A genomic perspective on human proteases as drug targets

Christopher Southan

Of the ~400 known human proteases, ~14% are under investigation as drug targets. Although the total is certain to rise during the finishing phase of the human genome project, the initial annotation of the ~30,000 human proteome set includes ~500 proteases. Bioinformatic analysis can now be performed on complete human protease families and will soon include comparisons with mice and fish. New sequences will require evaluation of their function in normal physiology and human disease. By revealing details such as splice variants and population polymorphisms, genomic sequence information will have a central role in the validation of protease drug targets.

Christopher Southan
Head of Computational Biology
Gemini Genomics (UK)
162 Science Park
Milton Road
Cambridge
UK CB4 0GH
tel: +44 (0)1223 435342
fax: +44 (0)1223 435301
e-mail: chris.southan@gemini-genomics.com

▼ The advantages of human proteases as potential drug targets are well established[1,2]. Reasons for their popularity include the accumulated data on their enzymology, high-throughput assays, biochemistry, physiology, pathology, 3D structures, small-molecule inhibitors and endogenous inhibitory proteins. In addition, the substantial body of work on the development of inhibitors of viral, bacterial and protozoan parasite proteases has an extensive methodological overlap with human proteases, especially where homology has been established (Box 1). Information regarding protease classification and nomenclature can be found in a recent compendium of protease families[3]. The number of known human proteases reported to be under investigation as drug targets has roughly doubled between 1998 and 2000. Expressed as a proportion of known human proteases, this represents an increase from 10% to 14% (Box 1).

By revealing all potential proteins, the Human Genome Project will have a major effect on both academic and pharmaceutical protease research. The recent landmark genome paper included the first compilation of an Integrated Protein Index (IPI) of 31,778 sequences[4]. This number is an approximate equal split between known mRNAs and predicted reading frames from genomic sequence for which there is supporting evidence. However, several processes are under way that will increase the quality of this proteome set and could increase the final number of proteases that can be detected by sequence similarity.

The first of these is the conversion of more data to the finishing standard, which stands at just over 40% in April 2001 (http://www.ebi.ac.uk/genomes/mot/). This extends genomic assembly and the assignment of known genes, and improves the prediction of novel genes. Second, several projects are under way to increase the level of mammalian mRNA transcript coverage. These include the large-scale production of human, mouse and bovine expressed sequence tags (ESTs) (including subtractive strategies to detect low-abundance transcripts) and full-length cDNAs. The literature and database entries demonstrate the use of both public and proprietary EST collections for detecting and cloning proteases. The availability of genomic data expedites this process by confirming the exon matches of human or other mammalian ESTs. Finally, the recently announced acceleration of both mouse and fish genomic sequencing will enhance comparative gene delineation between the three vertebrates.

Protease numbers

The following questions can be posed when considering the proteases in a genomic context. First, how many human proteases are present in the current transcript data? Second, how many will we be able to find when the genome is completed? Third, will the mechanistic class, domain combinations or sequence-family distribution shift significantly as we

1359-6446/01/$ – see front matter ©2001 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(01)01793-7

Box 1. Data sources for proteases[a]

Ensembl
Automated annotation of 25,790 homology-confirmed predicted gene products from all human genomic data up to Febuary 2001. Includes the full genomic context and InterPro matches for identified protease gene products (http://www.ensembl.org/).

Handbook of Proteolytic Enzymes
Detailed compendium of 569 protease families[b].

InterPro
Automated domain and family annotation, with graphic displays, for 30,773 human proteins from the human genome (http://www.ebi.ac.uk/proteome/).

Locus Link
Non-redundant collection of 15,909 human transcripts, most of which have links to the public genome map (http://ncbi.nlm.nih.gov/LocusLink/).

MEROPS
Expertly curated protease database with the internationally accepted family protease family classifications of 7337 sequences. Includes the mRNA, expressed sequence tag, protein and genomic links for 391 human proteases (http://www.merops.co.uk/).

Pharma Projects
Subscription drug pipeline database that lists >55 disclosed human proteases under pharmaceutical investigation. Includes descriptions of 219 inhibitors under active development (http://www.pjbpubs.co.uk/pharma/).

SwissProt
Index of entries for 214 human proteases with extensive annotation and links (http://www.expasy.ch/cgi-bin/lists?peptidas.txt).

[a]The numbers listed here were obtained from these sources between January and April 2001.
[b]Barret, A.J. *et al.*, eds (1998) Handbook of Proteolytic enzymes, Academic Press.

approach the genomic total? Fourth, what proportion of novel or known sequences will prove to be inactive homologues or pseudogenes? Fifth, what proportion will show splice variants that alter the protein sequence?

