Published online 18 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* **D351–D361**
doi:10.1093/nar/gkr1025

TopFIND 2.0—linking protein termini with proteolytic processing and modifications altering protein function

Philipp F. Lange[1,2,3,*], Pitter F. Huesgen[1,2,3] and Christopher M. Overall[1,2,3,*]

[1]Centre for Blood Research, [2]Department of Oral Biological and Medical Sciences and [3]Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, BC, Canada V6T 1Z3

Received August 26, 2011; Revised October 6, 2011; Accepted October 23, 2011

ABSTRACT

Protein termini provide critical insights into the functional state of individual proteins. With recent advances in specific proteomics approaches to enrich for N- and C-terminomes, the global analysis of whole terminomes at a proteome-wide scale is now possible. Information on the actual N- and C-termini of proteins *in vivo* and any post-translational modifications, including their generation by proteolytic processing, is rapidly accumulating. To access this information we present version 2.0 of TopFIND (http://clipserve.clip.ubc.ca/topfind), a knowledgebase for protein termini, terminus modifications and underlying proteolytic processing. Built on a protein-centric framework TopFIND covers five species: *Homo sapiens, Mus musculus, Arabidopsis thaliana, Saccharomyces cerevisiae* and *Escherichia coli* and incorporates information from curated community submissions, publications, UniProtKB and MEROPS. Emphasis is placed on the detailed description and classification of evidence supporting the reported identification of each cleavage site, terminus and modification. A suite of filters can be applied to select supporting evidence. A dynamic network representation of the relationship between proteases, their substrates and inhibitors as well as visualization of protease cleavage site specificities complements the information displayed. Hence, TopFIND supports in depth investigation of protein termini information to spark new hypotheses on protein function by correlating cleavage events and termini with protein domains and mutations.

INTRODUCTION

The function and morphology of a cell is by and large carried out by its proteome, likewise, that of a tissue by its cells and their proteomes. The functional state of the proteome, and hence a cell, is defined by the structures, interactions and post-translational modifications of the proteins present at any given time and place. Proteins are synthesized as single polypeptide chains that have a start represented by the amino (N-) terminus, and a carboxyl (C-) terminus forming its end. Protein termini are frequently modified, both cotranslationally and post-translationally, by a variety of chemical modifications, including acetylation, methylation, myristoylation and pyroglutamate formation as well as by proteolytic processing, which profoundly affect protein stability, localization and function (1) (Figure 1).

Limited proteolytic processing is a very common and important post-translational protein modification that plays a major role in almost all essential cellular processes ranging from cell cycle division and proliferation, to cell death (2). Exopeptidases trim the polypeptide chain by one or few amino acids from either the N-terminal or C-terminal ends and endopeptidases cleave polypeptide chains internally generating two or more new protein chains defined by the old and newly formed N- and C-termini, now referred to as neo-termini (3). Examples include instances where the N-terminus matches the end of a signal peptide or pro-peptide indicating maturation and activation, respectively. Neo-termini found within a protein or protein domain or between domains are often associated with loss of the associated functions such as protein–protein interactions or catalytic activity, in addition to alterations in protein stability and changes in cellular location (Figure 1). Therefore, termini not only define a protein chain, but are also characteristic of a protein's functional competence. Thus, the terminome is indicative of the competence and functional state of the

*To whom correspondence should be addressed. Tel: +1 604 822 3561; Fax: +1 604 822 7742; Email: philipp.lange@ubc.ca
Correspondence may also be addressed to Christopher M. Overall. Tel: +1 604 822 2958; Fax: +1 604 822 7742; Email: chris.overall@ubc.ca

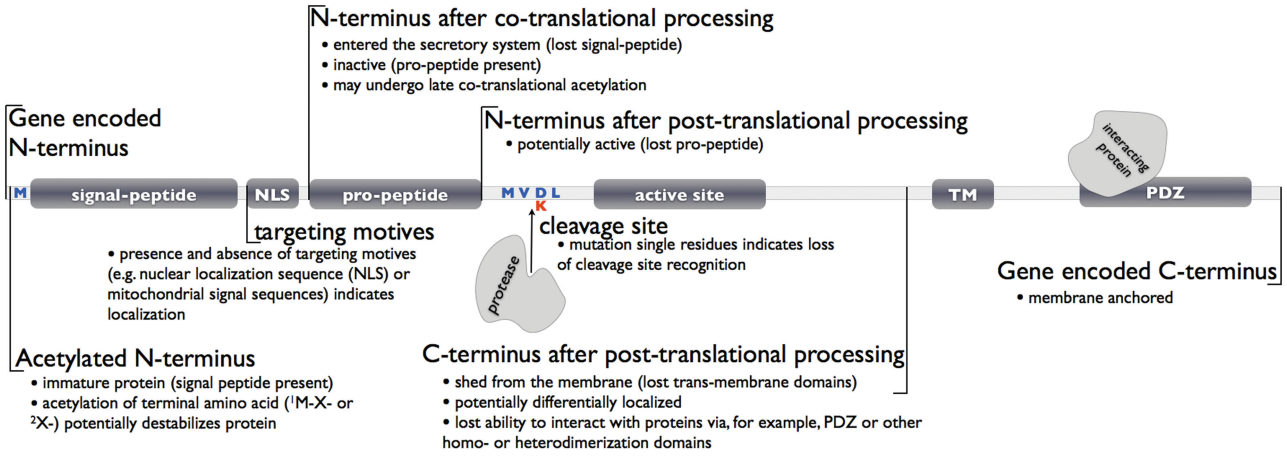


Figure 1. Functional state of a protein inferred from protein termini and cleavage events. A polypeptide chain is initially translated as encoded by the genome forming a gene encoded N- and C-terminus. Alternate start sites can occur as N-terminal isoforms. Subsequent processing leads to stable protein chains starting later or stopping earlier in the original polypeptide. Some of these processing events are part of the normal maturation of the protein during translation, such as methionine removal and signal peptide removal. In contrast, other cleavage events might occur post-translational and can be protein specific and necessary during normal protein function or are a result of encounters with a diverse array of proteases. Not shown here are nuclear targeting sequences or mitochondrial transit sequences that are proteolytically removed upon mitochondrial import. N-terminal modifications including acetylation and methylation or C-terminal modifications such as myristoylation can also occur and profoundly affect protein localization and function. Hence, identification of the termini and modifications from stable chains can indicate the processing state and functional potential of the protein. The acetylated N-terminus occurring after removal of the initiator methionine indicates an immature protein as its signal peptide is still present and suggest instability of this form as acetylation is known as a destabilizing modification under certain circumstances. N-termini located after the signal- or pro-peptide indicate successful targeting and activation respectively with the latter being indicative of an active protein. The depicted post-translationally created C-terminus removes the transmembrane and protein interaction domains from the protein chain indicating loss of membrane association and protein–protein interactions and hence a putative change in function.

proteome and the cell. Hence protein termini and their modifications are a rich information source that are underexploited for analytical, diagnostic, biomarker or drug discovery purposes in a wide range of diseases associated with aberrant proteolysis, such as Alzheimer's, colitis, arthritis and cancer.

