

# PROTEASE DEGRADOMICS: A NEW CHALLENGE FOR PROTEOMICS

Carlos López-Otín\* and Christopher M. Overall‡

Degradomics — the application of genomic and proteomic approaches to identify the protease and protease-substrate repertoires, or ‘degradomes’, on an organism-wide scale — promises to uncover new roles for proteases *in vivo*. This knowledge will facilitate the identification of new pharmaceutical targets to treat disease. Here, we review emerging degradomic techniques and concepts.

## PROTEASE

An enzyme that cleaves proteins by the catalysis of peptide-bond hydrolysis. On the basis of their catalytic mechanism, proteases belong to one of five classes (aspartic, cysteine, metallo, serine, or threonine).

## PROTEOLYTIC PROCESSING

Proteolysis that is distinct from degradation in that it represents highly specific and limited substrate cleavage, which results in a specific change of protein function.

\*Departamento de Bioquímica y Biología Molecular, Facultad de Medicina, Instituto Universitario de Oncología, Universidad de Oviedo, 33006 Oviedo, Spain.

‡Departments of Biochemistry and Molecular Biology, and Oral Biological and Medical Sciences, Protein Engineering Network of Centers of Excellence, University of British Columbia, Vancouver British Columbia, V6T 1Z3 Canada. Correspondence to C.M.O.  
e-mail: [chris.overall@ubc.ca](mailto:chris.overall@ubc.ca)  
doi:10.1038/nrm858

Our view of the proteolytic universe has expanded considerably in recent years. PROTEASES were initially characterized as nonspecific degradative enzymes that are associated with protein catabolism. However, it is becoming increasingly recognized that proteolysis represents another mechanism for achieving precise cellular control of biological processes in all living organisms, through the highly specific hydrolysis of peptide bonds<sup>1</sup>. This highly specific and limited substrate cleavage is termed PROTEOLYTIC PROCESSING. Proteases, through their ability to catalyse irreversible hydrolytic reactions, regulate the fate and activity of many proteins by controlling appropriate intra- or extracellular localization; shedding from cell surfaces; activation or inactivation of proteases and other enzymes, cytokines, hormones or growth factors; conversion of receptor agonists to antagonists; and exposure of cryptic neoproteins (which is when the proteolytic cleavage products are functional proteins with roles that are distinct from the parent molecule). Hence, proteases initiate, modulate and terminate a wide range of important cellular functions by processing bioactive molecules, and thereby directly control essential biological processes, such as DNA replication, cell-cycle progression, cell proliferation, differentiation and migration, morphogenesis and tissue remodelling, neuronal outgrowth, haemostasis, wound healing, immunity, angiogenesis and apoptosis<sup>1,2</sup>.

Considering the functional relevance of proteases for all living processes, including cell death, it is not difficult to understand that a deficiency, or a misdirected temporal and spatial activity, of these enzymes underlies several pathological conditions such as cancer, arthritis, neurodegenerative and cardiovascular diseases<sup>1,2</sup>.

Moreover, many infectious microorganisms, viruses and parasites use proteases as virulence factors, and animal venom commonly contains proteases to effect tissue destruction or to evade host responses. Accordingly, many proteases or their substrates are an important focus of attention for the pharmaceutical industry as potential drug targets.

Owing to the expanding roles for proteolytic enzymes, there has been an increasing interest in the identification and functional characterization of the many proteases that are present in various organisms, from bacteria to man. The near completion of several large-scale genome-sequencing projects has provided new opportunities to appreciate the complexity of protease systems. According to our survey of the human genome, more than 500 genes that encode proteases or protease-like molecules comprise our proteolytic labyrinth. The most recent release (17 December 2001) of the protease database — MEROPS version 5.7 — lists 461 proteases and homologues in man, with the metalloproteinases and serine proteases comprising the largest classes, having at present 159 and 140 members, respectively (TABLE 1). As unidentified genes are characterized and ‘hidden’ proteases, such as Rhomboid<sup>3</sup>, are recognized — some of which might have new catalytic machinery and highly specific activities — the total number of proteases in humans could grow beyond 500. The task of characterizing both the functions of all proteases (existing and newly identified), and their functional connections with other proteases and inhibitors in the various protease systems in humans, is therefore daunting, and model organisms will be invaluable for this. The murine protease map is similar to the human

**MATRIX METALLOPROTEINASES**  
A family of 23 endoproteases in humans that are encoded by 24 genes. These are characterized by a HEXXHXXGXXH zinc-binding motif, a cysteine-switch mechanism of proenzyme latency, an ability to cleave extracellular-matrix and bioactive molecules, and inhibition by tissue inhibitors of metalloproteinases (TIMPs).

Table 1 | **Current numbers of proteases in humans and model species**

Species	Total*	Catalytic class of protease				
		Aspartic	Cysteine	Metallo	Serine	Threonine
<i>Homo sapiens</i>	461	18	121	159	140	23
<i>Caenorhabditis elegans</i>	353	26	93	151	62	21
<i>Drosophila melanogaster</i>	513	38	59	157	225	34
<i>Mus musculus</i>	383	11	93	120	137	22
<i>Rattus norvegicus</i>	227	10	41	77	80	19

\*The total numbers are accurate as of the date of release of the most recent MEROPS database v 5.7 (17 December 2001), but will continue to grow as new proteases are discovered and characterized in the existing databases and in new sequence deposits.

map, although there are families such as the kallikreins, cathepsins or MATRIX METALLOPROTEINASES (MMPs) that have evolved somewhat differently<sup>4–6</sup> (TABLE 1). Surprisingly, *Drosophila melanogaster*, which has a gene content that is considerably lower than these vertebrate organisms, shows a similar number of protease genes as a result of the impressive expansion of a group of trypsin-like serine proteases in the fly genome<sup>7</sup>. However, further studies are needed to clarify the genetic and molecular bases of the evolutionary differences between the protease repertoires of these organisms. So, in addition to the universal proteolytic ‘routines’ that are common to all organisms, there are also specific functions that are carried out by unique proteases in different species. Finally, and in keeping with the observed protease complexity, recent studies are uncovering many diverse new substrates and new endogenous inhibitors that have the ability to balance protease activity in physiological and pathological conditions. Hence, the emerging pattern in the proteolytic world is one of diversity and complexity, which produces new layers of control — often of an exquisite nature — for many pivotal cell and tissue functions through precise and limited proteolysis rather than through nonspecific, general protein catabolism.

Despite these advances, the substrates and *in vivo* roles for newly identified proteases are unknown and, even for proteases that have been well characterized, their biological functions are often not fully understood. New techniques are urgently required to identify the protease

repertoire that is expressed and active in a cell, tissue or organism, as well as to identify all the natural substrates of each protease. This article addresses these crucial issues — the ‘omics’ language is adapted for proteases and the new terms are defined. We cover potential uses for global analyses of protease systems, which — as well as the limitations of existing strategies — will drive the development of new approaches and technologies for the system-wide analysis of proteases and their substrates. In particular, four classes of protease chips will be discussed, which can identify complete sets of proteases and their substrate repertoires at an organism, tissue or cellular level in a high-throughput manner. Other emerging techniques — such as the use of fluorogenic-labelled protease inhibitors, chemical proteomics and protease domains to identify new substrates — will also be presented to encourage the further development of these techniques. Finally, we discuss the impact that this new system-wide approach will have on the protease field in terms of our views of the roles of proteases *in vivo* in health and disease states, and in terms of the identification of new targets for the pharmaceutical industry.

