Proteases' prime targets revealed

Scott L Diamond & Doron Greenbaum

Mass spectrometry analysis of proteome-derived peptide libraries enables high-throughput determination of protease substrate specificity.

Much like the moving tumblers in a lock, the side chains in the active site of a protease interrogate proteins for cleavable peptide sequences. Most techniques for determining the substrate specificity of a particular protease reveal information only about residues on the non-prime side of the cleavage site, leaving the prime side unexplored. In this issue, Schilling and Overall[1] report a method, called proteomic identification of protease cleavage specificity (PICS), that uses proteome-derived peptide libraries for high-throughput determination of prime-side preferences of proteases (**Fig. 1a**). The authors apply the method to nine human proteases representing four mechanistic classes of proteases, leading to the identification of 3,691 cleavage sites. Rapid, large-scale identification of protease substrate specificity by PICS should substantially accelerate the discovery of physiological protease targets.

The PICS method rests on three innovations (**Fig. 1b**). The first is to use a protease digest of a cellular lysate as the source of peptide sequences, ensuring a broad diversity of potential substrates. The second is the deployment of protecting groups to mask primary amines and thiols such that terminal amines generated by the protease of interest can be selectively biotinylated. Third, the prime-side fragments are isolated by affinity chromatography and identified by tandem mass spectrometry. These fragments are then matched to their corresponding non-prime-side partners by automated database searching, allowing cleavage specificity on both sides of the scissile bond to be determined.

Knowledge of protease substrate specificity is critical to characterizing structure-function relationships, identifying substrates for high-throughput screening and designing inhibitors. For instance, using X-ray crystallography to determine critical prime-side binding sites for cathepsin K, a cysteine protease involved in bone resorption, Thompson *et al.* developed active site–spanning inhibitors that provided valuable leads for clinical candidates[2]. Similarly, substrate profiling identified clinically relevant caspase inhibitors[3]. Although PICS determines the breadth of protease substrate specificity and not the physiological targets of proteases, the results of PICS profiling could provide tractable lists of putative substrates of biological relevance.

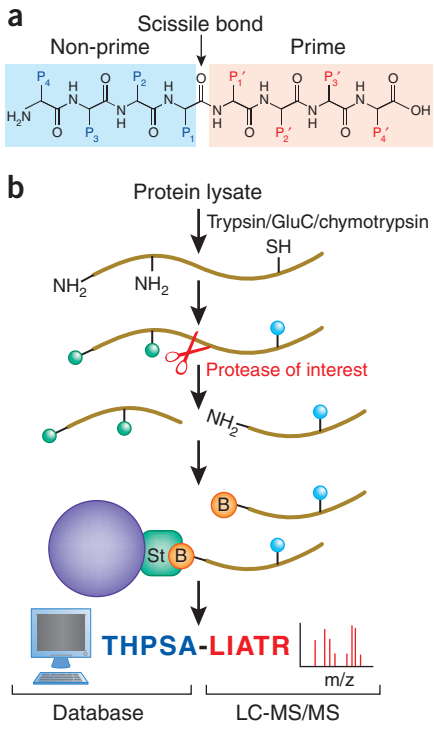

Figure 1 Proteomic identification of protease cleavage sites (PICS) determines the substrate specificity of a protease using diverse libraries of potential substrates derived from natural proteomes. (**a**) Numbering sequence for defining the non-prime (P_4-P_3-P_2-P_1) and prime (P_1'-P_2'-P_3'-P_4') amino acid residues adjacent to the cleavage site. (**b**) Peptide libraries (each containing $>10^6$ peptides) are generated from whole proteomes by nonspecific proteolysis of cell extracts, dimethylation of N termini and amino groups of lysine residues (green) and carboxyamidomethylation of sulfhydryl groups (blue). The peptide libraries are exposed to the protease of interest, and the new N termini are reacted with biotin (B). The tagged peptides are isolated by streptavidin (St) affinity chromatography and identified by liquid chromatography–tandem mass spectrometry (LC-MS/MS). The cognate non-prime-side sequence is deduced by database searching.

