Journal of Peptide Science
J. Peptide Sci. **6**: 453–458 (2000)

Assessing the Protease and Protease Inhibitor Content of the Human Genome

CHRISTOPHER SOUTHAN*

Department of Bioinformatics, Target Discovery, SmithKline Beecham Pharmaceuticals, Essex, UK

Received 20 June 2000
Accepted 25 June 2000

Abstract: The revealing of the entire complement of protease and protease inhibitor sequences by the Human Genome Project will be of great importance to both academic and pharmaceutical research. Although the finishing phase is not yet complete, a selection of secondary annotation sources and comparisons with completed model organism genomes already allow useful estimates to be made. Conservative extrapolation suggests a total of ~1.8% for human proteases. This is close to the figures for yeast (1.7%) and worm (1.8%) but lower than the fly (3.4%) which has a large trypsin-like protease content. Using estimates for the human proteome of between 40000 and 60000 genes would extrapolate to 700–1100 proteases, compared with ~360 currently represented as GenBank mRNAs. Preliminary comparisons between domain annotations for predicted human gene products and completed proteins suggest the genomic protease family and mechanistic class distributions will broadly reflect those in the current transcript data. The protease:inhibitor ratio at the mRNA level is currently ~9:1, but genome annotation data indicate that inhibitory domains are more widespread than this ratio would indicate.

Keywords: annotation; domain; genome; protease; proteases inhibitor

INTRODUCTION

The advantages of proteases as potential drug targets are well established [1,2]. Reasons for their popularity include the accumulated data on their enzymology, high-throughput assays, pathophysiology, three-dimensional (3D) structures, small molecule inhibitors and the characterization of many endogenous inhibitory proteins. The Human Genome Project, by having the potential to reveal the entire complement of primary structures, termed the proteome, will be of great importance to both academic and pharmaceutical protease research. Bioinformatic comparisons will then be able to include all human homologues (paralogues), selected mammalian homologues (orthologues), such as mouse and rat, but also homologues in non-mammalian model organisms such as fly, worm, yeast and fish. The same arguments apply to the endogenous protease inhibitors, not only because of their important roles in protease physiology, but also because they can form the basis of therapeutic protein agents.

However, the majority of the human genome has appeared initially in draft form. Difficulties associated with predicting protein reading frames in non-contiguous data and the imperfect performance of current *ab-inito* gene prediction means the protease and inhibitor inventory remains incomplete. Nonetheless, a variety of public sources, including completed model organism genomes and secondary annotation sources, have now developed to the point where we can make useful extrapolations and estimates of what total numbers we could expect. In addition, we can highlight which particular sequence families are expanding in current transcript databases and compare these with the first phase of human genome annotation. To limit the size of this article, an acquaintance with protease and inhibitor nomenclature is assumed [3,4].

* Correspondence to: Department of Bioinformatics, Smith Kline Beecham Pharmaceuticals, New Frontiers Science Park, Third Avenue, Harlow, Essex CM19 SAW, UK

Information Sources

Although the annotation associated with primary genomic and mRNA sequence data can be interrogated to track proteases and inhibitors, a number of secondary databases have been developed to add value to curated subsets extracted from the original raw data. They also provide querying tools and links. The main caveat to these secondary sources is (of necessity because of the work involved) the varying update frequency which determines the flow of new sequences from the primary databases into the secondary databases. Only a brief introduction can be given for the following sources that were found most useful in the preparation of this review. The web sites and/or literature descriptions should be consulted for background information.

