Chapter 6

Analysis of Masquerade: Case Study for Finding Function of Serine Protease Domains in Modular Proteins Involved in Patterning and Immune Response

6.1 Abstract

We describe here the structure-function relationship of the eukaryotic serine proteases, the most comprehensively studied enzyme family. A method to identify functional information from structures alone is described by analysis of structural properties of redundant data set of five sub families and a non-redundant data set of 43 protease structures derived from PDB using position specific information. Sequence information from SwissProt sequence database is used subsequently to derive 'consensus' functional residues. The analysis is then used to find the functional role of modular proteins containing functional and non-functional serine protease domains involved in early development, patterning and immune response with emphasis of masquerade sequence and its reported homologues. Masquerade is a protein reported for cell adhesion and contains C-terminal non-functional serine protease like domain. We have identified five chitin-binding motifs in N-terminal cysteine-knot domain in masquerade and related proteins. We propose the mechanism of binding and subsequent cleavage for the proteins having dual role in patterning and immune responses. Role of masquerade serine protease domain in early immune response is proposed. We also report two chitin-binding motifs for Drosophila *GRAAL* gene product identified from analysis of Drosophila genome and propose its role in patterning and immune response.

6.2 Introduction

6.2.1 The system

This chapter describes structure-function relationships in one of the largest and most comprehensively studied of all enzyme families, the serine proteases with trypsin fold and its application to other modular proteins containing this domain. It is also called the chymotrypsin fold, as it was the first member of the family to be crystallised. These enzymes function using catalytic triad made of His57, Asp102 and Ser195 (Lesk, 1981; Fersht, 1984). A similar catalytic triad has been observed in other proteases such as subtilisin (Kraut, 1971) and carboxypeptidases (Liao *et al*., 1992), as well as lipases (Dodson *et al*., 1992), but theses molecules have different structures. The crystal structure of α-chymotrypsin was reported in early days of crystallography in 1967 (Mathews *et al*., 1967). The 2Å-resolution structure followed soon (Birktoft and Blow, 1972). The structural information has grown enormously since then. Extensive investigations have elucidated the mechanism of enzymatic catalysis, folding pattern, inhibition, activation, and substrate specificity and have described the evolutionary variation in the family (Stroud, 1974; Neurath, 1975; Lesk and Fordham 1996).

Mammalian serine proteases (also called proteinases) participate in numerous physiological processes (Barret, 1977,1994; Horl, 1989; Bond, 1991; Twining *et al*., 1994); the best known are digestion, blood clotting (Davie *et al*., 1991), fertilisation (Baba *et al*., 1989), development (Gurwitz and Cunningham 1998), complement activation in the immune response in vertebrates (Reid *et al*., 1986; Goldberger *et al*., 1987) and insects (Kwon *et al*., 2000; Paskewitz *et al*., 1999). Their roles have also been suggested in signal transduction systems (Smirnova *et al*., 2001; Pendurthi *et al*., 2000). In several disease states, including emphysema (Watorek *et al*., 1988), tumor metastasis (Henderson *et al*., 1992), and arthritis (Froelich *et al*., 1993), the levels of proteases or inhibitors are elevated or out of balance. Proteins containing serine proteases domain are often modular (or mosaic) in nature. They sometimes contain multiple copies of different domains. Various proteins having protease domain are also known to contain Kringle (e.g plasminogen and apolipoprotein); Sushi (e.g. complement factor B and Limulus clotting factor); Apple (plasma kallikrein and coagulation factor XI); Growth factor and Ca^{2+} binding (coagulation factor VII and coagulation factor IX); and Finger (coagulation factor XII and t-plasminogen activator) domains (Barrett, 1994). Proteins like Sb-Sbd (Appel *et al*., 1992), snake (Delotto and Spierer, 1986) easter (Chasan and Anderson, 1989) and Limulus contain five repeats of disulfide motifs in N-terminus and serine proteases in C-terminus. These proteins are reported in taking part in patterning during development (Murugasu-Oei *et al*., 1995). There are reports of mosaic proteins having domains homologues to serine proteases that have mutations in their catalytic triad (Donate *et al*., 1994; Murugasu-Oei *et al*., 1995). Masquerade is a member of this category.

6.2.2 Description of Structure

The 1.55 release of **S**tructural **C**lassification **O**f **P**roteins (SCOP) database (Murzin *et al*., 1995) describes proteins having trypsin fold under all β class. It describes the fold as containing internal gene duplication with two closed barrels having 6 strands in each barrel and shear number of 8 as shown in Figure1. There are 3 invariant disulfide bridges (C42-C58, C168-C182 and C191-C220 for all family members (chymotrypsin numberings are used throughout in text). The other disulfide bridges differ. For example, trypsin of higher organisms has disulfide bridge formed by C128 and C132, which is, absent in thrombin. On the other hand thrombin has a disulfide bridge between C1 and C122, absent in trypsin. The superfamily of trypsin fold serine protease is divided in to four broad families. They are eukaryotic protease, prokaryotic protease, viral protease and viral cysteine protease of trypsin fold with respectively forty-eight, nine, four and three family members with different functional specificity. For example trypsin, chymotrypsin, thrombin, elastase, collegenase, coagulation factors etc. are classified as eukaryotic serine proteases family members. It is evident that a wealth of structural information is available for this particular domain. More than one structure is available for certain subfamilies as shown by SCOP records. For example crystal structures of trypsin (ogen) from various species like cow, pig, rat, human, atlantic salmon and mold are available. Further more there are 125 structures deposited alone for bovine (cow) trypsin, 21 structures from rat and 13 structures from pig. The reason being that each structure talks about different mutants or protease complexes with different chemical or protein inhibitors and how that might give important insights about the catalytic capacity or specificity of the proteases.

There has been a previous report of analysis of serine proteases of chymotrypsin family by Lesk and Fordham (Lesk and Fordham, 1996) describing in detail about basic structure, spatial relationship between the domains, mechanism of catalysis, packing of residues in individual domains, domain-domain interface, specificity pocket and similarity and divergence of proteases. Structural basis of substrate specificity (in terms of residues in S1-S4 positions and role of surface loop) and divergence has also been reviewed (Perona and Craik, 1995; 1997). There are no serious attempts to review the knowledge about the entire family afterwards best to our knowledge. Also, the data set used by Lesk and Fordham contained 13 structures only while the non-redundant data set used here is more than 3 times (43 structures).

