The proteins that inhibit peptidases are of great importance in medicine and biotechnology, but there has never been a comprehensive system of classification for them. Some of the terminology currently in use is potentially confusing. In the hope of facilitating the exchange, storage and retrieval of information about this important group of proteins, we now describe a system wherein the inhibitor units of the peptidase inhibitors are assigned to 48 families on the basis of similarities detectable at the level of amino acid sequence. Then, on the basis of three-dimensional structures, 31 of the families are assigned to 26 clans. A simple system of nomenclature is introduced for reference to each clan, family and inhibitor. We briefly discuss the specificities and mechanisms of the interactions of the inhibitors in the various families with their target enzymes. The system of families and clans of inhibitors described has been implemented in the MEROPS peptidase database (http://merops.sanger.ac.uk/), and this will provide a mechanism for updating it as new information becomes available.

Key words: clan, compound inhibitor, domain repeat, MEROPS database, peptidase inhibitor, protease inhibitor.

INTRODUCTION

Proteolytic enzymes (best termed peptidases) are essential for the survival of all kinds of organisms, and are encoded for by approx. 2% of all genes [4]. Despite their life-giving functions, enzymes that break down proteins are potentially very damaging in living systems, so their activities need to be kept strictly under control. Several distinct mechanisms exist for the control of excessive peptidase activity, important amongst which are the interactions of the enzymes with proteins that inhibit them. These proteins are the subject of the present review.

We have adopted a broad definition of ‘peptidase inhibitor’ in deciding what to include here. We think it likely that all of the proteins considered have the potential to attenuate the activities of peptidases both in vitro and in vivo by the formation of complexes with the enzymes, but we have not been able to apply any quantitative criterion to this assessment. Valuable proposals have been made as to how one can assess the physiological relevance of an inhibitor [5], but the appropriate data have often not been provided when an inhibitor was described.

The scientific study of the peptidase inhibitors is nearly as old as that of the peptidases themselves. Hundreds of protein inhibitors of peptidases are now known and they are the subjects of thousands of research communications. The research is driven by the many potential applications of knowledge about the inhibitors in medicine, agriculture and biotechnology. At the most fundamental level, an understanding of the mode of interaction of protein inhibitors with enzymes may suggest novel approaches to the design of synthetic inhibitors for use as drugs. Many naturally occurring inhibitors, such as the anticoagulant hirudin, are being used as the basis of engineered proteins for injection in their own right [6]. There are a number of inherited diseases that are attributable to abnormalities in peptidase inhibitors. These include forms of emphysema, epilepsy, hereditary angioneurotic oedema and Netherton syndrome [7–10]. Some such diseases may be susceptible to treatment with the inhibitors administered as drugs, with synthetic inhibitors that take over their function, or with the natural inhibitors made available by gene therapy. Excessive proteolytic activities may well contribute to a number of disease conditions and, again, gene therapy to introduce inhibitors is under consideration [11,12]. In agriculture, genetically modified crop plants expressing inhibitors of the digestive enzymes of their insect pests are already under study [13,14].

This active field of research generates a rapid flow of information, but the storage and retrieval of all the new information that is being obtained about the peptidase inhibitors are handicapped by difficulties of nomenclature. In what was perhaps the most significant review that has been written on the peptidase inhibitors, Laskowski and Kato [15] deplored the confusion of nomenclature that existed in the field in 1980. They pointed out that inhibitors are commonly discovered by their activity against readily available enzymes, most commonly trypsin, chymotrypsin or subtilisin, and then are named after the source organism or tissue, as ‘Streptomyces subtilisin inhibitor’ or ‘pancreatic trypsin inhibitor’. Such names give no clue to the relationships of the inhibitors, and make it difficult to know whether information that is available about the mechanism of action of one inhibitor can correctly be applied to another. It was evident to Laskowski and Kato [15] that peptidase inhibitors could best be classified in their homologous families, but the sequence information then available allowed only about a dozen families to be recognized.

The names used for peptidase inhibitors have not improved since 1980, but there is now a wealth of sequence data for these proteins, and the time seems right to make a new attempt at a systematic classification of them. In the text below, we first describe the methods we have used to classify the inhibitors in families and clans, and then describe and discuss the results we obtained.
METHODS

By way of preamble, we must mention that a major obstacle to the classification of peptidase inhibitors by their amino acid sequences is the fact that many of the proteins contain multiple homologous inhibitor domains in a single polypeptide chain. These domains are not identical to each other, but are functional inhibitors in their own right. The varying number of these repeats makes it impossible to assemble meaningful alignments of the entire amino acid sequences of the inhibitors with which to obtain measures of sequence similarity, as can be done for most other proteins. Because of this, the system of classification that we now propose is, strictly speaking, one for the inhibitory domains or ‘inhibitor units’ as we term them here, and not for the whole protein. It is only with the inhibitor units that one can quantify the structural similarities in the way that is necessary for the construction of dendrograms. That said, the great majority of inhibitors contain units from only a single family, and for these it is legitimate, for some purposes, to think of the whole protein as belonging to the given family.

Assembling peptidase inhibitor families

From the biochemical literature we assembled a list of known peptidase inhibitor proteins. The amino acid sequences of those containing 14 or more amino acid residues were obtained from the SWISS-PROT database or its supplement TrEMBL [16], downloaded in FASTA format [17] and copied into a library file. The program FASTA [17] (with the default BLOSUM50 matrix) was used in an all-against-all search of the library of inhibitors for sets of homologues. Homologous proteins were taken to be those for which the aligned amino acid sequences were related with an E value of 0.001 or less. With sequences as short as 14 residues it is potentially difficult to obtain matches at this level of significance, and the sunflower cyclic inhibitor was included in family 112 despite a less significant match on the basis of the study by McBride et al. [18]. The sets of homologous proteins formed provisional, or draft, families, and a well-studied inhibitor in each draft family was chosen as the ‘type-example’ for the family. The type-example was the representative member of the family to which all other members must be shown directly or indirectly (see below) to be homologous. If the provisional type-example was found to be a protein that contained more than one inhibitor unit (see below), one of these units was selected as the type-example. An ‘inhibitor unit’ was defined as the segment of the amino acid sequence containing a single reactive site (or bait region, for a trapping inhibitor) after removal of any parts that are known not to be directly involved in the inhibitory activity. A protein that contained only a single inhibitor unit was termed a simple inhibitor, and one that contained multiple inhibitor units was termed a compound inhibitor.