The first of these questions can be answered from the MEROPS peptidase database, which contained 405 curated sequences selected from the primary databases up to March 2001 (Ref. 5). The small number of published full-length patent sequences describing proteases that are in neither the GenBank patent nor the primate divisions would bring this total up to ~420. One approach to answering the second question would simply be to ratio these 405 sequences against the non-redundant human protein total, currently 15,712, whereby proteases would represent 2.9% of all human transcripts. However, most full-length mRNA entries in GenBank were actively selected for cDNA cloning, which consequently introduces a historical 'interest bias' towards proteases.

Box 2. GenBank accession numbers and genomic locations of recently reported human homologues of microbial proteases

- Sialoglycoprotease, two paralogues, gbAJ271669, 14q11.2 and gbAJ295148, 2q32.2
- Serine β lactamase, the mouse sequence is gbAF317900, the human orthologue is 15q22.2
- ATP-dependent metalloprotease, gbAJ132637, 10p14
- Metallo β lactamase, gbD83198, 19q13.2
- Polypeptide deformylase, gb AF322879, 16q23.1
- Bacterial heat-shock HtrA proteases, two paralogues gbY07921, 10q25.3 and gbAF141305, 2p.13
- Pyroglutamyl–peptidase, gbAJ278828, 19p13.2

A preliminary Genome Ontology annotation of an updated human proteome set of 30,585 sequences has classified 498 (1.6%) of these as proteases (Box 1). This is below the 1.8% estimate made recently, largely based on eukaryotic model organism data[6].

However, the yeast, worm and fly genomes have shown a slight but distinct upwards trend in the protease totals since their genomes were first annotated. The reasons for this include revisions of open reading frames, publications describing novel proteases and the application of more sensitive searching methods for homology assignments. A recent example of the last reason, using the PSI-BLAST algorithm, led to the discovery of a novel superfamily of predicted cysteine proteases from both eukaryotes and viruses that includes no less than five *Caenorhabditis elegans* and four human paralogues[7].

Another source of new human proteases is the discovery of eukaryotic representatives of protease families that had hitherto only been reported as microbial orthologues (Box 2). Novel proteases are also expected to be discovered as a consequence of structural genomics projects[8]. These high-throughput approaches to protein structure determination will expand the known fold space. Structure-versus-structure searches will then be able to recognize divergent relationships between known protease fold architectures that cannot be detected by sequence similarity. Even when this process approaches completion, it seems safe to predict that, of the 20–30% of proteins revealed by vertebrate genome projects that are functionally unclassifiable, at least some will turn out to be proteolytic enzymes that have novel structures, oligomeric complexes or catalytic

mechanisms. For example, it was suggested that presenilin-1 was homologous to microbial integral membrane proteases, and yet there is also experimental evidence that it functions as the core component of a new type of vertebrate proteolytic complex that is responsible for the γ-secretase amyloid precursor protein (APP) cleavage[9]. Notwithstanding the challenges of demonstrating independent function or unravelling the mechanism of such a complex, at least the availability of genomic sequence opens up the exon structure of any putative cofactors for comparative analysis. The trends described above are likely to push the final human protease total beyond 500 but we are not yet in a position to predict an upper limit. With regard to question four, there are indications that the relative size of genomic protease families might not be substantially different from the current transcript collection[6]. Despite the increase of ~180 new human protease mRNA entries between 1998 and 2000, there has been relatively little shift in the major mechanistic class distributions. These currently stand at 3% aspartic, 23% cysteine, 36% metallo and 32% serine. Some families have undergone major expansions over this time period, such as the S1 trypsins, the M12 adamalysin (ADAM) metalloprotease group and the C12/C19 ubiquitin-specific proteases, which (at the time of writing) stand at 105, 36 and 31 members, respectively. However, comparison with InterPro protein family annotation of the latest Ensembl output gives similar relative numbers; that is, there is no indication that the genomic data will selectively expand a limited number of families. The likelihood of new domain combinations is difficult to predict, as is the recognition of novel domains associated with known catalytic modules. Such domains will need expert analysis to be revealed, such as the recently reported protease-associated domain[10]. However, for those domains that are annotated, there are powerful protein database query tools available that will allow the direct interrogation of their combinations and cross-species comparisons[11,12].