This work has specifically focused on residues that renders specificity to the proteases or involved in structural changes that might be involved in inhibition or activation of proteases, giving a broader definition to functional residues. We will also discuss about the geometry of catalytic pocket and structural factors that renders them. It also demonstrates how biologically important information can be derived from data sets of ‘redundant’ (Please see Materials and Methods), and ‘non-redundant’ structures of eukaryotic trypsin fold serine proteases deposited in PDB (Berman *et al*., 2000), and sequences of Swissprot database (Bairoch and Apweiler, 2000) with available tools for structure and sequence analysis. Position specific properties of structures are used previously for detecting overall structure similarities and attempting fold prediction (Jones *et al*., 1992; Jones, 1999; Kelley *et al*., 2000) or homology detection (Gribskov *et al*., 1987; Karplus *et al*., 1998; Shi *et al*., 2001) or verification (Luthy R *et al*., 1992) but not function directly. This work reports position specific properties of loop regions spatially proximate to catalytic triad of serine proteases and adjoining secondary structures for predicting the functional residues. It also demonstrates increasing importance of well-curated and publicly available databases for biologists. We have also discussed some interesting evolutionary relationships with respect to SCOP classification at sub-family level.

The analysis is then applied to find out functional role of 'masquerade' and related proteins. Masquerade is a non-functional serine protease from *Drosophila melanogaster* reported to be functioning in the process of somatic muscle attachment during larval stages. Its homologues from other species like crayfish *Pacifastacus leniusculus* (Huang *et al*., 2000) and insect *Holotrichia diomphalia* larvae (Kwon *et al*., 2000) are also reported recently. Masquerade sequence is used to fish its homologues in Drosophila genome database. Chitin binding motifs has been identified over primary sequence of masquerade and other proteins. We suggest role of masquerade in both development and immune response in Drosophila. The putative function for GRAAL gene product and its mosquito homologue Sp22D (Danielli *et al*., 2000) were also examined. We have extensively worked with eukaryotic tryspsin fold serine proteases (with one exception of *Streptomyces griseus* trypsin) for reasons of ample availability of structures and sequences and also its immediate functional importance to disease biology. Hence forth, they will be referred to as 'serine proteases' only.

6.3 Materials and Methods

6.3.1 The Redundant Data Set

Using masquerade sequence (Swissprot accession number Q24019) a PSI-BLAST (Altschual *et al*., 1997) search ('NCBI gi' option ON) is made on Protein Databank (PDB) database (Berman *et al*., 2000) at RCSB web site. The resultant output essentially listed all the eukaryotic trypsin fold serine proteases in the database. All serine protease entries have been downloaded from structure explorer pages of PDB (http://www.rcsb.org/pdb/cgi/explore.cgi?pdbId=query id). Here query id is four-letter unique pdb code assigned to each unique structure. The key words- Title, Compound, Source, Primary Citation, Resolution, R-Value, Polymer chains and HET Groups were extracted from the pages. For all the entries keywords Compound and HET groups were analysed to seek for 'redundant structures' having different protein inhibitors and/or chemical or artificial inhibitors bound to them as described using Compound or HET Groups key words. The structures might also contain important mutations and can be from different species. The keywords Resolution and R-Value were used to choose high-resolution structure and also to prevent cases of 'tie'. Only high-resolution structures are used for the analysis. The resultant data set contained 23 structures of trypsin, 13 structures of thrombin, 7 structures of plasminogen activators and 6 structures of elastase and coagulation factors each. These families are selected also since there is plentiful information about structure-function relationships of these sub-families available in literature. The PDB files of resultant data set of structures were processed to remove all other protein chains, HET Group and water entries.

6.3.2 The Non-redundant Data Set

The PSI-BLAST (Altschul *et al*., 1997) output was also used for generating a non-redundant data set of structures with no two members sharing more than 95% identity as follows. The sequences reported in PSI-BLAST were extracted using the 'gi' identifiers from ENTREZ protein database using batch download option. An in-house program is written that discards one of the sequence (structure) of the pair with low quality (using keywords Resolution and R-Value) iteratively, using the pairwise identity from CLUSTALX output file (version 1.8; Jeanmougin *et al*., 1998). The resultant dataset has 43 structures. The data set includes the *Streptomyces griseus* trypsin as a protein under investigation, which is a prokaryotic serine protease. We will discuss the reasons for it in the discussion part.

6.3.3 Visulization and Analysis

The serine protease structures were viewed using RASMOL (Bernstein, 2000; Sayle and Milner-White, 1995). The loop regions spatially proximate to catalytic triad (please see introduction) were chosen and marked (Figure1). Loop nomenclature is adopted as described by Peisach and co-workers (Peisach *et al*., 1999). Loops are named from L1 to L15 (Figure 2) and loops L3, L5, L7, L9, L11, L12 and L14 with adjoining secondary structures are chosen for further examination.

6.3.3.1 Analysis Using Redundant Data Set

As mentioned before this set of proteins contains the proteins/chemical inhibitors or HET groups bound to proteases. A program is written to find out the interacting residues of proteins with its inhibitors using distance criterion of 4Å. The results were tabulated and plotted as shown in Figure 2.

6.3.3.2 Analysis Using Non-Redundant Data Set

The program H-BOND is used to calculate hydrogen bonding (Overington J., unpublished). H-BOND calculates hydrogen bonding of main chain to main chain (i.e. those responsible for secondary structure), side chain to main chain carbonyl, side chain to main chain amide and side chain to side chain (hetero-atoms). The unprocessed PDB files were used while running H-BOND. For solvent accessibility calculations the PDB files were processed and all non-protease chains and HETATM entries were removed. The calculation was done using PSA that implements algorithm of Lee and Richards (Lee and Richards, 1971). The 7% relative cut-off is applied (Hubbard and Blundell, 1987). Secondary structure and main chain conformation were calculated using SSTRUC (Smith D, Unpublished) that uses Kasbach and Sander definitions; Kasbach and Sander, 1983) It is also a part of the PROCHECK (Laskowski *et al*., 1993) suit of programs. The programs HBOND, PSA and SSTRUC are part of sequence-structure representation and analysis program JOY (Overington *et al*., 1990; Mizuguchi *et al*., 1998). The structure-based alignment of all 43 structures as shown in Figure 6.3 is prepared using STAMP (Russell and Barton, 1992) and manually edited to remove local misalignments. The alignment is annotated using JOY (Overington *et al*., 1990; Mizuguchi *et al*., 1998) to compare the structural properties. Perl scripts are written to process the JOY ‘tem’ output file to extract information about accessibility, hydrogen bonding, secondary structure, Ooi number for the segments under examination (loop region under investigation and adjoining secondary structures for each protein). The results for each property are plotted for loop regions under examination. The solvent accessibility results are the most interesting ones and reported here in Figure 6.4. The alignment was used to generate evolutionary tree using Neighbour-Joining method (Saitou and Nei, 1987; Figure 6.5) provided by CLUSTALX8.1 (Jeanmougin *et al*., 1998) and also using PHYLIP3.5 package (Felsenstein, 1985) that uses KITSCH algorithm. The protease domain of sequences of all five sub families were extracted from Swissprot database (Bairoch and Apweiler, 2000) using their function as keyword. Structure based alignment is used to guide the profile alignment of sequences using CLUSTALX8.1 profile alignment mode where the sequences are added to structural alignments. The resultant alignment was manually edited (Figure 6.6) and examined for conservation of functionally important residues among sub-families. Sequences of masquerade like proteins from different species were obtained from NCBI web site using Entrez search and retrieval system (http://www.ncbi.nlm.nih.gov:80/Entrez/) and aligned as mentioned above (with other proteins involved in patterning).