Each type-example-inhibitor-unit sequence was used to search the Entrez non-redundant protein sequence database by use of the BLASTP program [19] with the default BLOSUM62 matrix. Each sequence retrieved was considered to be a homologue, and a member of the same family as the type-example, if the E value of the alignment was 0.001 or less and the alignment spanned the inhibitor unit. A family was permitted to contain only a single member if no homologues were found. This process led to the identification of some proteins that were homologous to members of two draft families. Each such protein provided a link between the draft families and required them to be merged.

A number of well-known inhibitors did not show significant relationship directly to any type-example. For each of these, the sequence was run in a FASTA search against the entire collection of inhibitor sequences. When the result showed significant sequence relationship to one or more members of an inhibitor family other than the type-example, the query inhibitor was included in the family. A relationship such as this was described as a ‘transitive relationship’.

Dendrograms

A dendrogram for each family was prepared as follows. An amino acid sequence alignment for the family was prepared by use of CLUSTAL W [20], and a distance matrix of percentage identities was calculated. The distance matrix was converted into accepted point mutations according to the method of Schwartz and Dayhoff [21]. The tree was constructed by use of the Fitch-Margoliash algorithm with contemporary tips as implemented in the KITCH program from the PHYLP package [22].

Assembling of clans

We used the term ‘clan’ to designate a single evolutionary line of inhibitors defined by a single type of protein fold, very much as we did for peptidases [23]. A clan contains one or more complete families, since we assume that all members of a family have similar protein folds. We used definitions from the SCOP structural classification of proteins [24] to assemble clans of peptidase inhibitors. For these proteins, the superfamily level in the hierarchy of the SCOP database corresponds to our clan.

Terminology

Each clan of inhibitors was assigned an identifier (e.g. ‘IA’) formed from the letter ‘I’ or ‘J’ and an additional capital serial letter ‘A’ to ‘Z’. Once the series ‘IA’ to ‘IZ’ had been completed, the next clan was named ‘JA’, and the names ‘JB’ to ‘JZ’ are available for future use. The identifier assigned to each family of inhibitor units was formed from the letter ‘I’ and a serial number, e.g. ‘I25’. Many of the inhibitor units were assigned identifiers like ‘I01.001’, in which the first three characters represent the family identifier (padded with a zero when necessary) and the final three characters form a serial number. Inhibitors (or inhibitor units) that were found to have been characterized in some detail were assigned to the specific identifiers, and others were left as unassigned members of the given family. A compound inhibitor protein that contained several inhibitor units was termed a homotypic compound inhibitor, if the units were all from a single family, or a heterotypic compound inhibitor if they were from more than one family. The identifier for each homotypic compound inhibitor consisted of the letter ‘L’ followed by the family name of the inhibitor units (padded as necessary), a hyphen, and a three-digit serial number, e.g. ‘LI01-001’. For a heterotypic compound inhibitor, no single family name could be used in the construction of the identifier, so it was constructed as ‘L9’, followed by a one-digit serial number, a hyphen, and a three-digit serial number, e.g. ‘L190-001’.

RESULTS

Families of peptidase inhibitors

Our searches of the amino acid sequence databases led to the retrieval of 2500 sequences homologous to those of known peptidase inhibitors. From these, 48 families were built and were assigned identifiers (as described above). The families and the type-example for each are shown in Table 1, together with an indication of the families of peptidases that have been reported to be inhibited by proteins in the family.
The family for which the largest number of sequences was found is the serpin family, I4, with over 500 sequences. The serpins have been authoritatively reviewed by Silverman et al. [87]. Families I1, I2 and I25 each contain over 200 sequences, and the other families that have more than 100 members are I3, I7, I10.001) and sunflower cyclic inhibitor (I12.002), and units of less than 50 residues were found in families I7, I19, I31, I37 and I45.

Table 1 Families of peptidase inhibitors

The accession numbers, and for the residue numbers for the inhibitor units, are from the SWISS-PROT or TrEMBL databases. Where available, we also include the family assignment of each type-example-inhibitor unit in the Pfam database. Under 'families of peptidases inhibited' we list the MEROPS identifiers of the peptidase families of which one or more members have been reported to be inhibited by a protein from the inhibitor family, together with illustrative references.
not so well understood that we can predict with confidence that they are active inhibitors. For example, domains homologous to the family I1 (animal-type Kunitz) inhibitors occur in many multidomain proteins that are not known to be peptidase inhibitors. Again, the only known inhibitor in the large family I43 of immunoglobulin-like proteins is oprin, a snake venom metalloproteinase inhibitor from the Virginia opossum (Didelphis marsupialis) [75]. For these reasons, many uncharacterized homologues of inhibitors can be described only by the term ‘unassigned inhibitor homologues’.