For proteases the most comprehensive of these sources is the MEROPS peptidase database (bi.bbsrc.ac.uk/Merops/Merops.htm) [5]. The curation process classifies the peptidases into families on the basis of statistically highly significant sequence similarities between the catalytic domains. Also provided are interfaces to mRNA and tertiary structure data if known. The SwissProt peptidase collection (expasy.ch/cgi-bin/lists?peptidas.txt) also contains extensive annotations, links and analysis tools but lags behind primary data compared to MEROPS. Protease inhibitor sequences can be also retrieved from SwissProt and its update supplement TrEMBL [6]. Both proteases and inhibitors can be retrieved from GeneCards, a non-redundant collection of human proteins with extended links (bioinformatics.weizmann.ac.il/cards/) [7]. The National Center for Biological Information (NCBI) also has a resource, LocusLink (ncbi.nlm.nih.gov/LocusLink/), that facilitates query retrieval from a non-redundant transcript reference set that includes both human, mouse and rat mRNA sequences. InterPro (ebi.ac.uk/interpro/) is a new integrated resource for protein families, domains and functional sites. This merged annotation combines the advantages of the local alignments in PRINTS, regular expressions from PROSITE, extended Hidden Markov Model-based alignments from Pfam, and will soon include automated alignments from PRODOM. The links indicate which domains and functional sites are associated with particular proteases or inhibitors. Ensembl (ensembl.org/) is a joint project between the European Bioinformatics Institute and the Sanger Centre who have developed a software system for automatic annotation of the human genome data. The resulting database contains a unique set of partial and complete proteins predicted from all human genomic sequence so far deposited in GenBank, both finished and unfinished. These protein sequences can searched by BLAST or queried directly by Pfam family assignments for protease and inhibitors.

Many of the numbers below are derived from comparing between the different secondary database sources described above. Although the statistics quoted in this article for each database source are presumed to be calculated on the date posted, it should be born in mind that inter-database comparisons cannot be synchronized because they are all subject to 'snapshot' differences caused by the varying update and recompilation intervals with respect to the primary source sequences.

EUCARYOTIC PROTEASE CONTENT

Completed eucaryotic genomes

The complete genomes of three model organisms, fly, worm and yeast provide the first opportunity for a comparative assessment of their protease content and mechanistic class distribution. Table 1 shows a breakdown of this data with the incomplete human protein set included for comparison. The protein kinase family totals are included for comparison to another large enzyme group. A caveat with these comparisons is that, certainly within the metazoans, a low proportion of protease sequences can be detected that, as judged by the absence of critical residues, are likely to be non-catalytic.

One of the surprises from the recent fly genome was that it was possible for a complex metazoan to have a proteome only slightly more than twice the size of that in the yeast [8]. Another surprise was the large number of serine proteases in the fly, due to the expansion of the S1-trypsin family to approximately 200 proteins. The class distributions also indicate that worm and yeast have a lower serine and higher aspartyl protease content, with the fly having a lower cysteine protease content. Despite the increase in human protease mRNA entries, by $\sim$150 between 1998 and 2000, there has been relatively little shift in the mechanistic class distribution, except for an increase in the metalloprotease content from 32% to 38%.

Another trend that can be seen from inspection of Table 1 is an increase in the number of protease families with the transition for single celled to more complex metazoans. It will be of interest to see if the

Table 1 Eucaryotic Protease Content, Family Totals, Mechanistic Distribution, and Kinase Content

Organism	*Homo sapiens*	*Drosophila melanogaster*	*Caenorhabditis elegans*	*Saccharomyces cerevisiae*
Gene total	~40–60 000	13 601	18 424	6241
Protease total	360	467	270	107
Total proteases	~1.8%	3.4%	1.8%	1.7%
Aspartic	3%	5%	8%	7.5%
Cysteine	23%	12%	24%	25%
Metallo	38%	31%	47%	50%
Serine	33%	52%	14%	15%
Protein kinases	622 (~2%)	222 (3%)	363 (2%)	118 (1.8%)
Protease families	42	46	37	27

final human family number significantly exceeds that of the fly, i.e. if mammalian or vertebrate specific families will emerge. This complexity trend is reflected in the increase in the domain totals across the four organisms. As indexed by InterPro these, numbers are 1862, 1407, 1293 and 1054, for human, fly, worm and yeast, respectively. This increase of ~450 human domains would predict the likelyhood of more protease-associated domain permutations compared with fly proteases, for example, but this analysis will have to await the complete human proteome set.