The question regarding the number of pseudogenes and inactive homologues is important because these reduce the number of sequences that have to be evaluated as potential targets or considered for drug selectivity. In addition, if inactive paralogues show high sequence similarity to targets, they might even be able to bind the same compound classes. The current MEROPS release assigns 12% of the 405 sequences as probable inactive homologues by the criterion of residue substitutions within critical active-site regions. However, not all of the remaining proteases have been experimentally proven to be catalytically active. Current human gene loci annotations are running at ~7% pseudogenes (Box 1) but the finished data standard should make pseudogenes easier to recognize. Recently, protease pseudogenes have been reported within the kallikrein and tryptase–prostasin gene clusters on 19q13 and 16p13, respectively[13,14]. An example of a transcribed protease pseudogene (which might even be translated) is the gene for (pro)napsin B (Ref. 15). The mRNA is found exclusively in cells related to the immune system but lacks an in-frame stop codon and contains several polymorphisms, one of which replaces a catalytically crucial glycine residue with an arginine residue.

Last, but not least, the fifth question concerns the extent of transcript splicing. This is predicted to affect >35% of all human gene products and introduces another layer of complexity into protease biology and pathology[16]. Using angiotensin-converting enzyme (ACE) as an example,

Box 3. Information that can be gained from the bioinformatic analysis of proteases identified in the genomic data

- Detection of all homologues within the human genome (paralogues) with significant similarity scores to documented proteases.
- Identification of homologues in non-mammalian model organisms such as the fly, worm, yeast and fish.
- Elucidation of genomic structures, map locations, family clusters, syntenic positions and possible transcription control regions, via comparisons between human and mouse sequences.
- mRNA characterization including splice variants, differential polyadenylation and putative control elements in 5′ or 3′ untranslated region (UTR) sequences.
- Coverage of genetic variation by single nucleotide polymorphisms (SNPs), mutations, microsatellite repeats and microinsertions or microdeletions.
- Identification of physiological and pathological substrate-cleavage products from proteomic projects.
- Identification of new endogenous inhibitors.
- Identification of ancillary motifs associated with protease catalytic domains.
- *In silico* transcript verification and tissue abundance data obtained from public expressed sequence tag data
- Prediction of targeting and cellular-location features such as signal peptides, proregions, transmembrane domains, potential lipid-anchor sites and other possible post-translational modifications.
- Discrimination of probable active proteases from pseudogenes or inactive homologues.
- Tissue and cellular mRNA distributions from public microarray data.
- Homology modelling for sequences with >30–40% identity to known structures.
- Evolutionary analysis of multigene families giving clues to functional and/or regulatory divergence.

it is only recently that a physiological basis has been proposed to explain the expression of the enzymatically equivalent somatic and germinal isoforms[17]. Another example that also concerns an important protease drug target is the reported β-secretase ACE (BACE) mRNA, which lacks a 44 amino acid region from exon 3, located between the two catalytic aspartyl residues[18]. This is expressed as a pancreas-specific splice variant but is absent from the brain. These findings explain the previously observed paradox of high BACE transcription in pancreas with very low enzymatic activity but still leaves the question of what functional role this splice variant might have. Although the extent and functional significance of most protease splicing events is unknown, the availability of genomic data allows the sequences of alternative transcripts to be mapped to their exon combinations. These can then be used to design primers for the measurement of splice-variant expression *in vivo*.

Progression of current targets

The utility of genomic and transcript data for the investigation of new potential drug targets is obvious, but what value can genomic data add to known targets? Although these already have a perceived level of target likelihood that is high enough to justify running a high-throughput screen, it is always advantageous to accrue more data. This is especially important for those proteases that have been characterized more recently. In some cases, new data might contradict the initial target hypotheses if, for example, a mouse knockout phenotype might not support a predicted physiological context for a particular protease. However, it is as valuable to use genomic data to stop work on low-value targets as it is to select high-value targets. The human genomic structure has hitherto been a starting point to confirm the minimal amount of mouse gene structure necessary to design a knock-out or knock-in mouse model. The arrival of extensive mouse genomic sequence is certain to accelerate these types of genetic manipulation as routes to examining protease function *in vivo*.

