Metadegradomics
TOWARD IN VIVO QUANTITATIVE DEGRADOMICS OF PROTEOLYTIC POST-TRANSLATIONAL MODIFICATIONS OF THE CANCER PROTEOME*

Alain Doucet†‡, Georgina S. Butler, David Rodríguez, Anna Prudova, and Christopher M. Overall§

Post-translational modifications enable extra layers of control of the proteome, and perhaps the most important is proteolysis, a major irreversible modification affecting every protein. The intersection of the protease web with a proteome sculpts that proteome, dynamically modifying its state and function. Protease expression is distorted in cancer, so perturbing signaling pathways and the secretome of the tumor and reactive stromal cells. Indeed many cancer biomarkers are stable proteolytic fragments. It is crucial to determine which proteases contribute to the pathology versus their roles in homeostasis and in mitigating cancer. Thus the full substrate repertoire of a protease, termed the substrate degradome, must be deciphered to define protease function and to identify drug targets. Degradomics has been used to identify many substrates of matrix metalloproteinases that are important proteases in cancer. Here we review recent degradomics technologies that allow for the broadly applicable identification and quantification of proteases (the protease degradome) and their activity state, substrates, and interactors. Quantitative proteomics using stable isotope labeling, such as ICAT, isobaric tags for relative and absolute quantification (iTRAQ), and stable isotope labeling by amino acids in cell culture (SILAC), can reveal protease substrates by taking advantage of the natural compartmentalization of membrane proteins that are shed into the extracellular space. Identifying the actual cleavage sites in a complex proteome relies on positional proteomics and utilizes selection strategies to enrich for protease-generated neo-N termini of proteins. In so doing, important functional information is generated. Finally protease substrates and interactors can be identified by interactomics based on affinity purification of protease complexes using exosite scanning and inactive catalytic domain capture strategies followed by mass spectrometry analysis. At the global level, the N terminome analysis of whole communities of proteases in tissues and organs in vivo provides a full scale understanding of the protease web and the web-sculpted proteome, so defining metadegradomics. Molecular & Cellular Proteomics 7: 1925–1951, 2008.

The dynamic nature of complex proteomes, which differ between tissues, between cells, and within a cell during growth and development and over time, presents unique challenges that distinguish large scale proteomics efforts from full genome sequencing projects. Moreover the proteome is modified in response to external stimuli and hence proteolytic modifications, which occur after the synthesis of the polypeptide chain. Over 300 protein post-translational modifications have been identified, and they profoundly influence protein function, localization, activity, and structure. Common post-translational modifications include phosphorylation (1), glycosylation (2), conjugation of small proteins such as ubiquitin (3, 4) and SUMO (small ubiquitin-related modifier) (5), and acetylation of the N terminus (6) and lysine residues (7), but the proteolytic modification of proteins is often overlooked in its importance and ubiquity (8). Indeed many biomarkers of disease are stable proteolytic fragments in biological fluids. Pro1708/Pro2044 (the C-terminal fragment of albumin) (9), HER2 rb2 (the ectodomain of human epithelial growth factor receptor-2) (10, 11), and CYFRA 21-1 (a soluble fragment of cytokeratin 19) (12–14) are all potential cancer biomarkers that are generated by proteolysis.

Proteolysis, the irreversible hydrolysis of peptide and isopeptide bonds, affects every protein at some point in its life cycle and because of its ubiquity, sculpts whole proteomes. Arguably then, proteolysis may be more important than even phosphorylation, which affects only selected proteins. Proteolysis ranges from the more dramatic degradation-to-completion processing to single specific cleavages within a protein, including the almost imperceptible trimming of a few N- or C-terminal residues. This highly specific, kinetically efficient, and tightly controlled proteolysis, termed proteolytic processing, enables proteases to precisely modulate protein function. Biologically, proteolytic processing can be highly relevant: the removal of just two to four residues can convert a chemokine from a receptor agonist to an antagonist and hence abruptly change the cell migration patterns of immune and cancer cells (15–19). Like all post-translational modifications, if proteolysis is not considered in proteomics analyses, then considerable metadata will be lost, and the functional annotation of proteome components will be misguided or at worst wrong. Hence there is a critical need in proteomics analyses for recognition of the importance of identifying changes in the N and C terminomes, which often result from proteolysis, and
for the development of efficient approaches to identify and functionally annotate these overlooked tail ends of the proteome.

By precise proteolytic processing, proteases participate in every biological process including DNA replication and repair, cell cycle progression, cell proliferation, differentiation and migration, morphogenesis and tissue remodeling, neuronal outgrowth, hemostasis, wound healing, immunity, angiogenesis, and apoptosis (20). Indeed more than 53 specific hereditary diseases of proteolysis are recognized (21), and it is not surprising then that proteases are implicated in many pathologies, none more important than cancer. Hence proteases represent 5–10% of drug targets (20, 22–24) with protease inhibitor drugs being used to treat AIDS by blocking human immunodeficiency virus protease 1, cardiovascular disease by targeting angiotensin converting enzyme and renin, and multiple myeloma using the reversible covalent protease inhibitor bortezomib (25).

Degradomics is a specialized realm within proteomics that is defined as the identification and quantification of all proteases and their substrates on a system-wide scale. With six classes of proteases that account for ~2% of the genes of any organism (21) (see the peptidase database MEROPs for useful information (20)), proteases constitute the second largest enzyme family in man after ubiquitin ligases, totaling more than 567 proteases (Fig. 1) (21). One essential aspect of degradomics is the proteolytic repertoire, or protease degradome, of cells and tissues in homeostasis and diseases such as cancer. Another essential task is to determine the substrate repertoire of each protease, also called its substrate degradome, to determine the function of each protease (8). Deciphering the activity of a protease in vivo is rendered even more complicated as proteases do not act in isolation but rather form interactive cascades, circuits, and pathways, which are interlinked between protease families and classes, creating the “protease web” (Fig. 2) (26). The protease web is an intricate network composed of the proteases, their activators, and their inhibitors. Proteases and inhibitors are represented by nodes and form a network of interactions (black lines) composed of biological pathways (in green), regulatory circuits and feedback loops (in red), amplification cascades (in yellow), and other interactions (in blue). Many proteases are uncharacterized yet and others function alone (in orange).

In this paper, we review recently developed technologies to achieve this lofty goal. Together the techniques highlighted in
As a post-translational modification, proteolysis exerts powerful effects on the function of proteins, and characterizing proteolytic events is an important challenge in proteomics. In general, many of us think about proteolysis as a degradation-to-completion process. This is certainly an important aspect of proteolysis, and examples are shown in Table I. However, as well as degrading proteins, proteases perform highly specific processing that can affect protein structure, function, life span, and localization. By limited and specific cleavages, proteases can act as switches, turning protein activity on or off, or can modulate protein function in more complex ways, regulating vital processes (see examples in Table I). The power of proteolytic activation and inactivation is demonstrated by proteolytic cascades such as blood coagulation (29), complement activation (30), signaling in development (31), and apoptosis (32) where the result of on-off proteolytic switching is a rapid amplification of a signal with many potential points of regulation.

Proteolytic processing has distinct effects on protein function (Table I) that can be thought of as follows. (i) Switching occurs when the unprocessed and processed substrates act upon the same system but with different, often opposite effects. For example matrix metalloproteinase (MMP)\(^1\) and dipeptidyl-peptidase IV cleavage of the N-terminal residues of the monocyte chemoattractant proteins CCL 2, 7, 8, and 13 converts these chemokines from agonists to antagonists of their cognate receptors (16, 33), and MMP-2 cleavage of adrenomedullin, which induces vasodilation, generates a peptide causing vasoconstriction (34). (ii) Domain release from a functional parent molecule can occur. This can be a single cleavage that fragments the parent protein to release domains with biological roles, often affecting the same pathway but lacking the complete functionality of the full-length molecule, or to release a ligand (e.g. heparin-affin regulatory peptide (HARP) and connective tissue growth factor (CTGF) cleavage by MMPs to release VEGF, which can then stimulate angiogenesis (35)). Multiple cleavages of a multidomain protein can also occur to release constituent bioactive domains (e.g. processing of the growth factor proepithelin to release a number of ~6-kDa epithelins that have similar or opposing activities (36, 37)). In contrast, cleavage of viral polyproteins releases domains that gain function compared with the parent molecule (38–40). (iii) Shedding, that is the proteolytic release of the ectodomain of transmembrane proteins (growth factors and chemokines) and receptors or adhesion molecules by cell surface metalloproteinases such as a disintegrin and metalloproteases (ADAMs), MMPs, and some transmembrane serine proteinases (41–43), can occur. Proteolytic shedding of extracellular domains and release of intracellular domains alters signaling, cell adhesion, and protein localization as well as facilitating new compartment-specific functions (44, 45). Shed ectodomains can also function as soluble ligand-binding proteins acting as a ligand reservoir to sequester the ligand from biological activity (46) or conversely protecting their ligands from proteolysis, thus increasing their half-life or promoting agonist activity (47–49). A specialized form of shedding is the hydrolysis of peptide bonds within lipid bilayer.

\(^1\) The abbreviations used are: MMP, matrix metalloproteinase; 2D, two-dimensional; ABP, activity-based probe; ADAM, a disintegrin and metalloprotease; ADAMTS, a disintegrin and metalloprotease with thrombospondin motifs; COFRADIC, combined fractional diagonal chromatography; CTGF, connective tissue growth factor; Hsp90, heat shock protein 90\(\alpha\); ICDC, inactive catalytic domain capture; IL, interleukin; iTRAQ, isobaric tags for relative and absolute quantification; MD, multidimensional; MMPI, MMP inhibitor; MT1, membrane type 1; NHS, N-hydroxysuccinimide; PCPE, procollagen C-proteinase enhancer protein; PICS, proteomic identification of protease cleavage sites; SILAC, stable isotope labeling by amino acids in cell culture; TAILS, terminal amine isotope labeling of substrates; TAP, tandem affinity purification; TIMP, tissue inhibitor of metalloproteinase; TNBS, 2,4,6-trinitrobenzenesulfonic acid; VEGF, vascular endothelial growth factor; HtrA, high temperature requirement protein A; SPARC, secreted protein acidic and rich in cysteine; CUB, complement C1r/C1s, VEGF, Bmp1; CARD, caspase recruitment domain.
ers in a process known as RIP (for regulated intramembrane proteolysis) (50), an important aspect in protein maturation, transport, and clearance of transmembrane proteins. (iv) Gain of function is obviously also associated with protein maturation: initiator methionine and signal peptide removal and zymogen propeptide removal (51–53) as well as prohormone processing and the generation of bioactive peptides, such as neuropeptides (54, 55). (v) Specific proteolysis of proteins can also reveal molecules with functions completely different from the parent molecule, and these are termed cryptic or neoproteins, which might be thought of as "virtual gene" products. Examples include plasminogen (the zymogen of plasmin, a