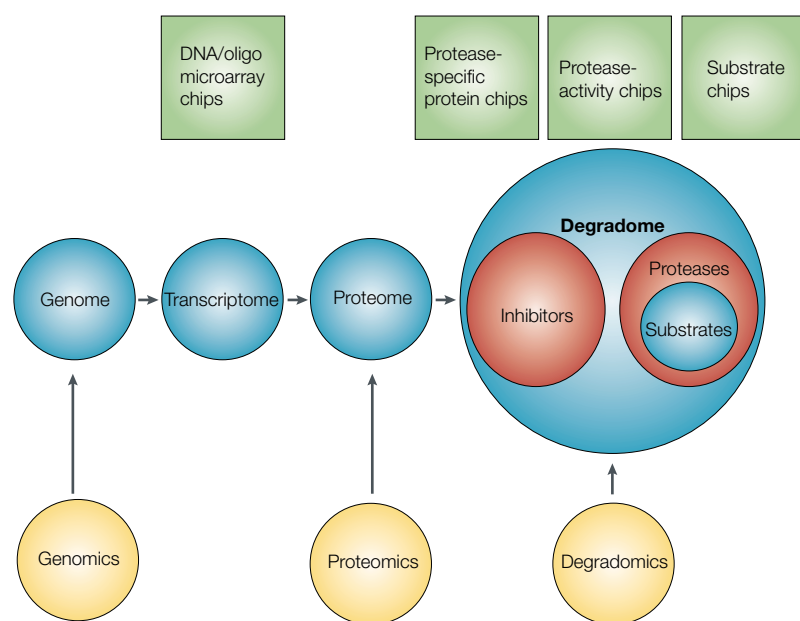
#### Degradomics and degradomes

Evidence for the increasing complexity and importance of the proteolytic systems that function in all organisms, and the ability to analyse systems in their entirety on genome- and proteome-wide scales (BOX 1), necessitates the introduction of new terms to clarify emerging concepts in this field. So, ‘degradomics’ was first coined to define the substrate repertoire of a protease on a proteome-wide scale<sup>8</sup>. Given the historical connection between proteases and general protein degradation, we propose that this is an appropriate and intuitive term, despite the more important, and now increasingly recognized, roles of proteases as processing enzymes. We extend this definition to one that describes all genomic and proteomic approaches for the identification and characterization of proteases that are present in an organism, including the substrates that are targeted by these proteases and their endogenous inhibitors (FIG. 1). We also propose the term ‘degradome’ to encompass two concepts. First, it is the complete set of proteases that are expressed at a specific moment or circumstance by a cell, tissue or organism. Second, the degradome of a protease is the complete natural substrate repertoire of that enzyme in a cell, tissue or organism (FIG. 1). The field of degradomics will be built using emerging, and new,

#### Box 1 | Genomics and proteomics: the parents of degradomics

Degradome and degradomics are two terms that have been adapted from the fields of genomics and proteomics, and these terms are defined below together with other ‘omics’ terms that appear in this Review.

- **Genome.** The entire collection of genes in the complete DNA sequence of an organism.
- **Transcriptome.** The complete set of mRNAs that are transcribed from the genome.
- **Proteome.** The expressed set of proteins that are encoded by the genome.
- **Genomics.** Investigations and techniques for identifying the genome.
- **Proteomics.** Investigations and techniques for identifying the proteome.
- **Degradomics.** All genomic and proteomic investigations and techniques regarding the genetic, structural and functional identification and characterization of proteases, and their substrates and inhibitors, that are present in an organism.
- **Degradomes.** The complete set of proteases that are expressed at a specific time by a cell, tissue or organism. The degradome of a protease is its substrate repertoire (FIG. 1).



**Figure 1 | Relationship of degradomics to the fields of proteomics and genomics, and of the degradome to the proteome and genome.** Degradomics is the application of genomic and proteomic approaches to identify the protease and protease-substrate repertoires — or ‘degradomes’ — on an organism-wide scale. A glossary of definitions is provided in BOX 1. Examples of protease chips that can be used for degradomics investigations are also shown in the figure (shaded green).

genomic and proteomic technologies to investigate and define both types of protease degradome.

Identifying the substrate degradomes of individual proteases will facilitate our understanding of their physiological and pathological roles and thereby point to new drug targets. This information, in conjunction with knowledge of the protease degradome of a cell, will increase our understanding of the biological roles of proteases in the cellular context with respect to cell function and pathology. Similar information on a tissue-wide scale should prove useful in the molecular diagnosis of disease, with the calibration of protease levels to disease severity or tumour grade enabling more accurate prognostic predictions to be made for patients. The protease degradome of an organism will be largely defined by genomic and bioinformatic analyses of gene sequences, with the expression pattern of protease messenger RNAs — the transcriptome (BOX 1) — showing tissue- or cell-specific expression levels that reflect proteolytic potential. However, functional degradomics is required to determine the actual proteolytic activity that is expressed at a particular time by a cell or tissue. Functional degradomics has two branches: the first is based on activity profiling of individual proteases, and the second involves determination of the net cleavage of target substrates. So, instead of defining individual contributions by specific proteases, this latter aim considers the protease degradome as a system that leads to substrate cleavage. The field of degradomics promises to uncover new proteases and physiological substrates, and to identify new and known regulatory pathways that are controlled by proteolytic processing. The regulation of these pathways might be

**MASS SPECTROMETRY**  
A technique that precisely measures sample mass from the analysis of mass-to-charge ratio ( $m/z$ ).

disrupted in disease states, or host proteases might be used by microorganisms for infection, and could therefore be therapeutically targeted.

### From global concepts to global approaches

Proteases do not operate in isolation — they are expressed and function in the context of a proteolytic system that is comprised of related and unrelated proteases, their substrates and cleavage products, inhibitors, cell receptors and binding proteins. It is often the case that one protease can cleave many substrates, and also that many proteases can cleave the same substrate, activate other members of a proteolytic cascade, undergo autolysis *in cis* and *in trans*, and degrade other members of the system. Only by considering individual proteases as a part of a system, and the proteolytic systems as a whole, can the impact of the protease degradome on the substrate degradome *in vivo* be understood and hence its perturbations recognized in disease. As systems hold considerably more information than their individual components, the analysis of proteases needs to move from characterization of individual proteins to the system-wide level. However, in the rapidly approaching post-proteomic era, the detailed biological characterization of individual proteases and drug targets will still be important in order to exploit the wealth of information that is generated by degradomics analyses.

The degradomics challenge is therefore significant — even for well-studied proteases, their biological roles and their relationships with other components of the protease systems are at present not fully understood. Specific activity and redundancy are crucial determinants of the role of a particular protease in a proteolytic system. The hierarchical importance of proteases within a system is also affected by expression levels, temporal/spatial distribution, activation, turnover and inhibition — properties that profoundly influence proteolytic potential *in vivo*. This is a crucial issue for drug development, as several proteases might cleave the same substrate *in vitro*, but, *in vivo*, substrate cleavage might be restricted to one, or only a few, of these proteases.

Although new protease substrates are being recognized, they are typically discovered by serial approaches that are time consuming, usually cumbersome, and have incomplete coverage. Moreover, substrate identification *in vitro* does not necessarily indicate that the protein will be a biologically relevant substrate *in vivo*. Considering the number and diversity of proteases that are present in different organisms, innovative approaches and tools are needed to profile the expressed protease repertoire and to screen for new substrates. Distilling this information will require many tools and iterative approaches.

Conventional methods of proteomics at present involve denaturation of all proteins in a mixture by two-dimensional electrophoretic methods or by tryptic digestion and serial liquid-chromatography (LC) steps before MASS SPECTROMETRY, but considerable information is lost as a result of denaturation or digestion. Fortunately, innovative and rapid advances in proteomics for native protein analysis have been made by new or improved sample preparation, labelling and

Table 2 | Overview of degradomics approaches

Degradomics approach	Analysis scale	Target	Limitations
DNA microarray chips	Transcriptome	Messenger RNA	Expression levels do not relate to protein abundance. Does not reflect recruitment of proteases <i>in trans</i> from other cells or tissues.
Protease-specific protein chips	Proteome	Protease protein	Abundance does not necessarily reflect activity. Lack of specific probes for all proteases.
Protease-activity chips	Proteome	Active protease	Activity does not necessarily indicate substrate cleavage. Measures absolute levels without considering protease turnover.
Substrate chips	Proteome	Protease substrate	Does not identify the active proteases. Difficult to obtain proteome-wide protein arrays. Substrate protein might not be in the correct three-dimensional biological conformation.
Two-dimensional gel tandem mass spectrometry	Proteome	Protease substrate Protease	Low-mass cleavage fragments (<~8 kDa) and fragments with high or low isoelectric points are not resolved on two-dimensional gels, and thereby exclude many bioactive mediators. Membrane proteins, such as receptors and adhesion molecules, are difficult to study. Low abundance proteins are not detected because of sensitivity limits. Protein masses are only inaccurately determined by electrophoresis before in-gel tryptic digestion. Substrates and large cleavage products might not resolve well.
Inhibitor-based electrophoretic profiling	Proteome	Active protease	Lack of specific probes for all proteases. Limited throughput. Denaturing method, so not suitable for proteases without a covalent acyl intermediate.
Inhibitor- or antibody-based liquid chromatography and tandem mass spectrometry	Proteome	Active protease	Non-quantitative.
Chemical proteomics	Proteome	Active protease	Pharmacokinetic problems might limit effective concentration <i>in vivo</i> . Toxicity concerns. Small amounts of active protease might not be inhibited. Due to the difficulty in developing protease-specific inhibitors, related proteases might also be inhibited.
Targeted isotope-coded affinity tags	Proteome	Active or total protease depending on tag selection	Quantitative, but difficulties arise in developing protease-specific probes. Development of covalent probes is a huge hurdle for proteases without a substrate-cleavage acyl intermediate. The mass of the intact protein is not measured by mass spectrometry.

fractionation techniques before mass spectrometric analysis. Among the emerging technologies, which can also be applied to the global analysis of proteases, are protease chips, which we envisage can be divided into four different formats (TABLE 2) — conventional DNA microarray chips, protease-specific protein chips, protease-activity chips and substrate chips. The global approaches that are available to degradomics will now be discussed in more detail.