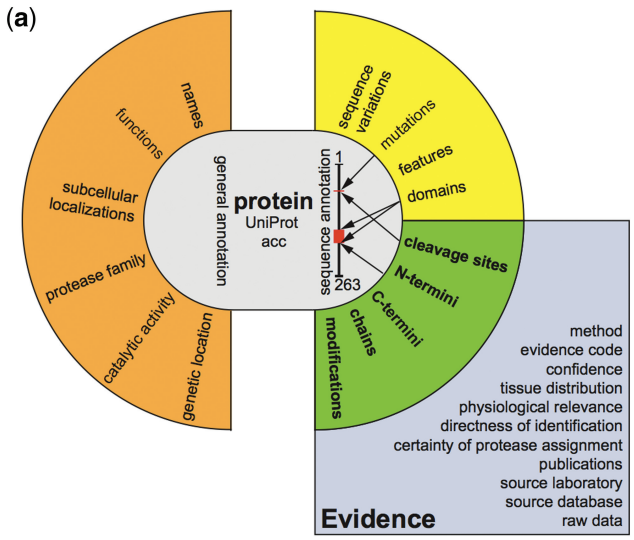
Recently a number of laboratories have made great efforts towards high-throughput mass spectrometry-based identification and characterization of protein N-termini (4–10) and C-termini (11–13). These proteomic methods now enable experimental determination of protein N- and C-termini and their modifications in complex *in vivo* samples, as demonstrated by investigations of the blood plasma N-terminome (14), the mitochondrial N-terminome (15) and the N-terminome during apoptosis (7,8,16). More than 50% of the protein termini identified in these unbiased N-terminome analyses were products of limited proteolysis, frequently occurring at unexpected positions (17), with unknown functional consequences. Hence, it is key to not only map the protein termini with existing protein annotation information and so collect information on their modification, but also to correlate this information with the proteolytic processes generating these chains and termini. In relation to the last, such events may form biomarkers for disease and offer new insight into pathological processes.

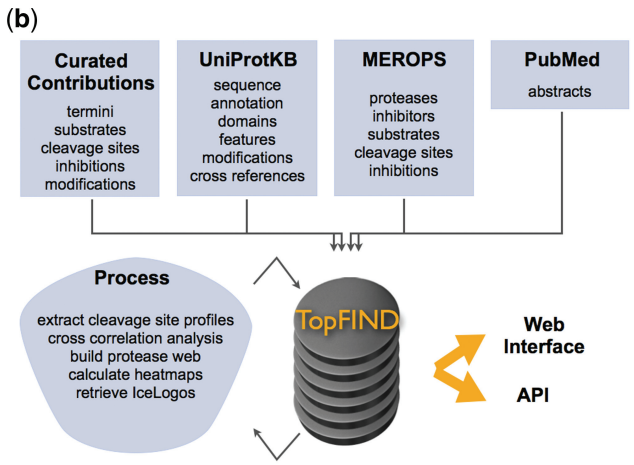
Many proteases have been characterized in depth by thorough biochemical analyses *in vitro*, and novel methodologies now enable the high-throughput identification of protease substrates and their precise cleavage sites in changing conditions *in vivo* (18,19). A number of databases like PMAP (20), CutDB (21) and the Degradome database (22) have been developed to make this detailed information on proteases and their substrates accessible, with MEROPS (23) being the gold standard and *de facto* authoritative resource. However, these resources are by design focused on the classification of proteases and their proteolytic activity towards a substrate and do not provide further knowledge integration and hypothesis generation. Furthermore, due to their static nature all these resources are not suited to explore the connections between different proteases, their inhibitors and substrates that are collectively known as the protease web (24). An understanding of the connectivity and the protease web as a whole is a crucial prerequisite for successful development of protease inhibitory drugs.

To address many of these issues we created the 'Terminus Oriented Protein Function Inferred Database' (TopFIND) in order to provide a central, user-friendly knowledgebase that integrates data on protein termini, their modifications and information on underlying proteolytic processes extracted from existing valuable resources and make this combined information accessible to a broad audience. TopFIND has allowed us to study the proteome-wide distribution and classification of N-termini across species as well as characteristics of N-terminal amino acid modifications (17). Here, we present version 2.0 of TopFIND that has been extended in content, species coverage and functionality.

TOPFIND FRAMEWORK AND CONTENT

TopFIND is designed as non-redundant, protein centric database (Figure 2a). Each protein is represented by its





(c)

total

	total	*H. sapiens*	*M. musculus*	*A. thaliana*	*S. cerevisiae*	*E. coli*
proteins	65,021	20,269	16,385	10,617	7,792	9,958
cleavages	11,000	6,662	3,679	43	190	426
N termini	83,362	29,221	23,483	10,837	9,401	10,420
C termini	73,815	24,867	19,790	10,722	8,198	10,238

annotation inferred

	total	*H. sapiens*	*M. musculus*	*A. thaliana*	*S. cerevisiae*	*E. coli*
N termini	68,618	22,101	17,122	10,773	8,307	10,315
C termini	66,172	20,630	16,626	10,722	8,204	9,990

cleavage inferred

	total	*H. sapiens*	*M. musculus*	*A. thaliana*	*S. cerevisiae*	*E. coli*
N termini	7,227	3,392	3,182	43	184	426
C termini	7,227	3,392	3,182	43	184	426

experimental evidence

	total	*H. sapiens*	*M. musculus*	*A. thaliana*	*S. cerevisiae*	*E. coli*
N termini	6,224	3,993	1,130	0	1,101	0
C termini	1,188	963	0	0	0	225

proteins with experimental evidence for one or more termini

	total	*H. sapiens*	*M. musculus*	*A. thaliana*	*S. cerevisiae*	*E. coli*
N termini	4,029	2,858	560	0	611	0
C termini	1,156	963	0	0	0	193

Figure 2. Framework and knowledgebase content. (**a**) TopFIND employs a protein-centric framework. General and sequence related annotation is linked to a protein entry identified by its UniProt accession. Sequence related annotations such as mutations, cleavage sites or protein termini are cross correlated based on their relative position. The focus of TopFIND, that is termini, cleavage events and chains, is associated with detailed qualitative and quantitative evidence information based on controlled vocabularies where defined. (**b**) Content is incorporated from curated community contributions, publications, UniProtKB, MEROPS and PubMED. Upon integration and data retrieval the data is processed further by TopFIND and made available through web and application programming interfaces. (**c**) Current TopFIND knowledgebase content by supported species. N- and C-termini are further split up into their origin.

genomic encoded polypeptide sequence and further detailed by general annotation such as alternative protein names, protein function and role in disease and by sequence-specific annotation as isoforms, domains and sequence variants. In addition to these commonly assembled protein characteristics TopFIND catalogs detailed information on protein chains, termini, modifications of the terminal amino acid and proteolytic processes. Additional information on chemical modifications of terminal amino acids is integrated from and referenced against the PSI-MOD ontology (25).