<table>
<thead>
<tr>
<th>Types of proteolysis</th>
<th>Biological function/consequences</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Proteolytic processing</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initiator methionine removal</td>
<td>Protein maturation and regulation of half-life (proteins with intact initiator methionine are shorter lived)</td>
<td>Methionine aminopeptidases ensure that the essential amino acid methionine is not sequestered in long lived abundant proteins</td>
</tr>
<tr>
<td>Signal peptide removal</td>
<td>Protein localization to endoplasmic reticulum</td>
<td>Signal peptidases remove signal peptide during protein translocation into endoplasmic reticulum</td>
</tr>
<tr>
<td>Localization signal removal</td>
<td>Protein targeting to specific compartments</td>
<td>Removal of nuclear, peroxisome, or mitochondrial localization signal upon protein translocation</td>
</tr>
<tr>
<td>Propeptide removal</td>
<td>Zymogen activation</td>
<td>Activation of proenzymes, e.g. MMPs and kallikreins</td>
</tr>
<tr>
<td>Domain shedding</td>
<td>Loss of receptor response, signaling</td>
<td>Soluble ectodomain of IL-1RII binds and inhibits caspase-1-mediated maturation of pro-IL-1β (195)</td>
</tr>
<tr>
<td>Other precise cleavage(s)</td>
<td>Signal amplification in proteolytic cascades</td>
<td>Serine proteases of blood coagulation (29), complement activation (30), and development (31); caspases in apoptosis (32)</td>
</tr>
<tr>
<td></td>
<td>Inactivation of protein function</td>
<td>Regulation of cell cycle progression by proteolytic inactivation of cyclin-dependent kinase inhibitors as well as cyclins (199)</td>
</tr>
<tr>
<td></td>
<td>Altered receptor binding upon ligand cleavage</td>
<td>Chemokine cleavage can result in agonist to antagonist conversion, binding to a different receptor, or increase/decrease in binding affinity (15, 16, 200)</td>
</tr>
<tr>
<td></td>
<td>Revealing cryptic or neoproteins and bioactive peptides</td>
<td>Generation of angiogenesis inhibitors: angiotatin from plasminogen, endostatin from type XVII collagen, and vasostatin from calreticulin (56, 57)</td>
</tr>
<tr>
<td></td>
<td>Induction of conformational changes</td>
<td>Cleavage of serpin reactive center leads to structural rearrangement and inhibition of serine proteinases (201)</td>
</tr>
<tr>
<td></td>
<td>Control of growth factor and cytokine bioavailability</td>
<td>Cleavage of IGFBPs to release IGFs (202), processing of CTGF and HARP to release VEGF (35)</td>
</tr>
<tr>
<td><strong>Degradation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete degradation</td>
<td>Provides free amino acids (as energy source or structural components)</td>
<td>Digestion of food proteins by pepsin, trypsin, and chymotrypsin (203); starvation-induced catabolism (204)</td>
</tr>
<tr>
<td></td>
<td>Elimination of misfolded, aged, and damaged proteins; termination of the function of a protein</td>
<td>Cellular &quot;quality control&quot; machinery consisting of lysosomal proteases and the proteasome-ubiquitin system (205)</td>
</tr>
<tr>
<td>Incomplete degradation</td>
<td>Immune surveillance</td>
<td>External antigens proteolyzed within the endosomal-lysosomal pathway are loaded onto MHC class II molecules (206); intracellular antigens from the ubiquitin-proteasome degradation pathway are displayed in complex with MHC class I molecules</td>
</tr>
</tbody>
</table>

**TABLE I** Examples of proteolysis and its biological consequences

IGFs, insulin-like growth factors; IGFBPs, insulin-like growth factor-binding proteins; MHC, major histocompatibility complex.
protease that initiates fibrinolysis), type XVIII collagen (a structural extracellular matrix protein), and calreticulin (a calcium-binding quality control molecular chaperone), which are each cleaved to release the angiogenesis inhibitors angiostatin, endostatin, and vasostatin, respectively (56, 57). Milk proteins such as casein, a dietary source of amino acids, release bioactive peptides with numerous functions upon digestion (58), and extracellular matrix protein degradation products are chemotactic for leukocytes (59, 60).

THE PROTEASE WEB

Inherent in this complexity is the fact that proteases do not operate in isolation but in interconnected pathways and amplification cascades where proteolytic information moves in a unidirectional flow or in regulatory feedback loops as depicted in Fig. 2. It is apparent that many pathways and cascades are bridged to form a larger global network, termed the protease web (26). This concept was seeded by recent quantitative mass spectrometry-based proteomics approaches, which revealed that many proteases and inhibitors are substrates for, or are modulated by, specific protease families with considerable cross-talk between protease classes and families (see examples in Fig. 4) (61–63). In addition, protease inhibitors such as secretory leukocyte protease inhibitor, α2-macroglobulin, and tissue inhibitor of metalloproteinase (TIMP)-3 can cross family boundaries.
boundaries to block diverse proteases, spreading connections in the protease web yet further.

In a recent quantitative proteomics study using ICAT labeling and MS/MS, we investigated the effects of treating membrane type 1 MMP (MT1-MMP)-transfected MDA-MB-231 breast cancer cells with prinomastat (AG3340), a metalloproteinase inhibitor used in phase II clinical trials (64). Thirty-seven proteases representing each of the five classes present in man were identified along with 10 protease inhibitors of metallo-, cysteine, and serine proteases. As well as anticipated effects on many MMP substrates, the levels of many of these proteases and inhibitors were modulated by treating these cells with prinomastat (Table II). Thus, ripples in the protease web due to therapeutic inhibition of MMPs suggest that MMPs are key nodal proteases. This may be at least in part due to MMP processing and activation of proteasezymogens and inactivation of a variety of protease inhibitors such as several serpins including secretory leukocyte protease inhibitor and cystatins, thereby indirectly regulating the activity of many proteases in other protease classes and families. MT1-MMP inhibition could affect all of the proteins connected in the interaction network shown in Fig. 4. Indeed levels of many of these were altered by the MMP inhibitor prinomastat (64). For example, blocking MMP cleavage of cystatin C by prinomastat would preserve the inhibition of cathepsin L, consequently preventing cleavage and activation of cathepsin L substrates such as IL-8, collagen XVIII (to form the angiogenesis inhibitor endostatin), and prourokinase type plasminogen activator. Thus, protease inhibition could induce unexpected actions by alterations in the net proteolytic potential of the system and not just through inhibition of the targeted protease. These findings suggest the existence of an architecture

<table>
<thead>
<tr>
<th>Protease class and peptidase EC no.</th>
<th>Protease</th>
<th>MMPI/vehicle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Medium</td>
</tr>
<tr>
<td>Metallo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.4.24.7</td>
<td>Interstitial collagenase; MMP-1</td>
<td>0.69</td>
</tr>
<tr>
<td>3.4.24.81</td>
<td>ADAM-10 endopeptidase</td>
<td>0.64</td>
</tr>
<tr>
<td>3.4.24.-</td>
<td>ADAMTS-7 endopeptidase</td>
<td>0.83</td>
</tr>
<tr>
<td>3.4.17.22</td>
<td>Carboxypeptidase D</td>
<td>0.57</td>
</tr>
<tr>
<td>Cysteine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1.2.15</td>
<td>Ubiquitin thioesterase; ubiquitin-specific protease 14</td>
<td>0.66</td>
</tr>
<tr>
<td>3.4.22.53</td>
<td>Calpain 2</td>
<td>0.59</td>
</tr>
<tr>
<td>3.4.22.34</td>
<td>Legumain</td>
<td>0.57</td>
</tr>
<tr>
<td>3.4.22.1</td>
<td>Cathepsin B</td>
<td>0.63</td>
</tr>
<tr>
<td>3.4.14.1</td>
<td>Cathepsin C</td>
<td>0.74</td>
</tr>
<tr>
<td>3.4.22.15</td>
<td>Cathepsin L</td>
<td>0.61</td>
</tr>
<tr>
<td>Aspartic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.4.23.5</td>
<td>Cathepsin D</td>
<td>0.34</td>
</tr>
<tr>
<td>Serine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.4.16.5</td>
<td>Cathepsin A</td>
<td>0.51</td>
</tr>
<tr>
<td>3.4.21.41</td>
<td>Complement subcomponent C1r</td>
<td>0.57</td>
</tr>
<tr>
<td>3.4.21.-</td>
<td>PCSK-9</td>
<td>0.28</td>
</tr>
<tr>
<td>3.4.21.-</td>
<td>Serine protease 23</td>
<td>0.36</td>
</tr>
<tr>
<td>3.4.21.-</td>
<td>Serine protease 11</td>
<td>0.09</td>
</tr>
<tr>
<td>3.4.21.73</td>
<td>U-plasminogen activator; uPA</td>
<td>0.50</td>
</tr>
<tr>
<td>Threonine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.4.25.1</td>
<td>Proteasome endopeptidase complex; proteasome subunit type 5</td>
<td>0.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inhibitor UniProt accession no.</th>
<th>Inhibitor</th>
<th>MMPI/vehicle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Medium</td>
</tr>
<tr>
<td>P01033</td>
<td>TIMP-1</td>
<td>0.60</td>
</tr>
<tr>
<td>P16035</td>
<td>TIMP-2</td>
<td>0.46</td>
</tr>
<tr>
<td>P02081</td>
<td>Calpastatin</td>
<td>0.74</td>
</tr>
<tr>
<td>P01034</td>
<td>Cystatin C</td>
<td>0.56</td>
</tr>
<tr>
<td>P50453</td>
<td>Serpin B9</td>
<td>0.63</td>
</tr>
<tr>
<td>P119957</td>
<td>Elafin</td>
<td>0.53</td>
</tr>
<tr>
<td>O13278</td>
<td>Kunitz-type protease inhibitor-1</td>
<td>0.31</td>
</tr>
<tr>
<td>P10646</td>
<td>Tissue factor pathway inhibitor</td>
<td>0.33</td>
</tr>
</tbody>
</table>


**TABLE II**

**Perturbation of the protease web by addition of a protease inhibitor drug to cultured cells**

The MMP inhibitor prinomastat or vehicle was added to MDA-MB-231 human breast cancer cells, and changes in protein levels over 48 h were quantified by ICAT labeling and MS/MS. A selection of proteases and inhibitors affected by treatment with prinomastat are shown with the averaged ICAT ratio for peptides from conditioned medium or membrane preparations. Full lists of peptides and ratios can be found in the supplemental information of Ref. 64. ICAT ratios <1.0 indicate reduced levels of peptides in the medium or membrane, whereas values >1.0 indicate accumulation in the presence of the inhibitor. Because proteases and their inhibitors are key regulatory elements of many processes and pathways, treatment will have a profound effect on the protease web, the degradome, and the cellular proteome. uPA, urokinase type plasminogen activator; PCSK-9, proprotein convertase subtilisin/kexin type-9.
where certain proteases act as key nodal elements regulating interconnected pathways that form the protease web.