Although various synthetic methods have been developed to create peptide diversity for the purpose of mapping protease active sites[4], and direct synthesis has been used to create libraries of individual peptide sequences or mixtures for positional scanning[5,6], the requirement for a fluorogenic leaving group limits these approaches to the study of the non-prime residues of the cleavage site. Two factors work against interrogation of the prime side of the scissile bond: first, the combinatorial explosion for the four positions that bracket the scissile bond (20^8 sequences) and, second, the use of a fluorogenic leaving group distal to the scissile bond. To get around these limitations, dual fluorescence labels are needed to generate a dequenched fluorescent fragment upon cleavage. As this synthesis is more challenging, the tendency is to use highly focused sequences that diversify only a few positions[7,8].

In contrast to these older methods, PICS allows deep coverage of peptide space without the limitations imposed by fluorescence detectors. A simple trypsin digest of a proteome provides $>10^6$ peptide sequences per proteome. As several protein lysates can be used and they can each be digested with either trypsin, chymotrypsin or *Staphylococcus aureus* protease V8 (endoproteinase Glu-C), the various limitations typical of MS-based proteomics, such as undersampling, can be reduced. Another bonus associated with sequences derived from digested lysates is the retention of cysteine (albeit chemically modified); cysteine is typically avoided in synthetic libraries to prevent peptide disulfide cross-linking. Phage display of substrate sequences[9] allows discovery of cleavable sequences, but this method is labor intensive, requiring iterative cycles of cloning and sequencing, and it does not identify the position of the scissile bond. For comparison, substrate phage display analysis of matrix metalloprotease-2 provided only 50 cleavage sequences, as compared with the >300 sequences provided by PICS[10].

Scott L. Diamond and Doron Greenbaum are in the Departments of Chemical Engineering and Bioengineering, University of Pennsylvania, 1024 Vagelos Research Laboratories, Philadelphia, Pennsylvania 19104-6383, USA. e-mail: sld@seas.upenn.edu

Other important advantages of the PICS method are the speed of analysis (~2 hours per sample for the tandem mass spectrometry analysis) and the ability to identify large numbers of cleavage sites. Consider, for instance, that whereas the MEROPS peptidase database contains <8,000 cleavage sites for >2,400 proteases, PICS analysis identified almost half as many cleavage sites for the nine proteases examined[1]. Moreover, in contrast with positional scanning, in which pooled peptide mixtures give average amino acid preferences, PICS can provide information about 'subsite cooperativity', whereby occupancy of one subsite by a particular amino acid changes the preferences of a neighboring subsite.

As PICS requires only small quantities of a purified protease, it should be especially useful for studying the vast number of uncharacterized proteases from viruses and parasites responsible for diseases such as malaria. Although this potential is not explored in the current paper[1], proteome-derived peptide libraries may provide a unique starting point for other methods of studying protein-protein interactions and protein-modifying enzymes. Peptide libraries subjected to chemical modifications distinct from those used in PICS could thus enhance our understanding of various aspects of signal transduction, kinome function, apoptosis and proteasome pathways.

1. Schilling, O. & Overall, C.M. *Nat. Biotechnol.* **26**, 685–694 (2008).
2. Thompson, S.K. *et al. Proc. Natl. Acad. Sci. USA* **94**, 14249–14254 (1997).
3. Thornberry, N.A. *et al. J. Biol. Chem.* **272**, 17907–17911 (1997).
4. Diamond, S.L. *Curr. Opin. Chem. Biol.* **11**, 46–51 (2007).
5. Backes, B.J. *et al. Nat. Biotechnol.* **18**, 187–193 (2000).
6. Harris, J.L. *et al. Proc. Natl. Acad. Sci. USA* **97**, 7754–7759 (2000).
7. Stennicke, H.R. *et al. Biochem. J.* **350**, 563–568 (2000).
8. Petrassi, H.M. *et al. Bioorg. Med. Chem. Lett.* **15**, 3162–3166 (2005).
9. Matthews, D.J. *et al. Science* **260**, 1113–1117 (1993).
10. Chen, E.I. *et al. J. Biol. Chem* **277**, 4485–4491 (2001).