Subfamilies in two families
The dendrograms for families I3 and I25 (results not shown) show deep divergences that justify the recognition of subfamilies. Each subfamily had been a separate draft family until transitive relationships were revealed by the BLAST searches. In family I3, the arrowhead protease inhibitor B (SWISS-PROT P07479) in subfamily I3B was shown to be homologous to members of subfamily I3A other than the type-example (e.g. SWISS-PROT Q39488). Family I25, containing the cystatins, has three subfamilies. The cysteine peptidase inhibitors are in subfamilies I25A (the type-1 cystatins) and I25B (the type-2 cystatins). The sequence of soya phytocystatin (SWISS-PROT Q39842) provided a transitive link between subfamilies I25A and I25B (e.g. to Q852N1). Amongst the significant relationships linking subfamilies I25B to I25C were that of chicken ovocystatin (P01038) to the first cystatin-like unit of bovine H-kininogen (P01044, residues 19–135). The domains in subfamily I25C are mostly not inhibitors of peptidases, but include inhibitors of a subtilisin homologue in peptidase family S8 [60] and of a snake-venom metalloendopeptidase in family M12 [61].

Compound inhibitors
Inhibitor units in eleven families were found to occur in proteins that contain two or more inhibitor units and are therefore defined as compound inhibitors (Table 2). For example, turkey ovomucoid contains three inhibitor units from the Kazal family, I1 (MEROPS; I01.001, I01.002, I01.003). All but a handful of the compound inhibitors are homotypic, i.e. contain units only from a single family. The number of inhibitor domains in the compound inhibitors ranges from 2 to 15. The few heterotypic compound inhibitors we encountered contained units from families I2 and I17; these were the Red Sea turtle chelolinian, human eppin [88] and the human WFIKKN and WFIKKNRP proteins [89,90]. As can be seen in Figure 1, the inhibitor units from the two families are arranged differently in chelolinian as compared with the mammalian proteins.

Assignment of families to clans
It is sometimes clear from similarities in tertiary structure that members of two families are distantly related to each other, and this justifies grouping the families in a clan. Similarities in protein fold are apparent between members of families I1, I5, I8 and I20, and these families are grouped in clan IA (Table 3). Similarly, families I2 and I52 are grouped in clan IB, and families I7 and I37 in clan IE. For 23 other families, three-dimensional structures show no relationship, and each family is assigned to a separate clan. Fifteen families cannot be assigned to any clan because no tertiary structure is available. The large proportion of single-family clans contrasts with the situation for peptidases, where most clans contain more than one family.

DISCUSSION
We set out on the task of establishing a system for the classification of peptidase inhibitors, expecting to use methods very similar to those we have developed for peptidases. But we became aware that there are at least three important differences between peptidases and their inhibitors, such that the inhibitors require special treatment. The crucial differences can be summarized as follows. (1) An effective inhibitor unit may contain as few as 14 amino acid residues, whereas the peptidase units commonly contain approx. 200 residues. (2) Reactive-site residues in inhibitor units are often not conserved in the way that active-site residues are in peptidases, so it is seldom possible to tell from the sequence whether or not an inhibitor homologue is likely to have inhibitory activity. (3) Inhibitor units in a dozen families have duplicated and reduplicated during evolution so that modern-day proteins exist that contain multiple divergent copies of them. In contrast, multiple peptidase units within a single polypeptide occur in only a very few peptidases (amongst which angiotensin-converting enzyme and metallocarboxypeptidase D are best known).

The style of identifiers for inhibitor families
The families of peptidases are divided into five main groups according to the chemistry of the catalytic sites of the enzymes [92]. In contrast with this, we have proposed here a single series for the families of inhibitors. The possibility was considered that the families could be usefully classified according to the catalytic types of the peptidases inhibited, in sets of serine peptidase inhibitors, cysteine peptidase inhibitors, and so on. Such classification has often been used in the past, and has been authoritatively advocated [93]. There are difficulties with this approach, however. As is shown in Table 1, a number of families contain inhibitors of peptidases of more than one catalytic type. For example, the proteins in family I3 (the plant Kunitz-type inhibitors) generally inhibit serine peptidases of family S1, but also include inhibitors of cysteine peptidases (C1) and the aspartic peptidase cathepsin D (A1) (Table 1) as well as, possibly, subtilisins [93a]. Similarly, family I4, the serpin family, has been thought of as a family of serine peptidase inhibitors, but it also contains CrmA and other important inhibitors of cysteine peptidases (Table 1). We concluded that it would not be feasible to classify and name the families of inhibitors according to the catalytic types of the peptidases that they inhibited, and simply numbering them in one series.

Classification at the level of inhibitor units
In the Results section, and specifically in Table 2, we drew attention to the fact that many of the proteins that inhibit peptidases contain multiple inhibitory domains. This clearly raised an issue of policy for the classification; were we to attempt a classification of the entire multidomain proteins, or only of the individual inhibitor units? It was clear that only the isolated inhibitor units could be handled by methods similar to those we have used successfully for peptidases, and that these should be the primary objects in the classification. This decision committed us to a kind of classification of the peptidase inhibitors that has not been previously used, as far as we are aware.

Notwithstanding the decision to make inhibitor units the objects that would populate the inhibitor level of the hierarchical classification of peptidase inhibitors, there remained a need for consistent identifiers for the compound inhibitor proteins. These must indicate the classification of the inhibitor units contained in
Table 2  Families of inhibitor units that are found in compound inhibitors

The inhibitor units that occur in compound inhibitors belong to the twelve families listed here. The examples include proteins that are simple inhibitors (number of units = 1) and homotypic compound inhibitors, which contain units from only one family. Square brackets contain the SWISS-PROT identifiers and MEROPS identifiers.