The extent of genome-wide single-nucleotide polymorphisms (SNPs) and other types of genetic variation within human populations is coming under increasing scrutiny for all drug targets[19]. A recent analysis of 24 kb from the ACE locus uncovered no less than 78 varying sites that could be resolved into 13 distinct haplotypes, some of which might confer increased susceptibility to cardiovascular disease[20]. Similar levels of population diversity (179 variants within 66 kb) have been reported for calpain 10 (Ref. 21). Evidence suggests that certain haplotype combinations for calpain 10 might affect susceptibility to type-2 diabetes in both Mexicans and Europeans, although the mechanism for such an effect is not yet clear. We can expect to find similar genetic diversity within other human protease loci. This emphasizes both the importance of disease associations for target validation and the methodological challenges of not only detecting such associations but also providing testable causative hypotheses. For targets selected for HTS, all variations in protein sequence at significant population frequency, especially those in the structural core regions, will have to be considered for any affects on protein stability, substrate turnover and/or inhibition kinetics. This type of variation might also have pharmacogenomic consequences if they are found to influence patient responses to drugs that are chosen for further development. The combination of disease associations, enzymatic variability and pharmacogenomic effects have made full gene structure analysis and population variant screening in genomic DNA an essential part of the validation process for both old and new protease targets.

Another consequence of having the genomic complement of proteases is the ability to make defined choices when inhibitor compounds have to be cross-screened to determine their selectivity. The need for this is increasing, as more target enzymes turn out to have nearest-neighbour paralogues identified in genomic sequence. Sequences that would have sufficient structural similarity within their inhibitor-binding pockets to warrant selectivity screening are expected to show an overall amino acid sequence identity above 30%. This makes them relatively straightforward to detect in genomic data but they would still require verification as enzymatically active translation products. Examples of such neighbours include the second ACE (ACE2) on Xp22 in addition to the ACE on 17q23 (Ref. 22), aggrecanases 1 and 2 on 1q23 and 21q21.3 (Ref. 23), and two paralogues of the Alzheimer's β-secretase enzyme, BACE/ASP2 on 11q22 and BACE2/ASP1 on 21q22.3 (Refs 24,25). There is, of course, always a caveat in that paralogues of known targets can turn out to have a role in the same, or related, pathological processes and each of the above examples is being investigated for this possibility. Two of the examples of microbial protease homologues (Box 3), sialoglycoprotease and the polypeptide deformylase, raise an interesting issue of crossreactivity because the microbial proteases have been targeted for inhibitor development as novel antibiotics. Although the sequence identity is low enough to make selective inhibition tractable, the human genome data have revealed homologies that were probably not suspected when these proteases were first chosen as targets on the basis of their essentiality for bacterial survival.

New drug targets

It was estimated that, as a consequence of the human genome project, the total number of drug targets actively

investigated by the pharmaceutical and biotechnology industries, currently ~500, would increase at least tenfold, but this projection was made before the confirmation of the unexpectedly low proteome number[26]. How many of these new targets are likely to be proteases? This question can only be answered when experimental findings on the biochemical characterization and possible disease relevance of novel proteases are released into the public domain. However, the data from genome annotation projects such as Ensembl (Box 1) at least give us numerical benchmarks to compare proteases with other drug target classes such as G-protein-coupled receptors (GPCRs), serine or tyrosine kinases and nuclear hormone receptors (NHRs). The Ensembl release for Febuary 2001, containing 25,790 confirmed genes, was interrogated by InterPro classifications to compare genomic numbers for the protein kinases (437), rhodopsin-type GPCRs (369) and NHR C4-type steroid receptor zinc fingers (35).

For reasons already discussed, these absolute genomic numbers might be revised upwards but the ratios for these target classes are likely to remain in the same relative order (i.e. proteases > kinases > GPCRs >> NHRs).