Human Total Estimates

The question arises as to what extent the totals can be extrapolated to the human case. Simple calculation of the representation of proteases as a proportion of all human transcripts, calculated from roughly concurrent releases of MEROPS and LocusLink (360/12940) is 2.7%. However, this is probably skewed upwards by the historical 'interest bias' towards proteases. Evidence will be presented later that the human genome does not contain the same proportionally large S1 family as *Drosophila*. If the fly totals are adjusted by scaling to the current human serine protease content of 33%, i.e. 'removing' ~150 S1 proteases, this brings the fly total down to 2.3%. Indirect evidence for this sort of number can be found from the protease total across all organisms. As more complete genomes are deposited this number becomes a representative per organism average as the historical 'interest bias' is diluted. From the MEROPS total of 6422 proteases (April 2000) this would represent ~1.8–2.0% depending on the exact non-redundant protein total at the time the proteases were extracted and collated. Without any prospect of more precise estimates at the time of writing, a conservative figure of 1.8% would seem appropriate for human proteases. However, it should be noted that, for the yeast and worm, there has been a slight but distinct upward trend in the protease totals since the genomes were completed. Reasons for this include the correction of some earlier protein reading frames, the application of more sensitive searching methods for updating homology assignments and the characterization of new proteases in the literature. There is no reason why this gradual increase should not continue and consequently push the human total towards 2% but this cannot be predicted. Despite the availability of 90% of the genome in draft form and the completion of chromosomes 21 and 22, there are still wide variations in estimates of the human proteome of between 35 K and 120 K [9]. Until the results for different methods converge, we are still faced with this uncertainty but the majority of independent estimates are converging at the low end of 40–60 K genes.

Using Expressed Sequence Tags for Abundancy Estimates

The assumption can be made that the 3 million human mouse and rat expressed sequence tags (ESTs) from approximately 380 tissue libraries in the public database, dbEST, represent a surrogate mammalian 'transcriptome' that broadly reflects mRNA levels across all the tissues sampled. Boolean queries indicate that 1.2% of these partial transcripts are assigned with protease similarity and 0.02% are assigned with protease inhibitor similarity, i.e. they represent both known or novel partial transcripts. This gives a global protease:inhibitor (P:I) abundancy ratio of ~6:1 but it must be remembered that these are proxy

estimates for the expression abundancy of these classes of gene products rather than gene numbers *per se*. The latter can be estimated from assembly databases where ESTs are collapsed to virtual mRNAs giving an approximation to gene numbers rather than abundance. The results of similar queries are 1.4% proteases and P:I ratio of ~10:1.

Comparative Protease Family Distributions

A comparison of protease family distributions between completed genomes can be made using either MEROPS or InterPro. In addition, it is now possible to compare the preliminary InterPro annotation of all human proteins included in SwissProt and TrEMBL, which roughly corresponds to the LocusLink mRNA transcript collection, with the current Ensembl annotation of 36299 partial and complete human genes (June 2000). For some protease families this can be done by interrogating the comparative proteome assignments in InterPro and comparing with the genomic Pfam assignments in Ensembl. However, for other families there is not a direct correspondence between the designations between InterPro, Pfam and MEROPS, because they have different classification schema. In particular, InterPro and Pfam merge some families that are differentiated by the more stringent clustering used in MEROPS. Despite these caveats, it is informative to compare a selection of protease families between human genome annotations, human transcripts and completed eucaryotic genome totals. The results of this survey are given in Table 2.

Several evolutionary processes could be causing the observed patterns of distribution. The first of these is ancient tetraploidy, which proposes that the vertebrate genome has resulted from two rounds of ancient duplication [10]. Simplistically, this predicts that a protease family in yeast would double in the worm or fly and show a four-fold expansion in humans. Although evidence from other gene families supports this hypothesis none of the groups in Table 2 fit this model. However, it has to be borne in mind that the genome annotation (the human genome (HG) column) does not yet represent the final total, i.e. some expansion could still occur. This is indicated by the detection of 723 kinases in Ensembl (the HG set in Table 2), compared with 622 in InterPro (the human protein (HP) set in Table 2). By extrapolation, therefore, we might see a final human S1 protease total of ~160 that exceeds the current MEROPS total of 94. Other processes include selective phylogenetic expansions or losses (it is difficult to discriminate between these alternatives until large numbers of organisms have been fully sequenced). Examples would include the S1 and M2 families in the fly and the M12A and M13 families in the worm where the human totals are almost certain to remain well below these numbers. Other families listed in Table 2, such as the M22 glycopeptidase family, seem to be fixed at low numbers and appear not have undergone extensive duplication in the vertebrate lineage. Between the MEROPS database releases of 1998 and 2000 many families expanded dramatically. These include trypsin S1, increasing from 25 up to 94, the M12 ADAMs reprolysins from 6 to 38, and the C19 ubiquitin-specific proteases from 3 to 25. Less dramatic but also significant was the

Table 2 Selected Comparisons Between Human Genome Annotations, Human Proteins and Completed Eucaryotic Genome Totals. The Families are Designated by Their InterPro Number and the Equivalent Pfam Number