We placed special emphasis on the idea that protein characteristics, such as its termini, can greatly differ among tissue types, physiological conditions and experimental systems. To accommodate this we incorporated a powerful evidence and metadata framework into TopFIND (Figure 2a) making this information accessible and usable for comparison of protein characteristics between specific conditions.

Each protein chain, terminus, terminus modification and cleavage event listed in TopFIND is accompanied by detailed records of specific evidence for its existence at a given time and location. The evidence information is structured by the following criteria: (i) Directness: a qualitative measure [direct, likely direct, likely indirect, indirect, unknown] describing if the reported cleavage or terminus (or modification etc.) has either been directly observed (e.g. an N-terminus identified by Edman sequencing or proteomics approaches such as ATOMS, or terminomics approaches such as TAILS or COFRADIC [for a review, see (19)] or is indirectly inferred (e.g. by inferring a C-terminus from an experimentally observed neo-N terminus). (ii) Physiological relevance: a qualitative measure of the likelihood that the identified terminus, cleavage or inhibition is present and relevant *in vivo* [physiologically relevant, likely physiologically relevant, likely no physiological relevance, no physiological relevance, unknown]. Examples that can be considered in this classification include whether the substrate was identified in tissue samples or only in biochemical assays, is native or not, whether the protease is normally present in a certain cell, tissue or development stage, or whether it is even present in a given species. (iii) Confidence value and type: a quantitative measure describing the confidence that the given observation is correct. For example, the results generated from proteomics data base search programs such as Mascot includes a peptide identification confidence score, which can be filtered upon by TopFIND. *Type* refers to the program or system used to generate these scores. This differentiation allows for the evaluation and filtering of confidence values from algorithms and methodologies that cannot necessarily be directly compared, like for example a peptide identification by mass spectrometry and sequence determination by Edman sequencing. (iv) Certainty of protease assignment: a qualitative measure that the protease assigned to a cleavage is correct. This reflects the possibility of additional proteolytic activities other than the one studied and reported being present in the experimental system: [I, no other proteolytic activities present (e.g. a simple in vitro cleavage assay); II,

proteolytic system present but abolished (e.g. denatured); III, proteolytic system present but impaired (e.g. inactivated by inhibitors); IV, proteolytic system present and active (e.g. cell culture or *in vivo* studies); or unknown]. (v) Evidence code: categorization [e.g. inferred from electronic annotation] according to the controlled evidence vocabulary from the Obo foundry (26). (vi) Tissue distribution: if relevant, the cells or tissues in which the observation was made according to the controlled UniProtKB tissue and cell line vocabulary. (vii) Method and proteases inhibitors: a dropdown list of the current main approaches [electronic annotation, COFRADIC, N-TAILS, C-TAILS, ATOMS, Edman sequencing, enzymatic biotinylation, chemical biotinylation, MS (gel based), MS (semi-tryptic peptide identification), MS (other), mutation based analysis]. As this is an emerging field with intense interest, the methods and approaches for identification of protein termini and protease processing are rapidly developing and change frequently. Hence, the methods used are also captured in free text. Next the experimental system [cell free, cell culture, organ or tissue culture, *in vivo* sample] and the protease activity and modulation thereof [natural expression, recombinant, over expression, knockdown (expression), genetic knockout, knockdown (functional, protease drug or blocking antibody or inhibitor used)] are classified. A free text field is used with prompts to describe the experimental set up. This ensures that also small but important differences are tracked. As new methods emerge and gain traction these will be included in updates. Due to the critical effect of the activity of any other proteases in the system on the data generated and its interpretation it is explicitly asked if protease inhibitors were used during sample preparation and if so which ones. (viii) Source Database: the database (e.g. MEROPS, neXtProt) from which the information has been integrated if applicable. (ix) Laboratory: the research group and corresponding author who reported the finding. (x) Publications: author, title, abstract and full text links for publications related to the identification. (xi) Raw data repository: links to the raw data supporting the finding and stored in external repositories.

Integration of existing resources

Protein entries are retrieved from UniProtKB/Swiss-Prot. The UniProt accession number is used as a unique identifier and reference point for all information stored or generated by TopFIND and for cross references to external resources (Figure 2b). Extended UniProtKB/SwissProt protein annotation information including protein function, sequence, isoform sequences and sequence variants, domains, modifications and cross-references are retained for each entry. Already established knowledge on predicted and experimentally verified protein N- and C-termini that can be derived from UniProtKB annotated chain start and end points as well as from annotations of co-translational methionine removal is retained in TopFIND. TopFIND evidence records details of these entries as 'inferred from electronic annotation' of unknown physiological relevance, directness and confidence with UniProtKB being the source database.

Information on proteases, their substrates, inhibitors and cleavage sites is integrated from MEROPS (23) and mapped on TopFIND protein entities based on the UniProt accession (Figure 2b). Supplementary data such as the method details, laboratory of origin and description listed in MEROPS is retained in the TopFIND evidence records where possible. In particular the method, laboratory and description are retained. The physiological relevance classification is retained or stated as 'unknown' for ambiguous classifiers and the directness and confidence are given as 'unknown'. The likelihood of protease assignment classification is set as 'unknown', since no comparable evidence codes corresponding to the MEROPS classifications 'theoretical' and 'experimental' are applied and publication information is attached where available.

Integration of published studies and user-contributed data

Evidence information for data integrated in bulk from existing resources is often limited by the lack of supporting information for the collected knowledge and the use of different or inconsistent non-controlled vocabulary. However, the full power of the TopFIND evidence framework can be employed during the integration of individual or multiple identifications obtained in a single carefully controlled and described study. Therefore, we have manually curated and integrated the results of several published large-scale terminome studies (7,9,13–16,27,28). In addition, we have enabled and strongly encourage data submission from the community through a user-friendly web interface described in greater detail below.

Data processing

In addition to direct observations, protein N- and C-termini can be derived from the vast amount of data on proteolytic processing. This is automatically inferred from cleavage information and from the likely extent of a protein chain and its termini (Figure 2c). Such inferred protein termini are accompanied with the non-standard evidence code 'inferred from cleavage', a directness of 'indirect' and described as 'unknown' in the physiological relevance and confidence entries.