**Protease DNA microarray chips.** Conventional DNA microarrays are based on complementary DNA or oligonucleotide-specific probes for all the different proteases of that species, as defined by whole-genome sequencing projects and bioinformatic analyses. Ideally, these chips would also contain probes that are specific for the different protease inhibitors that are produced by an organism. These chips can be used to obtain a general view of the protease and inhibitor transcriptomes in a normal or pathological tissue and, with developments in technology, will eventually be used to obtain a specific view at the single-cell level. However, mRNA expression levels of protease genes do not accurately reflect the protein expression levels of these enzymes, nor do they show proteases that are recruited from remote sources, such as

the serum or the adjacent tissue and infiltrating cells. DNA-based chips should therefore be complemented with protein-based chips<sup>9–11</sup>.

**Protease-specific protein chips.** In principal, we envisage that the second type of protease chip — protease-specific protein chips — would incorporate arrays of molecules on solid supports or in nanowells that are designed to capture and assay specific proteases from complex biological samples. Despite the promise of such techniques, as shown in other systems<sup>9–11</sup>, specific proof-of-concept papers that relate to proteases have yet to be published. However, protease-specific protein chips could be based on immobilized antibodies against different proteases or on protease-specific chemical reagents that have the ability to capture individual enzymes from complex mixtures. Retained proteases could be detected with antibodies or trypsin digested *in situ* and analysed by matrix-assisted laser desorption-ionization time-of-flight mass spectrometry (MALDI-TOF; BOX 2) and identified by PEPTIDE FINGERPRINTING (BOX 3) or by sequencing the tryptic peptides by mass spectrometry/mass spectrometry (MS/MS; BOX 3), which is also called tandem mass spectrometry.

**PEPTIDE FINGERPRINTING**  
A mass spectrometric technique of protein identification that matches tryptic peptide masses of an unknown protein with those that are generated *in silico* for all the proteins in a database.

**Box 2 | Mass spectrometry: a powerful tool for proteomics**

Mass spectrometry measures the mass of proteins or peptides from analysis of the mass-to-charge ( $m/z$ ) ratio. Mass spectrometers first ionize the sample, and then the ions are introduced into the mass analyser, which separates and detects the sample ions according to their mass. Fragmentation of the protein sample results in a collection of ions that have different masses, and the spectrometer measures the relative abundance of each ion, according to their  $m/z$  ratios, to obtain a spectrum of masses, hence the term mass spectrometry. Proteomic analyses typically use matrix-assisted laser desorption-ionization (MALDI) or electrospray ionization (ESI) sources followed by introduction into time-of-flight (TOF) or quadrupole (Q) mass analysers, which are described in more detail below.

**Ionization sources**

**Matrix-assisted laser desorption-ionization (MALDI).** Protein or peptide samples on solid matrices are ionized by a pulsed laser, entrained in the resulting ionized gas plume, and introduced into the mass analyser.

**Electrospray ionization (ESI).** Protein or peptide samples are passed through a fine needle to which a voltage is applied, which results in a fine spray of sample-containing droplets. Samples are delivered to the mass analyser after the breakup and evaporation of the droplets, which releases the protein or peptide samples to the gas phase. A low-flow-rate liquid chromatography system can be coupled to the needle to allow for protein fractionation of the samples before mass analysis.

**Mass analysers**

**Time-of-flight (TOF).** Time-of-flight analysers separate ions on the basis of their flight times over a known distance. The lower the mass of the ion, the greater the velocity and hence the shorter the flight time. Travel time from the ion source to the detector is transformed into the  $m/z$  ratio, from which the mass of the ionized sample can be calculated with extreme accuracy.

**Quadrupole (Q) mass filter.** A quadrupole mass filter consists of four parallel rods through which direct current and radio frequency electric fields are applied to sort the introduced ions. For each combination of voltages and frequencies, only ions with a specific  $m/z$  ratio pass undeflected through the quadrupole mass filter. Precise stepping of these settings therefore allows the quadrupole to be used as a mass analyser to scan for ions over a large  $m/z$  range.

Sensitivity of mass spectrometry, either of captured proteases on chips or from solution (as described later), however, is still an issue for obtaining complete coverage of the protease repertoire. To identify proteases that are involved in protein catabolism, this should not pose a problem, as the cell typically achieves this function by releasing relatively large quantities of proteases. For instance, this would apply to cathepsins, which are released into intracellular vacuoles or specific lysosomes that contain phagocytosed extracellular material that is destined for complete degradation, and to metalloproteinases such as the MMPs<sup>2,12</sup> and the related ADAMTS ('a disintegrin and metalloproteinase domain with thrombospondin modules') proteases<sup>13</sup>, which are secreted outside the cell with the potential to degrade extracellular matrix proteins in pathological tissue destruction. However, differential concentration and separation by LC will be needed to enhance the signature of low abundance proteases. For example, the highly selective and precise processing of intracellular signalling and regulatory molecules, or extracellular bioactive molecules, can be achieved by proteases that are expressed at low levels. These are under the tight control that is exerted by highly regulated transcription and expression pathways. Indeed, it is expected that the proteases

that are expressed at low levels — which initiate, modulate or terminate information cascades — are the very enzymes for which it is desirable to obtain this type of information.

A second, very important limitation of these techniques is that they focus on recording variations in protease-expression levels rather than on analysing their activity. This limitation, which is shared by all expression-based chips, is especially relevant in the case of proteases that are subject to several mechanisms of post-translational regulation, including zymogen activation, autocatalytic shedding of substrate- or inhibitor-binding domains, and inhibition.

**Activity profiling.** To advance from expression degradomics to functional degradomics, a third type of analysis — activity profiling — is required, which will lead to the development of protease-activity chips. Activity profiling is used to broadly survey the proteolytic activity of the different enzymes in complex samples. Importantly, it distinguishes active enzymes from their inactive-precursor or inhibitor-bound forms. The first attempts at functional degradomics have been directed at profiling serine- and cysteine-protease activities in crude protein samples<sup>14–16</sup>. As the catalytic mechanism of cysteine and serine proteases uses a covalently bound acyl intermediate, these approaches are based on the reactivity of these proteases towards tagged chemical probes in which the structure of a general inhibitor serves as a scaffold to which only active proteases can covalently bind. Active site-bound proteases are then identified by molecular weight and by fluorescent tag colour after electrophoretic migration past a fluorescence detector, or on western blot transfer to membranes<sup>16</sup>. An alternative approach, now in development in our laboratories, involves sequencing active proteases that have been captured by inhibitors, or other biotin-tagged affinity probes, and immobilized on affinity columns (FIG. 2a). After elution from avidin LC columns the captured proteases can then be identified by tandem mass spectrometry (BOX 3). This LC-MS/MS approach will be particularly useful for low abundance proteases, which can now be injected into the mass spectrometer in a more concentrated form.

**Protease-activity protein chips.** The next logical improvement of the protease-specific protein chip and protease-inhibitor LC approaches would be the immobilization of protease-specific inhibitors — either natural protein or chemical reagents — on chips to capture active proteases. Captured proteases could be identified in array readers using fluorescent-labelled antibodies or by quenching fluorescent-labelled capture probes, or, more directly and specifically, using MALDI-TOF or MALDI-quadrupole-TOF (MALDI-Q-TOF; BOX 2) tandem mass spectrometry after trypsin digestion *in situ*.

**Chemical proteomics and imaging.** The development of other types of tagged irreversible inhibitors, which are reactive against other classes of proteases, should provide an invaluable tool for the rapid detection, identification,

ADAM AND ADAMTS  
Cell surface (ADAM) or secreted (ADAMTS) metalloproteinases that are related to MMPs, but that have a different multidomain structure, which includes a cysteine-rich disintegrin domain and thrombospondin modules.