<table>
<thead>
<tr>
<th>Family</th>
<th>Number of inhibitor units</th>
<th>Example (source) [identifier]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family I1</td>
<td>1</td>
<td>elastase inhibitor (Anemonia sulcata) [P16895; I01.108]</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>bikazin (Canis familiaris) [P01002; LI01-003]; rhodniin (Rhodnius prolixus) [Q06684; I01-007]</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>ovumucoid (Melagris gallopavo) [P01004; LI01-001]</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>dipetalogastin (Dipetalogaster maximus) [O16790; LI01-006]</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>agCP6264 protein (Anopheles gambiae) [EA11963; putative]</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>ovoinhibitor (Gallus gallus) [P10184; LI01-002]</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>agrin (Rattus norvegicus) [P25304; putative]</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>serine protease inhibitor Kazal type 5 [Q0NQ38; LI01-004]</td>
</tr>
<tr>
<td>Family I2</td>
<td>1</td>
<td>aprotinin (Bos taurus) [P00974; I02.001]</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>α-1-microglobulin [P02760]; bikunin [P02760; LI02-001]; hepatocyte growth factor activator inhibitor 1 (all Homo sapiens) [O43278; LI02-004]</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>tissue-factor-pathway-inhibitor-1 (Homo sapiens) [P10646; LI02-002]; K10D3.4 protein (Caenorhabditis elegans) [CAA75670; putative]</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>ZC46.6 protein (Caenorhabditis elegans) [CAA49924; putative]; W01F3.3 protein (Caenorhabditis elegans) [F25517; putative]</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Y4S8B.3 protein (Caenorhabditis elegans) [F25517; putative]</td>
</tr>
<tr>
<td>Family I8</td>
<td>1</td>
<td>Ascaris chymotrypsin/elastase inhibitor [P07851; I08.001]</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>C09F9.2 protein (Caenorhabditis elegans) [CAB03861; putative]</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>HmEGFL-1 protein (Herdmania momus) [AAB60604; putative]</td>
</tr>
<tr>
<td>Family I12</td>
<td>1</td>
<td>sunflower cyclic inhibitor (Helianthus annuus) [I12.002]</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Bowman–Birk inhibitor (Gyneae majus) [P01005; LI12-001]</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Bowman–Birk trypsin inhibitor (Hordeum vulgare) [P12940; LI12-001]</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Bowman–Birk trypsin inhibitor P0037C04.14 (Drosophila virilis) [BAB55527; putative]</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Bowman–Birk trypsin inhibitor P0037C04.11 (Drosophila virilis) [Q09534; putative]</td>
</tr>
<tr>
<td>Family I15</td>
<td>1</td>
<td>hirustasin (Hirudo medicinalis) [P80302; LI15.001]</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>antistasin (Haementeria officinalis) [P15858; LI15-002]; ghilanten (Haementeria ghilianii) [P16242; LI15-003]</td>
</tr>
<tr>
<td>Family I17</td>
<td>1</td>
<td>elafin (Homo sapiens) [P19997; I17.002]</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>mucus proteinase inhibitor (Homo sapiens) [P03973; LI17-001]</td>
</tr>
<tr>
<td>Family I19</td>
<td>1</td>
<td>proteinase inhibitor II (Schistocerca gregaria) [Q04613; I19.001]</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>proteinase inhibitor LCM I (Locusta migratoria) [P89060; I19.001]</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>pacilastin-related peptide precursor (Schistocerca gregaria) [Q8MMK3; putative]</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>agCP9037 protein (Anopheles gambiae) [AA00733; putative]</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>pacilastin (Paestacoccus teniusculus) [P91776; LI19-001]</td>
</tr>
<tr>
<td>Family I20</td>
<td>1</td>
<td>proteinase inhibitor P11 (Solanum tuberosum) [P01079; putative]</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>proteinase inhibitor P12 (Solanum tuberosum) [P01090; LI20-001]</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>proteinase inhibitor II (Nicolitana tabacum) [Q04561; LI20-002]</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>proteinase inhibitor (Nicotiana alata) [Q95307; putative]</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>proteinase inhibitor (Nicotiana alata) [Q40378; putative]</td>
</tr>
<tr>
<td>Family I25</td>
<td>1</td>
<td>cystatin A (Homo sapiens) [P01040; I25.001]</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>metalloproteinase inhibitor (Bathypore trilobata) [Q9DG10; I25.026]</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>kininogen (Homo sapiens) [P01042; LI25-002]; multicycstin (Heilansis annuus) [BA956416; LI25-006]</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>multicycstin (Solanum tuberosum) [P37842; LI25-001]</td>
</tr>
<tr>
<td>Family I27</td>
<td>2</td>
<td>calpastatin Calp1 (Xenopus laevis) [CAB4944; putative]</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>calpastatin (Homo sapiens) [P20811; LI27-001]</td>
</tr>
<tr>
<td>Family I29</td>
<td>1</td>
<td>cystotoxic T-lymphocyte antigen (Mus musculus) [P12399]</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>salarin [91]</td>
</tr>
<tr>
<td>Family I31</td>
<td>1</td>
<td>MHC II invariant chain protein (P04233; I31.002)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>saxiphilin (Rana catesbeiana) [P31226; putative]</td>
</tr>
</tbody>
</table>

the compound inhibitor as far as possible, but not risk confusion with identifiers that truly fit into the hierarchical system. We adopted identifiers like ‘LI01-001’ in which the initial ‘L’ and the hyphen as the fifth character are common to all the identifiers for compound inhibitors. In this example, ‘I01’ indicates that this is a homotypic compound inhibitor containing units of family I1, and ‘001’ is simply the first in a sequence of serial numbers. The few heterotypic compound inhibitors we found all contain units from families I2 and I17, so there is no single family name that can be used to construct an identifier in exactly this way. We arbitrarily used ‘I90’ in place of the family identifier in constructing identifiers for the WFIKKN and WFIKKNRP.
The proteins as well as cheluinin and eppin, e.g. L190-003 for cheloinian.

**Disulphide-bond patterns**

Many peptidase inhibitors contain disulphide bonds, and in some families these may stabilize reactive-site loops, facilitating the resynthesis of the reactive-site bond after it has been cleaved by the target enzyme. Historically these disulphides have been used to confirm assignments to families when sequence similarities were low, and to reveal the repetition of inhibitor units in compound inhibitors [15]. The arrangements of the disulphide bonds in the various clans and families are shown schematically in Figure 2. The conservation of disulphide-bonding patterns within clans is not striking, and we suggest that the many crystallographically determined structures now available demonstrate distant evolutionary relationships much more clearly.