MEROPS description	InterPr	Pfam	HG	HP	CE	DM	SC
S26 leader peptidase	000223	00461	2	1	1	1	2
M41 FtsH ATPase	000642	01434	3	5	1	3	3
S1 trypsin	001254	00089	154	112	13	205	1
C1 papain/cathepsins	000668	00112	18	11	27	10	0
M12A astacin/reprolysin	001590	01421	18	9	36	13	0
M22 glycopeptidase	000905	00814	1	1	1	2	2
C14 caspase	001309	00656	11	23	7	10	0
M2 ACE	000130	01401	2	2	2	6	0
M13 neprylysin/ECE	000718	01431	5	2	33	19	0
A1 pepsin	001461	00026	5	8	16	13	7

HG refers to the Pfam annotations of Ensembl human genome data. HP refers to the human proteins annotated by InterPro. CE = *C. elegans*, DM = *D. melanogaster*, SC = *S. cerevisiae*.

aspartyl protease A1 expansion by four new members (although one may be a pseudogene) from 7 to 12. A tentative conclusion from Table 2 and additional queries against the Ensembl data indicate there may not be any dramatic expansions of individual families but more likely an incremental increase across all the larger families. These comparisons between human and complete metazoan proteomes will become easier when the Ensembl input includes a higher proportion of finished sequence and the output is processed by InterPro (R. Apweiler, personal communication).

PROTEASE INHIBITORS

Transcript Assessment

In the absence of a public curation resource for protein inhibitors, it is more difficult to assess the human transcript numbers. There is an additional problem of terminology for deciding which gene products are classified as protease inhibitors. Historically, the majority of proteins that appeared to have an intact protease catalytic domain have been primarily classified as a protease, regardless of whatever additional functional domains were on the same polypeptide. Sometimes other functionalities have become common terminology, e.g. the use of the terms bone morphogenic protein-1 and procollagen-C proteinase-1 for the same protein. The situation for protease inhibitors is different in that they have been classified if they were discovered as soluble proteins with inhibitory domains that have in most cases been experimentally verified. However, many proteins, including some proteases, contain inhibitor domains but have their primary functional annotation associated with another domain in the same protein, e.g. the collagens and amyloid proteins that contain kunitz domains. In addition, potentially new serine inhibitor canonical loop structures that have not yet been annotated as inhibitor domains can be found in a variety of secreted proteins [11]. Interrogation of GeneCards by keywords retrieves a total of 40 human protein entries that are annotated as protease inhibitors. As the GeneCards version, at 10286 proteins (May 2000), was close in date to the MEROPS release of 360 proteases (April 2000) this gives a P:I ratio of 9:1. This ratio is quite close to that determined from EST abundancies in the previous section.

Genomic Assessment of Inhibitors

For reasons given above, querying by InterPro or Pfam domains as a means of assessing transcript or genomic distribution of protease inhibitory domains will not discriminate between independent inhibitors or inhibitor domains in proteins assigned to other functional categories by their primary annotation. However, it is informative to make such an assessment as is shown in Table 3. This includes three categories, the first is the annotated inhibitor proteins as indexed in the GeneCards database, the second category lists the same domain for transcribed proteins, the third category list the number of genomic proteins indexed as containing the domain (this is extracted from the same 36299 data set used for the proteases in Table 2). There is a consistent pattern that more inhibitor domains are being detected in genomic data than are annotated as individual protease inhibitors, the discrepancy being as high as a factor of 10 for the Kazal type.

CONCLUSIONS AND IMPLICATIONS FOR DRUG DISCOVERY

From the genomic numbers discussed above the figure of 1.8% is a reasonable estimate for the protease content of the human genome with a conservative P:I estimate of 10:1. Assuming a final human proteome content of between 40 K and 60 K, therefore, results in broad estimates of 700–1100

Table 3 Protease Inhibitor Assessment in Genomic and Transcript Data

Inhibitor type	Genomic (Ensembl)	Trans. (InterPro)	GeneCards
Serpin	43	35	20
Kazal	20	7	2
Kunitz	18	23	3
WAP/4-disulphide	7	6	2
Cystatin	23	20	9
Tissue metalloprotease	20	4	4

proteases and 70–110 protease inhibitors. However, it will take some time before genomic data has reached the finished quality that is necessary for optimal processing by secondary annotation projects to provide more solid numbers. As this proceeds, we will certainly see the expansion of known families of proteases and inhibitors, but it seems unlikely that there will be radical changes in family sizes or major shifts from the current mechanistic class proportions. We will also see novel examples of mammalian homologues of microbial proteases and new combinations of protease catalytic modules with other domains. It can also be speculated that 20–30% of the currently functionally unclassifiable proteins being revealed by the genome projects at least some will turn out to be new proteases and inhibitors, either as remote homologues of known families or with novel structures and mechanisms.