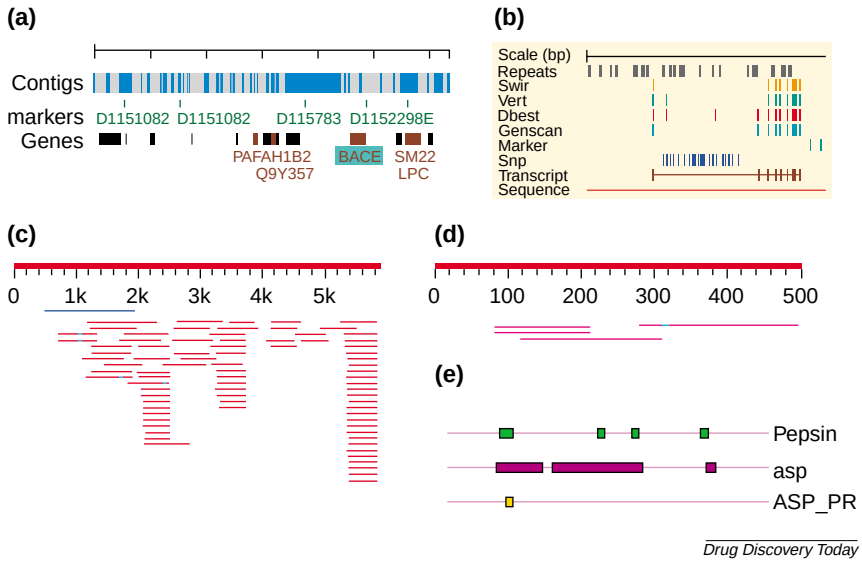

Figure 1. Examples of genomics-related information available for β-secretase angiotensin-converting enzyme (BACE)/ASP2. **(a)** The genomic context of the gene encoding BACE/ASP2 (brown on green) as annotated by the Ensembl project. This shows the position of the genomic markers (green), the known genes in the vicinity (brown) and the unknown genes (black). **(b)** A zoom view from **(a)** that includes the intron–exon structure of the gene encoding BACE/ASP2 and the primary database matches [SwissProt (Swir), GenBank vertebrate division (Vert) and expressed sequence tags (ESTs)] that confirm the exons of the transcript prediction. **(c)** The EST distribution across the mRNA. Most of these match on the 3′ side of the protein reading-frame (blue) and the discrete stacks identify sites of alternative polyadenylation. **(d)** The result of a translation search of the BACE/ASP2 protein against the partial genomic data from *Tetraodon*, the freshwater puffer fish (pink), which aligns with >80% of the BACE/ASP2 sequence (red). **(e)** The InterPro domain analysis of the BACE/ASP2 protein. The green sections are PRINTS alignments for aspartyl proteases, the purple sections matches to Pfam and the yellow section shows a match to the PROSITE aspartyl protease active site. Ensemble has just made a major update of the genome viewer as well as the feature displays shown in (a) and (b).

Across the pharmaceutical and biotechnology industries these and other target classes compete against each other for resources. Many factors influence the balance that any particular drug discovery organization arrives at for the proportion of proteases in their target portfolio. Key considerations include history-based judgements of the likelihood of success, the perceived level of target validation within major therapeutic areas, the internal success rates of protease ultra-high-throughput screens, the chemical diversity of the compound collections and the ability to use combinatorial scaffolds to progress rapidly from screening hits to selective and developable leads. It is outside the scope of this article to compare these factors but the relative genomic numbers certainly suggest that proteases will retain their major position in post-genomic target portfolios.

Bioinformatic analysis

In the post-genomic era, bioinformatics has become central to the process of selecting potential new drug targets and prioritizing them alongside known targets[27]. A genome-wide assessment of proteases will yield the type of *in silico* information described in Box 3. An example of the data available is illustrated in Fig. 1, using the BACE/ASP2 Alzheimer's β-secretase. Soon after the mRNA sequences were published, the unfinished genomic sequence was available from chromosome 11. Although this was initially unordered sequence, the current Ensembl output has identified both the known and unknown neighbouring genes on the overlapping genomic sequences. In addition, the exon–intron structure, confirmed by other database matches, is shown with the position of both local SNPs and sequence-tagged site genomic markers. The mRNA search shows the ability not only to perform an 'electronic northern blot' by assessing which tissue libraries the 285 ESTs came from but also to identify alternative polyadenylation patterns by characteristic EST 'stacks' in the 3′ untranslated region[28]. The InterPro entry for BACE/ASP2 gives an informative graphical description of the domain organization.

Searching the protein sequence against the genomic end-read division of GenBank facilitates the assembly of ~80% of the probable orthologous sequence from the freshwater puffer fish. The data in Figure 1 could be collected in a matter of minutes and only represent an excerpt of what is available from bioinformatic web-based resources.

If data have been published for at least one member of a protease family, it is logical to extrapolate that biochemical function to orthologous sequences in humans, mice, rats or fish. However, caution must be exercised when extending the extrapolation to physiological functions between species. An unusual genomic example is the kallikrein locus. In humans, this contains 15 genes but is present as a genomic duplication in the mouse, which thus contains 30 genes[29]. These combinations of protein orthologues and pseudogenes make it difficult to decide which sequences are fulfilling distinct physiological roles or showing functional redundancy. Extrapolation of function from paralogous relationships, even between nearest-neighbour sequences, is more hazardous because the persistence of paralogues after the ancestral duplication events implies functional and/or regulatory divergence.

Substrate identification for orphan proteases

Sequences that have been in the public domain for some time, and might even have expression and/or biochemical characterization data, can still be considered to be orphan if no direct evidence has been obtained for their physiological function. The BACE/ASP2 enzyme is a paradoxical example that has been solidly validated as a target for aberrant APP turnover[30]. However, despite extensive published biochemical data, including a 3D structure and a mouse knockout, an understanding of its normal physiological role remains elusive[31,32].