### Inhibitors and their homologues

Clearly the crucial property of all the proteins we are seeking to classify here is their inhibition of peptidases, and it is therefore appropriate to consider to what extent their grouping in families and clans reflects their inhibitory activities and mechanisms. First it has to be noted that there are many homologues of the known inhibitors for which there are no published experimental data as to their possible inhibitory activity. Nor are we able to predict such activity. For nearly all families of peptidases, conserved active-site residues are known that appear to be essential for catalysis. This means that a homologue of a known peptidase that contains all of these residues can reasonably be termed a putative peptidase, and any that lacks one or more of these residues can confidently be termed a non-peptidase homologue. In the MEROPS database the non-peptidase homologues are listed separately from the active, or potentially active, peptidases. But the situation is very different for the inhibitors. The reactive-site residues of the inhibitors are

#### Table 3 Clans of peptidase inhibitors and inhibitor units

<table>
<thead>
<tr>
<th>Clan</th>
<th>Type structure for clan (source; Protein Data Bank accession code)</th>
<th>Families included</th>
<th>Fold description for family</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA</td>
<td>ovomucoid domain 3 (Coturnix coturnix; 1OVD)</td>
<td>11</td>
<td>disulphide-rich small proteins with α and β</td>
</tr>
<tr>
<td>IB</td>
<td>aprotinin (Bos taurus; 4PTI)</td>
<td>11</td>
<td>disulphide-rich small proteins with α and β</td>
</tr>
<tr>
<td>IC</td>
<td>soybean trypsin inhibitor (Glycine max; 1AVU)</td>
<td>15</td>
<td>disulphide-rich small proteins nearly all β</td>
</tr>
<tr>
<td>ID</td>
<td>α2-proteinase inhibitor (Homo sapiens; 1ATU)</td>
<td>16</td>
<td>disulphide-rich small proteins with α and β</td>
</tr>
<tr>
<td>IE</td>
<td>trypsin inhibitor MCTI-1 (Momordica charantia; 1F2S chain I)</td>
<td>120</td>
<td>disulphide-rich small proteins with α and β</td>
</tr>
<tr>
<td>IF</td>
<td>Bowman–Birk plant trypsin inhibitor (Vigna angularis; 1P12)</td>
<td>12</td>
<td>disulphide-rich small proteins with α and β</td>
</tr>
<tr>
<td>IG</td>
<td>eglin C (Hirudo medicinalis; 1E59)</td>
<td>13</td>
<td>α- and β-sandwich</td>
</tr>
<tr>
<td>IH</td>
<td>ovocystatin (Galus gallus; 1CEW)</td>
<td>14</td>
<td>α- and β-core; helix packs against coiled antiparallel β-sheet</td>
</tr>
<tr>
<td>U</td>
<td>rati seed trypsin-α-amylose inhibitor (Eleusine coracana; 1B1U)</td>
<td>17</td>
<td>all-helical (folded leaf)</td>
</tr>
<tr>
<td>IK</td>
<td>metalloproteinase inhibitor Erwinia (Erwinia chrysanthemi; 1SMP chain I)</td>
<td>138</td>
<td>closed β-barrel</td>
</tr>
<tr>
<td>IL</td>
<td>α2-macroglobulin (Homo sapiens; 108)</td>
<td>139</td>
<td>knottin</td>
</tr>
<tr>
<td>IM</td>
<td>hirudin (Hirudo medicinalis; 4HTC chain I)</td>
<td>114</td>
<td>knottin</td>
</tr>
<tr>
<td>IN</td>
<td>ecotin (Escherichia coli; 1EC2)</td>
<td>115</td>
<td>α-sheet</td>
</tr>
<tr>
<td>IO</td>
<td>antistasin inhibitor unit 1 (Haementeria officinalis)</td>
<td>117</td>
<td>knottin</td>
</tr>
<tr>
<td>IP</td>
<td>ellafin (Homo sapiens; 1F1E chain I)</td>
<td>118</td>
<td>knottin</td>
</tr>
<tr>
<td>IQ</td>
<td>baculovirus p35 caspase inhibitor (Spondoptera litura nucleopolyhedrovirus; 113S)</td>
<td>125</td>
<td>Greek key β-sandwich</td>
</tr>
<tr>
<td>IR</td>
<td>ascaris pepsin inhibitor Pi-3 (Ascaris suum; 1F32)</td>
<td>135</td>
<td>disulphide-rich small proteins with α and β</td>
</tr>
<tr>
<td>IS</td>
<td>leech carbonpeptidase inhibitor (Hirudo medicinalis; 1BD7 chain I)</td>
<td>146</td>
<td>OB-fold</td>
</tr>
<tr>
<td>IT</td>
<td>TIMP-1 (Homo sapiens; 1UFA chain B)</td>
<td>136</td>
<td>Greek key β-sandwich</td>
</tr>
<tr>
<td>IU</td>
<td>Streptomyces metalloproteinase inhibitor (Streptomyces nigrescens; 1BH1U)</td>
<td>138</td>
<td>disulphide-rich small proteins with α and β</td>
</tr>
<tr>
<td>IV</td>
<td>BIRC-3 protein (Homo sapiens; 1E31)</td>
<td>132</td>
<td>metal-bound small proteins with α and β</td>
</tr>
<tr>
<td>IW</td>
<td>proteinase inhibitor LCMI I (Locusta migratoria; 1GL1)</td>
<td>119</td>
<td>disulphide-rich small proteins all β</td>
</tr>
<tr>
<td>IX</td>
<td>MHC II invariant-chain p41 form (Homo sapiens; 1ICF)</td>
<td>131</td>
<td>intertwined trimer of identical 3-helical subunits</td>
</tr>
<tr>
<td>IY</td>
<td>subtilisin inhibitor (Streptomyces altogriseolus; 3SS1)</td>
<td>115</td>
<td>α- and β-sandwich</td>
</tr>
<tr>
<td>IZ</td>
<td>triabin (Triatoma pallidipennis; 1AVG)</td>
<td>159</td>
<td>open or closed β-barrel (lipocalin-like)</td>
</tr>
<tr>
<td>JA</td>
<td>protease-A-inhibitor-3 (Saccharomyces cerevisiae; 1DPJ)</td>
<td>134</td>
<td>non-globular all-α</td>
</tr>
</tbody>
</table>
Evolutionary families of peptidase inhibitors