A BLASTP (Altschul *et al*., 1997) search is made using masquerade as query sequence on Drosophila genome and hits were examined using disulfide bridge conservation as a criterion (see results) to identify its homologues.

6.4 Results

6.4.1 Finding the Functional Residues

The interacting residues of structures in redundant data set is listed and their occurrence is converted in to frequencies by dividing their occurrence to number of the structures used for the analysis. They are then plotted according to their positions as bars as shown in Figure 6.2. Surprisingly all resides that are reported as utilised in binding inhibitors or HET groups falls on loop regions and a very few on neighbouring secondary structures.

Highest number of structures analysed (23 structures) are of trypsins, but the variation reported is larger in case of thrombin (13 structures). The reported residues on a linear sequence starts from residue number 34 (reported as interacting residue once and thrice respectively; chymotrypsin numberings used throughout the 'Results' and 'Discussion' sections), but for trypsin last residue reported in binding is 228 and for thrombin it is 245 (reported once each). For elastase, co-agulation factors and plasminogen activators the first and last residues reported interacting are 35, 41, 36 and 224, 228, 217 respectively (each reported once). While deciding for functional residues, it is important to put a cut-off value on data interpreted from frequency of binding. After studying the behaviour of all plots a cut off value of 0.4 was used to discriminate between commonly used residues and very specific residues for each protease or a false positive (see discussion). As it is evident from graphs corresponding to loop 9 where no peaks cross the cut off of 0.4. All the reported residues where checked for published reports of functionally mutations for all 5 types of proteases under study. A frequency value of 0.5 is decided as 'interacting' value by studying behaviour exhibited by His 57 residue of thrombin. Cases where the cut off values fall between 0.35 and 0.5 (as exhibited by behaviour of loop 14 for thrombin) were solved by studying the behaviour of analogues segments from other proteases under study and literature. This problem may arise because of the segment movements in reported structures.

The residues that come as interacting residue from loop3 are residues 40, 41 and 42 of trypsin and residue 41 of elastase. However, residues 38 and 40 of thrombin fall under investigating range. Catalytic His 57 (Loop 5) comes out as predominant residue for all of the proteases except co-agulation factors. Residues 60A-60F of thrombin are insertion relative to chymotrypsin (and other proteases). These residues are not reported as interacting residues in proteases under investigation and may be a unique insertion for thrombin. Residues 97 (investigating range), 99 of trypsin, 97A, 98, 99 of thrombin, 97, 98, 99 of co-agulation factors and 99 of elastase and plasminogen activators comes as interacting residues. The residues in loop 9 are well below cut off value and they seem to be examples of false positives. Loop11 contributes residue 174 in thrombin, co-agulation factors and plasminogen activators but in case of trypsin residue 175 is reported. Loop12 lines the S_i-S_i' residues and it is the most important loop for catalysis. It harbours Ser195. The residues 189 to 195 are reported as binding residues (except for residue 189 for elastase) in all sub families. The most important were results for loop 14. The residues 213-217, 219, 220 and 226 are reported as interacting residues in trypsin and plasminogen activators. Thrombin structures lack report of residue 216, coagulation factors lack report of reside 214 and elastase structures lack report of residues 213 and 214 but they have addition of residue 218 reported as interacting residue.

The alignment was analysed for conservation of interacting residues among sub-families. Interacting residues were mapped on to loop residues as shown in Figure 6.6 for trypsin and thrombin. The variation is mapped for regions in loop3, loop5, loop7, loop 12 and loop 14 on the alignment containing structures and sequences of all five sub-families as shown in the figure 6.6. Residues 40-42, 57, 60, 60A-F (for thrombin), 99, 174, 189-195 and 213-218 and 226 are identified as residues rendering functional specificity to respective proteases. Hence, the residues needed for specificity resides in loop regions and barrel structures are simply providing the scaffold needed. The experimental evidences supporting the above mentioned residues are discussed later. It should be noted that the evidences are not available for all residues for all five sub-types under examination.

6.4.2 Solvent Accessibilities and Hydrogen Bonding of Functional Regions

Mean side chain accessibilities of 43 proteases at each position of functional loops are plotted. The error bars are the variation observed from the mean value. The gap positions of each protease were assigned an arbitrary value of 100. Therefore positions showing accessibility values near 100 with very low error bars are actually gap regions. The side chains of residues in beta stands β2, β4, β7, β10, and β11 are buried in side the barrel structures. Catalytic His57 is adjacent to β4, catalytic Asp102 is adjacent to β7 and catalytic Ser195 is harboured by loop region of β10 and β11. The hydrogen bonding properties (main chain to main chain, side chain to main chain carbonyl and side chain to main chain amide H-bonding) of each loop region under examination were also plotted for each positions for each of 43 proteases. The H-bonding is found to be conserved (data not shown) as evident from the JOY alignment of structures shown in Figure 6.3. These results suggest the catalytic triad residues to be very rigidly held by the adjacent secondary structures to maintain the geometry of catalytic triad. It is also evident from Figures 6.3 and 6.4 that side chains of catalytic His57 and Ser195 are only moderately solvent accessible (20-40%), but these residues are coming out as interacting residues in analysis with redundant data set. These results also support the fact that the catalytic triad is rigidly held in all proteases. The side chain of Asp102 is reported not at all solvent accessible. This fact is well known from previous studies and is also supported by the fact that Asp102 is not reported as interacting residues in analysis with redundant data set. The region of residues 215 to 226 shows highest variation in solvent accessibilities at each position with the fact that we are investigating family properties. This region is highly conserved among all known serine proteases. This points to the fact that it is the most highly mobile region among the structures under investigation and can be of functional importance.