REGULATION OF PROTEOLYSIS

As might be expected for a fundamentally important and irreversible process such as proteolysis, regulation is imperative. Protease expression and secretion are tightly regulated (65), and the production of proteases as zymogens is a major protective strategy to restrict unwanted proteolysis (66). Specific endogenous protein inhibitors exist for various classes of proteases. For instance the serpins and cystatins for serine and cysteine proteases, respectively (67, 68), and TIMPs for MMPs (69), whereas some inhibitors such as α2-macroglobulin are inhibitory for most endoproteases (70). The balance of protease and inhibitor controls the degree to which proteolysis occurs. The uncontrolled degradation of intracellular proteins by the proteasome would be devastating for the cell, but proteasomal degradation is tightly controlled resulting in distinct half-lives for different proteins (71). Deubiquitinating enzymes constitute the largest family of proteases in the degradome with 104 members (21). They cleave isopeptide bonds to not only recycle ubiquitin and related molecules but, by dynamically regulating protein conjugation status, precisely control the proteasomal degradation of many proteins, including important transcription factors, signaling molecules, and even other deubiquitinating enzymes (72).

Factors determining which proteins are protease substrates and which are not include physical features such as the size and structure of the protease active site; each protease has a “preference” for particular residues at and adjacent to the scissile bond that is determined by interacting subsites and residues within the enzyme active site (73–75) and the presence of exosites, regions or motifs distinct from the active site that enhance substrate binding and hydrolysis by making substrates more accessible to the active site (76). There are numerous other strategies for regulating proteolysis including substrate-induced conformational change in the active site, masking the protease active site, and regulation of substrate entry by regulatory subunits and ubiquitination such as for the proteasome (71). This variety of control mechanisms presents challenges for protease quantification by activity-based probes (see below). Other factors regulating proteolysis include masking the substrate cleavage site with interacting proteins, sequestering the cleavage site within the protein itself pending conformational change, and compartmentalization of proteases and their substrates. Many of these cannot be replicated in vitro assays thus necessitating analysis of proteolysis in the cellular context or in tissues. Hence protease substrates identified in biochemical cleavage assays should only be considered as candidate substrates in vivo, requiring validation to address the statement “just because it can does not mean it does.” However, choosing a system that encompasses biological relevance also increases complexity, requiring new approaches for their analysis.

DEGRADOMICS: THE FUNCTIONAL ANNOTATION OF THE PROTEOME

The functional annotation of each proteome requires degradomics approaches to identify the proteases and their activity state to map the key nodal proteases that interconnect in the protease web as well as to identify the substrate repertoire of individual proteases. Subsequently by identifying interactors of the proteases and inhibitors the interconnections at the interface of the protease web with the proteome can be mapped. Together this system-wide analysis can reveal the level at which each protease acts in biological pathways and so define their activity on signaling pathways, other bioactive mediators, structural proteins, and hence their effects on cell behavior.

The first step to shed light on the protease web is to identify and quantify the proteases in biological samples from a defined location and under specific conditions and to determine the level of activity of those proteases. Recently two DNA microarray technologies, the Hu/Mu ProtIn chip and the CLIP-CHIP®, were developed (28, 77). These platforms facilitate the relative quantification of mRNA of proteases, protease inhibitors, and protease interactors extracted from biological samples. However, transcript levels only indicate which proteases can potentially be found in a sample because mRNA levels do not always correlate with protein abundance and are devoid of information regarding protease activity. Hence identification and quantification of protease and protease activity at the protein level are needed. Proteomics techniques that use MS/MS coupled to fractionation of the proteome to increase the coverage of proteins and hence proteases can accomplish this to some extent. However, these general techniques are not designed to specifically and comprehensively target proteases and their activity state. It is not in the scope of this review to cover the basis of proteomics, and we invite readers to read review articles for more information (78–80). Because of particular properties of proteases, specialized and dedicated techniques are required to profile the protease degradome.

ACTIVITY-BASED PROBES

Identification and quantification of active proteases found in biological samples can be achieved by combining the use of activity-based probes (ABPs) and mass spectrometry analysis. ABPs are molecules derived from mechanism-based inhibitors, and they target a specific protease class based on their enzymatic mechanism. ABPs irreversibly bind to active proteases but not to inactive zymogens (81, 82) or an inhibited enzyme (83). ABPs are composed of a chemically reactive group, often termed a “warhead,” attached to a tag moiety via a spacer molecule (84). Selectivity of the probe can be increased by the addition of a specificity-enhancing module. Upon binding of the ABP to the protease active site, the chemically reactive group reacts irreversibly with the active site nucleophilic residue. The tag of ABPs allows for either
visualize (fluorophore or radioactive probe) or isolation (affinity tag) of the ABP-protease complex.

ABPs targeting serine proteases (81, 82), cysteine proteases (85–88), threonine proteases (89, 90), aspartyl proteases (91), and metalloproteases (92, 93) have been developed. Metallo- and aspartic proteases pose difficulty in the design and application of ABPs due to the absence of a covalent acyl intermediate in the catalytic mechanism. The nucleophile of metallo- and aspartic proteases is an activated water molecule instead of an amino acid. Hence ABPs for these two protease classes do not directly react and attach to a catalytic residue, limiting the use of these noncovalent inhibitors to non-denaturing analytical techniques. To address this, UV light-activated cross-linking ABPs targeting metallo- and aspartic proteases have been developed (91–93) but suffer from low yield and hence low sensitivity.

More recently, quenched near-infrared fluorescent ABPs have been developed for specific labeling and visualization of the papain family of lysosomal cysteine cathepsins in live mice (94). After in vivo labeling of the cathepsins in human breast cancer cell lines implanted in mice, proteins from the solid tumor were extracted and separated by SDS-PAGE, and the labeled cathepsins were visualized using the fluorescent ABP. In-gel digestion and MS/MS identification of the labeled proteases is possible, but interestingly it is surprisingly difficult to identify the ABP-conjugated active site peptide by MS/MS. Isotope-coded ABPs have yet to be produced, but these might allow for protease mass spectrometric quantification.

The production of ABPs necessitates extensive organic chemistry synthesis efforts, and some probes also suffer from stability problems and poor bioavailability. Limitations in accurate quantification may be presented by mechanisms involving active site masking and conformational changes as discussed under “Regulation of Proteolysis.” Nevertheless the use of ABPs is currently the only method allowing the identification and quantification of the active, non-inhibited form of a protease in a biological sample.

PROFILING PROTEASE SUBSTRATE SPECIFICITY

The other side of the degradomics coin is the substrate degradome. The substrate degradome of a protease is defined by its biologically relevant substrates in vivo. Indeed without a substrate, we can rarely understand protease function. Limited knowledge of the full repertoire of substrates of a protease can be profoundly detrimental to an antiproteolytic drug regime because unpredicted and undesirable effects (side effects) can result. For example, causal relationships between the overexpression of MMPs and tumor progression prompted the development of MMP inhibitors (MMPIs) as anticancer therapeutics more than 20 years ago (95). However, these broad spectrum MMPIs were unsuccessful in clinical trials (65, 96–98) when the extent of the MMP substrate degradome was greatly underestimated (26). Many beneficial activities of proteases in disease have led these proteases to be classified as drug antitargets that must be therapeutically spared during treatment.

To assist in the prediction of potential protease substrates, high throughput, peptidomic techniques were developed to characterize protease active site specificity and hence identify preferred cleavage sites. Methods based on screening of phage-displayed peptides (99) and synthetic peptide libraries (100–105) have been used to characterize protease specificity, but degenerate synthetic peptide libraries cannot be analyzed by mass spectrometry. In addition, phage display requires hundreds of phage DNA sequencing reactions and only identifies cleavable sequences and not the precise scissile bond. Methods based on peptide libraries also involve massive peptide synthesis and sequencing efforts and typically only provide information on the N-terminal side of the cleavage site. Thus, methods that provide more information on the precise cleavage site and its surrounding sequence in a high throughput manner are required.

PROTEOMIC IDENTIFICATION OF PROTEASE CLEAVAGE SITES (PICS)

Our laboratory recently developed a method termed PICS to harness the power of mass spectrometry for protease active site mapping (75). In this method, biologically relevant peptide libraries are generated from a complex proteome by endoproteolysis with trypsin, chymotrypsin, or Glu-C, thus eliminating the need for solid-state peptide synthesis. These natural peptide libraries are digested with the protease of interest, generating peptides displaying new N termini (neo-N termini) not found in the original libraries. Peptides containing a neo-N terminus are affinity-purified and identified by mass spectrometry. Efficient discrimination between the primary amine group of the neo-N termini and the N termini of the library (and the amine groups of lysine side chains) is achieved by reductive dimethylation of all primary amine groups during initial preparation of the peptide libraries. When these amino-blocked libraries are exposed to the protease of interest, peptide fragments with neo-N termini free amine groups are generated. These are then incubated with a cleavable biotin tag using N-hydroxy-succinimide ester as an amine-reactive group and are isolated by affinity chromatography. The isolated neo-N termini peptide fragments are released from the biotin tag, and because these peptides originate from the proteome of a sequenced organism they can be analyzed by LC-MS/MS, validated by the Trans Proteomic Pipeline (106), and then identified by database searching. The identified peptides constitute the prime side C-terminal fragments of the peptides that were cleaved by the protease. A search of this peptide in the database allows the identification of the parent protein and hence the N-terminal nonprime side of the peptide. Hence PICS uniquely permits the active site mapping of both the C- and N-terminal (prime and nonprime side, respectively) preference of the protease in a single experiment.
Using a library generated from the human fibrosarcoma cell line HT1080, Schilling and Overall (75) mapped the active site specificity of thrombin, neutrophil elastase, cathepsin G, human immunodeficiency virus protease 1, cathepsin K, caspases 3 and 7, Glu-C, and MMP-2. In total, 3691 cleavage sites were identified. To put this into perspective, MEROPS, the peptidase database, lists less than 8000 cleavage sites for more than 2400 proteases.

Protease active site mapping technologies indicate the peptide sequence most likely to be cleaved by a protease and can help in the identification of protease substrates. However, simple bioinformatics searches of protein databases for cleavage sites as a guide to new substrates is beset with problems that result in too many false positives to be practically useful. Predicted cleavage sites may not be cleaved in vivo as other factors such as protein structure and exosite binding strongly influence protease specificity. Many proteases also cleave at kinetically nonpreferred sites. So techniques for protease substrate identification that use full-length, folded native proteins preferably from a biological system where interactors and native conditions prevail are required.