**Box 3 | Protein identification by mass spectrometry**

Proteins can be identified by mass spectrometers by using two main methods — peptide fingerprinting and sequencing. Matrix-assisted laser desorption-ionization time-of-flight (MALDI-TOF; BOX 2) mass spectrometry (MS) precisely measures the masses of the tryptic fragments of a protein. To identify the protein, this peptide fingerprint is compared with predicted tryptic peptides generated *in silico* from all the proteins in a database. If protein identity cannot be established by peptide fingerprinting (because of ambiguity), the peptides can then be sequenced by MS/MS, which is also called tandem mass spectrometry. Two separate stages of mass analysis are carried out on tandem mass spectrometers using two integrated mass analysers (BOX 2), which are coupled together in the same machine. First, tryptic-digest samples are ionized either by MALDI or by nanoelectrospray ionization (BOX 2). The first mass spectrometer is used to select for peptide ions of known mass, the selected ions are then fragmented, and the second mass analyser measures the masses of the daughter ions from which the sequence can be derived.

Tandem mass spectrometry can be carried out using an ion-trap mass spectrometer, in which all tryptic peptide ions, except the ions of a selected mass, are expelled from the trap. Increasing the energy of the trap then fragments this peptide, with the mass detector measuring the masses of these fragments for peptide identification. Other examples of tandem mass spectrometers include triple quadrupole mass spectrometers — in which the second quadrupole is an argon gas collision cell — and the quadrupole-TOF mass spectrometer. The latter state-of-the-art instrument uses a quadrupole mass filter to select for ions of a desired mass, which are then fragmented in a gas collision cell, with the masses of these daughter fragments then being measured extremely accurately by a TOF mass analyser to derive their sequence.

isolation and even imaging of active enzymes that are present in cell and tissue proteomes. In what has been termed ‘chemical proteomics’<sup>16</sup>, the use of protease-specific or broad-spectrum protease inhibitors to achieve chemical knockouts of proteases promises to be a powerful new approach to uncover the biological roles of proteases in culture or *in vivo*. Moreover, fluorescent dyes coupled to inhibitors can be used to localize the cellular and tissue distribution of active proteases in histological sections<sup>16</sup>. In mouse models — and in humans in the future — the use of specialized labels coupled to protease inhibitors can be used for INTRAVITAL IMAGING of active proteases in tissues and in pathological lesions such as tumours. For example, using fluorescence-reflectance imaging, fluorescence-mediated tomography and near-infrared fluorochromes, active MMP-2 (gelatinase A), cathepsins and caspases have been localized in living tissues<sup>17–19</sup>. Other techniques for intravital imaging of active proteases, such as positron-emission tomography and multiphoton confocal microscopy, will also be possible<sup>19</sup>, some of which are based on using quenched fluorescent synthetic substrates that generate a signal only on cleavage.

There are significant challenges to these approaches. The metalloproteinases do not have a covalent-bound substrate transition state, and there are no known covalent inhibitors that can be chemically developed for denaturing approaches. Innovative native strategies will be required for these enzymes — potentially based on biotin-tagged hydroxamate peptidic inhibitors — which could be used to separate these proteases from mixtures under mild native conditions. Elution and identification by LC-MS/MS would then follow. Although specificity might be built into these probes for proteases with unique inhibition or cleavage properties, if we consider

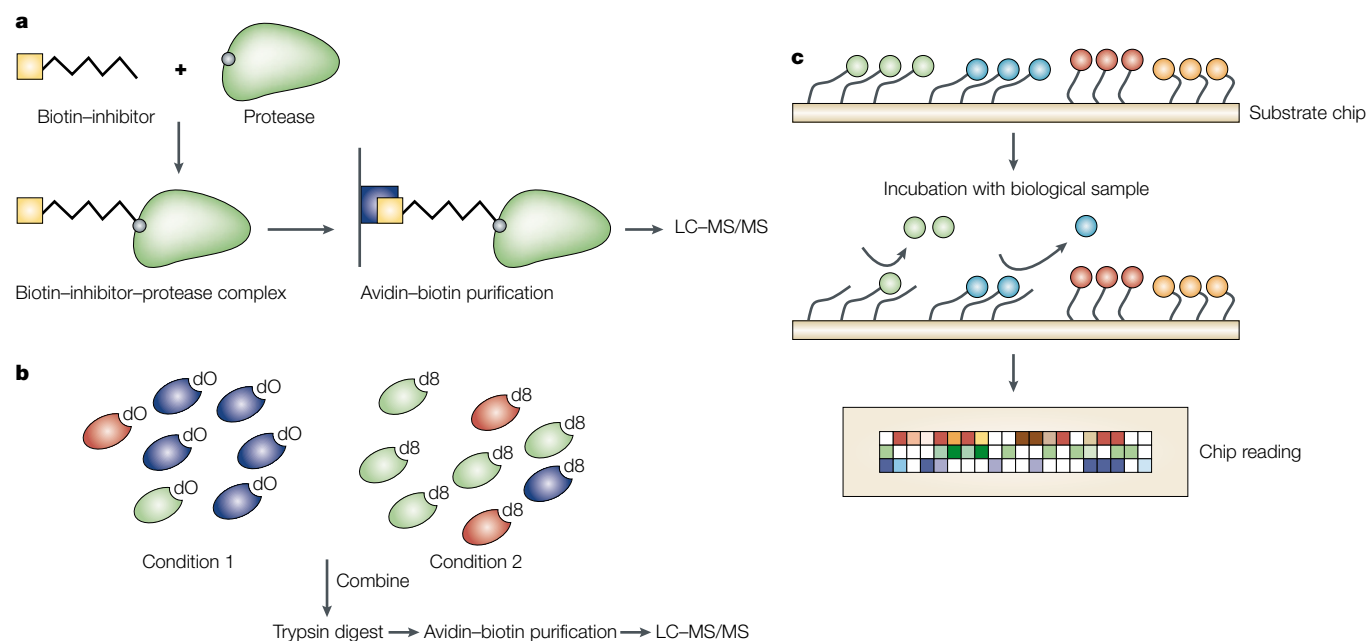
the robustness of proteolytic systems, with their multiple overlapping substrate specificities, it is unlikely that it will be possible to fully cover the proteolytic landscape with protease-specific inhibitor probes. A potential variant of this might be to use broad-spectrum-family- or subfamily-specific inhibitors to profile proteases on a system-wide basis, and to use LC-MS/MS to then identify the inhibited proteases. However, this approach is not necessarily quantitative, so the relative abundance of individual proteases would not be reliably determined.

**Protease isotope-coded affinity tag labelling.** A recent technique that does provide quantitative data is based on protein labelling by ISOTOPE-CODED AFFINITY TAGS (ICAT)<sup>20</sup>. Potentially, protease ICAT will prove useful for mass-spectrometric identification and quantitation of proteases that are differentially expressed by cells or tissues in two distinct physiological or pathological conditions, such as normal and arthritic cartilage, or tumours (FIG. 2b). In this approach, an isotope-coded affinity tag — with either eight deuteriums (d8) or eight hydrogens (d0) and a biotin moiety — is used to label reactive cysteines in reduced protein samples or, potentially, to label the active-site residue in cysteine proteases. Identical proteins in two samples can be quantitated and compared by labelling one sample with the ‘heavy’ tag (d8) and the other sample with the d0 tag. After tryptic digestion, avidin-bound peptides are eluted from a nanoscale capillary affinity column, with identical peptides co-eluting and entering the mass spectrometer simultaneously. The relative abundance of the labelled proteases in the two samples is quantitated by mass spectrometric measurement of the 8-Da mass difference that is imparted by the heavy tag. Tandem mass spectrometry is then used to sequence and identify the eluted proteins (FIG. 2b). How can this technique be specifically adapted for proteases? The specificity of ICAT analyses might be improved by substituting covalent protease inhibitors for the cysteine tag to enable active proteases in biological samples to be quantitated by targeted ICAT. Hence, isotope-coded inhibitors that are based on irreversible inhibitors could be used for cysteine and serine proteases, for example.

**Substrate chips.** As the net cleavage of a particular substrate in a cell or tissue determines the biological outcome, valuable information can be gained from analysis of the net proteolytic potential of the entire functional degradome towards a particular substrate, without discriminating between the individual contributions of different proteases. At a system level, it is as important to quantitate whether a particular substrate is cleaved at a particular location, time and rate, as it is to identify the individual proteases that are present. So, a new fourth class of chip that we call ‘substrate chips’, which are based on protein, protein-analogue or peptide-cleavage and the removal of fluorescent end-labelled tags, can be envisaged (FIG. 2c). The development of protease-substrate chips is highly feasible, with the technology being adapted from similar applications for other systems. For example, protein-kinase chips have been developed to assess the ability of kinases to phosphorylate immobilized

**INTRAVITAL IMAGING**  
Visualization of biological processes in intact animals or organ systems.

**ISOTOPE CODED AFFINITY TAGS (ICAT).** ICAT probes have different masses, but are chemically identical. They incorporate a reactive cysteine, a biotin moiety, and eight deuteriums in place of eight hydrogens, and they are used to specifically label, by mass-difference, identical proteins in two separate samples for the identification and semiquantitative comparison of abundance.