Mechanisms of inhibition

Having assigned the inhibitor units to families, one can ask to what extent the mechanisms of inhibition are constant within a family, and how it is that some families contain inhibitors of very different peptidases? Seven of the families and subfamilies in Table 1 (I3A, I4, I8, I12, I16, I25C, I39) can be seen to contain proteins that inhibit peptidases of more than one catalytic type, and four more (I13, I25B, I35, I50) contain inhibitors of peptidases from more than one family. This implies that either structurally dissimilar peptidases can be inhibited by a single mechanism, or structurally similar inhibitors inhibit by different mechanisms. The ways in which the inhibitors interact with their target enzymes vary enormously, but two general types of interaction can be recognized: irreversible ‘trapping’ reactions, and reversible tight-binding reactions.

1. Trapping reactions

The kind of interaction that depends most directly on the peptidase activity of the target enzyme is that which can be described as ‘trapping’. This kind of reaction is specific for endopeptidases because it depends upon the cleavage of an internal peptide bond in the inhibitor that triggers a conformational change. Any catalytically inactive form of a peptidase, such as anhydrotrypsin, fails to enter into a trapping reaction, although it may well bind tightly to an inhibitor from one of the reversible classes. Trapping reactions are never truly reversible because unmodified inhibitor is not reformed, so the inhibitor can also be described as a suicide inhibitor. The three families that show trapping reactions are I4, I39 and I50. In these families there is some flexibility as to which peptide bond is cleaved to ‘close the trap’, and this broadens their inhibitory spectra. In families I4 and I50 the enzyme–inhibitor complex is normally covalent, and covalent complexes may also be formed by the macroglobulins in family I39.

In the large and widespread serpin (I4) family, the cleavage of an appropriate peptide bond in the reactive-site loop of the inhibitor triggers a dramatic conformational change, which is so rapid that catalysis proceeds only to the formation of an acyl enzyme and release of the C-terminal part of the reactive-site loop [29]. The N-terminal part of the loop inserts into a β-sheet, carrying the enzyme molecule still attached as the acyl enzyme to the opposite pole of the inhibitor molecule. This violent event disrupts the structure of the enzyme molecule and its catalytic site, so that hydrolysis of the acyl enzyme does not proceed and the covalent complex persists. The formation of an acyl enzyme is crucial to the serpin interaction, and the reaction is specific for the serine and cysteine peptidases that form acyl enzymes, whereas metallopeptidases may simply turn over the inhibitor without formation of any complex [95]. The reaction of serpins with cysteine endopeptidases of the papain family is apparently similar to that for serine peptidases [31]. Most serpins are inhibitors either of serine peptidases or cysteine peptidases, but not both. However, it has been shown that both antithrombin [32] and the mouse serpin SQN-5 are inhibitors of both serine and cysteine peptidases [31]. The use of non-identical but overlapping reactive sites in the serpin family is reviewed by Al-Khunaizi et al. [31], and the way in which the serpin architecture supports inhibition of cysteine peptidases has been systematically investigated by Irving et al. [96].

The second family of inhibitors that mediate trapping reactions is the macroglobulin family, I39. In these proteins, as illustrated by α2-macroglobulin, cleavage of the inhibitor in the highly susceptible ‘bait region’ triggers the conformational change by which the target enzyme is trapped within the large re-folded inhibitor molecule [72]. The enzyme–inhibitor complex is stabilized largely by steric effects, although highly reactive thiol-ester groups can also form covalent links [97]. The target enzyme must be an endopeptidase to cleave the bait region of the inhibitor, and not too large a molecule to be enclosed by the macroglobulin, but any catalytic class of endopeptidase can be susceptible. Because there is no persistent interaction with the active site of the enzyme, the complexed enzyme molecule continues to hydrolyse accessible substrate molecules; these include artificial substrates and other small molecules, but not proteins. The exceptionally broad specificity of inhibition by α2-macroglobulin, embracing endopeptidases of four catalytic types, is due in part to the fact that a great variety of bonds in the bait region of the molecule can be cleaved to trigger the trapping reaction [98].

Family I50 contains the baculovirus protein p35, which blocks the apoptosis of host cells by inhibition of the caspases. Cleavage of the caspase-sensitive bond in the reactive-site loop again leads to a conformational change that stabilizes the acyl-enzyme in an irreversible complex [81,99]. Protein p35 also inhibits gingipain K (family C25), and the cleavage site for inhibition of the gingipain is Lys-94, seven residues C-terminal to the caspase inhibitory site [82].

---

**Figure 2 Patterns of disulphide bonds in type-examples from the inhibitor families**

The distances between half-cystine residues in the polypeptide chains are not drawn to scale. Data are from the SWISS-PROT records cited in Table 1. The red circles indicate the positions of reported reactive-site residues, but do not exclude the possibility that there may be others elsewhere (because the disulphide-bonding pattern is not conserved between the two units in the example from family I12, both units are shown).

by no means strictly conserved, and indeed have been found to be amongst the most variable residues in several families [15,94]. Because of this, we are unable to distinguish non-inhibitory homologues from putative inhibitors simply by sequence, and can only describe a related protein that has not been demonstrated to have inhibitory activity as an ‘unassigned inhibitor homologue’.