**QUANTITATIVE PROTEOMICS TO IDENTIFY PROTEIN SUBSTRATES**

Proteomics techniques such as those utilizing multidimensional (MD) LC or two-dimensional gel electrophoresis and mass spectrometry are highly sensitive and unbiased, and their high content outputs have revolutionized protease substrate discovery (107). Proteomics has the potential for identification of all proteins present in a sample, even those of low abundance, without preconceptions and thus is capable of identifying protease substrates directly from cells, tissues, and body fluids. By incorporation of isotopic labels into proteins or tryptic peptides, identical peptides from multiple samples can be identified and quantified by mass spectrometry and by detecting changes in protease processing, so identify substrates. For a highly sensitive degradomics screen, samples are best derived from systems (cultured cells or tissues from transgenic animals) differing only in activity of the protease of interest so that changes in substrate processing are emphasized. This can be protease-expressing versus null (protease-transfected versus vector-transfected cells) (61), control versus protease knock-out cells (62), protease versus small interfering RNA-treated cells or samples from wild-type versus protease knock-out mice (108), cells expressing active versus inactive protease mutants (62), or inhibitor-treated versus vehicle-treated cells or animals (64).

**SUBSTRATE IDENTIFICATION FROM QUANTITATIVE PROTEOMICS ANALYSES**

After peptide identification and analysis for false positives using the Trans Proteomic Pipeline (106), quantitative data are obtained as ratios of sample 1/sample 2. To interpret these ratios, hypotheses must be generated regarding expected effects on levels of protease substrates and binding partners (Fig. 5). For example in the case of increased protease activity (protease/null) one would expect increased clearance of soluble proteins and increased shedding of cell surface protein ectodomains and pericellular proteins. Thus, in the secretome harvested from the conditioned medium, the ratio for soluble substrates would decrease (<1), and shed proteins would increase (>1). Accordingly levels of shed proteins would decrease in membrane preparations (<1). Conversely where protease activity is reduced by RNA interference or inhibitory drugs (reduced protease/protease), uncleaved soluble substrates would accumulate (>1), whereas shed proteases would decrease (<1) in the secretome but increase in membrane preparations (>1). Indeed these hypotheses have held up remarkably well (35, 61–64), so setting the stage for routine use of these approaches by many laboratories to investigate proteolysis in the cellular context.

It should be noted that altered ratios may result from indirect effects: levels of proteins that are bound to the substrate, such as ligands of cleaved and shed receptors or proteins bound to cell surface proteoglycans, would be modulated as described for the chemokine KC, which binds to syndecan-1 (109), and peptidyl-prolyl cis-trans isomerase B (cyclophilin B), which binds heparan sulfate proteoglycans (110, 111). These ratios would increase, although these proteins are not protease substrates, but this provides valuable interactomics data. Inhibited proteases might also act as dominant negative substrate traps, titrating substrates and interacting proteins. The effects on cascades in the protease web also need to be considered: for example, activation of a second protease by
MT1-MMP, such as MMP-13 (112) or MPP-2 (113, 114), which then cleaves substrate proteins detectable by altered ratios. Lastly ratios may be altered as a result of changes in protein expression that occur as a result of proteolytic cleavages in the cellular signaling environment.

ICAT® LABELING

ICAT labeling (115) is a technique for the differential labeling of proteins in two samples that, followed by MD LC-MS/MS, allows relative quantification and identification of the proteins therein. In ICAT, two samples are reduced and labeled by reductive alkylation of cysteines using chemically identical biotin-tagged reagents that differ in isotopic composition and hence mass (¹³C and ¹²C). The labeled protein samples are then combined and thereafter treated identically, enabling quantitative comparisons to be made. Following trypsin digestion, biotin-tagged cysteine-containing peptides are enriched by avidin column chromatography and subjected to MD LC-MS/MS. Spectral peak analysis in single MS mode of the isotopically resolved peptides (9-Da mass difference) from the two sources enables relative quantification of protein levels, and the differentially expressed proteins are then identified by MS/MS sequencing and database searching. The use of ICAT reagents allows for a significant reduction in sample complexity by biotin pullout enrichment of the cysteine-labeled peptides. This improves statistical chances for the identification of lower abundance proteins. Also the analysis of cysteine-rich potential substrates (e.g. many extracellular matrix proteins and cytokines) may benefit from enrichment for cysteine-containing peptides by ICAT. However, the main limitation of ICAT is this dependence on cysteine labeling because cysteines are scarce in proteins: 7% of all proteins have no cysteine and so will not be labeled, and proteins containing only one cysteine residue (35% of all proteins) will have limited coverage.

Despite these issues, we have used ICAT very successfully for protease substrate discovery in the cellular context (35, 62–65). Substrates of MT1-MMP were identified that were shed from the cell membrane or pericellular environment to the conditioned medium of cultured human breast cancer cells transfected with MT1-MMP compared with vector or an inactive MT1-MMP mutant (61, 64). Expression of MT1-MMP compared with vector resulted in increased MT1-MMP/vector ICAT ratios for many proteins in conditioned medium, indicating MT1-MMP-mediated shedding (see Fig. 5 and Table III). Interleukin-8, secretory leukocyte protease inhibitor, death receptor-6, and CTGF were biochemically confirmed as novel substrates of MT1-MMP (61).

We also utilized an MMPI drug to block MT1-MMP shedding activity in the breast cancer cells in comparison with the drug vehicle (63, 64). For many proteins, the MMPI/vehicle ICAT ratio for conditioned medium was reversed (low) compared with the MT1-MMP/vector ratio, indicating that MT1-MMP-mediated shedding of these proteins was blocked by the inhibitor (Table III). In the same study, we also used ICAT to label proteins in membrane preparations. The reduction in shedding to the medium in the presence of the MMPI was often complemented by the accumulation of the protein in the plasma membrane as indicated by high MMPI/vehicle ICAT ratios (Table III). Carrying out both overexpression and inhibition studies in the same cellular system provides a strong indication for biologically relevant substrates and led to the identification of many undescribed MT1-MMP substrates, 20 of which we biochemically validated in vivo experiments, including DJ-1, galecin-1, Hsp90α, pentraxin 3, progranulin, Cry61, peptidyl-prolyl cis-trans isomerase A, and dikkopf-1. A different cell system that is expected to result in a higher signal to noise ratio was reported using Mmp2−/− cells compared with Mmp2−/− cells transfected with Mmp-2 (62).

**TABLE III**

<table>
<thead>
<tr>
<th>Protein</th>
<th>MT1-MMP/vector, medium</th>
<th>MMPI/vehicle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fibromodulin</td>
<td>4.22</td>
<td>0.58</td>
</tr>
<tr>
<td>EGF-containing fibulin-like extracellular matrix protein 1</td>
<td>3.90</td>
<td>0.25</td>
</tr>
<tr>
<td>Quiescin Q6</td>
<td>2.16</td>
<td>0.72</td>
</tr>
<tr>
<td>Elafin</td>
<td>1.85</td>
<td>0.53</td>
</tr>
<tr>
<td>IGFBP-7</td>
<td>1.54</td>
<td>0.28</td>
</tr>
<tr>
<td>Niemann-Pick disease, type C2 variant</td>
<td>1.42</td>
<td>0.36</td>
</tr>
<tr>
<td>MMP-1</td>
<td>1.85</td>
<td>0.71</td>
</tr>
<tr>
<td>CD59</td>
<td>1.67</td>
<td>0.41</td>
</tr>
<tr>
<td>uPAR</td>
<td>—</td>
<td>0.58</td>
</tr>
<tr>
<td>Axl receptor tyrosine kinase</td>
<td>—</td>
<td>0.24</td>
</tr>
<tr>
<td>RCC2 protein</td>
<td>—</td>
<td>0.57</td>
</tr>
<tr>
<td>Multifunctional protein ADE2</td>
<td>—</td>
<td>0.71</td>
</tr>
</tbody>
</table>

MDA-MB-231 human cells were transfected with MT1-MMP or vector, and changes in protein levels were quantified by ICAT labeling and MS/MS. Shedding was increased by MT1-MMP expression, and the ICAT ratios (MT1-MMP/vector) show increases of peptides in the medium. Administration of the inhibitor prinomastat (compared with vehicle) to the MT1-MMP-expressing cells decreased shedding, reversing the ICAT ratios in the medium, as well as increasing previously shed proteins in the cell membrane. A dash (—) indicates that no peptides were detected. This dual analysis was found to be a strong predictor of MT1-MMP substrates (64). EGF, epidermal growth factor; uPA, urokinase type plasminogen activator.
the protease “naïve” proteome was compared with the same proteome exposed to the active enzyme. Shed substrates of MMP-2 were identified in conditioned media by detecting secretome differences between Mmp2−/− murine fibroblasts and those rescued by MMP-2 transfection (35). As well as identifying many known MMP-2 substrates, a large variety of novel substrates were identified and validated, including CTGF, HARP (pleiotrophin), insulin-like growth factor-binding protein 6, follistatin-like 1, and cystatin C. The function of MMP-2 as related to these substrates was dissected. Precise processing of HARP and CTGF inactivated these growth factors and mobilized proangiogenic VEGF from stable inhibitory HARP-VEGF and CTGF-VEGF complexes, thus facilitating angiogenesis. Also cleavage of HARP itself generated fragments with distinct functions. The N-terminal domain increased cell proliferation, whereas the HARP C-terminal domain was antagonistic and decreased cell proliferation and migration. These examples illustrate the need to analyze the proteolytic state of a proteome so that biological interpretations of the proteomics results are accurate and rational hypotheses are formulated. We have used ICAT to discover MMP substrates, but these degradomics screens can easily be adapted for dissection of the proteolytic function of other proteases in complex and dynamic biological contexts.

**ISOBARIC TAGS FOR RELATIVE AND ABSOLUTE QUANTITATION (ITRAQ™)**

Rather than exclusively labeling cysteine residues in paired samples, iTRAQ labels the N terminus (and lysine side chains) of tryptic peptides following reduction and alkylation (116). Samples are then combined and subjected to MD LC-MS/MS. The iTRAQ tag consists of an amine-specific N-hydroxysuccinimide (NHS) group linked to the isobaric tag that consists of a reporter group and a balance group. Currently the simultaneous quantification of peptides from up to eight different samples is possible with reporters of 113.1, 114.1, 115.1, 116.1, 117.1, 118.1, 119.1, and 121.1 Da. The mass of the tags is maintained at 305 Da by varying the balance portion to ensure identical behavior of tagged peptides during chromatography and in MS1 mode (117). A labeled peptide present in two or more samples appears as a single precursor ion by MS due to the consistent mass of the iTRAQ tag, thus simplifying spectra and analysis compared with ICAT and stable isotope labeling by gmnino acids in cell culture (SILAC) (see below). Quantification is achieved following MS/MS when the reporter ion fragments with the peptide, thereby allowing relative quantification and identification of the peptide from each sample. Importantly the identification and quantification occur at the same point in the mass spectrometer cycle, leading to higher confidence peptide identification and quantification. The isotopic reporter groups are detected as low mass fragment ions in a “quiet range” of the spectrum, and their intensities are used for relative quantification, whereas the rest of the fragment ions derived from the peptide are used for peptide identification.