6.4.3 Phylogeny and SCOP

We have used here NJ method (Saitou and Nei, 1987; Figure 6.5) provided by CLUSTALX8.1 package (Jeanmougin *et al.*, 1998) to generate and visualise (NJview) the evolutionary tree of 43 structures of non-redundant database. As expected proteases with similar function are clustered together in the evolutionary tree. Here we discuss those examples where the proteases come together forming single node in a cluster but are assigned to different sub-family by SCOP (Murzin *et al.*, 1995). The occurrence of the pair of proteins at a single node is also confirmed using PHYLIP3.5 (Felsenstein, 1985) that uses KITSCH algorithm. Trypsin from *Fusarium oxyparium* (1try-; the last character of PDB ID in each case shows the chain identifier) and *Streptomyce*s *griseus* (prokaryotic protease; 1sgt-) forms a pair different from rest of the trypsins. Human leukocyte elastase (1ppfe) gets clustered different from other elastases (1qnja, 1elt, and 1brup) but it pairs up with myeloblastin (1fuja). Tissue type plasminogen activators 1a5ha and 1a5ia groups together with other plasminogen activators (1ejna, 1ddja) with different functional specificity but they are also assigned to different sub-families by SCOP (Murzin *et al.*, 1995). Chymotrypsinogen C (1pytd) forms a pair with porcine pancreatic elastase (1brup) and cluster together with other elastase structures showing that it is related to elastase sub-family but it is grouped with chymotrypsin(ogen) by SCOP (Murzin *et al.*, 1995).

1try- is grouped with other trypsin(ogen) in SCOP database (Murzin *et al.*, 1995) and 1sgt- is grouped with prokaryotic proteases. These two proteins are similar in many respects. 1try- shares respectively 42% and 38% identity with 1sgt- and 5ptp- (bovine trypsin) but 5ptp- shares only 30% similarity with 1sgt-. However, R.M.S deviations (a measure of structural similarity) between 1try- and 1sgt-, 1try- and 5ptp-, and 1sgt- and 5ptp- are 1.16Ǻ, 1.21Ǻ and 1.21Ǻ respectively. This shows that the backbones of all structures are similar. As shown in the alignment of Figure6.6 1sgt- and 1try- (with trypsin of lower organisms) also lacks disulfide bridges by C22-C157 and C128-C232 reported for trypsins from higher organisms. However, 1sgt- shares very less similarity with other prokaryotic serine proteases (for example ~20% with 1hpga, a glutamic-acid specific protease).

1ppfe is grouped with 1qnja, 1elt- and 1brup under elastase sub-family by SCOP (Murzin *et al.*, 1995). While, 1fuja is defined as a lone member of myeloblastin sub-family. 1ppfe and 1fuja shares 55% sequence identity and are equally similar to other members of elastase sub-family (~36%). The R.M.S. deviation of the pair of 1ppfe and 1fuja is 0.68 Ǻ, a virtually identical backbone. While that of 1ppfe and 1fuja with other members of elastase sub-family is ~1.19 Ǻ and ~1.26 Ǻ. These numbers are suggestive of the fact that they are sub-family members (for example, see the R.M.S. deviation reported for 1try- and 5ptp-). The alignment in Figure6.6 shows that the functional residues are very similar for 1fuja and 1ppfe but different than other elastases. This protein is also known to degrade elastin, fibronectin, laminin, vitronectin, and collagen types 1, 3, and 4 (Swissprot entry P24158 at www.expasy.ch). It is also known that genes for human neutrophile elastase (1ppfe), myeloblastin (1fuja) or proteinase3 (PR3) and azurocidin are organized as single genetic locus and are homologues (Zimmer *et al.*, 1992).

Chymotrypsinogen C (1pytd) which also known as caldecrin or elastase4 shares 63% identity with porcine pancreatic elastase (1brup) but only 30% with leukocyte elastase (1ppfe). It also shares 41% identity with α-chymotrypsinogen. The R.M.S deviations of 1pytd-1brup, 1pytd-1ppfe and 1pytd-4cha are 1.02 Ǻ, 1.37 Ǻ and 1.29Ǻ respectively. This suggests that 1pytd is equally related to both the subfamilies. Indeed 1pytd is reported to have characteristics of both elastase and chymotrypsin sub-families. As described it is sequentially more related to elastase sub-family but it shares disulfide bridge pattern and catalytic specificity of chmotrypsins (Gomis-Ruth *et al.*, 1995). Caldecrin are known to be expressed in pancreas but TPCK doesn't inhibit them, suggesting that it is different than chymotrypsins (Yoshino-Yasuda *et al.*, 1998). The enzyme comission numbers for elastase, caldecrin and chymotrypsin sub-families are 3.4.21.36, 3.4.21.2 and 3.4.21.1 respectively suggesting difference in the sub-families. Thus 1pytd should be assigned a different sub-family than both elastases and chymotrypsins.

Tissue type plasminogen activators 1a5ha and 1a5ia share 78% identity and the R.M.S. deviation for this pair is 0.78 Ǻ. They are assigned the same Enzyme Commission number (EC 3.4.21.68) suggesting that the catalytic specificity and other features are same. There has been studies in past reporting that above 70% sequence similarities the function can be reliably transferred for sequences under consideration (Devos and Valencia, 2000; Wilson *et al.*, 2000). Hence, this pair should also be considered as part of same sub-family.

6.4.4 Finding Function for Masquerade and others

Masquerade is a secreted molecule encoded by Drosophila *masquerade* (*mas*) gene. It is 1047 amino acid long and reported to contain an N-terminal domain containing five disulfide knotted motifs and C-terminal serine protease like domain (Murugasu-Oei *et al.*, 1995).

BLASTP search (Altschul *et al*., 1997) of PDB database (Berman *et al*., 2000) and NR database at NCBI (www.ncbi.nlm.nih.gov) reports trypsins (~ 35% identity with bovine trypsin 2tld-) and thrombins (~ 32% identity with thrombin 1ucyk) as the closest homologues of masquerade. It is well known that masquerade is a non-functional serine protease due to mutation of catalytic serine residue to glycine at position 195 (Murugasu-Oei *et al*., 1995). The other proteins containing five repeats of disulfide motifs in arthropod serine proteases includes Sb-Sbd (Appel *et al*., 1992), snake (Delotto and Spierer, 1986) easter (Chasan and Anderson, 1989) and Limulus (Muta *et al*., 1990). Drosophila *GRAAL* or *Tequila* gene product and its *Anopheles gambiae* homologue Sp22D, which is associated with hemocytes and hemolymph, are also reported to have cysteine rich N-terminal domains (Danielli *et al*., 2000) and a functional C-terminal serine protease domain. In addition, both *GRAAL* gene product and Sp22D is shown to have poly-threonine stretches also common to masquerade. Sp22D is shown to be a chitin binding (adhesive) protein and it is suggested to serve as sentinel to detect exposed chitin, and then trigger appropriate physiological, developmental or immune response as chitin is also found on the surface of invading agents (Danielli *et al*., 2000). In drosophila the pathways involving toll ligand is implicated both in patterning drosophila embryo and generating early immune response and it involve serine proteases like Nudel, Gastrulation defective (GD), Snake and Easter (Lemosy *et al*., 2001; Levashina *et al*., 1999). Furthermore, the defensive prophenoloxidase cascade in *Manuuca sexta* is reported to be initiated by proteolytic processing (Jiang *et al*., 1998). The prophenoloxidase activating factor-1 (PPF1) in hemolymph of *Holotrichia dimphalia* larve (Lee *et al*., 1998) and *Anopheles gambiae* (Paskewitz *et al*., 1999) is shown to be a homologue of drosophila easter protease. Recently drosophila masquerade like proteins were reported from coleopteran *Holotrichia dimphalia* and (gi 10697070) *Tenebrio molitor* larvae (gi 10697178) and shown to be necessary for prophenoloxidase activity (Kwon *et al*., 2000). Analysis of sequences shows that they contain only one N-terminal disulfide knot and serine protease domain catalytic serine mutated to glycine. A cell adhesion protein containing multiple disulfide-knotted motif (total 7) and serine protease domain catalytic serine mutated to glycine is reported in crayfish *Pacifastacus lenisculus*.