---

CLAN IN (I11)
CLAN IK (I38)
CLAN IC (I3A)
CLAN IB (I36)
CLAN IM (I14)
CLAN IR (I33)
CLAN IA (I12)
CLAN IE (I7)
CLAN IW (I19)
CLAN IZ (I39)
CLAN IH (I25B)
CLAN IA (I12)
CLAN IB (I12)
CLAN IC (I3A)
CLAN IE (I7)
CLAN IY (I16)
CLAN IU (I36)
CLAN IN (I11)
CLAN IK (I38)
CLAN IC (I3A)
CLAN IB (I36)
CLAN IM (I14)
CLAN IR (I33)
CLAN IA (I12)
CLAN IE (I7)
CLAN IW (I19)
CLAN IZ (I39)
CLAN IH (I25B)
CLAN IA (I12)
CLAN IB (I12)
CLAN IC (I3A)
CLAN IE (I7)
CLAN IY (I16)
CLAN IU (I36)
CLAN IN (I11)
CLAN IK (I38)
CLAN IC (I3A)
CLAN IB (I36)
CLAN IM (I14)
CLAN IR (I33)
CLAN IA (I12)
CLAN IE (I7)
CLAN IW (I19)
CLAN IZ (I39)
CLAN IH (I25B)
CLAN IA (I12)
CLAN IB (I12)
CLAN IC (I3A)
CLAN IE (I7)
CLAN IY (I16)
CLAN IU (I36)
CLAN IN (I11)
CLAN IK (I38)
CLAN IC (I3A)
CLAN IB (I36)
CLAN IM (I14)
CLAN IR (I33)
CLAN IA (I12)
CLAN IE (I7)
CLAN IW (I19)
CLAN IZ (I39)
CLAN IH (I25B)
CLAN IA (I12)
CLAN IB (I12)
CLAN IC (I3A)
CLAN IE (I7)
CLAN IY (I16)
CLAN IU (I36)
CLAN IN (I11)
CLAN IK (I38)
CLAN IC (I3A)
CLAN IB (I36)
CLAN IM (I14)
CLAN IR (I33)
CLAN IA (I12)
CLAN IE (I7)
CLAN IW (I19)
CLAN IZ (I39)
CLAN IH (I25B)
CLAN IA (I12)
CLAN IB (I12)
CLAN IC (I3A)
CLAN IE (I7)
CLAN IY (I16)
CLAN IU (I36)
CLAN IN (I11)
CLAN IK (I38)
CLAN IC (I3A)
CLAN IB (I36)
CLAN IM (I14)
CLAN IR (I33)
CLAN IA (I12)
CLAN IE (I7)
CLAN IW (I19)
CLAN IZ (I39)
CLAN IH (I25B)
2. Reversible tight-binding interactions

Inhibitors in this group make high-affinity interactions with the active site of the target enzyme. The mechanism that has been studied in greatest detail is that termed the ‘standard’ mechanism by Laskowski and Qasim [100]. The inhibitory unit of a standard mechanism inhibitor has a single reactive-site peptide bond, and inhibition is caused by the binding of the inhibitor to the enzyme in a substrate-like fashion. The intact ‘virgin’ form of the inhibitor molecule exists in the complex in equilibrium with the ‘modified’ form of the inhibitor, in which the reactive-site peptide bond is cleaved, and the complex can dissociate to yield either virgin or modified inhibitor. The standard mechanism has been demonstrated conclusively only for inhibitors of serine peptidases.

We recognize 19 families of such standard mechanism inhibitors. The reactive-site bonds occur in structurally similar reactive-site loops, and the conformation of the peptide containing the reactive site, which is the same in all of the inhibitors, is described as ‘canonical’ [46,100]. Crystallographic structures of enzyme–inhibitor complexes show that standard mechanism inhibitors are contained in families I1, I2, I3, I7, I8, I10, I11, I12, I13, I15, I16, I17, I18, I19, I20, I36 and I40. In addition, the structures of the uncomplexed ascidian trypsin inhibitor from Halocynthia (family I5; [101]) and the corn Hageman-factor inhibitor (family I6; [102]) suggest that they too probably inhibit by the standard mechanism.

The 19 families of standard mechanism inhibitors fall into 13 different clans because of their different protein folds, and it therefore seems certain that the standard mechanism of peptidase inhibition has evolved on many separate occasions. Clearly, reactive-site loops that can adopt the canonical conformation, and thus inhibit serine endopeptidases by the standard mechanism, can be built on to a wide variety of supporting scaffold structures.

The families (or subfamilies) I3A, I8, I112 and I16 that generally inhibit serine peptidases by the standard mechanism each contain reversible inhibitors of peptidases of other catalytic types, but little is yet known about the structural basis for this. Intuitively, it seems likely that different reactive sites and mechanisms are responsible. However, the small Kunitz-type (I3) inhibitor of Prosopis juliflora has been reported to have overlapping inhibitory sites for trypsin and papain [27]. Also, there is evidence that in family I13 a single site is responsible for the inhibition of two very different types of serine peptidases, chymotrypsin (S1) and subtilisin (S8), by an inhibitor from wheat [103].

Inhibitors of thrombin have evolved in a wide variety of animals that feed on vertebrate blood and therefore need agents that can prevent it from clotting. The control of blood coagulation is also a medically important topic, so the natural inhibitors of thrombin have been the subjects of intense study. Hirudin and haemadin in family I14 are clearly not standard mechanism inhibitors. They enhance the specificity and affinity of their inhibition of thrombin by binding not only in the active site, but also to distant exosites [104,105]. The use of exosites is further extended by triabin (family I59) which interacts with thrombin exclusively via its fibrinogen-recognition exosite [86], and the complex that is inactive against fibrin retains activity on small-molecule substrates.