Like ICAT, iTRAQ has been used effectively to identify novel substrates of MMP-2 in the secretome from cultured Mmp2−/− murine fibroblasts transfected to express low levels of active MMP-2 compared with a catalytically inactive MMP-2 mutant at different time points (62). As before, identification of known substrates of MMP-2 validated this technique, and many novel MMP-2 substrates were discovered including the CX3CL1 chemokine fractalkine, osteopontin, galectin-1, and Hsp90α. The latter illustrates the point that quantitative proteomics is an unbiased technique and as such can identify substrates that otherwise would probably not even be contemplated. This is the case for Hsp90α, which despite being annotated as an intracellular molecular chaperone has been recently shown to function also extracellularly as a regulator of wound healing (118, 119) and as such may be regulated by proteolytic processing.

The utility of ICAT- versus iTRAQ-based strategies for substrate discovery was tested within the same cellular context by Dean and Overall (62). iTRAQ enabled identification of 9-fold more proteins in total, 8-fold more known substrates, 4-fold more protease inhibitors, and 31-fold more proteases compared with ICAT, consistent with the cysteine chemical bias of ICAT (62). The increased number of peptides identified per protein by iTRAQ labeling of the N termini and lysines allows “peptide mapping”. Mapping of multiple peptides within the protein sequence in combination with their relative abundance ratios can reveal distinct partitioning of iTRAQ ratios, the location of the cleavage site can be inferred, and domains of substrates that are cleaved and released can be predicted (Fig. 6) (62). For example, in the MMP-2 degradomics screen, there were four iTRAQ-labeled tryptic peptides from the C-terminal region of procollagen C-proteinase enhancer protein (PCPE) with high iTRAQ ratios (−4.0 to >30, the latter representing a peptide singleton) in the conditioned medium compared with five peptides near the N terminus that showed no change (mean ratio of 1.2), suggesting a cleavage site between the two sets of mapped peptides (62). As confirmed by biochemical analysis, the portion of PCPE with the high iTRAQ ratios represented the C-terminal fragment of PCPE that was solubilized by MMP-2 proteolytic activity, whereas the peptides that did not change were from the N-terminal complement C1r/C1s, VEGF, Bmp1 (CUB-1) domain that remained bound pericellularly to procollagen. Similarly all three peptides identified in the N-terminal chemokine domain of CX3CL1 (fractalkine), which was found to be shed by MMP-2, showed high iTRAQ ratios in conditioned medium that increased over time, whereas a peptide from the mucin-rich stalk, which is retained on the cell membrane, had a low ratio of 1.6 (Fig. 6). Peptide mapping allowed the identification of six other proteins, including perlecan, calgizzarin, dystroglycan, and hepatoma-derived growth factor, where the proteolytic release of the N, C, or both domains by MMP-2 was predicted (62).
Fig. 6. Peptide mapping. Where distinct partitioning of peptide iTRAQ ratios is observed in a protein when multiple tryptic peptides are mapped, the location of the cleavage site can be inferred, and domains of substrates that are cleaved and released can be predicted. These can be confirmed by mass spectrometry, Western blotting, and N-terminal sequencing. Peptide mapping is illustrated by ratios obtained for CX3CL1 (fractalkine) in conditioned medium from Mmp2−/− cells transfected with wild-type MMP-2 compared with an inactive MMP-2 mutant. The figure was adapted from Dean et al. (35) using an image from FirstGlance in Jmol. The N-terminal chemokine domain containing peptides 1–3 (which have high iTRAQ ratios; white arrows) is indeed shed by MMP-2 cleavage (green arrows), whereas the stalk (gray) containing peptide 4 (which has a lower iTRAQ ratio; yellow arrow) is retained on the plasma membrane.

Utilizing ICAT or iTRAQ is well suited to extracellular and cell membrane protease substrate discovery because fractionation of the cleaved protein to the culture medium or the membrane is inherent to protein shedding. This can be exploited in sample preparation to increase the probability of robustly satisfying the hypotheses described earlier that will identify substrates from the altered peptide ratios. These techniques could also be used for analysis of intracellular proteolysis if there is some partitioning that allows discrimination between cleaved and uncleaved substrates, such as degradation to completion of the cleaved proteins or intracellular membrane shedding, provided that organelle preparative techniques are available. However, the scissile bond is rarely identified in ICAT and iTRAQ approaches as the semitryptic peptides generated after protease cleavage are not specifically enriched. Another limitation is where proteolytic processing leads to only one peptide being charged with no effect on clearance: for example limited processing near the N or C termini of a protein might not alter the protein abundance and only changes one peptide. Hence the ratios of all tryptic peptides other than the terminal peptides are predicted not to be altered significantly. The commercial availability of ICAT and iTRAQ labeling kits means that almost any mass spectrometry facility can perform these analyses, so facilitating the widespread adoption of these approaches by many laboratories. These approaches are therefore a good entry point for many groups to commence substrate degradomics.

SILAC

Highly abundant proteins, such as serum albumin, are detrimental to ICAT and iTRAQ labeling protocols because they are also labeled, thus leaving considerably less free label for low abundance proteins, and necessitate serum-free cell culture. Although it is desirable to reduce serum proteins as much as possible to reduce ion suppression this limitation can be overcome by SILAC, which accommodates exogenously added factors such as serum but otherwise follows the same principles as described above (120). In SILAC, test and control cell cultures are grown in media containing isotopically labeled amino acids (e.g. 12C- or 13C-labeled leucine, arginine, or lysine) but otherwise deficient in that amino acid so that a heavy or light label is incorporated into all the proteins synthesized by the cells (120–123). After harvesting, samples containing equivalent amounts of protein are combined and can be analyzed and quantified in MS1 mode. Peptides containing isotopically labeled amino acids can be quantified. The main advantage of SILAC is that it allows combination of the labeled samples at an early stage of the work flow, thus reducing experimental error in the ratio estimations, but its use is limited to samples that can be metabolically labeled. Also in SILAC, 100% incorporation of the labeled amino acid is observed, whereas chemical labeling such as ICAT and iTRAQ never proceeds to completion and leaves about 5% of the targeted residues unlabeled. SILAC has been applied to identify shed substrates of the snake venom metalloprotease atrolysin A (124). The protease or serum-free medium was added exogenously to cultured fibroblasts, and proteins were separated by one-dimensional SDS-PAGE and subjected to in-gel trypsin digests and MS/MS. Enrichment of proteolytically processed N-terminal peptides (described below under “Positional Proteomics”) and MD LC fractionation can simplify the proteome and improve coverage as we showed for the identification of MMP-2 substrates from the conditioned medium of human breast cancer cells transfected with MMP-2 versus vector or inactive mutant and Mmp2−/− fibroblasts exposed or not to exogenous MMP-2 (125).

TWO-DIMENSIONAL (2D) DIFFERENCE GEL ELECTROPHORESIS (DIGE)

2D gel electrophoresis utilizes widely available technology to compare proteins in two samples on separate 2D gels. Spot intensities are compared using gel image analysis software, and then protein spots of altered intensity are excised and subjected to in-gel trypsin digestion and mass spectrometry. This technique was used for the identification of novel substrates of MT1-MMP by addition of the MT1-MMP catalytic domain to plasma (126). However, reproducibility is an issue when comparing 2D gels, and they have a reputation for poor resolution of proteins with extremes of pl or mass and highly hydrophobic (membrane) proteins. Optimization, such as altering detergent compositions to include lipid (127) and performing a third dimension on excised gel sections where
resonance was poor, is reported to mitigate many of these problems, and good results can be achieved. Quantification can be introduced by labeling each sample with a fluorescent dye (such as Cy3 and Cy5) in DIGE and co-electrophoresing them on a single gel, which greatly improves reproducibility. 2D DIGE has been used to identify substrates of the intracellular serine protease granzyme B where the enzyme, inhibited enzyme, or vehicle were added to cell lysates (128). By enriching for glycoproteins among conditioned medium proteins from cells overexpressing ADAM-17 or from cells treated with a metalloproteinase inhibitor, substrates shed by ADAM-17 were revealed (129). Also in vivo substrates of the gelatinases MMP-2 and MMP-9 were identified in bronchoalveolar lavages fluid from Mmp2−/−/Mmp9−/− mice versus wild-type mice (108), and calpain 3 substrates have been identified using DIGE in samples from calpain 3-overexpressing transgenic mice versus wild-type counterparts (130).

Comparative studies suggest that a combination of gel-free and gel-based proteomics approaches may be the best option to maximize proteome coverage and thus to optimize the discovery of novel protease substrates (131, 132). Other techniques, such as analysis of all gel slices of an electrophoresed proteome by conventional one-dimensional SDS-PAGE, performed for various applications including analysis of ubiquitination and proteolysis also offer an easy entry point to degradomics, albeit at the cost of missing large portions of the proteome and small trimming of the protein termini and at the cost of less reliable label-free quantitation (133).

**SELECTION AND VALIDATION OF CANDIDATE SUBSTRATES**

In non-biased quantitative proteomics approaches, criteria must be set to differentiate significant ratio changes induced by protease activity from experimental variations and systematic errors. Initially thresholds tended to be set in a rather arbitrary manner, such as a 1.5-fold increase and 0.5-fold decrease for ICAT (61). For proteases where a significant number of substrates are known, “cutoffs” can be based upon ratios for known substrates (62, 64). For instance, in home on MT1-MMP substrates, the ICAT ratio “cutoff” was based around ICAT ratios measured for known substrates identified in a breast cancer cell system. Changes in the ratios for known substrates were much less extreme than one might set using arbitrary cutoffs, but they reflect the levels of change that occurred in a complex system (64). Thus setting higher cutoffs would have missed a number of substrates: there is a balance between finding all the true substrates and false discovery of substrates.

For proteases where there are no known substrates, it might be possible to identify potential substrates based on redundancy. For instance many proteins are substrates of more than one protease in a family due to homology. Likewise molecules from families where other members are processed by the protease (or family members of the protease) are strong candidate substrates. Statistical tests are becoming the norm to predict the false negative and false discovery rates of peptides and proteins and to increase confidence in the isotope ratios (p values are quoted by the Protein Pilot analysis software; q values were used in a study by Mayor et al. (134) and the Trans Proteomic Pipeline (106)). Ratio cutoff values must be statistically chosen to reflect true changes and bona fide substrates. Also, there must be validation that 1) the ratios are correct, typically done by comparative Western blotting, and 2) the proteins that change are true substrates, generally carried out by in vitro assays initially followed by in vivo studies. Validation can also be achieved by combining approaches that have opposite effects, i.e. where ratios for substrates would be reversed: for example expression of MT1-MMP versus vector increased substrate shedding (61), whereas shedding of these substrates was decreased in MT1-MMP-transfected cells incubated with inhibitor versus vehicle, and thus ICAT ratios for substrates were reversed (63, 64).