A variety of approaches for identifying the substrates of orphan proteases have appeared in the literature, but a clear distinction must be made between surrogate substrates that can be turned over under specified conditions by the enzyme *in vitro* and 'real' substrates shown to be cleaved *in vivo*. Surrogate substrate determinations are most often made by screening phage display or combinatorial peptide libraries[33,34]. If a restricted set of peptide P–P′ residue preferences are found, it might be possible to use the specificity fingerprint to narrow down the candidate substrates *in silico*. However, these preferences cannot be searched with any precision and there are no general findings to suggest that physiological substrates are optimized at the primary sequence level. The main use of surrogate peptide substrates is to develop a fluorogenic HTS assay. These assays can then produce tool inhibitor compounds that might be specific enough to probe the function of the protease in cellular systems or animal models. A possible consequence of the genomic harvest of orphan proteases is that pharmaceutical or biotechnology operations might consider industrializing both the expression of recombinant proteins and surrogate substrate screening, especially because the former is already being set up for structural genomics projects. The next stage could be extended to automated screening of a tractable subset of orphan or novel proteases, especially if these were prioritized with bioinformatic filters, such as secreted protein likelihood and similarity to known targets. Screening significant numbers of novel or orphan enzymes without pre-existing target indications would be a speculative resource commitment. However, as a direct route to produce tool inhibitors and the consequent possibility of accelerated target validation, it is a logical way to exploit the genome data.

Physiological substrate determination for an orphan protease presents a much more difficult problem. One reason for this is that many proteases show broad and overlapping specificities, particularly between close paralogues, and the concept of 'one protease, one tissue location, one endogenous inhibitor, one substrate' is the exception rather than the rule. In addition, substrate specificity might not be an inherent property of the enzyme but can be derived from the co-localization of active protease and substrate in particular tissues, cell types, subcellular compartments (such as membranes) or developmental stages. It might also involve activation pathways in which propeptide cleavage of the zymogen is a prerequisite for full activity. A promising approach to identifying protease substrates with at least a modest throughput is using proteomics with 2D gel electrophoresis[35]. This technology can detect precursor–product polypeptide size changes provided that the products are larger than ~5 kDa. This strategy is becoming more effective with the increase in human genomic data because the probability of definitively identifying the cleavage product on a gel by MS is now high. However, it will still require detailed biochemical experiments to identify directly which proteases are associated with particular substrate turnover under particular physiological or pathological conditions.

Target validation

The literature shows that one of the most important experimental findings linking proteases to disease is the identification of a substrate where the extent of turnover is causatively related to pathological processes. The next key step of validation is to show if this can be ameliorated, without undesirable side effects, by inhibition of the specific protease. Additional aspects need to be considered when the pathology arises from 'inappropriate' protease

activity on substrates not cleaved under normal conditions. Two of the proteases in Box 2 provide examples of this problem. The transcription of human HtrA1 serine protease has been identified as being upregulated in osteoarthritis and the proteolytic activity of HtrA2 (also termed Omi) is upregulated in the mouse kidney following ischemia–reperfusion[36,37]. HtrA2 has also been shown to be involved in the mammalian cellular stress response and contains domain sections that can interact with presenilin-1 under experimental conditions[38]. These observations are suggestive of pathological associations that could present therapeutic targets. However, in neither case has it yet been possible to link these enzymes that show similar *in vitro* characteristics to the bacterial heat-shock proteases, with defined human substrates.

Although fewer in number, there are instances of the reverse problem: of disease-associated substrates for which the protease is unknown. A long-standing example of this is the so-called CD23ase, the proteolytic activity that produces soluble CD23 and thereby contributes to pathological IgE production in asthma[39]. Although medicinal chemistry efforts have progressed to metalloprotease inhibitor compounds that modulate IgE production, the identity of the enzyme at the sequence level remains unclear. Even assuming that selective inhibitor profiling can pin down the mechanistic class involved, proving the identity of the transmembrane metalloproteinase responsible for the proteolytic release or 'shedding' of physiologically important cell-surface proteins is notoriously difficult, especially because these are likely to be involved in multiple shedding events[40]. The successful identification of the tumour necrosis factor α converting enzyme (TACE or ADAM-17) was a notable exception[41]. Similar biochemical approaches using compound labelling could be applied to CD23ase and the recent expansion in ADAM protease discovery in genomic data will make it easier to identify the right sequence. Another example of an orphan pathological substrate is myelin-associated glycoprotein (MAG). In this case, proteolysis has been implicated in demyelinating diseases[42]. Using a variety of experiments including peptidolysis assays, the candidate protease was narrowed down to an extracellular, cystatin-C-inhibitable, cathepsin-L-like enzyme. However, definitive validation would mean obtaining data to pin down which of the 15 known human C1 cathepsins, and any additional members of this family that can be detected in genomic data, could be an effective therapeutic target.