**Figure 2 | Degradomics approaches: activity profiling. a** | Protease-specific inhibitors that are coupled to biotin are used to probe for active proteases in complex biological mixtures. After affinity purification on avidin columns the captured proteases can be identified after trypsinization by liquid chromatography–mass spectrometry/mass spectrometry (LC–MS/MS). Alternatively, inhibitors that are arrayed on chips can capture active proteases, which can be identified by matrix-assisted laser desorption-ionization quadrupole time-of-flight (MALDI-Q-TOF) mass spectrometry after *in situ* trypsinization on the chip. **b** | Targeted isotope-coded affinity tag (TICAT) identification of proteases that are expressed by cells or tissues under two different conditions, such as normal and inflamed tissue, tumour and peritumour stroma tissue, or cells in culture that are treated with different growth factors. The active-site cysteine in the cysteine proteases, or elsewhere in other proteases, is reduced and alkylated using deuterium-8 (d8)-labelled tags and is compared with samples in the other tissue conditions that are similarly labelled, but not deuterium tagged (d0). The different proteases in the samples are represented by different colours in the figure. After mixing the two samples, the proteins are digested with trypsin and the biotin-tagged peptides are purified on avidin columns. Sample analysis is therefore greatly simplified by reducing the number of peptides for separation and analysis. For co-eluting, and therefore identical, peptides, the relative abundance of the d0- and d8-labelled peptides is quantified by nanoscale LC–MS, with nanoscale capillary LC–MS/MS used to identify the protease by peptide molecular weight and amino-acid sequence. **c** | Substrate chips could analyse net proteolytic activity towards specific scissile bonds that are present in synthetic peptides, protein substrates or protein analogues that are arrayed on chips. Substrates are fluorescent end-labelled and incubated with the samples, or smaller combinatorial peptides can be labelled by quenched fluorescent fluorochromes. Loss or gain of signal, respectively, at the corresponding array position measures activity toward specific substrates. Although this approach could be successful using a single fluorescent dye for all the arrayed substrates, as new dyes become available, higher density chips will be possible, with the added potential of multiple labelling of proteins, for example, labelling individual subunits.

substrates<sup>21,22</sup>. A system-wide approach should prove valuable for the molecular diagnosis of patients and the molecular analysis of both the catabolic and processing actions of proteases. For example, the degradation of extracellular matrix proteins might be important for the diagnosis of cancer grade, or for the severity of arthritis and other inflammatory diseases. Determination of the tissue or cellular propensity for the proteolytic activation of bioactive molecules, such as *Fas ligand*<sup>23</sup>, *α-defensin*<sup>24</sup>, interleukin (IL)-8 and tumour-necrosis factor- $\alpha$ <sup>12</sup>, or the inactivation of factors such as the chemokine immune mediators<sup>8,25</sup> by MMPs, would provide valuable information on the role of proteases and proteolytically processed factors in the initiation, development and resolution of many types of pathology.

*What benefits will be gained from these studies?*  
Degradomics analysis of the protease systems that are

active in biological samples on a system-wide scale using activity profiling and substrate chips will provide a new level of information that is not available at present. System-wide analysis of whole families of proteases will provide data on both the synergy and functional redundancy of different proteases, and on their relative roles in different tissues or diseases. This should provide insights into how a cell or tissue responds in terms of initiating protease action in different physiological and pathological processes. So, different tissues, or the same tissue in different diseases, might achieve similar net substrate cleavage using different sets of proteases. Understanding this will prove extremely valuable in the design of therapeutic strategies to modulate proteolytic activity *in vivo*. For example, a system-wide analysis that identifies one or a few proteases that are responsible for processing a target substrate would logically direct drug development along highly specific lines to avoid side effects.

Alternatively, these analyses might point to large numbers of proteases that as a group are pathologically crucial, but that are individually redundant. In these instances, such as in the metastatic spread of cancer, it would be more appropriate to develop and administer a broad-spectrum inhibitor, provided that any side effects are manageable.

The proposed approaches to degradomics based on protease chips still have many of the limitations that are common to any approach that is derived from protein analysis by array technology (TABLE 2). These include problems with surface chemistries, nonspecific adsorptions, array instability owing to loss of native protein conformation, and the lack of robust and quantitative detection methods<sup>26</sup>. Despite these limitations, all of the proposed protease chips that have been discussed have either already been tested experimentally or can be adapted from similar applications in other systems. These preliminary results, as well as rapid advances in this field<sup>16,27,28</sup>, show great promise for the widespread application of these approaches to degradomics.

#### Proteases in search of substrates

In addition to strategies that are aimed at profiling the protease complement of each organism or tissue at a given time, there is also a requirement for new approaches to examine the other side of the protease world — the identity of the substrates that are targeted by proteases. Several strategies have proven useful to address this problem over the past decade, and new advances have been reported recently. These approaches can be classified into two broad categories — genetics-based and proteomics-based — although there is significant overlap between them.

There are several recent studies in which *in vivo* protease substrates have been identified genetically<sup>2,24,29,30</sup>. These studies are based on the detection of non-processed substrates that have accumulated in the tissues or body fluids of mutant mice that are deficient in specific proteases. Complementary genetic approaches aimed at generating mutant mice that express noncleavable target substrates have also been reported<sup>31</sup>. However, the generalization of model-system-based genetic approaches for substrate identification might be hampered by the frequent occurrence, in all organisms, of robust proteolytic systems with redundant and compensatory enzymes. Moreover, directly ascribing a specific role to a protease can be problematic if proteolytic cascades are initiated by a target protease that has downstream effector proteases, which, in turn, are responsible for the cleavage of the substrate.

Nevertheless, knockout technology has been successfully used to determine the physiological roles of many proteases. For example, the use of gene-disrupted animals has provided interesting information on the role of caspases in important developmental and homeostatic programmes, including cardiac and neuronal development<sup>32</sup>. *Caspase-1*- and *caspase-11*-null mice do not process and activate IL-1 $\beta$ , IL-1 $\alpha$ , IL-18 and  $\gamma$ -interferon<sup>33</sup>, which indicates a role for these enzymes in cytokine activation. Similarly, gene-targeting of lysosomal

cysteine proteases has confirmed the relevance of *cathepsin S* and *cathepsin L* in antigen presentation<sup>34–36</sup>, and has uncovered new functions for some of these proteases, including a role for cathepsin L in hair-follicle morphogenesis and cycling<sup>37</sup>. Likewise, analysis of *cathepsin-C*-deficient mice has shown an *in vivo* function for this enzyme in the activation of several serine proteases, such as *cathepsin G*, *granzyme A* and *granzyme B*, *neutrophil elastase*, and *chymase*<sup>38</sup>, and, from the analysis of *cathepsin-B*-knockout mice<sup>39</sup>, this protease has been implicated in intracellular trypsinogen activation and the onset of acute pancreatitis.

**Current methods for substrate identification.** The biochemical identification of substrates has been typically carried out in a serial manner using defined substrates; however, the identification of the key *in vivo* substrates for many proteases will require more general proteomics-based methods. With the advent of mass spectrometry, cleaved substrates can be identified on a much larger scale by comparative analysis after the separation — using two-dimensional gel electrophoresis — of complex cell- or tissue-protein extracts that have been exposed to proteases, but with the limitations that were discussed earlier. Other global screening methods include phage-displayed peptide libraries, combinatorial fluorogenic substrate libraries or positional scanning libraries coupled with highly selective affinity labels<sup>40–43</sup>. Peptide-library screens, combined with database searching, provide a means for the rapid identification of potential protein substrates with consensus cleavage sites. Combinatorial peptide libraries, either end-labelled or quench-fluorescent labelled and arrayed on specialized substrate chips, might provide the ability to profile the amino-acid and scissile-bond preferences of individual proteases. However, considerable technological hurdles have to be overcome in the synthesis of combinatorial peptide libraries on chips *in situ*. The preference of active sites for particular amino-acid sequences can also be determined by combinatorial peptide-based inhibitor libraries<sup>42</sup>. Compared with peptide-substrate libraries, a particular advantage of this approach is that specific inhibitors are also identified that, if they are tagged with fluorescent dyes, can be used for activity profiling and histological<sup>16</sup> or intravital imaging. Inhibitors that are coupled to biotin can also be used for affinity purification and LC-MS/MS identification of proteases in biological samples, as discussed above. However, the identification of preferred scissile bonds does not identify substrates, and identification of substrates *in vitro* does not necessarily mean that the substrate is biologically relevant *in vivo*. Temporal and spatial localization of a protease with its substrate is essential for substrate recognition and cleavage *in vivo*. Structural conformation and exposure of potential scissile bonds to the protease are clearly also important for successful catalysis. Substrate-binding sites — called EXOSITES — which are located on specialized domains or modules outside the catalytic domain, and which can be crucial for substrate recognition and cleavage<sup>12</sup>, are also neglected by these methods.