To extract positional and relational information we integrate the accumulated data further and perform a positional cross correlation analysis. The position of termini and cleavage sites is correlated with the positional information of protein domains and features like mutations or active sites. Termini and cleavage sites that match a point feature or coincide with the start or end of an extended protein feature are specifically marked as affected features and feature boundaries. This is particularly useful if the function of particular protein domains or sequences are known, as hypotheses can then be made on the function of the cleavage fragments versus the intact protein (see below).

Current knowledgebase content

Following the initial release, in version 2.0 TopFIND has been expanded and now covers five major organisms ranging from prokaryotes to fungi to plants and mammals, including *Escherichia coli*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Mus musculus* and *Homo sapiens*. More are being added over time. The human and murine TopFIND termini datasets contain significant contributions from experimental N-terminome determinations, whereas protein termini inferred from existing UniProt/Swissprot annotations dominate the datasets for the remaining species (Figure 2c). However, this is certain to change as recently established proteomics based high-throughput methods become more widely adopted. The framework provided by TopFIND enables complete collection and consistent annotation of the emerging data.

INFORMATION RETRIEVAL

We have developed several access routes to the information in TopFIND to accommodate a great variety of questions. A web interface provides a powerful, convenient and intuitive way to study information on individual proteins. It is however limited in throughput and difficult to integrate into existing workflows. Therefore, we further developed two dedicated Application Programming Interfaces (API). These enable automated querying and retrieval of TopFIND stored knowledge and facilitate the integration with other resources and analysis platforms.

Application programming interfaces

PSICQUIC API. This API is based on the popular PSI common query interface (PSICQUIC) for standardized querying of molecular interaction data across multiple sources (29). Proteolytic cleavage processes, as well as the inhibition thereof are a specific form of protein–protein interaction. Although providing important functional information, protein cleavage events are currently underrepresented with only 300 out of 16 000 000 entries found in 14 molecular interaction resources (http://www.ebi.ac.uk/Tools/webservices/psicquic/view). To add TopFIND knowledge to this information pool, all TopFIND cleavage events and inhibitions are exposed to PSICQUIC using the reference implementation (29). The query interface is publicly available at http://clipserve.clip.ubc.ca/topfind-psicquic-search/query/* and the SOAP interface can be accessed at http://clipserve.clip.ubc.ca/topfind-psicquic-ws. An example search query retrieving the first 10 entries is http://clipserve.clip.ubc.ca/topfind-psicquic-search/query/*:*?firstResult=0&maxResults=10. This not only allows users to retrieve TopFIND information using PSIQUIC enabled software (e.g. in Cytoscape) but also to query information on protein cleavage events in parallel to general protein–protein interaction data from all other PSICQUIC enabled resources (e.g. DrugBank, IntAct, STRING) using the PSICQUIC View provided by the European Bioinformatics Institute at http://www.ebi.ac.uk/Tools/webservices/psicquic/view.

XML API. PSICQUIC is limited to retrieval of the protein cleavage information. We therefore developed an additional XML-based API to make the extensive additional information found in TopFIND accessible to automated retrieval (http://clipserve.clip.ubc.ca/topfind). In particular, the positional cross correlation analysis is made accessible to external analysis and annotation systems. A simple query notation, listed in detail on the TOPFIND help webpage, allows retrieval of lists with matching proteins, cleavage events, N- or C-termini. If the first query was not aimed directly at proteins, detailed information on the corresponding proteins including all termini, proteolytic processing and substrates can then be retrieved in a second query using the protein accession delivered by the first query. The returned information is XML encoded for automated post-processing using standard computational procedures.

Web interface

We created a web interface to make the information compiled by TopFIND available to a broad scientific audience. The interface puts protein centric information at center stage and provides detailed background information on amino acid modifications, a comprehensive documentation and a step-by-step data contribution system.

Searching. The user is presented with a simple search interface that allows searching by protein accession number or name, alternate protein or gene names and, in the case of proteases, by MEROPS family identifiers. An advanced search interface can be used to restrict the search by species, genomic location, presence of specific terminal amino acid modifications, number of reported C- or N-termini, or restricted to proteases or their inhibitors. This allows for complex queries, for example searching for N-terminally acetylated human proteins encoded on a specific chromosome.

The output of a successful search for a protein of interest is presented with a detailed protein overview page composed of seven sections. Some of these may also be exported as a tab-delimited spreadsheet for offline use.

Protein annotation and domain structure. General protein information, the sequence and isoform sequences are accompanied by MEROPS protease classification if applicable and more detailed protein annotation derived from UniProtKB. This overview is followed by a graphical depiction of the sequential arrangement of protein domains, features, termini and cleavage events. This allows for easy assessment how termini and cleavage events relate to each other as well as to protein domains and features (Figure 3). Evidence information for termini and cleavage events is directly accessible in the overlay linked to their graphical representation in the domain overview.

Protein termini and processing. Proteolytic processing of the protein, its substrates and known protein termini are listed in separate sections. Each terminus or cleavage event (of the protein itself and, in the case of proteases, or one of