Conclusions

As the finishing phase of the Human Genome Project proceeds, new proteases with significant sequence similarity to known proteases are already being indexed in secondary databases, such as MEROPS, InterPro and Ensembl. The first phase of annotation suggests that there might not be significant changes in the relative family sizes, domain combinations or mechanistic classes. The genomic annotations will need bioinformatic analysis to compare them with mRNAs, ESTs, polymorphisms, 3D structures, expression data and species orthologues.

Although there will be an increase in the number of proteases under active investigation as drug targets, there is a compelling need for the experimental verification of catalytic activity *in vitro* and the assignment of physiological and/or pathological substrates *in vivo*. These are the basic prerequisites for evaluation as potential therapeutic targets. It will be of interest to see which of the new technologies for high-throughput biology could open up these bottlenecks. It is also open to speculation whether the new proteases already annotated in genome data will have lower and/or more specific patterns of tissue expression, which could make them more attractive as drug targets. At the end of the day, however, despite the burgeoning protease knowledge base, the biggest impetus to bringing new, post-genomic representatives forward into the drug discovery process will be successful outcomes from the pre-genomic 'declared' protease targets in producing marketable drugs.

References

1 Beeley, L.J. *et al.* (2000) The impact of genomics on drug discovery. *Prog. Med. Chem.* 37, 1–43
2 Leung, D. *et al.* (2000) Protease inhibitors: current status and future prospects. *J. Med. Chem.* 43, 305–341
3 Barrett, A.J. *et al.*, eds (1998) *Handbook of Proteolytic Enzymes*, Academic Press
4 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysisi of the human genome. *Nature* 409, 860–921
5 Rawlings, N.D. and Barrett, A.J. (2000) MEROPS: the peptidase database. *Nucleic Acids Res.* 28, 323–325
6 Southan, C. (2000) Assessing the protease and protease inhibitor content of the human genome. *J. Pept. Sci.* 6, 453–458
7 Makarova, K.S. *et al.* (2000) A novel superfamily of predicted cysteine proteases from eukaryotes, viruses and *Chlamydia pnemoniae*. *Trends Biochem. Sci.* 25, 50–52
8 Russell, R.B. and Eggleston, D.S. (2000) New roles for structure in biology and drug discovery. *Nat. Struct. Biol.* 7 (Suppl.), 928–930
9 Wolfe, M.S. and Haass, C. (2001) The role of presenilins in gamma-secretase activity. *J. Biol. Chem.* 276, 5413–5416
10 Mahon, P. and Bateman, A. (2000) The PA domain: a protease associated domain. *Protein Sci.* 10, 1930–1934
11 Southan, C. (2000) Website review: InterPro the integrated resource of protein functional domains. *Yeast* 17, 327–334
12 Apweiler, R. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* 29, 37–40
13 Gan, L. *et al.* (2000) Sequencing and expression analysis of the serine protease gene cluster located in chromosome 19q13 region. *Gene* 257, 119–130

14 Caughey, G.H. *et al.* (2000) Characterization of human gamma-tryptases, novel members of the chromosome 16p mast cell tryptase and prostasin gene families. *J. Immunol.* 164, 6566–6575

15 Tatnell, P.J. *et al.* (1998) Napsins: new human aspartic proteinases. Distinction between two closely related genes. *FEBS Lett.* 441, 43–48

16 Gravely, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17, 100–107

17 Kessler, S.P. *et al.* (2000) Physiological non-equivalence of the two forms of angiotensin converting enzyme. *J. Biol. Chem.* 275, 2659–2664

18 Bodendorf, U. *et al.* (2001) A splice variant of beta-secretase deficient in the amyloidogenic processing of the amyloid precursor protein. *J. Biol. Chem.* 276, 12019–12023

19 Grey, I.C. (2000) Single nucleotide polymorphisms as tools in human genetics. *Hum. Mol. Genet.* 9, 2403–2408

20 Rieder, M.J. (1999) Sequence variation in the human angiotensin converting enzyme. *Nat. Genet.* 22, 59–62

21 Horikawa, Y. *et al.* (2000) Genetic variation in the gene encoding calpian-10 is associated with type-2 diabetes mellitus. *Nat. Genet.* 26, 163–175