Serine proteases like domain of masquerade sequences from drosophila, crayfish and coleopteran larves were aligned with the 43 proteins of non-redundant database using profile alignment mode of CLUSTALX8.1 (Jeanmougin *et al*., 1998). The resultant evolutionary tree has shown the sequences to be equally related to trypsins and thrombins (not shown). Noticeably the proteins contain disulfide bridge C1-C122 like thrombins which is absent in trypsins and C136-C201 like trypsins which is absent in thrombins. A BLASTP search (Altschul *et al*., 1997) is made on drosophila genome database at NCBI web site (ncbi.nlm.nih.gov) using serine protease like domain of masquerade as a query sequence (total 226 reported hits). The hits were analysed and grouped as masquerade homologue using disulfide bridge as criterion till the hit annotated having different function was found (first 34 hits). The gene product CG4998 is found to be among top hits with catalytic serine mutated to glycine. We hypothesise a masquerade like function for this gene product. The other hits with serine as catalytic residues are gene products CG5390, CG8586, CG8738, Tequila (GRAAL), CG13318, CG6639, CG2105, CG3117, CG14990, Nudel and CG18557. The GRAAL gene product, its mosquito homologue (Sp22D) and Nudel gene product were taken for further examination.

An alignment of Nudel, Snake, Easter (drosophila proteases involved in patterning), 5ptp-, 1c1uh, *GRAAL*, Sp22D and masquerade like proteins from drosophila, crayfish and coleopteran larvae serine protease like domain is shown in the Figure 6.7. The functional residues identified before are marked in the alignment. It is clear from the figure that the functional residues are mutated to random in all four masquerades like proteins with catalytic serine mutated to glycine. Hence, catalytic serine is not the only mutated functional residue. Thus suggesting that they can not function as a competitive antagonist of serine proteases.

Five cysteine rich repeats are identified by careful analysis of masquerade N-terminal domain. Similar repeats were also found in other masquerade like proteins (not shown). An alignment of the repeats with ‘chitin binding motif’ reported by Suetake *et al* (2000) shows each repeat contained one chitin binding motif as shown in Figure 6.8. Chitin binding motifs of GRAAL and Sp22D as identified are also shown in the figure 6.8. Thus showing masquerade having a chitin binding or very similar activity.

6.5 Discussion

We have identified positions 40, 41 and 42 (loop3), 57, 60, and 60A-F (loop5), 97 and 99 (loop7), 174 (loop11), 189-195 (loop12) and 213-219 and 226 (loop14) as residues rendering specificity and catalysis to serine proteases. Catalytic Asp102 was not reported as it is completely buried and doesn't interact directly with substrate or inhibitors. The barrel structures are shown only as supporting structures for loops to carry out a particular function. The allosteric sites were not identified by this study since all the HETATM (mostly ions) entries other than specific inhibitors were removed from the PDB files. We than searched for literature reports of structure function relationships for the proteases.

Indeed, there has been a wealth of information of mutagenesis studies in chosen sub-families of serine proteases, while others are uncharacterised in terms of function (Maxwell *et al*., 1999). New members of the family are constantly being added due to genome sequencing efforts (for example, 29 out of 34 top hits (from Drosophila genome database) examined in this work are not annotated). Hence, searching for "consensus" positions providing functional specificity is important for fast characterisation of known sequences. For reports correspond to catalytic triad and description of catalytic mechanism we recommend analysis by Lesk and Fordham, a review by Perona and Craik and literature cited within (Lesk and Fordham, 1996; Perona and Craik, 1995).

The diversity of substrate specificity among the chymotrypsin like proteases rests upon small differences in structure of the substrate-binding cleft composed of two juxtaposed β-barrel domains, with catalytic residues bridging the barrels (Figure1; Kraut, 1977; Steitz and Shulman, 1982; Bazan and fletterick, 1990). Position 189, located at the base of S1 pocket, is highly conserved as an Asp in enzymes with trypsin like specificity. It is found as Ser or other small amino acid in chymotrypsin and elastase like enzymes.

Position 190 extends into the base of the pocket as well as plays an additional role to modulate specificity profile. Amino acids at positions 216 and 226 are usually Gly in both trypsin and chymotrypsin-like enzymes; larger amino acid side chains at positions partially or fully block access of larger substrate side chains to the base of the pocket (Craik *et al*., 1985; Wilke *et al*., 1991). Accordingly elastases possess larger, usually non-polar residues at this positions, providing a platform for interaction with small hydrophobic substrates (Perona and Craik, 1995). Consequently, C-terminal sequence is postulated to encode function for serine proteases (Stroud, 1974; Maxwell *et al*., 1999). A view supported by the fact in protease structures as the C-terminal end is approached, the surface area containing the substrate increases sharply. The residue 192 is shown to be important for blood coagulation and fibrinolytic systems but not tissue type plasminogen activators and have different roles in other sub-families (Zhang *et al*., 1999). The residues 189-220 in C-terminal sequences were found to account for >95% of the area around the specificity pocket S1 and catalytic His57 and >70% of the area of around specificity sites S2 and S3 (Maxwell *et al*., 1999). Role of residue 172 for trypsin substrate specificity is also known (Hedstrom *et al*., 1994).

But this view is not entirely correct as N-terminal residues are identified as the functional residues. Role of residues 60B-F (loop5), Trp96 and effect of charge reversal of Arg93, 97 and 99 (loop7) for thrombin has been reported (DiBella and Scharaga, 1998; He *et al*., 1997). Glu39 of thrombin is reported to play part in P3 specificity (Le Bonniec *et al*., 1991). Role of residue 99 of factor Xa, activated protein C and thrombin is shown to exhibit P2 specificity (Rezaie, 1997). Role of loop5 (loop60) is reported to be instrumental in S1′ specificity for trypsin (Kurth *et al*., 1997). But these reports has been scattered, information is derived from largely biochemical analysis and discussed about role of particular residue for only particular sub-family. We have identified these residues on the basis of structural and sequence analysis and hypothesise the role of the consensus residues in rendering specificity in all the characterised and yet to be characterised sub-families.