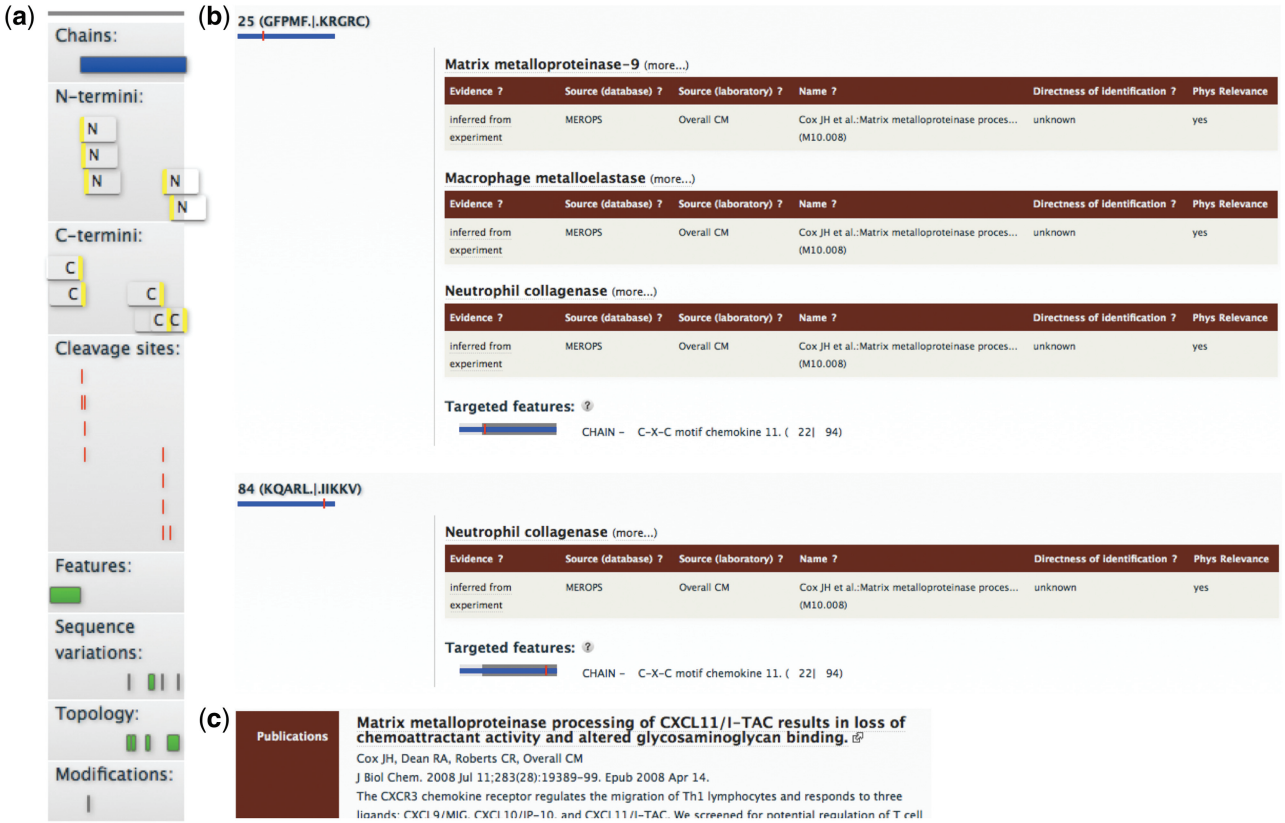


Figure 3. A substrate example. Chemokine function is altered by proteolytic processing. (**a**) Domain structure, cleavage sites and termini of human C-X-C motif chemokine 11 (CXCL11) as depicted by the web interface. (**b**) Tabular information on proteolytic processing at position 25 and 84 by different proteases along with evidence information and affected features. (**c**) Abstract of the publication describing the cleavage events depicted in b as it is displayed along with other evidence information by the web interface.

its substrates) is listed along with the position, the local sequence context and supporting evidence. The comprehensive evidence information including the methodology, a description and abstracts from related publications is accessible through links opening as overlays. In addition all results of the cross correlation analyses are displayed. This section makes it easy to identify potential physiological implications as detailed below.

All termini and cleavage site listings can be exported as tab-delimited spreadsheets to allow further offline analysis and visualization.

Protease web. An interactive network graph visualizes the relation of the queried protein to multiple proteases, their inhibitors and shared as well as unique substrates (Figure 4). The dynamic network representation is realized using Cytoscape Web (30) and allows the user to browse the protease web. Clicking on a node of interest opens the corresponding protein page and re-focuses the protease web around that protein. Use of the evidence filtering mechanism described below enables to study complex questions such as differences in connectivity and redundancy in various species, cell lines or conditions.

Protease cleavage specificity. The cleavage site sequence specificity for proteases is calculated from their associated substrate cleavage sites and displayed both as a heatmap (absolute amino acid occurrence percentages) and as an iceLogo, which corrects for the natural amino acid distribution and tests for statistic significance (31) (Figure 5).

Online documentation and help. Throughout the page small question marks provide access to short explanations parameters or entry fields. In addition an extensive online documentation section provides in depth explanation of all options and parameters from searching over data interpretation to data submission.

Data submission. We strongly encourage the scientific community to batch upload relevant experimental data on protein termini and protease cleavage sites directly to TopFIND. A detailed guide including screenshots of every step and provided as downloadable PDF file guides the user through the submission process. The submission interface is password protected to prevent abuse by automated systems, but registration is free and requires only a valid e-mail address and choice of a password. After login, data can be uploaded as a comma separated file describing the reported termini or cleavage events and following a concise format described in detail on the submission site. A complementing electronic form asks for detailed descriptions of the experimental system and employed methodology automatically linking the provided information

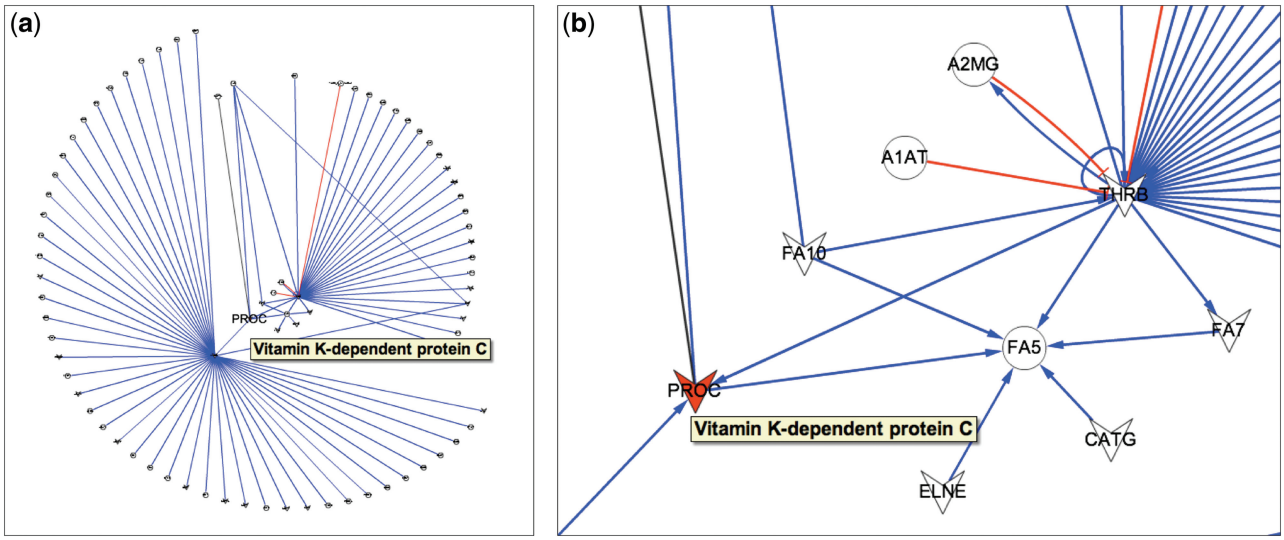


Figure 4. Browsing the protease web. (**a**) The protease web centered on vitamin K-dependent protein C as visualized in the TopFIND web interface. (**b**) Enlarged section of the network depicted in a highlighting the strong interconnectivity with proteases (V shapes) cleaving (blue arrows) normal proteins (circles) and other proteases themselves (circular arrow for THRB). The inhibitors involved are also shown (red bars).

to controlled vocabularies. Uploaded and submitted data will be reviewed for completeness and data integrity by the TopFIND curators before it is made available to the public.