Of course, ratios may be correct, but the protein might still not be a substrate because of downstream effects as discussed above under “The Protease Web.” Although not useful for substrate discovery, these show the multitude of changes that can be wrought on a cell by a protease and have great potential for dissecting the proteome web. Because 5–10% of drug targets are proteases, pharmacoproteomics screens can be applied for preclinical assessment and validation of protease inhibitor drugs. We showed that ICAT interfaced with MD LC-MS/MS is a generally applicable, fast, and effective technique to evaluate preclinical drug effects and toxicity as well as to identify novel substrates and potential new drug targets (63, 64). The quantitative approaches described here can confirm desired effects and reveal adverse consequences from interactions with off-target molecules including previously unrecognized countetargets (23).

These quantitative methods allow identification of substrates from a complex cellular system but not the cleavage site. Occasionally the cleavage site can be localized by peptide mapping (62), but for the most part, cleavage site information is not readily available. Validation of cleavage and determination of cleavage sites must be carried out on each novel substrate if so desired. Therefore other techniques, which can identify proteolytic cleavage fragments easily, particularly from complex biological systems, are required.

**POSITIONAL PROTEOMICS**

Proteome samples always contain proteolytically modified proteins. Peptide solutions for mass spectrometry often present semitryptic peptides (those not presenting the tryptic cleavage specificity at both ends because of previous proteolytic processing or representing the N- and C- terminal peptides of the protein) that are frequently ignored in conventional MS/MS analyses but confer significant information on the functional state of the proteome. Therefore, proteomics methods need to be developed to identify cleaved proteins from
the N terminome, both from the standpoint of functional annotation of the proteome and to understand the underlying biology of protease activity.

High complexity of an average biological proteome in general and its wide dynamic range result in undersampling of lower abundance proteins preventing detection of many protein domains or peptides that have undergone proteolytic processing. Therefore, identifying potential substrates in a complex proteome still heavily resembles looking for a proverbial needle in a haystack. As discussed above, ICAT reduces sample complexity due to selection of cysteine-containing peptides only. Enrichment of proteome samples for the N-terminal portion(s) of each protein also results in sample complexity reduction and yet offers a far more relevant context for protease substrate discovery because the very act of proteolysis yields a neo-N terminus representing the prime side of the cleavage site. Thus, N-terminal peptide analysis allows for identification of both candidate substrates and their exact cleavage sites (27). In the sections below we give an overview of recent advances in the contemporary area of protease substrate discovery using N-terminal enrichment approaches (as summarized in Table IV).

Combined Fractional Diagonal Chromatography (COFRADIC)

The first study to use N-terminal enrichment for the identification of protein proteolytic processing was based on COFRADIC (135). In this strategy, acetic anhydride is used to acetylate primary amino groups of proteins (N termini and lysine side chains). The proteome is then digested with trypsin and separated by reversed phase HPLC. The resulting 12 fractions are then chemically modified by 2,4,6-trinitrobenzenesulfonic acid (TNBS), which reacts with the N termini of internal tryptic peptides to form very hydrophobic trinitrophenyl derivatives. TNBS labeling is then followed by a second reversed phase fractionation of each of the 12 primary fractions. The trinitrophenyl-modified internal peptides now elute in later fractions because of their higher hydrophobicity and are discarded, whereas acetylated peptides representing the original N-terminal portion of each protein are eluted at the same point in the elution gradient as before, collected, and then subjected to MS/MS analysis.

A modified COFRADIC technique incorporating stable isotope labeling by digesting the samples with trypsin in the presence of $^{16}$O-water or $^{18}$O-water allows for quantification (136). Thus, trypsin-catalyzed incorporation of two $^{16}$O atoms at the C terminus of the newly cleaved peptide results in a 4-Da mass difference compared with the same peptide in the presence of light $^{16}$O-water (137). Using anti-Fas antibody to induce apoptosis in human Jurkat T lymphocytes, 93 cleavage sites from 71 proteins (of 1834 proteins in total) were detected mostly at caspase consensus sites (136). The few cleavages showing other than aspartate P1 specificities represent either unlikely non-canonical caspase cleavages, additional protease classes activated during apoptosis, or false positive peptide identifications. More recently, quantitative COFRADIC was used to identify the substrates of an apoptosis-activated mitochondrial serine protease high temperature requirement protein A2 (HtrA2/Omi) in a cell-free system. The analysis yielded 1162 total protein identifications (from 1964 peptides) and determined 50 cleavage sites in 15 proteins (138).

Hence COFRADIC is a powerful approach that allows for significant sample simplification and 70–80% N-terminal peptide enrichment and therefore enables identification and quantification of proteolytic processing in a biological sample. However, the experimental design does not allow the analysis of in vivo modified (i.e. acetylated) and unblocked protein N termini in a single experiment. This limits its use for characterization of N-terminal post-translational modifications in eukaryotes (where up to 80% of proteins are estimated to be acetylated (6)). With two chemical labeling steps, 13 HPLC fractionation runs, and 96 LC-MS/MS analyses per experiment, this technique requires significant instrumentation time and labor and might suffer from protein losses because of multiple handling steps.

Negative Selection of N Termini

A different N-terminal enrichment approach that is better suited for annotation of N-terminal post-translational processing in eukaryotic samples also utilizes protein acetylation as the first step followed by tryptic digestion (139). The newly formed internal tryptic peptides are coupled to an NHS ester derivative of biotin at their tryptic N termini and removed using immobilized streptavidin beads. Following this negative selection, both naturally blocked and chemically acetylated protein N-terminal peptides remain and are analyzed to yield information on the basal proteolysis in the sample. This technique was tested on the soluble protein fraction of mouse skeletal muscle and a more complex mixture of soluble proteins from mouse liver yielding information on N-terminal processing, such as removal of the initiator methionine, the signal peptide, or the propeptide (139). Qualitative comparison of MALDI-TOF spectra of an unfractionated peptide mixture with or without the N-terminal enrichment step indicated a significant simplification of spectra and enabled assignment of the highest intensity signals to true N-terminal peptides in the enriched samples.

More recently, this protocol was modified to decrease the number of steps and therefore to increase sample recovery. The biotinylation step was replaced by a direct coupling and removal of internal peptides using amino-reactive NHS-activated Sepharose (140). Thus, in the final protocol the internal peptides can be removed directly after the digest with the flow-through being analyzed without further treatment. The protocol was tested using LC-MS/MS with soluble proteins
### TABLE IV

**Summary of current techniques for protease substrate discovery by enrichment of N-terminal peptides**

RP, reversed phase.

<table>
<thead>
<tr>
<th></th>
<th>Negative selection</th>
<th>Positive selection</th>
<th>In silico enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>COFRADIC</strong></td>
<td>NHS-Sepharose pullout</td>
<td>TAILS</td>
<td></td>
</tr>
<tr>
<td><strong>Major steps</strong></td>
<td>1. Acetylation of protein N termini and lysines</td>
<td>1. iTRAQ(^3) or isotope labeled formaldehyde(^2) labeling of protein N termini and lysines</td>
<td>1. Lysine guanidinylation</td>
</tr>
<tr>
<td></td>
<td>2. RP HPLC</td>
<td>2. Trypsin digest</td>
<td>2. Biotin coupling of protein N termini</td>
</tr>
<tr>
<td></td>
<td>3. Trypsin digest with C-terminal isotopic labeling of tryptic peptides</td>
<td>3. Biotin coupling and pullout of neopeptides resulting from trypsin digest or their pullout using NHS-activated Sepharose</td>
<td>2. iTRAQ labeling of protein N termini</td>
</tr>
<tr>
<td></td>
<td>5. RP-HPLC second round of separation; internal peptides are discarded</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6. LC-MS/MS (135, 136)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Quantification</strong></td>
<td>Not reported</td>
<td>MS1 with formaldehyde labeling; MS/MS with iTRAQ reagents</td>
<td>Not reported</td>
</tr>
<tr>
<td><strong>Comments</strong></td>
<td>Powerful; labor- and equipment-intensive with 96 LC-MS/MS analyses; does not distinguish between chemically and in vivo acetylated protein N termini in a single experiment</td>
<td>Fast and robust analysis of N terminome; with quantitation could be adapted for analysis of proteolysis</td>
<td>Allows analysis of only unblocked protein N termini (acylated and unblocked) and proteolysis-derived neo-N termini in a single experiment; multiplex analysis (up to eight samples using iTRAQ)</td>
</tr>
</tbody>
</table>
from *Escherichia coli* and identified ~300 proteins by their N termini with relatively few internal peptides. This negative selection protocol is fast, simple, and efficient for monitoring basal N-terminal proteolytic processing but is not suitable for identification and quantification of a specific protease activity. Labeling with $^{18}\text{O}$ at the C terminus during trypsin digestion, as now used in COFRADIC (136), or with acetic anhydride at the N terminus could be used for comparative quantification of specific protease activity between two samples, but this has yet to be demonstrated.

A further technique for substrate discovery termed “terminal amine isotope labeling of substrates” (TAILS) is in development in our laboratory. In this approach, the sample is enriched for N-terminal peptides of each protein, both original and neo-N termini. The proteomes of two samples containing active and inactive protease (control) are first reduced, alkylated, and labeled with amino-reactive isotope-containing reagents such as formaldehyde for quantification in MS1 mode or with iTRAQ for MS/MS quantification. Such labeling selectively modifies lysine residues and protein N termini. Following trypsin digestion, the newly created and therefore unblocked internal peptides are selectively removed by a highly derivatized and highly soluble amine-scavenging polymer by negative selection. The remaining N terminome fraction (i.e. all N-terminal peptides of a given proteome) is then analyzed by LC-MS/MS. For instance, when the iTRAQ reagents 114.1 and 115.1 are used to differentially label samples containing active and inactive protease, respectively, this results in protease-cleaved neo-N termini identified as spectra with singletons (i.e. containing only one isotopic signature, 114.1). In contrast, the original N-terminal peptides of these proteins will be represented by 115.1 singletons as these peptides can only be found in control sample, and non-cleaved peptides would exhibit both isotopic signatures, 114.1 and 115.1, at a 1:1 ratio. Therefore, MS/MS sequencing identifies protease substrates and defines the sequence of the cleavage site. In addition, iTRAQ labeling of lysines allows for characterization and quantification of original acetylated N termini as many of these harbor lysine residues in their sequence.

Using two alternative labels with the same approach provides different benefits. Reductive dimethylation with formaldehyde is very inexpensive. On the other hand, more costly iTRAQ labeling allows for multiplexing up to eight samples in one experiment and does not result in doubling of spectra complexity in the MS1 mode (as observed with any other non-isobaric labeling techniques such as ICAT, SILAC, acetylation, or reductive dimethylation).

In summary, the TAILS approach utilizes the power of multiplex isotope labeling and negative selection by use of a highly soluble, highly derivatized polymer that minimizes peptide losses. Thus, it enriches for protease-cleaved neopeptides and natural N termini (acetylated and non-acetylated) allowing determination of both protease substrates and their cleavage sites as well as annotation of N-terminal post-translational proteome processing in a single experiment.