Loop12 and Loop14 are longest loops in the serine proteases that take part in the catalysis and substrate specificity. Loop 12 and loop14 are also highly conserved among all the proteases. Interestingly the movement of the loop12 is shown to be restricted by the secondary structures from both the sides. Here it should be noted that the loop is glycine rich and is reported in limited amount of conformational change forming the oxyanion hole for catalysis by mediating changing in hydrogen bonding state of residue 194 (Lesk and Fordham, 1996; Peisch *et al*., 1999). Loop14 however on the contrast shows the highest variability in the solvent accessibility analysis. This simply suggests that this loop is highly mobile in all 43 structures used for the analysis. According to the expectations, loop14 of proenzyme domain of plasminogen is reported to have an entirely different loop conformation than active conformation of trypsin and chymotrypsingoen showing W214 side chain blocking the S1 specificity pocket in a foot in mouth mechanism of inactivation (Peisach *et al*., 1999). Residues 214-220 of chymotrypsinogen that makes up the opposite of S1 subsite is reported to narrow down active-site pocket. Such change is not reported for trypsinogens and plasmins (Parry *et al*., 1998). Thus analysis solely based on structural properties is capable of identifying the functional sites in proteins. Though it requires large amount of structural information which we hope to be common in future with high number of redundant structures deposition and current rate of growth of PDB database (http://www.rcsb.org/pdb/holdings.html).

Homology of *Streptomyces griseus* trypsin (1sgt-) with other eukaryotic proteases has been reported and attributed to gene transfer from eukaryote to the bacterium as early as in 1970 (Hartely, 1970). It is grouped differently as the proteases are grouped on the basis of the source unlike other proteins in SCOP database (Murzin *et al*., 1995). We propose that 1sgt- should be grouped with other eukaryotic trypsins. The tissue type plasminogen activators 1a5ha and 1a5ia shows identical functional residues identified by our analysis and other similarities described above and hence should be grouped under single sub-family. The neutrophile elastase (1ppfe) and myeloblastin (1fuja) shows conserved functional residues among the pair but different than other elastases at positions 41, 60, 189, 190, 191 and 220 leading to the suggestions that these pair should be classified as differently than other elastases. We also propose that chymotrypsinogen C to be considered in a different sub-family than both elastase and chymotrypsin sub-families. These reported differences with the SCOP classification however points to an interesting question of definition of function. There is no clear measure for functional similarity. The definition of function itself is often vague. For example, all the proteins under consideration in this paper serve function of a protease and cleave the scissile bond. But they have been divided or grouped as ‘different’ on the basis of the substrate specificity or catalytic mechanism. It is also evident form the case of chymotrypsinogen C that functional similarity can not be inferred always with confidence on the basis of sequence similarities. Hence, such differences should be considered subjective and waiting for clearer structure-function relationships.

At last we apply our analysis of serine protease domain and hits identified from the drosophila genome to a cell adhesion protein masquerade and its homologues reported functioning in adhesion as well as immune responses. The closest functional serine protease domain from Drosophila genome shares 34% identity with that of masquerade. As shown previously all the functional residues of masquerade and its homologues have been mutated randomly. Thus, suggested role of masquerade acting, as antagonist of a seine protease domain is very unlikely. Instead we have identified 5 chitin binding (adhesive) motifs in masquerade N-terminal cysteine-knotted repeats. Such motifs are also identified for GRAAL gene product and Sp22D protein of Anopheles in this study. As mentioned before many proteases simultaneously involved in patterning and immune response possess such motifs. We propose that these proteins carry out their both the roles with common mechanism of adhesion (to chitin of invading pathogen or a tissue for degradation) and subsequent protease activity. We strongly attribute this role to GRAAL gene product and Sp22D but also in general all the proteins involved in patterning and immune response (like snake, easter etc). The role of Sp22D during early immune response is demonstrated (Danielli *et al*., 2000). In addition to this we propose the involvement of GRALL and Sp22D in development. We suggest that the C-terminal domain of masquerade like proteins are acting as prophenoloxidase response factor or serving a signalling role during the early development reported for HGF/SF like proteins with non-functional serine protease domain (Thery *et al*., 1995). This suggestion is strengthen by the fact that total loss of function of *mas* gene is embryonic lethal (Murugasu-Oei *et al*., 1995) and masquerade for its homologues with prophenyloxidase activity contain only one improper cysteine-knot motif (unpublished results).

6.6 Conclusions

In this study the consensus residues involved in rendering in substrate specificity in eukaryotic serine proteases has been identified by analysis of redundant and non-redundant data set of structures and sequence information. We have predicted functional sites on the basis of structural properties like solvent accessibility and hydrogen bonding (not shown) analysis of non-redundant data set and showed that the secondary structures adjacent to catalytic triad residues are immobile and maintain rigid geometry required for efficient catalysis. The results were applied to proteins 'masquerading' its real function. Chitin binding motifs have been identified in masquerade, GRAAL and Sp22D and multiple roles of during adhesion process, immune response and development has been suggested for the proteins.

6.7 References

Andersen, N. H., Cao, B., Rodriguez-Romero, A., and Arreguin, B. (1993). Hevein: NMR assignment and assessment of solution-state folding for the agglutinin-toxin motif. *Biochemistry* **32**, 1407-22.

Appel, L. F., Prout, M., Abu-Shumays, R., Hammonds, A., Garbe, J. C., Fristrom, D., and Fristrom, J. (1993). The Drosophila Stubble-stubbloid gene encodes an apparent transmembrane serine protease required for epithelial morphogenesis. *Proc Natl Acad Sci U S A*, **90**, 4937-41

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-402.

Baba, T., Watanabe, K., Kashiwabara, S., Arai, Y. (1989). Primary structure of human proacrosin deduced from its cDNA sequence. *FEBS Lett* **44**, 296-300.

Baker, B. M., and Murphy, K. P. (1997). Dissecting the energetics of a protein-protein interaction: the binding of ovomucoid third domain to elastase. *J Mol Biol*, **268**, 557-69.

Barrett, A. J. (1977). Editor of Proteinases in Mammalian Cells and Tissues. North-holland, amesterdam.

Barrett, A. J. (1994). Editor of Proteolytic Enzymes: Serine and Cysteine Peptidases. Methods Enzymol, 244, Academic Press, New York.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., and Sonnhammer, E. L. (2000). The Pfam protein families database. *Nucleic Acids Res*. **28**, 263-6.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*, **28**, 45-8.