#### EXOSITE

A substrate-binding site that lies outside the active-site cleft of a protease and that is located on specialized substrate-binding modules/domains. Exosites generate diversity in substrate specificity and produce increased rates of cleavage.

#### EXOSITE SCANNING

On the basis of the hypothesis that proteins that bind protease exosites might be substrates, this is a technique for substrate identification that uses recombinant exosite domains as bait (see REF 8).

**INACTIVE CATALYTIC DOMAIN CAPTURE (ICDC).** A technique used for the identification of protease substrates. This uses mutated, proteolytically inactive catalytic domains to capture potential substrates without cleavage and subsequent release.

**Exploiting exosites.** Exosites are essential for the efficient cleavage of complex substrates, such as native triple-helical collagens by MMPs<sup>12</sup>. Exosites have also been found to be crucial for the binding and presentation of simpler molecules — such as chemokines — to the active site of MMPs for cleavage<sup>8,25</sup>. ‘EXOSITE SCANNING’ is a recent substrate-screening method that recognizes the importance of exosites in proteolytic function. The initial strategy that was reported was a new application of the yeast two-hybrid system, which used these ancillary protease exosite domains as bait to screen cDNA libraries for potential binding partners<sup>8</sup> (FIG. 3). The

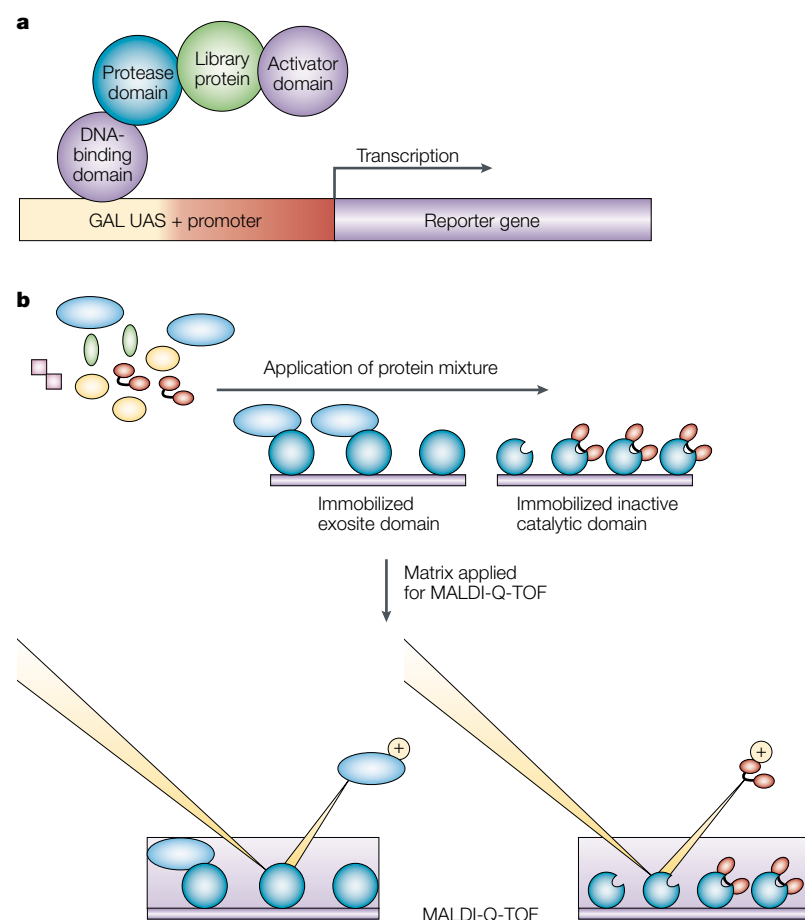
strategy is based on the hypothesis that proteins that bind exosites might be substrates of the protease. Exosite scanning with disulphide-bond-containing domains of extracellular proteases as bait for extracellular substrates was first shown with the identification of unexpected cytokine substrates for well-known proteases, such as for MMP-2, and the discovery of potential new roles for MMPs in the regulation of essential processes like inflammation<sup>8</sup> and HIV infection<sup>25</sup>. This approach should be very useful in the characterization of the growing number of multidomain proteases, including members of the ADAM, ADAMTS and membrane type serine proteinase (MTSP) families. These proteases contain many diverse ancillary domains (mostly of unknown function), and their specific substrates remain largely unknown<sup>13,44</sup>.

It cannot be expected that all exosite domains will fold properly in yeast host cells, nor can it be expected that all substrates will be expressed and form functional complexes with exosite domains in the yeast nucleus, particularly if post-translational modifications are important in substrate recognition. However, protein substrates that are naturally present at low levels *in vivo* — such as growth factors and some cell-surface molecules — can be potentially identified by genetic methods such as this, more readily than by proteomic screens, in which abundance sets sensitivity limits. This strategy has also been adopted for proteomic screens of complex protein mixtures in which immobilized exosite-domain proteins are used to bind potential substrates that are identified by nanoscale capillary LC-MS/MS (FIG. 3). For those proteins that are not expressed effectively in yeast cells, this offers an alternative approach and is actively in development in our laboratories.

**Inactive catalytic domains as substrate baits.** Although our recent data indicate that exosites are crucial for the efficient turnover of certain substrates, or in some cases select against the cleavage of exposed peptide sequences that are rapidly cleaved when present in synthetic peptides, many substrates do not require exosite assistance. Therefore, in what we term ‘INACTIVE CATALYTIC DOMAIN CAPTURE’ (ICDC), we are also using inactive catalytic-domain mutants as baits in yeast two-hybrid screens and as immobilized native-protein baits for genetic and proteomic profiling, respectively, to identify new substrates (FIG. 3). In the absence of cleavage, proteins that are bound to the active site can then be identified. Known MMP substrates — such as laminin, galectin-3 and collagen — and some previously unknown cytokine substrates, have been cloned using genetic ICDC of membrane type 1-MMP (MT1-MMP) in yeast two-hybrid screens. Similarly, MMP-2 ICDC proteomic screens of cell and tissue extracts have also confirmed the utility of this technique.

#### Changing views of proteases

Approaches such as those that are described above and new technologies that are now being developed in our laboratories and by others, which use variants of ICAT and innovative techniques for sample preparation before



**Figure 3 | Exosite scanning and inactive catalytic domain capture (ICDC): substrate scanning by yeast two-hybrid screens and immobilization techniques.** **a** | In exosite scanning and ICDC, the yeast two-hybrid baits are the protease exosite domain or inactive catalytic domain, respectively (labelled ‘Protease domain’ in the figure), which are expressed in yeast as a fusion protein with the GAL4 DNA-binding domain (‘DNA-binding domain’). Co-transfected into the cells is a second plasmid that expresses the GAL4 activation domain (‘Activator domain’) to which cDNA from a library is fused (‘Library protein’). On interaction of the protease exosite with a library protein in the yeast nucleus, the DNA-binding and -activation domains are brought into proximity, which activates the reporter gene. In yeast two-hybrid ICDC, substrates that are bound to the mutated active site of the protease catalytic domain cannot be cleaved, which means that a positive transcription signal can be initiated. Interactor proteins that are identified are subsequently assessed biochemically for binding and cleavage by the protease. **b** | Exosite scanning and ICDC can also be used in proteomics screens, in which the exosite domain or inactive catalytic domain are immobilized on columns, solid matrices or chips to capture potential protein substrates. The column-bound substrate proteins can be eluted, trypsin digested and identified by tandem mass spectrometry or, for solid matrices and chips, digested *in situ* and identified directly by MALDI-Q-TOF mass spectrometry. MALDI-Q-TOF, matrix-assisted laser desorption-ionization quadrupole time-of-flight; UAS, upstream-activating sequence.

tandem mass spectrometry, promise to uncover many new and unexpected substrates and functions for proteases. Even now, the identification of new protease substrates is redefining our outlook on the *in vivo* roles of certain proteases. Over the past 10 years, the identification of an enormous array of cytoplasmic and nuclear-protein substrates of the intracellular cysteine caspases<sup>33</sup> has provided convincing evidence for the exquisite degree of control that can be executed by proteolytic processing in pivotal cellular processes, such as apoptosis. In addition to the activation of downstream caspases, the highly selective scissile-bond specificity that is shown by caspases is manifested by limited proteolysis and constitutive activation of several kinases, including several protein kinase C isoforms and the mitogen-activated protein kinase/extracellular-signal regulated kinase (MAPK/ERK) kinase kinase **MEKK-1**. Other targets, however, are inactivated by caspase processing, including **p53** and retinoblastoma susceptibility protein (**pRb**), focal adhesion kinase (**FAK**), the transcription factors nuclear factor- $\kappa$ B and signal transducer and activator of transcription-1 (**STAT-1**), and the cytoplasmic proteins  **$\beta$ -catenin** and the anaphase-promoting complex (APC) gene product<sup>33,45</sup>.