**Positive Selection of N Termini**

Another approach utilizing biotin reagents for N-terminal enrichment was recently demonstrated (141). Here the proteins are first denatured, reduced, and alkylated, and then the amino groups of lysine side chains are protected by lysine-specific guanidinylation. In the next step, the N termini of the proteins are selectively labeled with NHS-biotin. Following tryptic digestion biotin-coupled peptides are selectively immobilized on a streptavidin column, then liberated by reduction using dithiothreitol, and analyzed by LC-MS/MS. Because the protocol does not involve isotopic labeling, it can only be used to describe the proteolytic state of the sample.

This approach was applied to samples of different biological origin and yielded proteome coverage from ~350 peptides in serum to ~500 peptides in *E. coli*, yeast, and mouse tissues and ~1000 peptides in a 293A human embryonic kidney cell line. Currently there is considerable contamination from internal peptides that cannot be ascribed to known proteolytic modifications. This may result from incomplete lysine guanidinylation in the beginning of the protocol leading to undesired biotin coupling to lysine residues or side reactions with serine, threonine, or histidine residues also occurring during this modification that result in other peptides being pulled out and analyzed as false positives.

In contrast to COFRADIC and negative selection N-terminal enrichment strategies where naturally acetylated N termini are automatically included in the analyte mixture, positive selection results in retention of only the N termini of unblocked proteins. Thus, such positive selection excludes from the analysis up to 80% of total natural N termini in eukaryotic samples and may have an advantage of further concentrating on the very few that are being proteolytically processed (and/or not blocked). On the other hand, retaining naturally modified (i.e. acetylated) N termini helps to minimize sample loss and has an additional bonus of higher confidence protein identification as it is then based on a positionally anchored original N-terminal peptide (139). Although the current protocol does not allow for quantification of proteolytic events, incorporating stable isotope-labeled biotin for N-terminal labeling or C-terminal trypsin-dependent $^{18}\text{O}$ exchange, similar to COFRADIC, is a possible future development.

**In Silico Enrichment of N Termini**

In an alternative approach for substrate discovery, the sample is virtually “enriched” by filtering for N-terminal fragments from a complete protein digest at the sample analysis stage.
by MALDI-TOF/TOF (142). Briefly the protease-treated and control proteins are first guanidinylated to block their lysine amino groups and then labeled with two different iTRAQ reagents at the protein N termini. Following mixing of two proteomes, the sample is digested by trypsin, fractionated by reversed phase C18 chromatography, and spotted on MALDI plates. During analysis, the sample is first surveyed for the presence of peptides with the iTRAQ reporter ion in the spectra to form a data-dependent inclusion list. Then these pre-selected peptides with the iTRAQ tags are fragmented for identification and quantification. In this work flow, the iTRAQ-bearing peptides represent the original N termini of the proteins as well as protease-generated neo-N termini with iTRAQ ratios allowing discrimination between the two. Thus, the original N termini equally present in both samples will have an iTRAQ ratio of ~1, whereas protease-dependent neo-N termini will be singletons or have an iTRAQ signal ratio >1.

This approach was validated using seven purified E. coli proteins containing putative human caspase cleavage sites treated with wild-type or catalytically inactive caspase 3. A total of 12 cleavage sites in six proteins were identified compared with five cleavage fragments identified by SDS-PAGE and 8–10 indicated by Western blotting analysis. Analysis of a complex proteome, i.e. a cell-free apoptosis model using HEK293 hypotonic extracts, identified only 20 different cleavage sites; this can be explained by ion suppression because of the physical presence of so many internal peptides in the sample. Thus, this virtual N-terminal enrichment technique is more suitable for detection and quantification of protease cleavage sites in defined protein mixtures rather than complex proteomes and can only be done using MALDI-TOF/TOF mass spectrometers.

**PROTEASE INTERACTOMICS**

The characterization of the molecular partners of proteases and protease inhibitors is essential for the analysis of their biological functions and can be extremely useful for mapping the signaling pathways that depend upon proteolytic regulation or that are regulated by proteolysis. Moreover the intersection of aberrantly expressed proteases and hence perturbed proteolytic pathways with the proteome leads to pathological alterations such as cancer and arthritis. Regulation of protease activity is greatly dependent on dynamically regulated interactions with intra- and extracellular protein partners. Transient, stable, or irreversible complex formation mediates activation and inhibition as well as oligomerization, compartmentalization, localization, and clustering by controlling proteolytic processing of substrates. The study of protein-protein interactions at the system-wide level, a component of functional proteomics, is known as "interactomics" (143, 144).

In recent years, several methodological breakthroughs, spanning from the introduction of the yeast two-hybrid system in 1989 to the modern mass spectrometry-based techniques (143–147), have facilitated the achievement of large scale protein-protein interaction analyses at different levels. Despite this, protease protein-protein interaction screenings have been largely prevented by the intrinsic difficulty of detecting binding partners in the presence of an active protease catalytic site because of proteolytic cleavage and release. It is likely that most of the protease interactors, including substrates, remain to be discovered. Because the technology to reveal them has been lacking, new protein-protein interaction screening techniques like the ones described below are needed to address this issue.

**Exosite-mediated Interactions**

Mosaicism is an important feature of many proteases, which bear one or more ancillary (non-catalytic) domains or modules distinct from the catalytic domain. These domains can harbor one or more binding sites, known as exosites (148), that play important roles as mediators for a great diversity of interactions, most of them involving protein binding. Such is the case of the penta-EF-hand in the members of the calpain-caspsatin system (149) or the caspase recruitment domain (CARD) domains present in caspases that are responsible for protease oligomerization and in the latter case also for recruitment to the apoptosome and activation (150). Similar roles have been also described for the interaction of the complement CUB and the epidermal growth factor modules in C1s and C1r complement proteins in the C1 complex assembly (151). In extracellular metalloproteases like the ADAMs, cell surface localization through integrin binding is mediated by disintegrin domains, whereas adhesion to proteoglycans depends on cysteine-rich domains (152). Closely related proteins from the a disintegrin and metalloprotease with thrombospondin motifs (ADAMTS) family interact with extracellular matrix components through their thrombospondin 1 domains (153).

Exosite interactions can reach high levels of sophistication in MMPs upon the formation of the MMP-2 cell surface activation complex in concanavalin A-induced cells (154) where TIMP-2 forms a bridge that tethers pro-MMP-2 to cell surface MT1-MMP (114). In this complex, the N-terminal domain of TIMP-2 binds to and inhibits the active site of MT1-MMP, whereas the C-terminal domain binds to the hemopexin C domain of pro-MMP-2 (155). The quaternary complex is completed after clustering with a second MT1-MMP to cleave the prodomain of MMP-2. Fully active MMP-2 is finally generated by in trans autolytic cleavage. Additional interactions of MMP-2 and MT1-MMP hemopexin C domains and the MMP-2 fibronectin module type II-like collagen binding domain to extracellular matrix components like fibronectin, heparin, and collagen can modulate the activation process (148, 156–158).

Exosites also play a pivotal role in substrate recognition and cleavage in a broad range of protease families. This is the case for PDZ domains that have been proposed to induce
proteolytic activity in the HtrA family upon binding of hydrophobic residues from misfolded proteins (159) or for the apple domains in the plasminogen-related proteins that carry binding sites for factor XIIa, kininogen, factor IX, and heparin (160, 161). Other examples include the cooperative activity between the C-terminal thrombospondin 1 repeats and the CUB domains of ADAMTS-13 reported to be crucial for recognition of the von Willebrand factor under flow conditions (162). This is akin to the C-terminal domains of the bone morphogenic protein-1/procollagen C-proteinase that have been proposed to be important for substrate recognition, control, and restriction of the proteolytic activity (163). In the MMP family, 21 of 23 enzymes present a C-terminal hemopexin domain with the two gelatinases (MMP-2 and MMP-9) carrying an additional collagen binding domain inserted at the N-terminal side of the active site. Both types of domain have been thoroughly characterized, showing essential exosite binding roles for many extracellular matrix components like secreted protein acidic and cysteine rich (SPARC), fibronectin, elastin, or fetuin (148, 156, 164). These domains are also proposed to be responsible for the triple helicase activity that unwinds the native structure of collagen before cleavage (148, 165).

**Yeast Two-hybrid-based MMP Protein-Protein Interaction Analyses**

**Exosite Scanning**—Specialized yeast two-hybrid applications have been pioneered for protease interactomics. Since exosites are pivotal elements in protease protein-protein interactions, they can be harnessed as tools for interactomics screens to search for new protease substrates and interactors. We developed an exosite scanning technique (15, 76) where the isolated human MMP-2 hemopexin C domain was used as bait in a yeast two-hybrid screen to interrogate a cDNA library from human gingival fibroblasts for potential binding partners (Fig. 7A). This genomics approach identified monocyte chemoattrac-
tant protein-3, a tissue-derived CC chemokine that recruits monocytes and other leukocytes in inflammation and osteosarcoma, as an interactor that was subsequently confirmed by in vitro binding and cleavage assays as a novel MMP-2 substrate (15). The removal of just four N-terminal residues results in a switch from agonist to antagonist behavior, again emphasizing the need to annotate the proteome for cleavage events to understand the biology of the system.

**Inactive Catalytic Domain Capture (ICDC)**—Many substrates do not require exosite assistance for cleavage, and so a different approach is needed to identify such proteins. In what we termed ICDC, catalytically inactive mutant protease domains, which do not cleave and therefore do not release substrates, are used as baits in yeast two-hybrid screens (28) (Fig. 7B). Known substrates for MMP-2 and MT1-MMP, such as laminin-5, galectin-3, and collagen, as well as previously undescribed cyto-
kine substrates like the WNT1-inducible signaling pathway protein-2 (WISP2) were detected using this approach (28, 107).

**Caveats of Yeast Two-hybrid Screening**—Although the yeast two-hybrid system has been successfully used in the specialized applications described to discover extracellular protease substrates and has been the standard technique for large scale protein-protein interaction screenings, it suffers from several drawbacks. It is slow, but the main limitation is the significant number of both false positives and negatives generated that has been thoroughly assessed after analysis of data from two large scale high throughput yeast two-hybrid screens (166, 167). The overlap among both large scale experiments was only around 10% of the total, whereas 90% of known interactions remained unde-
tected (167). The performance of the yeast two-hybrid sys-
tem has additional limitations, namely its binary nature (in-
ability to identify complexes containing more than two elements), the lack of compartmentalization, and the exclu-
sion of protein-protein interactions that depend on post-
translational modifications. Finally the method is not applicable for analyses of dynamic interactions (e.g. drug treatment versus control) because of its non-quantitative nature.