Bazan, J. F., and Fletterick, R. J. (1990). Structural and catalytic models of trypsin like viral proteases. *Semin Virol*. **1**, 311-22.

Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., and Westbrook, J. (2000). The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol*,**7**, 957-9.

Bernstein, H. J. (2000). Recent changes to RasMol, recombining the variants. *Trends Biochem Sci*, **25**, 453-5.

Birktoft, J. J., and Blow, D. M. (1972). Structure of crystalline -chymotrypsin. V. The atomic structure of tosyl--chymotrypsin at 2 A resolution. *J Mol Biol.* **68**, 187-240.

Bond, J. S. (1991). Plasma-membrane proteases-introductory remarks. *Biomed. Biochem. Acta*, 50, 775-780.

Boggon, T. J., Shan, W. S., Santagata, S., Myers, S. C., and Shapiro, L. (1999). Implication of tubby proteins as transcription factors by structure-based functional analysis.

Chasan, R, and Anderson, K. V. (1989). The role of easter, an apparent serine protease, in organizing the dorsal-ventral pattern of the Drosophila embryo. *Cell*, **56**, 391-400.

Chothia, C., and Janin, J. (1975). Principles of protein-protein recognition. *Nature*, **256**, 705-8.

Craik, C. S., Largman, C., Fletcher, T., Roczniak, S., Barr, P. J., Fletterick, R., and Rutter, W. J. (1985). Redesigning trypsin: alteration of substrate specificity. *Science*, **228**, 291-7.

Danielli, A., Loukeris, T. G., Lagueux, M., Muller, H. M., Richman, A., and Kafatos, F. C. (2000). A modular chitin-binding protease associated with hemocytes and hemolymph in the mosquito Anopheles gambiae. *Proc Natl Acad Sci U S A*, **97**, 7136-41.

DeLotto, R., and Spierer, P. (1986). A gene required for the specification of dorsal-ventral pattern in Drosophila appears to encode a serine protease. Nature, 323, 688-92.

Devos, D., and Valencia, A. (2000). Practical limits of function prediction. *Proteins*, **41**, 98-107.

DiBella, E. E., Scheraga, H. A. (1998). Thrombin specificity: further evidence for the importance of the beta-insertion loop and Trp96. Implications of the hydrophobic interaction between Trp96 and Pro60B Pro60C for the activity of thrombin. *J Protein Chem*, **17**, 197-208.

Dodson , G. G., Lawson, D. M., and Winkler, F. K. (1992). Structure and evolutionary relationships in the lipase mechanism and activation. *Faraday Discuss*. **93**, 95-105.

Donate, L. E., Gherardi, E., Srinivasan, N., Sowdhamini, R., Aparicio, S., and Blundell, T. L. (1994). Molecular evolution and domain structure of plasminogen-related growth factors (HGF/SF and HGF1/MSP). *Protein Sci.* **3,** 2378-94.

Felsenstein, J. (1985). *Evolution*, **39**, 583.

Fersht, A. (1984). *Enzyme Structure and Mechanism*, 2nd edit., W.H. Freeman, San Fransisco, CA.

Froelich, C. J., Zhang, X., Turbov, J., Hudig, D., Winkler, U., Hanna, W. L. (1993). Human granzyme B degrades aggrecan proteoglycan in matrix synthesized by chondrocytes. *J Immunol*, **151**,7161-71.

Goldberger, G., Bruns, G. A., Rits, M., Edge, M. D., and Kwiatkowski, D. J. (1987). Human complement factor I: analysis of cDNA-derived primary structure and assignment of its gene to chromosome 4. *J Biol Chem*, **262**, 10065-71.

Gomis-Ruth, F. X., Gomez, M., Bode, W., Huber, R., and Aviles, F. X. (1995). The three-dimensional structure of the native ternary complex of bovine pancreatic procarboxypeptidase A with proproteinase E and chymotrypsinogen C. *EMBO J*, **14**, 4387-94.

Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*, **84**, 4355-8.

Gurwitz, D., Cunningham, D.D. (1988). Thrombin modulates and reverses neuroblastoma neurite outgrowth. *Proc Natl Acad Sci U S A*, **85**, 3440-4.

Hartley, B. S. (1970). Homologies in serine proteinases. *Philos Trans R Soc Lond B Biol Sci*, **257**, 77-87.

He, X., Ye, J., Esmon, C. T., Rezaie, A. R. (1997). Influence of Arginines 93, 97, and 101 of thrombin to its functional specificity. Biochemistry 1997, 36, 8969-76.

Hedstrom, L., Szilagyi, L., and Rutter, W. J. (1992) Converting trypsin to chymotrypsin: the role of surface loops. *Science*, **255**, 1249-53.

Hedstrom, L., Perona, J. J., and Rutter, W. J. (1994). Converting trypsin to chymotrypsin: residue 172 is a substrate specificity determinant. *Biochemistry*, **33**, 8757-63.

Henderson, B. R., Tansey, W. P., Phillips, S. M., Ramshaw, I. A., Kefford, R.F. (1992). Transcriptional and posttranscriptional activation of urokinase plasminogen activator gene expression in metastatic tumor cells. *Cancer Res*, **52**, 2489-96.

Horl, W. H. (1989). Proteinases: potential role in health and disease. In *Design of Enzyme Inhibitors as Drugs* (Sandler, M & Smith, H. J., eds), pp.573-581, Oxford University Press, Oxford.

Huang, T. S., Wang, H., Lee, S.Y., Johansson, M. W., Soderhall, K., and Cerenius, L. (2000). A cell adhesion protein from the crayfish Pacifastacus leniusculus, a serine proteinase homologue similar to Drosophila masquerade. *J Biol Chem*, **275**, 9996-10001.

Hung, S. H., and Hedstrom, L. (1998). Converting trypsin to elastase: substitution of the S1 site and adjacent loops reconstitutes esterase specificity but not amidase activity. *Protein Eng*, **11**, 669-73.

Hubbard, T. J., and Blundell, T. L. (1987). Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling.

Hubbard, S. J., Eisenmenger, F., and Thornton, J. M. (1994). Modeling studies of the change in conformation required for cleavage of limited proteolytic sites. *Protein Sci*, **3**, 757-68.

Iengar, P., and Ramakrishnan, C. (1999). Knowledge-based modeling of the serine protease triad into non-proteases. *Protein Eng*, **12**, 649-56.

Janin, J., and Chothia, C. (1976). Stability and specificity of protein-protein interactions: the case of the trypsin-trypsin inhibitor complexes. *J Mol Biol*, **100**, 197-211.

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G., and Gibson, T. J. (1998). Multiple sequence alignment with Clustal X. *Trends Biochem Sci*, **23**, 403-5.