Evidence based filtering. A key feature of TopFIND is the opportunity to filter search results by several detailed quantitative and qualitative evidence categories for each entry, including directness, physiological relevance, confidence value and type, evidence code, tissue distribution, experimental method, source database, source laboratory and source publications as defined above.

By default, all information linked to a query is displayed on the search result page. This can be customized through several filter options accessible on the top right side of the web interface or by adding specific parameters to an API query call. After filtering, only data that is backed with evidence matching the chosen filter settings is returned. This will affect the listed termini, cleavage sites and substrates. Also the new network view will only display connections matching the specified criteria, and for proteases the displayed cleavage site preference will be re-calculated. This filter system provides a powerful tool for deeper analysis of the provided data.

Terminus centric data retrieval. In addition to the protein centric approach described above, information on protein termini can be searched for and retrieved in a terminus centric approach. Queries return a list of matching protein C-termini or N-termini including sequence excerpts, their position in the protein sequence, modifications and a link to the corresponding protein page and evidence information. A flexible interface enables a variety of searches by amino acid sequence or sequence pattern (using a regular expression syntax described in detail in the online help), modification, position, species, or specific proteins. The results can be exported as comma separated spreadsheet for further offline analysis and visualization.

USAGE AND DATA MINING

In the following paragraphs we illustrate how TopFIND can be interrogated to extract new and useful information.

One prominent example of how minor differences in the length and start of a protein chain can profoundly affect protein function are chemokines (32). For example identification of a N-terminus at position 25 of human C-X-C motif chemokine 11 (CXL11, UniProtKB accession O14625) can spark interest in the functional state of this chemokine. TopFIND reveals that a number of cleavage events cluster shortly after the start and before the end of the annotated mature chemokine chain (Figure 3a). One of those cleavage events matches the position of the N-terminus under investigation. The detailed information on this cleavage and similarly the other cleavage events clustered at the N- and C-terminal ends of the chain lists a number of different proteases that have been shown to process CXL11 at these positions (Figure 3b). The abstract of the related publication can be pulled up for each of these, often providing important information on the physiological role of the cleavage (Figure 3c). In this example the abstract reveals that a N-terminus at position 25 represents a processed form that lost the agonistic properties of the mature active form of CXL11. It also indicates that for full functional characterization the precise C-terminus of the chain has to be determined as it directly affects glycosaminoglycan binding and antagonistic activity (33).

A query for murine α-enolase (UniProtKB accession P17182) using the web interface displays the stored information overlaid onto the protein sequence and predicted protein domains. This includes a long list of cleavage sites

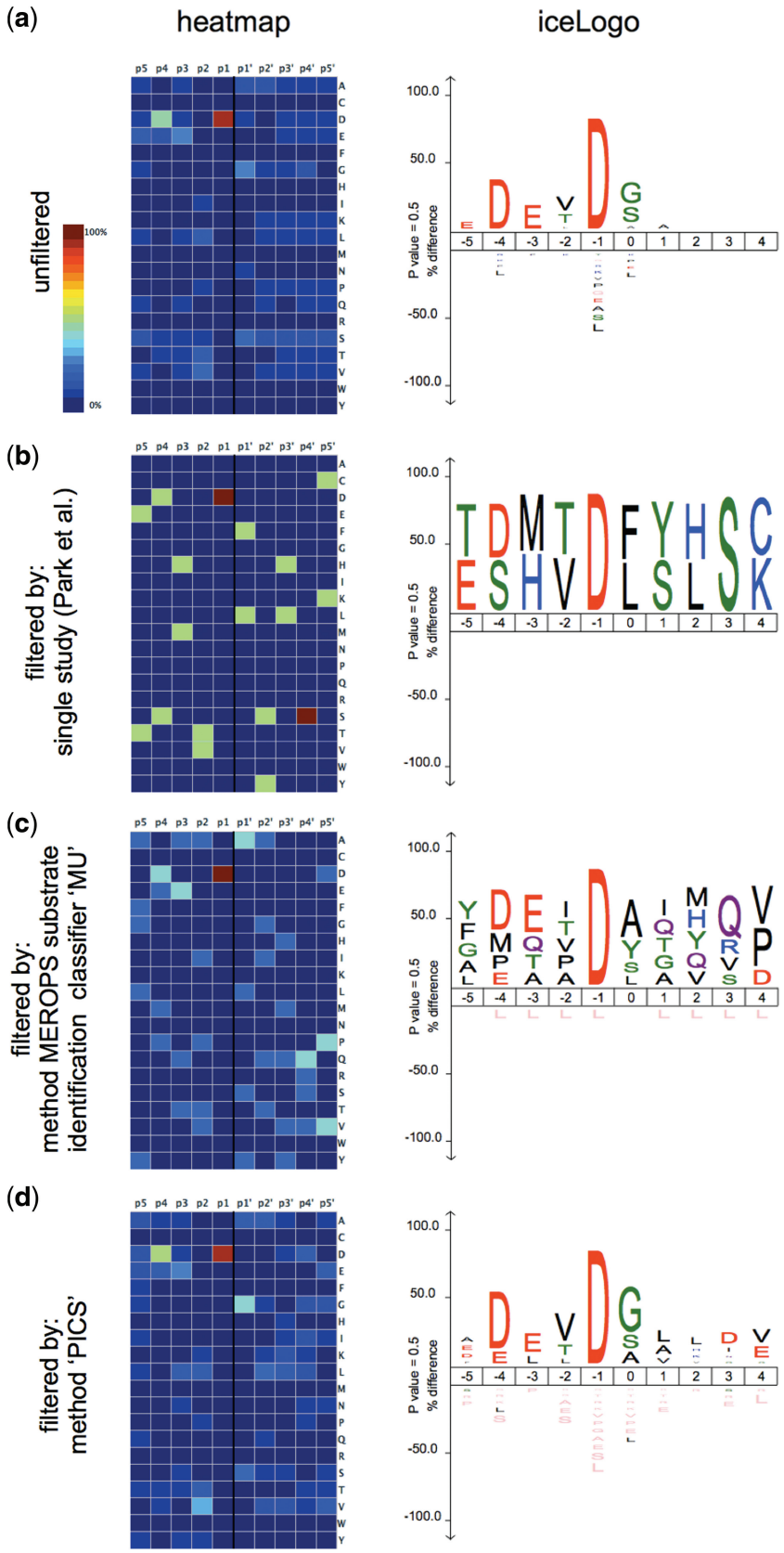


Figure 5. Protease sequence specificity with different filter settings. Heatmap and iceLogo representations of the sequence specificity of human caspase 3 as displayed by the TopFIND web interface using different filter settings. (**a**) Unfiltered, all cleavage events site sequences are taken into account ($n = 410$). (**b**) Filtered for the sequences reported in a single study by Park *et al.* ($n = 2$). (**c**) Filtered by methodology selecting all sequences tagged with the MEROPS method classifier of 'MU' representing substrate identification by mutation ($n = 5$). (**d**) Filtered by sequences reported using the method 'PICS' (41) ($n = 36$).

and termini, often clustered and separated only by one or a few amino acids. However, many of the termini are inferred from proteolytic processing, which may in part represent degradative events that do not result in stable, functional protein fragments. A user may therefore question whether any termini have been directly observed, and so can select for these by setting the 'directness' filter option to 'direct'. This greatly reduces the number of N- and C-termini to only those that have been experimentally verified. Interestingly, several of these N-termini are distributed across the whole polypeptide, indicating that several different stable protein chains are generated by proteolytic processing. This corresponds well with the general annotation derived from UniProtKB and displayed in the annotation section describing α-enolase as multifunctional enzyme involved in various processes at locations as different as the cytoplasm and the cell surface (34).