**Mass Spectrometry-compatible Interactomics Approaches**

**Affinity Purification-based Techniques**—To quantitatively and qualitatively improve the search for protease interactors, we have developed mass spectrometry-based exosite scanning and ICDC techniques. The power of mass spectrometry as a sensitive, fast, and reliable method for protein identification has been the driving force for the evolution of proteomics (78, 168), which includes the analysis and identification of protein-protein interactions (143, 146, 147). Mass spectrometry-based identification and characterization of protein complexes usually relies on a range of affinity purification-based techniques that can be used in two different general approaches: the specific in vivo isolation of endogenous complexes or the alternative in vitro pullout experiments using surface-immobilized recombinant proteins as baits (143).

**Co-immunoprecipitation**—The general biochemical approach used for the identification of protein-protein interactions is the antibody-based co-immunoprecipitation method, which can be readily adapted for mass spectrometry analysis. Here the complexes are precipitated from the sample proteome using an immobilized antibody specific for a known component of the complex or for a tag when recombinant epitope-tagged proteins are used. After washing away the nonspecific binders, the interacting proteins can be separated by gel electrophoresis followed by band extraction, tryptic digestion, and peptide recovery prior to MS/MS- or LC-MS/MS-based identification. HPLC fractionation can be used as an alternative to gel electrophoresis. The main drawback of co-immunoprecipitations is the problematic identification of bona fide interactors because of significant background generated by nonspecific binding and cross-reactions that is amplified by the exquisite sensitivity of mass spectrometry. The use of more stringent washes or binding conditions can disrupt the real complexes and prevent the identification of weak interactors. Protein cross-linking can be used to stabilize the complexes (169), but cross-linking conditions must be carefully determined to avoid spurious interactions and alterations in the tertiary structure of the proteins that can prevent epitope recognition by the antibody. The use of different chemical cross-linking strategies in mass spectrometry-based interactomics, including isotopically labeled cross-linkers and protein interaction reporters, has been reviewed elsewhere (146, 147).

**Protein Tagging Techniques**—The most widely used affinity purification- and mass spectrometry-based studies exploit the general recombinant protein tagging techniques. The bait protein of interest is easily fused in frame with a C- or N-terminal tag, enabling affinity pullout of the protein along with its binding partners without the need for specific optimization of the purification process. Several tagging systems that have been successfully used involve either short peptides such as polyhistidine tags, FLAG tags, calmodulin or streptavidin binding peptides, and the hemagglutinin epitope or small full-length proteins like GST, β-galactosidase, thioredoxin, or maltose-binding protein (170).

An evolution of these tagging techniques is the tandem affinity purification (TAP) system, which facilitates high levels of enrichment in two sequential purification steps. TAP is based on a dual tagging system composed of two different epitopes, protein A and a calmodulin binding peptide separated by a tobacco etch virus cleavage sequence (171, 172). In the first purification step, the complexes are bound to IgG-Sepharose through the protein A tag. After non-stringent washing steps and elution by tobacco etch virus cleavage, the sample is bound to an immobilized calmodulin column in the presence of calcium, washed, and finally eluted with EGTA (Fig. 8A). Initially created for yeast, several different TAP systems have been developed for higher eukaryotes and mammalian cells (173–176). The two-step procedure allows for efficient removal of contaminant proteins through several washing steps and specific elutions, thus yielding highly enriched complexes. However, despite the use of mild binding and washing conditions, loosely bound interactors can be lost during the procedure. This should be considered when deciding between one-step and two-step purification systems as recovery levels are generally 3–5 times higher with single step purification at the cost of significantly higher background noise (143).

**Affinity purification-mass spectrometry approaches** have been successfully used to analyze large scale interaction networks in *Saccharomyces cerevisiae* using either genome-wide FLAG or TAP tagging by homologous recombination techniques (177–179). Each of the two most recent studies achieved roughly 2000 purifications and describe 547 (179) and 491 (178) putative complexes. Gavin *et al.* (177, 179) based their de novo identifications on the development of a *socioaffinity index* formed by the number of times that two proteins were found together in relation to what would be expected from their frequency in the data set. In the second report, Ho *et al.* (178) used a hand-curated reference database of protein complexes as a "gold standard" training set to assign probabilities to pairwise interactions through machine learning followed by optimization by means of a Markov graph-clustering algorithm. Although the vast amount of information on in vivo protein-protein interactions obtained in these experiments requires further analysis and validation, these studies have shown the feasibility and power of affinity purification-mass spectrometry for the analysis of interaction networks at almost any scale.

**Quantitative Affinity Purification-Mass Spectrometry**—The recent advances in quantitative proteomics, mainly based on
isotopic labeling techniques, have improved and expanded the use of affinity purification-mass spectrometry in the analysis of protein-protein interactions. Quantification of the mass spectrometry data enables the identification of nonspecific interactions, the assessment of protein-protein interactions dynamics, and the determination of complex stoichiometries. For example, SILAC has been used for protein-protein interactions studies using “quantitative immunoprecipitation combined with knockdown” (QUICK) for the detection of interacting partners and the dynamics of complex formation using β-catenin and Cbl as bait proteins (180). In this approach, the authors knocked down β-catenin or Cbl in SILAC “light” cells by means of short hairpin RNA induction or small interfering RNA transfection, respectively, leaving “heavy” cells untreated. After immunoprecipitation of cell lysates with the appropriate monoclonal anti-β-catenin or anti-Cbl, precipitates from the light and heavy cells were combined, washed, eluted, digested with trypsin, and analyzed by LC-MS/MS. In both cases, most of the proteins detected showed peptide ratios of 1:1 for the heavy/light forms, corresponding to nonspecific contaminant proteins that precipitated irrespective of the presence or absence of the bait protein in the cells. Peptides derived from proteins that were enriched upon specific binding to the bait proteins showed relative intensity heavy/light ratios close to 1, whereas specific binding partners are enriched upon interaction with the bait protein, hence showing heavy/light > 1. TEV, tobacco etch virus.
ratios for the heavy/light forms that were significantly higher than 1 (Fig. 8B). As expected, both β-catenin and Cbl were several times more abundant in the heavy than in the light form. In addition, three and four proteins with significantly increased abundance ratios were found in the β-catenin and the Cbl eluates, respectively. In the former case, the three hits were well known β-catenin interactors, whereas in the second experiment, three of the four hits were already reported members of the Cbl interactome, results that validated the efficiency of the technique.

In a combined approach termed “quantitative analysis of tandem affinity purified in vivo cross-linked (X) protein complexes” (QTAX), Guerrero et al. (181) used SILAC labeling, in vivo formaldehyde cross-linking, and TAP to capture and identify all the components of the 26 S yeast proteasome along with 64 potential proteasome interactors. The accuracy, sensitivity, and power of affinity purification–quantitative mass spectrometry have located this technology at the forefront of interactomics research. At the tissue level, ascribing a particular cleavage event to one protease is inherently difficult even when redundancy or rather by utilizing different proteases to achieve similar outcomes but in different physiological or pathological states can be debated. Regardless of the contributing proteases, the biological importance of their activity is in the extent of proteolysis of a key substrate and whether this reaches a tipping point in the system to change the biological outcome. System robustness will buffer against such changes to a certain extent. However, “overflow” of protease activity can result in excessive and irreparable changes thus leading to pathology. Hence identification of the processed substrate(s) and the responsible protease(s) is of paramount importance for drug target and antitarget selection.

The field of proteomics faces several serious technical challenges. First, the sheer number and diversity of gene products expressed in a single cell at any given time make their accurate measurements or even detection difficult. Second, an extremely wide dynamic range of protein levels often results in undersampling of lower abundance species. Therefore, sophisticated proteome fractionation techniques and higher sensitivity mass spectrometers are continuously being developed to overcome these obstacles. Although such technological advances have created a wealth of knowledge about the composition of different proteomes, they provide little understanding of their functional state. This information is often encoded in various post-translational modifications, which can escape detection in conventional proteomics experiments but are directly targeted by phosphorytoproteomics, redox proteomics, and degradomics. These newer, functional sub-divisions of proteomics together with the emerging field of metabolomics provide ultimate and invaluable insight into protein function, activity, interactors, and even stability.

Additional levels of complexity arise from compartmentalization in living cells as well as from the interactions of the protease web with the proteome: other proteins, inhibitors, receptors, substrates, and proteolytic products can modify proteolytic activity and potential. It is this interconnection between the protease web and the proteome that provides the system with new layers of information that conventional proteomics analyses often overlook. Hence specific N and C terminome analyses need to be considered as a new way forward to annotate the proteome in the context of the very important proteolytic post-translational modifications outlined in this review. At the tissue level, ascribing a particular cleavage event to one protease is inherently difficult even when comparing tissues between protease knock-out and littermate controls. The global analysis of proteolysis in tissues by communities of proteases on a system-wide level, what we introduce as “meta-degradomics” aims to identify the multi-
tude of proteolytic events by multiple proteases in entire tissues or organs by N- and C-terminomeric approaches. By monitoring these changes during development and physiological homeostasis, pathogenesis, and post-treatment, the scope of proteolysis as a driving force of change in the proteome can be characterized as well as identifying new biomarkers of disease.

The metadegradomics challenge is enormous but so too are the rewards that can be realized from achieving a deeper functional understanding of the proteome rather than from solely increasing its coverage. In an era of systems biology, one can envision any biological object of study as an enormous iceberg of information to be undiscovered. Here the tip of the iceberg is formed by the genomic content of the DNA and the underlying RNA layer. Proteomics data comprise the vast submerged body of the iceberg and share the same plane as complementary “omics” that cover different biological molecules (e.g. lipidomics, glucidomics, and metallomics). In this hierarchy of systems biology disciplines, functional proteomics including degradomics occupies a yet deeper layer of information followed only by metabolomics. Like icebergs where the physics of the submerged body largely dictates their floatation behavior, in biology, functional proteomics and metabolomics provide information that carries the most weight and predictive power on the overall outcomes of a biological system and are therefore the best suited techniques for biomarker discovery. Hence the information content of the protease-sculpted proteome generated through metadegradomics analyses promises to play an increasingly significant role in the functional annotation of the proteome in health and disease.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

† To whom correspondence may be addressed. Tel.: 604-822-2958; Fax: 604-822-7742; E-mail: adoucet@interchange.ubc.ca.

‡ To whom correspondence may be addressed. Tel.: 604-822-2958; Fax: 604-822-7742; E-mail: chris.overall@ubc.ca.

REFERENCES

CLIP-CHIP, a dedicated DNA microarray of all human proteases and inhibitors. *Biol. Chem.* **385**, 493–504


Quantitative Degradomics of the Cancer Proteome


type IV collagenase. Isolation of the activated form of the membrane metalloproteinase. J. Biol. Chem. 270, 5331–5338


Molecular & Cellular Proteomics 7.10 1949

Quantitative Degradomics of the Cancer Proteome
clusters of hemopexin modules III and IV. J. Biol. Chem. 274, 4421–4429


187. Pfieger, D., Junger, M. A., Muller, M., Rinner, O., Lee, H., Gehrig, P. M.,