22 Tipnis, S.R. *et al.* (2000) A human homologue of angiotensin-converting enzyme. Cloning and functional expression as a captopril-insensitive carboxypeptidase. *J. Biol. Chem.* 275, 33238–33243

23 Caterson, B. *et al.* (2000) Mechanisms involved in cartilage proteoglycan catabolism. *Matrix Biol.* 19, 333–344

24 Hussain, I. *et al.* (1999) Identification of a novel aspartic protease (Asp 2) as beta-secretase. *Mol. Cell. Neurosci.* 14, 419–427

25 Hussain, I. *et al.* (2000) ASP1 (BACE2) cleaves the amyloid precursor protein at the β-secretase site. *Mol. Cell. Neurosci.* 16, 609–619

26 Drews, J. (2000) *Quo vadis*, biotech (Part 1). *Drug Discov. Today* 5, 547–553

27 Searls, D.B. (2000) Using bioinformatics in gene and drug discovery. *Drug Discov. Today*. 4, 135–143

28 Southan, C. (2000) Bioinformatic analysis of mRNA heterogeneity in the putative Alzheimer's beta-secretase, Asp2. *Biochem. Soc. Trans.* 28, p84, abstract No. 63

29 Eleftherios, P. *et al.* (2000) The new human kallikrein gene family: implications in carcinogenesis. *Trends Endocrinol. Metab.* 11, 54–60

30 Howlett, D.R. (2000) In search of an enzyme: the β-secretase of Alzheimer's disease is an aspartic proteinase. *Trends Neurosci.* 23, 565–570

31 Hong, L. *et al.* (2000) Structure of the protease domain of memapsin 2 (β-secretase) complexed with inhibitor. *Science* 290, 150–153

32 Luo, Y. *et al.* (2001) Mice deficient in BACE1, the Alzheimer's beta-secretase, have normal phenotype and abolished beta-amyloid generation. *Nat. Neurosci.* 4, 231–232

33 Harris, J.L. *et al.* (2000) Rapid and general profiling of protease specificities by using combinatorial flourogenic substrate libraries. *Proc. Natl. Acad. Sci. U. S. A.* 97, 7754–7759

34 Deng, S-J. *et al.* (2000) Substrate specificity of human collagenase 3 assessed using a phage-displayed peptide library. *J. Biol. Chem.* 275, 31422–31427

35 Celis, J.E. *et al.* (2000) Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Lett.* 480, 2–16

36 Hu, S-I. *et al.* (1998) Human HtrA, an evolutionarily conserved serine protease identified as a differentially expressed gene product in osteoarthritic cartilage. *J. Biol. Chem.* 273, 34406–34412

37 Faccio, L. *et al.* (2000) Characterization of a novel human serine protease that has extensive homology to bacterial heat shock endoprotease HtrA and is regulated by kidney ischemia. *J. Biol. Chem.* 275, 2581–2588

38 Gray, C.W. *et al.* (2000) Characterization of human HtrA2, a novel serine protease involved in the mammalian cellular stress response. *Eur. J. Biochem.* 267, 5699–5710

39 Mayer, R. *et al.* (2000) Inhibition of CD23 processing correlates with inhibition of IL-4-stimulated IgE production in human PBL and hu-PBL-reconstituted SCID mice. *Clin. Exp. Allergy* 30, 719–727

40 Brown, M.S. *et al.* (2000). Regulated intramembrane proteolysis, a control mechanism conserved from bacteria to humans. *Cell* 100, 391–398

41 Killar, L. (1999) Adamalysins. A family of metzincins including TNF-alpha converting enzyme (TACE). *Ann. New York Acad. Sci.* 878, 442–452

42 Stebbins, J.W. (1998) Characterization of myelin-associated glycoprotein (MAG) proteolysis in the human central nervous system. *Neurochem. Res.* 23, 1005–1010

Drug Discovery Today will protect your identity!

Here is an unrivalled opportunity to get off your chest those things that really irritate you!

…and to be able to tell the relevant people what really irritates you without them knowing it's you!

Send in your letters and get some real debate going!

Please send all contributions to Dr Rebecca Lawrence
e-mail: rebecca.lawrence@drugdiscoverytoday.com

Publication of letters is subject to editorial discretion

Tripos

repeat pdf from DDT vol 6, No. 8, page XII

OR

No. 11, page XI