As another example of a change in the view of proteases, the MMPs have traditionally been thought to be the effectors of extracellular matrix catabolism in normal remodelling and pathological tissue destruction. This view was modified about 10–15 years ago, when the influence of MMPs on cell behaviour was recognized and ascribed to the disruption of the pericellular matrix and cognate integrin contacts that results in alterations in cellular signalling<sup>2</sup>. MMP production by cancer cells was also thought to lead to tumour-cell invasion and metastasis. However, it is now widely recognized that most MMP expression in cancer is from the reactive tissue stroma. Indeed, a recent example of a new role for MMPs in cancer is where selective MMP production by cancer cells drives selection for tumour-cell resistance to apoptosis<sup>23</sup>. This example is indicative of the marked revision in our concepts of the roles of MMPs in cancer and other pathologies. Moreover, the growing awareness that MMPs cleave a wide range of bioactive substrates (see online link to **Bioactive MMP substrates**) *in vivo*<sup>8,24,29</sup>; the relatively minor phenotypic alterations in normal matrix turnover in most MMP-knockout mice; the general absence of effects on normal connective-tissue remodelling during long-term exposure to synthetic MMP inhibitors in animal models and in clinical trials with MMP-inhibitors; and the existence of efficient intracellular pathways of extracellular matrix turnover<sup>46</sup> indicate that MMPs are not important for normal matrix remodelling as generally assumed. Indeed, the general lack of *in vivo* evidence reported for extracellular-matrix-protein cleavage by MMPs, even in pathology, indicates other important roles for these proteases. We suggest that bioactive substrate processing by MMPs, with its resulting regulatory effects on the cellular responses of the tissue and host defence cells, in physiological and pathological processes will prove to be more important overall than their potential catabolic roles.

Hence, as new substrates are described for all proteases, it is anticipated that the biological roles for several protease families will be viewed quite differently in the years ahead, with many showing new potential as drug targets to treat disease.

### Outlook

The completed genome projects have provided a good overview of the composition and organization of the proteolytic machinery that is used by different organisms. However, not all proteases have been discovered yet, nor have their *in vivo* substrates and functions been identified. Indeed, many human protease genes that have been predicted from genome computer searches remain to be cloned and the proteolytic activity of their gene products confirmed. An important challenge for the future is to solve the structure of, and to assign physiological and pathological functions to, the proteins that are encoded by these new protease genes. Other proteins that have significant identity to proteases but that lack catalytic residues (such as several members of the ADAMs family) — which we propose might function as natural highly specific antagonists by binding substrates without cleavage, thereby masking scissile bonds — will also need to be characterized to determine whether they contribute to the fine-tuning of protease activity *in vivo*. Furthermore, new structural designs for proteases that would not be detected by homology-based scanning or cloning methods might remain hidden in our genome<sup>3</sup>.

Genomic-based approaches to degradomics will be essential for determining the basis of the increasing number of genetic diseases that are recognized as being caused by mutations in protease loci. These genomic approaches will also be essential for the identification of putative single-nucleotide polymorphisms in protease genes that might confer increased susceptibility, or resistance, to complex diseases<sup>47</sup>. In parallel, the development of strategies for intravital imaging of active proteases, and the generation of new animal models with a gain or loss of protease function, should contribute to the evaluation of the role of individual proteases or protease families in health and disease. Chemical proteomics will be a powerful complement to determine the roles of protease families as a proteolytic system or as specific proteases individually in culture or *in vivo*. It will also be essential to define the regulatory and functional intersections between proteolytic systems and other important networks, such as those involved in signalling, and to identify new protease circuits that function in distinct cells and biological processes.

No single screening technology will provide complete genomic or proteomic coverage to identify the entire protease degradome or the substrate degradome of a protease, and each technology has its own specific advantages and niche. However, emerging technologies for high-throughput assays, including protease chips, will greatly assist in generating the necessary data for profiling proteases in proteomes and for the identification of substrates and inhibitors, thereby defining new therapeutic targets. Similarly, determining the three-dimensional structures of proteases by new high-throughput

X-ray crystallographic approaches<sup>48</sup>, or by using a combination of homology modelling and *ab initio* predictions that aim to predict the three-dimensional structure of a protein from its amino-acid sequence<sup>49</sup>, will be valuable resources for rational inhibitor design. Again, there are many challenges ahead, especially in the case of membrane proteases that are responsible for processes

such as regulated intramembrane proteolysis<sup>50</sup>. By combining our knowledge at present with new and creative strategies, degradomics promises to provide answers to questions about the expression, control and *in vivo* roles of proteases — a large group of enzymes that profoundly influence cell behaviour, survival and death in all living organisms with precision.