Similarly, the positional cross-correlation of mutations or sequence variants in a protein of interest with observed cleavage sites can indicate that the physiological effect of these mutations might be triggered by modifying or abolishing cleavage at a given site. Using TopFIND, several somatic mutations in the tumor suppressor protein p53 were found to coincide with a caspase-3 cleavage site, which is abolished by mutation. Hence, lack of caspase cleavage at this site during apoptosis potentially diminishes the cancer-protective functions of p53 (17).

TopFIND is especially valuable for researchers interested in proteases and their inhibitors, as it provides a unique tool to investigate the multiple interdependencies between proteases, substrates and their inhibitors collectively known as the protease web (35). For example, a query for human vitamin-K dependent protein C (protein C, UniProtKB accession P04070), a serine protease that acts as a major anticoagulation factor (36), shows a complicated network of interacting proteases, inhibitors and substrates (Figure 4a) that can be studied more closely by zooming in (Figure 4b). The displayed interactions show that protein C is a substrate of furin and thrombin, which reflect limited processing during secretion by furin and activation from its zymogen precursor by thrombin. Hence, this identifies coagulation factors V and VIII as protein C substrates. The network view further illustrates other proteases cleaving the same substrates, proteases and inhibitors affecting upstream located proteases such as thrombin, and other substrates of proteases that interact with protein C. A simple click on any of the displayed proteins opens the associated TopFIND result page and thereby focuses the network on this next level of interactions.

TopFIND can also be used to infer protease cleavage site preferences (17). The query result for human caspase-3 (UniProtKB accession P42574) contains a large number of substrate cleavage events that are also summarized as an iceLogo. The logo clearly shows the well known strong cleavage site preference for Asp at P1 (37) directly preceding the cleavage site, and an additional preference for Asp at P4. TopFIND filter options can be used to compare cleavage events reported in different publications, derived from different resources or obtained by different methods (Figure 5). This and similar searches may be used to query the substrate lists and displayed cleavage logos and to identify potential biases introduced by chosen experimental conditions.

CONCLUSION AND FUTURE DEVELOPMENTS

TopFIND provides a central repository and access point for knowledge on protein termini, their modifications and proteolytic processes in general. More importantly, by displaying this positional correlated information TopFIND assists researchers in intuitively creating new, experimentally testable hypotheses for the role of proteins and protein domains and their modifications, including proteolytic processing, in health and disease.

In addition to regular updates of the knowledgebase content we have identified three focus areas for the future development of TopFIND. First, the ongoing effort of integration with other existing or new resources and analysis pipelines is core to provide even better access to the stored data. Hence we will improve integration with UniProt and MEROPS, not only on the user level but also heavily on the level of data exchange. We also plan to improve the import and visualization of the network knowledge by Cytoscape.

Second, we regard the full integration of protein isoforms as well as spatial-temporal and structural information as the next important step forward. To this end we are working on the integration of resources like functional gene expression data (38), ProteinAtlas (39) and structural mapping of termini and cleavage sites.

Third, TopFIND is fully supporting the Human Proteome Project (HPP) (40). As part of this international effort to map the complete human proteome, its modifications and spatiotemporal variations in healthy subjects and eventually across diseases and disease states, TopFIND will act as central hub for all knowledge generated by the terminome subprojects. Bidirectional data exchange with participating research laboratories and the central HPP knowledge hub neXtProt (http://www.nextprot.org/) is already one important aspect of this. More importantly however, TopFIND will integrate the assembled knowledge to guide researchers in identification of proteins, tissues and disease states that need further investigation to reach completeness.

We believe that the many unique features of TopFIND, in particular the ability to gather seemingly disparate evidences on events and modifications that can alter protein termini or generate new termini, make TopFIND a powerful new tool for assigning protein function. By such data integration, information from terminomics and other experiments can be converted into knowledge. We hope that TopFIND will be a foundation for new hypotheses and exciting new concepts on protein function that spring from knowledge of the alterations made at the end of the protein chain and not just the end of a protein's life.

ACKNOWLEDGEMENTS

The authors thank Anna Prudova for stimulating discussions.

FUNDING

Canadian Institutes of Health Research; Cancer Research Society; British Columbia Proteomics Network, a Canada Research Chair in Metalloproteinase Proteomics and Systems Biology (to C.M.O.); Michael Smith Foundation for Health Research (to P.F.L. and P.F.H.); Breast Cancer Society of Canada, Alexander von Humboldt Foundation and the German Federal Ministry of Education and Research (to P.F.L.); German Academic Exchange Service (to P.F.H.). Funding for open access charge: Canadian Institutes of Health Research.

Conflict of interest statement. None declared.