In the best-known family of cysteine peptidase inhibitors (I25), ovocystatin inhibits papain by binding to sites on either side of the active site, which becomes blocked. There is no interaction with the cysteine nucleophile but, instead, residues 8 and 9 interact with the S2 binding pocket in a substrate-like manner [5,58]. Remarkably, ovocystatin has a second independent inhibitory site for the non-papain-like cysteine peptidase, legumain (family C13, clan CD) [59]. The structure of the p41 II fragment (family I31) bound to cathepsin L [106] shows a three-loop arrangement of the p41 fragment reminiscent of the inhibitory edge of cystatins, but also suggests the reason why inhibitors based on the thyroglobulin type-1 fold inhibit much more selectively than the cystatins [5].

The inhibitors in family I29 are homologous to propeptidases of several cathepsins including cathepsin L, and the crystal structures of the cathepsin proenzymes show how the inhibitors are likely to interact with the enzymes. The propeptidases run through the active-site cleft in the reverse direction to substrates, blocking the active site and maintaining the catalytically inactive state of the proenzymes [5]. A further example of an inhibitor binding backwards in the active-site cleft of a cysteine peptidase is seen in XIAP (X-linked inhibitor of apoptosis protein) and its homologues in family I32. This tight-binding, reversible interaction has been reviewed by Stennicke et al. [99].

The crystal structure of staphostatin B (family I57), in complex with staphostatin, shows that the inhibitor runs through the active-site cleft in a substrate-like way, but the P1 Gly98 residue (conserved throughout the family) has a strained backbone conformation that would be sterically forbidden to any other residue [107].

Few protein inhibitors of aspartic peptidases are known, but Ascaris PI-3 (family I33) inhibits pepsin by blocking part of the substrate-binding site. The inhibitor molecule makes interactions with the pepsin ‘flap’ (subsites S1–S3) such that its three N-terminal residues occupy the S1 ‘to’ S3 ‘subsites, thus preventing substrates from binding [66]. Remarkably, the yeast IA3 inhibitor (I34) seems to lack secondary structure until it is bound to its natural target enzyme, saccharopepsin [67].

In family I35, the molecule of the tissue inhibitor of metalloproteinase inhibitors, TIMP-1, is wedge-shaped, and the long edge occupies the active-site cleft of stromelysin-1 (family M10), binding either side of the catalytic zinc [68]. It seems probable that the mechanism of inhibition of ADAM (a disintegrin and metalloproteinase-like)-type (M12) metzincin metallopeptidases by TIMP-3 and other members of family I35 are similar [69].

The Streptomyces metalloproteinase inhibitor (family I36) is thought to inhibit thermolysin by the standard mechanism [70], but this awaits confirmation.

In family I38, the N-terminus of the Pseudomonas metalloproteinase inhibitor molecule binds to the peptidase along the active-site cleft in an extended conformation with the N-terminal serine acting as a zinc ligand, displacing the activated water molecule and thus inhibiting the peptidase [71].

Although it has a protein fold similar to that of the standard mechanism serine peptidase inhibitors from Cucurbitaceae (family I7), the potato carboxyproteinase inhibitor of family I37 (also in clan IE) inhibits by a very different mechanism. It binds the metallocarboxyproteinase A in a substrate-like fashion with its four C-terminal residues inserted into the active-site cleft. The C-terminal Gly-39 is slowly split off, but remains bound in the S1 subsite where it is effectively buried by the remainder of the inhibitor. The complex has been described as ‘an enzyme–product intermediate in the catalytic mechanism’ [46]. Leech carboxypeptidase inhibitor (I46) interacts in a similar way, despite having a different protein fold [77].

Distribution of families amongst organisms

We concluded from the large number of clans of peptidase inhibitors, that inhibitors have arisen on many different occasions during the evolution of living organisms. From the data presently available, it seems that most of these events have occurred in the course of the evolution of eukaryotes. Figure 3 shows how the families of inhibitors are distributed in the three superkingdoms of cellular organisms, and in viruses. We found reports of peptidase
Figure 3  Distribution of inhibitor families throughout the superkingdoms of cellular organisms

The figure summarizes data for the distribution of the families that can be found in the MEROPS database. Although, for clarity, no intersection is shown between viruses and the other groups, it should be noted that family I4 occurs in all four groups.

inhibitors from all kinds of organisms, but the known families are, at present, numerous only in eukaryotes. Only three families are known so far from Archaea, two of which [I4 (serpins) and I42 (chagasin)] are present in all three superkingdoms; I4 is the most widespread of all, being found even in viruses. Distinct inhibitors, however, are most numerous in eukaryotes. None of the prokaryote genomes we examined contains more than six genes encoding members of the families of peptidase inhibitors so far recognized, whereas all of the eukaryotic genomes contain tens or hundreds of them (precise counts being shown in the MEROPS database).

Conclusions

In conclusion, we have assigned the inhibitory domains, termed inhibitor units, of the proteins that inhibit peptidases to families on the basis of their sequence relationships. We have also grouped together some families of distantly related inhibitors in clans. It was necessary to work at the level of the inhibitor units rather than the whole proteins, because many of the proteins contain multiple inhibitor units. A simple system of terminology for reference to the clans, families and inhibitor units within a hierarchical system is introduced. The complete compound inhibitors that contain multiple inhibitor units cannot be classified in any simple way, but a set of identifiers is used for these too. Not unexpectedly, we find that inhibitors in a single family tend to inhibit peptidases of a single catalytic type by a single kind of mechanism. But some families contain inhibitors that target peptidases in more than one family, or even catalytic type, and the mechanisms of these reactions may well be different, although few of them have yet been studied in detail. The new system of classification of the peptidase inhibitors has allowed them to be added to the MEROPS database, where additional details will be found.

We thank Dr Alex Bateman and our colleagues in the Pfam database team for helpful discussions, and the Medical Research Council and the Wellcome Trust for financial support.

REFERENCES


© 2004 Biochemical Society


Evolutionary families of peptidase inhibitors