- Barrett, A. J. *et al.* (eds) *Handbook of Proteolytic Enzymes* (Academic, London, 1998).  
**A 'must-have' volume for any protease lab that comprehensively lists and describes numerous proteases.**
- Sternlicht, M. D. & Werb, Z. How matrix metalloproteinases regulate cell behavior. *Annu. Rev. Cell Dev. Biol.* **17**, 463–516 (2001).
- Urban, S., Lee, J. R. & Freeman, M. *Drosophila* rhomboid-1 defines a family of putative intramembrane serine proteases. *Cell* **107**, 173–182 (2001).
- Balbin, M. *et al.* Identification and enzymatic characterization of two diverging counterparts of human interstitial collagenase (MMP-1) expressed at sites of embryo implantation. *J. Biol. Chem.* **276**, 10253–10262 (2001).
- Deussing, J. *et al.* Identification and characterization of a dense cluster of placenta-specific cysteine and peptidase genes and related genes on mouse chromosome 13. *Genomics* **79**, 225–240 (2002).
- Yousef, G. M. & Diamandis, E. P. The new human tissue kallikrein gene family: structure, function, and association to disease. *Endocrine Rev.* **22**, 184–204 (2001).
- Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
- McQuibban, G. A. *et al.* Inflammation dampened by gelatinase A cleavage of monocyte chemoattractant protein-3. *Science* **289**, 1202–1206 (2000).  
**The first description of exosite scanning and the use of the yeast two-hybrid system to identify new extracellular protease substrates.**
- MacBeath, G. & Schreiber, S. L. Printing proteins as microarrays for high-throughput function determination. *Science* **289**, 1760–1763 (2000).  
**This study shows the usefulness of protein-based microarray chips for functional analyses.**
- De Wildt, R. M. T., Mundy, C. R., Gorick, B. D. & Tomlinson, I. M. Antibody arrays for high-throughput screening of antibody-antigen interactions. *Nature Biotechnol.* **18**, 989–994 (2000).
- Fung, E. T., Thulasiraman, V., Weinberger, S. R. & Dalmasso, E. A. Protein biochips for differential profiling. *Curr. Opin. Biotechnol.* **12**, 65–69 (2001).
- Overall, C. M. Matrix metalloproteinase substrate binding domains, modules and exosites: overview and experimental strategies. *Meth. Mol. Biol.* **151**, 79–120 (2001).
- Cal, S. *et al.* Cloning, expression analysis and structural characterization of seven novel human ADAMTSs, a family of metalloproteinases with disintegrin and thrombospondin-1 domains. *Gene* **283**, 49–62 (2002).
- Liu, Y., Patricelli, M. P. & Cravatt, B. F. Activity-based protein profiling: the serine hydrolases. *Proc. Natl Acad. Sci. USA* **96**, 14694–14699 (1999).  
**Together with references 15 and 16, this paper shows the feasibility of activity-based profiling of proteases that are present in complex mixtures.**
- Greenbaum, D., Medzihradsky, K. F., Burlingame, A. & Bogyo, M. Epoxide electrophiles as activity-dependent cysteine protease profiling and discovery tools. *Chem. Biol.* **7**, 569–581 (2000).
- Greenbaum, D. *et al.* Chemical approaches for functionally probing the proteome. *Mol. Cell. Proteomics* **1**, 60–68 (2002).  
**A key paper in the development of active-site functional probes for cysteine proteases. This shows that BODIPY-dye-conjugated protease inhibitors can be used to specifically label and identify proteases in mixtures without prior separation. This also shows their use in histological localization, and in the identification of active-site peptide-bond specificity.**
- Weissleder, R., Tung, C. H., Mahmood, U. & Bogdanov, A. Jr. *In vivo* imaging of tumors with protease-activated near-infrared fluorescent probes. *Nature Biotechnol.* **17**, 375–378 (1999).
- Bremer, C., Tung, C. H. & Weissleder, R. *In vivo* molecular target assessment of matrix metalloproteinase inhibition. *Nature Med.* **7**, 743–748 (2001).
- Weissleder, R. Scaling down imaging: Molecular mapping of cancer in mice. *Nature Rev. Cancer* **2**, 1–8 (2002).  
**A comprehensive review of the current status of various new and emerging techniques for intravital imaging of cancer proteases that use labelled substrate analogues or inhibitors.**
- Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.* **17**, 994–999 (1999).  
**One of several key papers from the Abersold laboratory on a new mass-spectrometric proteomic technique that identifies and semiquantitatively measures differences in the abundance of the same protein in two samples.**
- Williams, D. M. & Cole, P. A. Kinase chips hit the proteomic era. *Trends Biochem. Sci.* **26**, 271–273 (2001).
- Zhu, H. *et al.* Analysis of yeast protein kinases using protein chips. *Nature Genet.* **26**, 283–289 (2000).
- Fingleton, B., Vargo, T., Crawford, H. C. & Matrisian, L. M. Matrilysin expression selects for cells with reduced sensitivity to apoptosis. *Neoplasia* **3**, 459–468 (2001).
- Wilson, C. L. *et al.* Regulation of intestinal  $\alpha$ -defensin activation by the metalloproteinase matrilysin in innate host defense. *Science* **286**, 113–117 (1999).
- McQuibban, G. A. *et al.* Matrix metalloproteinase activity inactivates the CXc chemokine stromal cell-derived factor-1. *J. Biol. Chem.* **276**, 43503–43508 (2001).
- Mitchell, P. A perspective on protein microarrays. *Nature Biotechnol.* **20**, 225–229 (2002).  
**Together with references 27 and 28, this paper presents the current status, recent developments and limitations of protein- and peptide-microarray protein chips.**
- Heng, Z. *et al.* Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105 (2001).
- Houseman, B. T., Huh, J. H., Kron, S. J. & Mrksich, M. Peptide chips for the quantitative evaluation of protein kinase activity. *Nature Biotechnol.* **20**, 270–274 (2002).
- Liu, Z. *et al.* The serpin  $\alpha$ 1-proteinase inhibitor is a critical substrate for gelatinase B/MMP-9 *in vivo*. *Cell* **102**, 647–655 (2000).
- Pendas, A. M. *et al.* Defective prelamin A processing and muscular and adipocyte alterations in Zmpste24 metalloproteinase-deficient mice. *Nature Genet.* **31**, 94–99 (2002).
- Zhao, W., Byrne, M. H., Wang, Y. & Krane, S. M. Osteocyte and CD4 T cell selection in the thymus. *Science* **280**, 450–453 (1998).
- Nakagawa, T. Y. *et al.* Impaired invariant chain degradation and antigen presentation and diminished collagen-induced arthritis in cathepsin S null mice. *Immunity* **10**, 207–217 (1999).
- Turk, V., Turk, B. & Turk, D. Lysosomal cysteine proteases: facts and opportunities. *EMBO J.* **20**, 4629–4633 (2001).
- Roth, W. *et al.* Cathepsin L deficiency as molecular defect of *furless*: hyperproliferation of keratinocytes and perturbation of hair follicle cycling. *FASEB J.* **14**, 2075–2086 (2000).
- Pam, C. T. N. & Ley, T. J. Dipeptidyl peptidase I is required for the processing and activation of granzymes A and B *in vivo*. *Proc. Natl Acad. Sci. USA* **96**, 8627–8632 (1999).
- Halangck, W. *et al.* Role of cathepsin B in intracellular trypsinogen activation and the onset of acute pancreatitis. *J. Clin. Invest.* **106**, 773–781 (2000).
- Deng, S. J. *et al.* Substrate specificity of human collagenase 3 assessed using a phage-displayed peptide library. *J. Biol. Chem.* **275**, 31422–31427 (2000).
- Harris, J. L. *et al.* Rapid and general profiling of protease specificity by using combinatorial fluorogenic substrate libraries. *Proc. Natl Acad. Sci. USA* **97**, 7754–7759 (2000).  
**Together with reference 43, this paper describes the identification of preferred protease scissile bonds in iterative peptide-library-based approaches.**
- Nazif, T. & Bogyo, M. Global analysis of proteasomal substrate specificity using positional-scanning libraries of covalent inhibitors. *Proc. Natl Acad. Sci. USA* **98**, 2967–2972 (2001).
- Turk, B. E., Huang, L. L., Piro, E. T. & Cantley, L. C. Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. *Nature Biotechnol.* **19**, 661–667 (2001).
- Hooper, J. D., Clements, J. A., Quigley, J. P. & Antalis, T. M. Type II transmembrane serine proteases. Insights into an emerging class of cell surface proteolytic enzymes. *J. Biol. Chem.* **276**, 857–860 (2001).
- Nicholson, D. W. & Thornberry, N. A. Caspases: killer proteases. *Trends Biochem. Sci.* **22**, 299–306 (1997).
- Sodek, J. & Overall, C. M. In *The Biological Mechanisms of Tooth Eruption and Root Resorption* (ed. Davidovitch, Z.) 303–311 (EBSCO Media, Birmingham, Alabama, 1988).
- Zhu, Y., Spitz, M. R., Lei, L., Mills, G. B. & Wu, X. A single nucleotide polymorphism in the matrix metalloproteinase-1 promoter enhances lung cancer susceptibility. *Cancer Res.* **61**, 7825–7829 (2001).
- Thornton, J. Structural genomics takes off. *Trends Biochem. Sci.* **26**, 88–89 (2001).
- Bonneau, R. & Baker, D. *Ab initio* protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 173–189 (2001).
- Brown, M. S., Ye, J., Rawson, R. B. & Goldstein, J. L. Regulated intramembrane proteolysis: a control mechanism conserved from bacteria to humans. *Cell* **100**, 391–398 (2000).

#### Acknowledgements

C.M.O. is supported by a Canada Research Chair in Metalloproteinase Biology, and by grants from the Canadian Institutes for Health Research, the National Cancer Institute of Canada, the Protein Engineering Network of Centers of Excellence and the Canadian Arthritis Network of Centers of Excellence. C.L.-O. is supported by the Comision Interministerial de Ciencia y Tecnologia, Spain, and by the European Union.

#### Online links

##### DATABASES

The following terms in this article are linked online to:

**Flybase:** <http://flybase.bio.indiana.edu/>  
Rhomboid  
**LocusLink:** <http://www.ncbi.nlm.nih.gov/LocusLink>  
 $\alpha$ -defensin | Fas ligand | galectin-3 | IL-1 $\alpha$  | IL-1 $\beta$  | IL-8 | IL-18 |  $\gamma$ -interferon | MEK-1 | MMP-2 | MT1-MMP | tumour-necrosis factor- $\alpha$   
**Swiss-Prot:** <http://www.expasy.ch/>  
Caspase-7 | caspase-11 |  $\beta$ -catenin | cathepsin B | cathepsin C | cathepsin G | cathepsin L | cathepsin S | chymase | FAK | GAL4 | granzyme A | granzyme B | neutrophil elastase | p53 | pRb | STAT-1

##### FURTHER READING

**Bioactive MMP substrates:** <http://www.clip.ubc.ca/mmp.shtm>  
**Chris Overall's laboratory:** <http://www.clip.ubc.ca>  
**Encyclopedia of Life Sciences:** <http://www.els.net/>  
confocal microscopy | liquid chromatography  
**MEROPS version 5.7:** <http://www.merops.ac.uk>  
**Access to this interactive links box is free online.**