REFERENCES

1. Meinnel,T. and Giglione,C. (2008) Tools for analyzing and predicting N-terminal protein modifications. *Proteomics*, **8**, 626–649.
2. Puente,X.S., Sánchez,L.M., Overall,C.M. and López-Otín,C. (2003) Human and mouse proteases: a comparative genomic approach. *Nat. Rev. Genet.*, **4**, 544–558.
3. Overall,C.M. and Blobel,C.P. (2007) In search of partners: linking extracellular proteases to substrates. *Nat. Rev. Mol. Cell Biol.*, **8**, 245–257.
4. 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.
5. McDonald,L., Robertson,D.H.L., Hurst,J.L. and Beynon,R.J. (2005) Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. *Nat. Methods*, **2**, 955–957.
6. Timmer,J.C., Enoksson,M., Wildfang,E., Zhu,W., Igarashi,Y., Denault,J.-B., Ma,Y., Dummitt,B., Chang,Y.-H., Mast,A.E. *et al.* (2007) Profiling constitutive proteolytic events in vivo. *Biochem. J.*, **407**, 41–48.
7. 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.
8. Xu,G., Shin,S.B.Y. and Jaffrey,S.R. (2009) Global profiling of protease cleavage sites by chemoselective labeling of protein N-termini. *Proc. Natl Acad. Sci. USA*, **106**, 19310–19315.
9. 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.
10. Doucet,A. and Overall,C.M. (2011) Broad coverage identification of multiple proteolytic cleavage site sequences in complex high molecular weight proteins using quantitative proteomics as a complement to edman sequencing. *Mol. Cell Proteomics*, **10**, M110.003533.
11. Dormeyer,W., Mohammed,S., Breukelen,B.V., Krijgsveld,J. and Heck,A.J.R. (2007) Targeted analysis of protein termini. *J. Proteome Res.*, **6**, 4634–4645.
12. Van Damme,P., Staes,A., Bronsoms,S., Helsens,K., Colaert,N., Timmerman,E., Aviles,F.X., Vandekerckhove,J. and Gevaert,K. (2010) Complementary positional proteomics for screening substrates of endo- and exoproteases. *Nat. Methods*, **7**, 512–515.
13. Schilling,O., Barré,O., Huesgen,P.F. and Overall,C.M. (2010) Proteome-wide analysis of protein carboxy termini: C terminomics. *Nat. Methods*, **7**, 508–511.
14. Wildes,D. and Wells,J.A. (2010) Sampling the N-terminal proteome of human blood. *Proc. Natl Acad. Sci. USA*, **107**, 4561–4566.
15. Vögtle,F.-N., Wortelkamp,S., Zahedi,R.P., Becker,D., Leidhold,C., Gevaert,K., Kellermann,J., Voos,W., Sickmann,A., Pfanner,N. *et al.* (2009) Global analysis of the mitochondrial N-proteome identifies a processing peptidase critical for protein stability. *Cell*, **139**, 428–439.
16. Van Damme,P., Martens,L., Van Damme,J., Hugelier,K., Staes,A., Vandekerckhove,J. and Gevaert,K. (2005) Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis. *Nat. Methods*, **2**, 771–777.
17. Lange,P.F. and Overall,C.M. (2011) TopFIND, a knowledgebase linking protein termini with function. *Nat. Methods*, **8**, 703–704.
18. Doucet,A., Butler,G.S., Rodríguez,D., Prudova,A. and Overall,C.M. (2008) Metadegradomics: toward in vivo quantitative degradomics of proteolytic post-translational modifications of the cancer proteome. *Mol. Cell Proteomics*, **7**, 1925–1951.
19. auf dem Keller,U. and Schilling,O. (2010) Proteomic techniques and activity-based probes for the system-wide study of proteolysis. *Biochimie*, **92**, 1705–1714.
20. Igarashi,Y., Heureux,E., Doctor,K.S., Talwar,P., Gramatikova,S., Gramatikoff,K., Zhang,Y., Blinov,M., Ibragimova,S.S., Boyd,S. *et al.* (2009) PMAP: databases for analyzing proteolytic events and pathways. *Nucleic Acids Res.*, **37**, D611–D618.
21. Igarashi,Y., Eroshkin,A., Gramatikova,S., Gramatikoff,K., Zhang,Y., Smith,J.W., Osterman,A.L. and Godzik,A. (2007) CutDB: a proteolytic event database. *Nucleic Acids Res.*, **35**, D546–D549.
22. Quesada,V., Ordóñez,G.R., Sánchez,L.M., Puente,X.S. and López-Otín,C. (2009) The Degradome database: mammalian proteases and diseases of proteolysis. *Nucleic Acids Res.*, **37**, D239–D243.
23. Rawlings,N.D., Barrett,A.J. and Bateman,A. (2010) MEROPS: the peptidase database. *Nucleic Acids Res.*, **38**, D227–D233.
24. Overall,C.M. and Kleifeld,O. (2006) Tumour microenvironment - opinion: validating matrix metalloproteinases as drug targets and anti-targets for cancer therapy. *Nat. Rev. Cancer*, **6**, 227–D239.
25. Montecchi-Palazzi,L., Beavis,R., Binz,P.-A., Chalkley,R.J., Cottrell,J., Creasy,D., Shofstahl,J., Seymour,S.L. and Garavelli,J.S. (2008) The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol.*, **26**, 864–866.
26. Smith,B., Ashburner,M., Rosse,C., Bard,J., Bug,W., Ceusters,W., Goldberg,L.J., Eilbeck,K., Ireland,A., Mungall,C.J.C.J. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.
27. 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.
28. Prudova,A., auf dem Keller,U., Butler,G.S. and Overall,C.M. (2010) Multiplex N-terminome analysis of MMP-2 and MMP-9 substrate degradomes by iTRAQ-TAILS quantitative proteomics. *Mol. Cell Proteomics*, **9**, 894–911.
29. Aranda,B., Blankenburg,H., Kerrien,S., Brinkman,F.S.L., Ceol,A., Chautard,E., Dana,J.M., De Las Rivas,J., Dumousseau,M., Galeota,E. *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods*, **8**, 528–529.
30. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.
31. 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.
32. McQuibban,G.A., Gong,J.H., Tam,E.M., McCulloch,C.A., Clark-Lewis,I. and Overall,C.M. (2000) Inflammation dampened by gelatinase A cleavage of monocyte chemoattractant protein-3. *Science*, **289**, 1202–1206.

33. Cox,J.H., Dean,R.A., Roberts,C.R. and Overall,C.M. (2008) Matrix metalloproteinase processing of CXCL11/I-TAC results in loss of chemoattractant activity and altered glycosaminoglycan binding. *J. Biol. Chem.*, **283**, 19389–19399.
34. Butler,G.S. and Overall,C.M. (2009) Proteomic identification of multitasking proteins in unexpected locations complicates drug targeting. *Nat. Rev. Drug Discov.*, **8**, 935–948.
35. Overall,C.M. and López-Otín,C. (2002) Strategies for MMP inhibition in cancer: innovations for the post-trial era. *Nat. Rev. Cancer*, **2**, 657–672.
36. Esmon,C.T. (2003) The protein C pathway. *Chest*, **124**, 26S–32S.
37. Thornberry,N.A., Rano,T.A., Peterson,E.P., Rasper,D.M., Timkey,T., Garcia-Calvo,M., Houtzager,V.M., Nordstrom,P.A., Roy,S., Vaillancourt,J.P. *et al.* (1997) 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**, 17907–17911.
38. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
39. Uhlen,M., Oksvold,P., Fagerberg,L., Lundberg,E., Jonasson,K., Forsberg,M., Zwahlen,M., Kampf,C., Wester,K., Hober,S. *et al.* (2010) Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.*, **28**, 1248–1250.
40. The call of the human proteome. (2010) *Nat. Methods*, **7**, 661.
41. Schilling,O. and Overall,C.M. (2008) Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nat. Biotechnol.*, **26**, 685–694.