53[19] and this information was easily retrieved through TopFIND.

Stable cytoplasmic C-terminal fragment of Bap31. Searching for '6C6-AG', which is part of an alternative name for B-cell receptor-associated protein 31 (Bap31), quickly pulls up the protein page in TopFIND. The domain overview (**Supplementary Fig. 6a**) shows one known chain from amino acids 2-246, but lists several more identified termini. Jumping to the termini section, four N termini and three C-termini are listed. The first indicates a protein start at Ser2 and the sequence shows a methionine as the preceding amino acid indicating initiator methionine removal. The listed evidence clearly shows that this is based on UniProtKB annotation (Source: UniProtKB, Evidence code: inferred from electronic annotation). Together with the absolute protein C terminus also listed as being referred from UniProtKB, these make up the main protein chain.

Several more N and C termini are listed with unknown physiological relevance as coming from a cleavage that has been reported in MEROPS. These can be considered putative as they would be the protein termini resulting from the stated cleavage, but without having been observed *in vivo* the stability of the cleavage products and existence of the termini in cells or tissues is unknown.

Two N termini (Gly164 and Asn171) however should catch the user's attention as they are indicated by TopFIND as having been directly observed by experimental data (**Supplementary Fig. 6b**). We can access the full information on the evidence that led to these observations by following the TopFIND "evidence" link. The detailed view for evidence on the N terminus Gly164 shows two pieces of evidence inferred from caspase 1 and caspase 8 cleavage and derived from MEROPS. The next two items of evidence indicate that the N-terminus Gly164 has been observed using a N terminal enrichment procedure from normal as well as apoptotic Jurkat cells. Both observations have been published[2] and the abstract is displayed in TopFIND along with a link to the full text on the journal website. In this example the publication does not give any further insight into the physiological role that a chain with this terminus might possess, but this observation might stimulate the user to hypothesis-driven research based on this and the following information that TopFIND also pulls in. The general annotation on the Bap31 protein page associates Bap31 protein function with anterograde ER Golgi transport and caspase mediated apoptosis. The schematic domain representation (**Supplementary Fig. 6a**) puts the protein having an N terminus at Gly164 into a cytoplasmic C terminal domain following three transmembrane domains.

A N terminus of Gly164 is also inferred from two cleavages, which provide further insight. Cleavage by caspase 1 and 8 has been shown to be important for apoptotic membrane blebbing and cytochrome c



Supplementary Figure 6 | Information on Bap31 provided by TopFIND (a) Domain and feature representation outlining the position of chains, termini and cleavage sites in relation to protein domain, sequence variations and topological features as transmembrane domains and cytosolic stretches. (**b**) Overview for N terminus Gly164. The initial five amino acids are stated, followed by a schematic showing the position of the N terminus (red vertical line) relative to the protein backbone (blue bar). A summary for individual pieces of evidence is given in tabular format. Protein features that coincide with the N terminus position are listed including a schematic, the feature type and short description.

release from mitochondria[21]. A recent study has shown that Bap31 establishes a platform for induction of apoptosis by bridging the mitochondria-ER interface[21]. While it has been shown that following caspase cleavage the N terminal fragment p20Bap31 (2-163) remains anchored to the ER membrane and is crucial for transmitting the apoptosis signal[22] the role of the cytoplasmic C terminus is not clear. Prior to caspase cleavage it has been shown to play a role in promoting vascular trafficking of a range of proteins including cellubrevin[23] and major histocompatibility complex class I molecules[24] as well as in protein quality control of cystic fibrosis transmembrane conductance regulator[25].

To date no role has been shown for the C terminal fragment (164-238) upon release following caspase cleavage. The identification of a protein with N terminus Gly164 from Jurkat cells in two independent conditions[26] makes it highly likely that this fragment is indeed stable *in vivo*. The presence of a leucine-zipper like domain[27] suggests a possible role in transcription regulation upon caspase dependent release. Hence, TopFIND has collected previously dispersed information to form a coherent picture from which several hypotheses easily can come to mind that might be tested if so desired.

CONCLUSIONS

We report the creation of a publicly available knowledgebase to combine information on protein termini, proteolytic processing and general protein information obtained from database annotations, biochemical experiments and recent high throughput terminomics and degradomics techniques. TopFIND is also designed with the specific needs of the Human Proteome Project in mind. The chromosome-centric approach is seamlessly implemented and TopFIND is ready to function as a reference repository for protein termini identified by the HPP.