An assessment of inhibitor compounds for human proteases reported as having entered the early phases of pharmaceutical development up to 1998 indicated that approximately 10% of the then known 220 human proteases were being actively pursued as drug targets. As we can safely predict a tripling of the 1998 protease number, this could imply at least doubling of potential drug targets. What is not clear is the extent to which genomic data might shift the balance between proteases and other pharmaceutical target classes such as *G*-protein coupled receptors and kinases[1]. Whatever these totals turn out to be it is certain that protease genomic annotations, when analysed together with mRNAs, ESTs, polymorphisms, 3D structures, expression data and species orthologues, will greatly enhance the *in-silico* biology that is an essential adjunct to experimental target validation [1]. However, there will be increasing bottlenecks for the experimental verification of catalytic activity *in vitro* and assignments of physiological or pathological substrates *in vivo*. It is also interesting to speculate if the next 360 proteases to be discovered will have lower or more specific patterns of tissue expression, which could make them more attractive as targets. The necessity for cross-screening will also increase as more target enzymes turn out to have close paralogues. Examples include the discovery of a second angiotensin converting enzyme on Xp22 in addition to the ACE on 17q23, aggrecanases 1 and 2 on 1q23 and 21q21.3 and two paralogues of the Alzheimers beta-secretase enzyme, BACE1/ASP2 on 11q22 and BACE2/ASP1 on 21q22.3 [12].

The impact of an expanded collection of protease inhibitors on pharmaceutical research is not so easy to predict. Only two, pancreatic elastase inhibitor and antithrombin III are currently marketed therapeutic proteins but additional products based on or derived from protease inhibitors may find applications in the treatment of deficiency syndromes or other disease conditions. An impact is also likely where 3D structures of new macromolecular protein inhibitors complexed with proteases might reveal new avenues for the rational design of small molecule inhibitors.

REFERENCES

1. Beeley LJ, Duckworth M, Southan C. The impact of genomics on drug discovery. *Prog. Med. Chem.* 2000; **37**: 1–43.
2. Leung D, Abbenante G, Fairlie DP. Protease inhibitors:current status and future prospects. *J. Med. Chem.* 2000; **43**: 305–341.
3. Barrett AJ, Rawlings ND, Woessner JF (eds). *Handbook of Proteolytic Enzymes*. Academic Press: London, 1998.
4. Roberts RM, Mathialagan N, Duffy JY, Smith GW. Regulation and regulatory role of proteinase inhibitors. *Crit. Rev. Eukaryot. Gene Expr.* 1995; **5**: 385–436.
5. Rawlings ND, Barrett AJ. MEROPS: the peptidase database. *Nucleic Acids Research* 2000; **28**: 323–325.
6. Bairoch A, Apweiler R. The Swiss-Prot protein database and its supplement TrEMBL in 2000. *Nucleic Acids Research* 2000; **28**: 45–48.
7. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics* 1998; **14**: 656–664.
8. Rubin GM, *et al*. Comparative genomics of the eukaryotes. *Science* 2000; **287**: 2204–2215.
9. Aparicio SAJR. How to count human genes. *Nature Genetics* 2000; **25**: 129–130.
10. Sidow A. Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* 1996; **6**: 715–722.
11. Jackson RM, Russell RB. The serine protease inhibitor canonical loop conformation: examples found in extracellular hydrolases, toxins, cytokines and viral proteins. *J. Mol. Biol.* 2000; **296**: 325–334.
12. Hussain I, Powell D, Howlett DR, Tew DG, Meek TD, Chapman C, Gloger IS, Murphy KE, Southan CD, Ryan DM, Smith TS, Simmons DL, Walsh FS, Dingwall C, Christie G. Identification of a novel aspartic protease (Asp 2) as beta-secretase. *Mol. Cell. Neuroscience* 1999; **6**: 419–427.