Stable transgenes bear fruit

Ajay Kohli & Paul Christou

Analysis of the transgenic papaya genome sequence suggests that transgenes generally stay put following integration and can achieve stable expression level from generation to generation.

Enhanced understanding of the mechanisms of transgene insertion and rearrangement in plant chromosomes is essential not only for the routine production of transgenic loci, but also ultimately to spur the development of directed transgene integration approaches. The availability of the complete genome sequence of the SunUp variety of papaya recently reported in *Nature* by Ming *et al.*[1] represents a major step in this regard. This is the third complete genome sequence of a multicellular plant to be published (*Arabidopsis* and rice were the other two) and the first ever genome sequence of a transgenic organism. From a biosafety perspective, the papaya sequencing project also provides the first definitive molecular evidence against *in situ* transgene rearrangements, one of the main suspected causes of 'transgene instability'. Although no comparative sequencing over multiple generations is available, the fact that the transgene remains structurally and functionally intact in this distant descendent of the original integration event is convincing proof that transgenes generally become fixed elements of the plant genome and can achieve a consistent and predictable expression level from generation to generation.

Papaya is a tropical fruit that was almost wiped out in Hawaii by the papaya ringspot virus a decade ago. The onslaught of the virus on the island prompted pioneering efforts in the early 1990s to create a transgenic variety of papaya resistant to ringspot using ballistic methods. These efforts resulted in the creation of two virus-resistant transgenic cultivars 'SunUp' and 'Rainbow'. The former produces a red-fleshed fruit that expresses the coat protein gene of an attenuated mutant of the virus, conferring resistance via post-transcriptional gene silencing.

Ajay Kohli is at the Institute for Research on Environment & Sustainability (IRES), Devonshire Building, University of Newcastle upon Tyne, Newcastle, NE1 7RU, UK. Paul Christou is at ICREA, Universitat de Lleida, PVCF, Av Alcalde Rovira Roure, 191, E-25198, Lleida, Spain.
e-mail: christou@pvcf.udl.es

Ming *et al.*[1] used a whole genome shotgun approach to facilitate the sequencing of >90% of the euchromatic papaya genome, including 92.1% of previously identified expressed sequence tags, 92.4% of known genetic markers and the genomic sequences surrounding the integrated transgene DNA. The sequences flanking the transgene appear very similar to those around 'natural' DNA integration events, such as the occasional integration of chloroplast DNA fragments into the papaya genome, as reported in the same publication[1], and tobacco genome, as reported previously[2]. This supports research on both direct DNA transfer (mostly particle bombardment) and *Agrobacterium*-mediated transformation, which has shown that transgene integration is the consequence of a natural process carried out by enzymes involved in DNA break and repair[3].

The SunUp genome sequence lends support to many aspects of the current model for transgene integration in plants (**Fig. 1**) and suggests that identical methods are involved in artificial and 'natural' DNA integration events. These include the tendency for exogenous DNA sequences to undergo recombination, rearrangement and truncation before or during integration (but rarely after integration), the tendency for exogenous DNA to integrate at AT-rich sites resembling topoisomerase recognition sequences, and the tendency for junction regions to show evidence of microhomology and nontemplated DNA synthesis (filler DNA)[3]. All these processes occur regardless of the transformation method because they rely on enzymes endogenous to the nucleus of the host cell.

The initial stages of transgene integration are characterized by complex rearrangements of input sequences resulting from a combination of homology-dependent and homology-independent processes stimulated by the presence of a high concentration of free DNA ends (**Fig. 1a**). This often results in the formation of transgene concatemers that may comprise any number of genes, from fewer than 10 to more than 100, some of which are complete and some truncated (**Fig. 1b**). These structures compete with individual genes for integration sites, which are thought to reflect the positions of naturally occurring DNA nicks and breaks[4]. DNA repair complexes in the vicinity of such breaks are thought to incorporate exogenous DNA into the repair, generating complex structures that need to be resolved before the DNA duplex can be sealed. The frequent occurrence of topoisomerase sites near both integrated transgenes and natural integration sites suggests that this enzyme plays an important role in the