Analysis of the data provided by TopFIND and comparison to existing resources shows a great underestimation and underrepresentation of internal processing and termini by current resources. Moreover, current databases lack extensive annotation of protein termini modifications that can profoundly alter protein function. TopFIND also shows that current large scale termini identification projects mostly lack coverage of modifications beyond terminal acetylation. This highlights the need to implement newer N terminomics approaches such as TAILS[28], which enriches all N termini, whether blocked or not, to fill in these features.

The presented test cases highlight that by putting seemingly unrelated information into context, TopFIND enables the user to not only quickly derive relevant information, but to also enable new hypotheses to be quickly formed that can in turn be tested by experiments. Hence, TopFIND is an extremely useful and easy to use knowledgebase for the community.

REFERENCES

1. Gevaert, K. et al. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat Biotechnol* **21**, 566-9(2003).

2. Mahrus, S. et al. Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. *Cell* **134**, 866-76(2008).

3. Kleifeld, O. et al. Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nat Biotechnol* **28**, 281-8(2010).

4. Dormeyer, W. et al. Targeted analysis of protein termini. *J Proteome Res* **6**, 4634-45(2007).

5. Van Damme, P. et al. Complementary positional proteomics for screening substrates of endo- and exoproteases. *Nat Methods* **7**, 512-5(2010).

6. Schilling, O. et al. Proteome-wide analysis of protein carboxy termini: C terminomics. *Nat Methods* **7**, 508-511(2010).

7. Wildes, D. & Wells, J. a Sampling the N-terminal proteome of human blood. *Proc Natl Acad Sci U S A* **107**, 4561-6(2010).

8. Vögtle, F.-N. et al. Global analysis of the mitochondrial N-proteome identifies a processing peptidase critical for protein stability. *Cell* **139**, 428-39(2009).

9. Van Damme, P. et al. Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis. *Nature Methods* **2**, 771-777(2005).

10. The call of the human proteome. *Nature Methods* **7**, 661-661(2010).

11. Keller, U. auf dem & Schilling, O. Proteomic techniques and activity-based probes for the system-wide study of proteolysis. *Biochimie* **92**, 1705-14(2010).

12. Vizcaíno, J.A. et al. A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* **9**, 4276-83(2009).

13. Arnesen, T. et al. Proteomics analyses reveal the evolutionary conservation and divergence of N-terminal acetyltransferases from yeast and humans. *Proc Natl Acad Sci U S A* **106**, 8157-62(2009).

14. Hwang, C.-S., Shemorry, A. & Varshavsky, A. N-terminal acetylation of cellular proteins creates specific degradation signals. *Science* **327**, 973-7(2010).

15. Hook, V. et al. Alternative pathways for production of beta-amyloid peptides of Alzheimer's disease. *Biol Chem* **389**, 993-1006(2008).

16. Stephan, A. et al. Mammalian glutaminyl cyclases and their isoenzymes have identical enzymatic characteristics. *FEBS J* **276**, 6522-36(2009).

17. Vousden, K.H. & Prives, C. Blinded by the Light: The Growing Complexity of p53. *Cell* **137**, 413-31(2009).

18. Hua, G. et al. Ignition of p53 bomb sensitizes tumor cells to granzyme K-mediated cytolysis. *J Immunol* **182**, 2152-9(2009).

19. Sayan, B.S. et al. p53 is cleaved by caspases generating fragments localizing to mitochondria. *J Biol Chem* **281**, 13566-73(2006).

20. Colaert, N. et al. Improved visualization of protein consensus sequences by iceLogo. *Nat Methods* **6**, 786-7(2009).

21. Nguyen, M. et al. Caspase-resistant BAP31 inhibits fas-mediated apoptotic membrane fragmentation and release of cytochrome c from mitochondria. *Mol Cell Biol* **20**, 6731-40(2000).

22. Iwasawa, R. et al. Fis1 and Bap31 bridge the mitochondria-ER interface to establish a platform for apoptosis induction. *EMBO J* **30**, 556-568(2010).

23. Breckenridge, D.G. et al. the BAP31 complex at the endoplasmic reticulum. *PNAS* (2002).

24. Annaert, W.G. et al. Export of cellubrevin from the endoplasmic reticulum is controlled by BAP31. *J Cell Biol* **139**, 1397-410(1997).

25. Paquet, M.-E. et al. Bap29/31 influences the intracellular traffic of MHC class I molecules. *J Immunol* **172**, 7548-55(2004).

26. Wang, B. et al. BAP31 interacts with Sec61 translocons and promotes retrotranslocation of CFTRDeltaF508 via the derlin-1 complex. *Cell* **133**, 1080-92(2008).

27. Kettenbach, A.N., Rush, J. & Gerber, S. a Absolute quantification of protein and post-translational modification abundance with stable isotope–labeled synthetic peptides. *Nature Protocols* **6**, 175-186(2011).

28. Mukasa, T. et al. Preliminary structural studies on the leucine-zipper homology region of the human protein Bap31. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **63**, 297-9(2007).

SUPPLEMENTARY METHODS

Database schema

The database schema is designed cognizant of the interrelationship between biological features of interest (see the **Entity-relationship diagram** for an excerpt). In data modeling, entities are distinct, uniquely identifiable building blocks, which are further described by their attributes and relationships to other entities. In TopFIND, molecules and molecular processes such as a cleavage event are considered as entities.

A protein entry, identified by its unique UniProtKB/Swiss-Prot accession code, is the single point of reference against which all other information is ultimately mapped and can therefore be used as a reliable reference to external resources.

A stable continuous stretch of amino acids with known N and C termini comprising only a part of the full-length protein, as annotated in the UniProtKB/Swiss-Prot database, is referred to as a chain and forms a one-to-one relation to its parent. Both termini are associated in a one to one relationship with the chain and can thereby referred to by each other.

Isoform and chain entities can have the same relations as protein entities except that they have to be linked directly (isoform or chain) or through any number of relations to other chains (only chain entities) to a single protein entity representing the canonical protein (as identified by UniProtKB/Swiss-Prot).

Each protein can have many N or C termini. This can be due to processing to remove the initiator methionine, signal peptide or propeptide, as well as alternate start sites, alternate splicing and cleavage to produce proteins or stable chains. A protein may be connected with many other proteins through a cleavage event. A cleavage is a directional relation with the first protein being referred to as protease cleaving the second protein, being referred to as substrate. Self referencing of one protein on itself through a cleavage event is allowed to encompass self-cleaving proteins. A given cleavage is specific to a position within the substrate amino acid sequence and is linked to the respective N termini and C termini of the fragments formed by such a split of the primary substrate sequence. Due to its central role, the cleavage event is considered an entity.

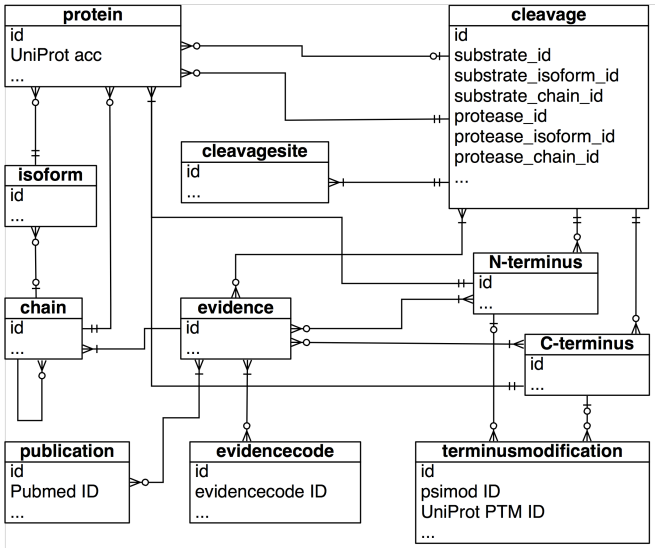
Similarly a protein (protease) can be linked to any other protein (inhibitor) in a many-to-many association by which its proteolytic activity is inhibited.

All entities describing biological components (protein, chain, isoform) or events (cleavage, inhibition) can be associated with` an evidence entity in a many to many fashion which in turn holds the associations to tissues, publications, supporting documents and evidence codes, as a controlled ontology classifying evidence e.g. ‘inferred from database annotation’ or ‘direct assay result’.

Implementation

The database is implemented in the format of a web resource to allow for maximum accessibility using a web interface designed to be compatible with all recent web browsers.

The data is stored in a relational database while data processing functionality is implemented in Ruby. Html files are generated through Ruby on Rails and can be delivered to the web browser by any Ruby compatible webserver like Apache. To improve data visualization and navigation, Javascript, AJAX technology and CSS-based styling is applied on the client side.



Entity-relationship diagram showing an extract of the TopFIND database | The relationship between the central protein, cleavage, C terminus, N terminus and related entities is displayed using the Crow’s Foot notation. A foot with three toes represents a cardinality of Many, a straight line a modality of 1, whereas a circle represents a modality of 0.

The current public database (http://clipserve.clip.ubc.ca/topfind) is running Mac OS X Server 10.6 (http://www.apple.com), MySQL 5.0.91 server (http://www.mysql.com), Ruby 1.8.7 (http://www.ruby-lang.org), Rails 2.3.8 (http://www.rubyonrails.org).

Data integration

General protein information, protein sequences, features and variations are imported from UniProtKB[1,2] by parsing UniProtKB/Swiss-Prot data sets as retrieved in the text (*.dat*) format from http://www.uniprot.org/downloads. Isoform sequences are retrieved by parsing the UniProtKB splice variant dataset provided in *fasta* format.

Where available, information on protease and inhibitor classification, substrates and cleavage sites is retrieved from MEROPS[3] using the provided SQL database dump (http://merops.sanger.ac.uk/download_list.shtml). MEROPS cleavage entries are read from the ‘cleavage table’ and mapped to TopFIND protein entries through the UniProt accession (acc) as it is stored in the ‘uniprot_acc’ field. For information that has been experimentally obtained by positional scanning synthetic-type libraries[4,5] or proteome-derived peptide libraries using PICS[6,7], which means these sites can not necessarily be mapped to native substrates, only protease and cleavage site entries are generated, while the stored substrate information is omitted. Each entry imported from MEROPS is accompanied by its own evidence-holding information from the ‘cleavage_evidence’, ‘cleavage_type’ and ‘cleavage_notes’ fields. Where available the MEROPS identifier is parsed from the ‘Ref’ field and converted into a PubMed ID by parsing the return from ‘http://merops.sanger.ac.uk/cgi-bin/refs?id=’ which is in turn used to create a link to the appropriate publication entry in TopFIND. Similarly, information on inhibitors is read from the ‘peptidase_inhibitor_complex’ table and entries are mapped through the UniProt accession stored in the ‘alternative_id’ table.

The information retrieved from UniProtKB/Swiss-Prot and MEROPS is periodically updated to the latest releases and version

numbers are stated in the web interface. Note, free form data is not automatically imported from UniProtKB and so some information in such fields will be absent in TopFIND, for example, if alternate start sites are manually listed here then this information will be missed in the parsing mode but can be added on a case by case basis when found.

Evidence information is stored using standardized ontologies where possible. The type of evidence is based on the Evidence Code ontology from the Open Biological and Biomedical Ontologies Foundry (http://www.obofoundry.org). Tissues and cell lines are cross-referenced with the UniProtKB tissue ontology. As the field and methodology are currently evolving very quickly the method used is stored using free form text. However, this will be replaced by a standardized ontology as soon as one covering the majority of methods becomes available. A probability score of correctness of the evidence, like for example a MASCOT score for the identification fo a peptide can be assigned to the evidence.

Modifications of the N terminal amino acid are classified according to the UniProtKB ontology for posttranslational modifications. Additional annotation is retrieved from the corresponding entries of the PSI-MOD ontology[8]. Related entries are grouped by their Keyword and a top level category entry carrying the name of the keyword is created (e.g. 'Acetylation' grouping, 'N-acetylmethionine' and 'N-acetylalanine').

Data visualization

Network visualization is carried out through CytoscapeWeb[9] integration. All substrates, proteases, inhibitors and inhibited proteases are retrieved recursively up to two levels deep. In addition, protein interactions are retrieved from UniProtKB annotation data. Multiple connections are collapsed into one and the data is sent to CytoscapeWeb by *graphml* inline with the main html file.

Protease cleavage site amino acid preference is visualized in two ways: For the average amino acid distribution in the 10 amino acids spanning the cut site (P5 to P5'), all cleavage site entries matching the site, protease and evidence filter settings are retrieved. Due to cleavages occurring close to a protein terminus or for methodological reasons, a cleavage site can have positions without amino acid information. To accommodate this, the percentage of occurrence is calculated individually for each position and is represented in the form of a heatmap. This representation is intended to visualize the raw data therefore no statistics and no compensation for the natural abundance of amino acids is applied.

Secondly the cleavage site consensus sequence is visualized as an iceLogo[10]. Utilizing the SOAP interface provided at http://iomics.ugent.be/icelogoserver/services/icelogo, cleavage site sequences are corrected for the natural amino acid abundance of the given species and significantly over and under represented amino acids ($p > 0.5$) are reported by their single letter code with the letter size proportional to the percentage. The result is retrieved as .svg image, stored and cached locally and displayed in the web interface.

Dataset import

A total of thirteen datasets derived from eight publications[11-18] describing direct identification of protein N or C termini from a variety of proteomes and species are currently integrated into TopFIND. More will be added as they become increasingly available.

N termini reported from (non-) apoptotic Jurkat cells are included as reported by in the supplementary tables 2 and 3 of Mahrus et. al.[12]. Protein N and C termini from K562 and PC3 cell lysates are imported as reported by Van Damme et al.[19], mitochondrial N termini from yeast have been identified by Voegtle et al.[14] and the human blood N terminome by Wildes et al.[13]. For quantitative studies comparing native and recombinant protease treated cell culture supernatants N termini with a ratio between 0.75 and 1.5 are considered not to be affected by the protease and imported into TopFIND[11,16,17]. N termini of LPS induced lung inflammation are extracted from the supplementary table 9 of Kleifeld et al.[11] and *E. coli* C termini have been reported by Schilling & Overall[18].

The identifiers used in the publications are converted to UniProt Accession numbers according to the UniProtKB cross-reference table. Entries that can not be unambiguously assigned to a UniProtKB Protein entry are omitted.

REFERENCES

1. Jain, E. et al. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* **10**, 136(2009).

2. Consortium, T.U. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* **39**, 214-219(2010).

3. Rawlings, N.D., Barrett, A.J. & Bateman, A. MEROPS: the peptidase database. *Nucleic Acids Res* **38**, D227-33(2010).

4. Turk, B.E. et al. Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. *Nat Biotechnol* **19**, 661-7(2001).

5. Thornberry, N.A. et al. A combinatorial approach defines specificities of members of the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis. *J Biol Chem* **272**, 17907-11(1997).

6. Schilling, O. & Overall, C.M. Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nat Biotechnol* **26**, 685-94(2008).

7. Schilling, O. et al. Characterization of the prime and non-prime active site specificities of proteases by proteome-derived peptide libraries and tandem mass spectrometry. *Nature Protocols* **6**, 111-120(2011).

8. Montecchi-Palazzi, L. et al. The PSI-MOD community standard for representation of protein modification data. *Nat Biotechnol* **26**, 864-6(2008).

9. Lopes, C.T. et al. Cytoscape Web: an interactive web-based network browser. *Bioinformatics* **26**, 2347-8(2010).

10. Colaert, N. et al. Improved visualization of protein consensus sequences by iceLogo. *Nat Methods* **6**, 786-7(2009).

11. Kleifeld, O. et al. Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nat Biotechnol* **28**, 281-8(2010).

12. Mahrus, S. et al. Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. *Cell* **134**, 866-76(2008).

13. Wildes, D. & Wells, J. a Sampling the N-terminal proteome of human blood. *Proc Natl Acad Sci U S A* **107**, 4561-6(2010).

14. Vögtle, F.-N. et al. Global analysis of the mitochondrial N-proteome identifies a processing peptidase critical for protein stability. *Cell* **139**, 428-39(2009).

15. Van Damme, P. et al. Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis. *Nature Methods* **2**, 771-777(2005).

16. Keller, U. auf dem et al. A statistics-based platform for quantitative N-terminome analysis and identification of protease cleavage products. *Mol Cell Proteomics* **9**, 912-27(2010).

17. Prudova, A. et al. Multiplex N-terminome analysis of MMP-2 and MMP-9 substrate degradomes by iTRAQ-TAILS quantitative proteomics. *Mol Cell Proteomics* **9**, 894-911(2010).

18. Schilling, O. et al. Proteome-wide analysis of protein carboxy termini: C terminomics. *Nat Methods* **7**, 508-511(2010).

19. Van Damme, P. et al. Complementary positional proteomics for screening substrates of endo- and exoproteases. *Nat Methods* **7**, 512-5(2010).