Jiang, H., Wang, Y., and Kanost, M. R. (1998). Pro-phenol oxidase activating proteinase from an insect, Manduca sexta: a bacteria-inducible protein similar to Drosophila easter. *Proc Natl Acad Sci U S A*, **95**, 12220-5.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86-9.

Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol*, **287**, 797-815.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-637.

Karplus, K, Sjolander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L., and Sander, C. (1997). Predicting protein structure using hidden Markov models. *Proteins*, Suppl **1**, 134-9.

Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol*, **299**, 499-520.

Kraut, J. (1971). Subtilisin: X-ray structure. In *The Enzymes* (Boyer, P. D., ed.), vol. 3, pp. 547-560, Academic Press, New York and London.

Kraut, J. (1977). Serine proteases: structure and mechanism of catalysis. *Annu Rev Biochem*, **46**, 331-58.

Krem, M. M., Rose, T., Di Cera, E. (1999). The C-terminal sequence encodes function in serine proteases. *J Biol Chem*, 274, 28063-6.

Kurth, T., Ullmann, D., Jakubke, H. D., and Hedstrom, L. (1997). Converting trypsin to chymotrypsin: structural determinants of S1' specificity. *Biochemistry*, **36**, 10098-104.

Kwon, T. H., Kim, M. S., Choi, H. W., Joo, C. H., Cho, M. Y., Lee, B. L. (2000). A masquerade-like serine proteinase homologue is necessary for phenoloxidase activity in the coleopteran insect, Holotrichia diomphalia larvae. *Eur J Biochem*, **267**,6188-96.

Laskowski, R. A., Moss, D. S., and Thornton, J. M. (1993). Main-chain bond lengths and bond angles in protein structures. *J Mol Biol*, **231**, 1049-67.

Levashina, E. A., Langley, E., Green, C., Gubb, D., Ashburner, M., Hoffmann, J. A., and Reichhart, J. M. (1999). Constitutive activation of toll-mediated antifungal defense in serpin-deficient Drosophila. *Science*, **285**, 1917-9.

Lee, B., and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*, **55**, 379-400.

LeMosy, E. K., Tan, Y. Q., and Hashimoto, C. (2001). Activation of a protease cascade involved in patterning the Drosophila embryo. *Proc Natl Acad Sci U S A*, 98, 5055-60.

Lesk, A.M. (1981). *Introduction to Physical Chemistrey*, sections 18-10, Prentice-Hall, Inc., Englewood Cliffs, NJ.

Lesk, A. M., and Fordham, W. D. (1996). Conservation and variability in the structures of serine proteinases of the chymotrypsin family. *J Mol Biol* **258**, 501-37.

Liao, D. I., Breddam, K., Sweet, R. M., Bullock, T., and Remington, S. J. (1992). Refined atomic model of wheat serine carboxypeptidase II at 2.2-A resolution. *Biochemistry*, **31**, 9796-812.

Luthy, R., Bowie, J. U., and Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83-5.

Matthews B. W., Sigler, P. B., Henderson, R., and Blow, D. M. (1967).Three-dimensional structure of tosyl-alpha-chymotrypsin. *Nature*, **21**, 4652-6.

McLachlan, A. D., and Shotton, D. M. (1971). Structural similarities between alpha-lytic protease of Myxobacter 495 and elastase. *Nat New Biol*, **229**,202-5.

Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S., and Overington, J. P. (1998). JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617-23.

Murugasu-Oei, B., Rodrigues, V., Yang, X., and Chia, W. (1995). Masquerade: a novel secreted serine protease-like molecule is required for somatic muscle attachment in the Drosophila embryo. *Genes Dev.* **9**, 139-54.

Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol.*, **247**, 536-40.

Muta, T., Hashimoto, R., Miyata, T., Nishimura, H., Toh, Y., and Iwanaga, S. (1990). Proclotting enzyme from horseshoe crab hemocytes. cDNA cloning, disulfide locations, and subcellular localization. *J Biol Chem.* 265, 426-33.

Neurath, H. (1985). Proteolytic enzymes, past and present. *Fed Proc.* **44**, 2907-13.

Overington, J., Johnson, M. S., Sali, A., and Blundell, T. L. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc R Soc Lond B Biol Sci*., **241**, 132-45.

Parry, M. A., Fernandez-Catalan, C., Bergner, A., Huber, R., Hopfner, K. P., Schlott, B., Guhrs, K. H., and Bode, W. (1998). The ternary microplasmin-staphylokinase-microplasmin complex is a proteinase-cofactor-substrate complex in action. *Nat Struct Biol*, **5**, 917-23.

Paskewitz, S. M., Reese-Stardy, S., Gorman, M. J. (1999). An easter-like serine protease from Anopheles gambiae exhibits changes in transcript abundance following immune challenge. *Insect Mol Biol*, **8**,329-37.

Peisach, E., Wang, J., de los Santos, T., Reich, E., and Ringe, D. (1999). Crystal structure of the proenzyme domain of plasminogen. *Biochemistry*, **38**, 11180-8.

Pendurthi, U. R., Allen, K. E., Ezban, M., Rao, L. V. (2000). Factor VIIa and thrombin induce the expression of Cyr61 and connective tissue growth factor, extracellular matrix signaling proteins that could act as possible downstream mediators in factor VIIa x tissue factor-induced signal transduction. *J Biol Chem* **275**, 14632-41.

Perona, J. J., and Craik, C. S. (1995). Structural basis of substrate specificity in the serine proteases. *Protein Sci*, **4**, 337-60.

Perona, J. J., and Craik, C. S. (1997). Evolutionary divergence of substrate specificity within the chymotrypsin-like serine protease fold. *J Biol Chem*, **272**, 29987-90.

Reid, K. B. M., Bentley, D. R., Campbell, R. D., Chung, L. D., Sim, R. B. Kristensen, T. and Tack, B. F. (1986). Complement system proteins which interact with C3B or C4B. *Immunol. Today*, **7**, 230-234.

Russell, R. B., and Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309-23.

Smirnova, I. V., Citron, B. A., Arnold, P. M., Festoff, B. W. (2001). Neuroprotective signal transduction in model motor neurons exposed to thrombin: G-protein modulation effects on neurite outgrowth, Ca(2+) mobilization, and apoptosis. *J Neurobiol*, **48**,87-100.

Sali, A., and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, **234**, 779-815.

Sali, A., and Blundell, T. L. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol*, **212**, 403-28.

Sayle, R. A., and Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem Sci*, **20**, 374.

Saitou, N, and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**, 406-25.

Shapiro, L., and Harris, T. (2000). Finding function through structural genomics. *Curr Opin Biotechnol*, **11**, 31-5.

Shi, J, Blundell, T. L., and Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, **310**, 243-57.

Srinivasan, N., and Blundell, T. L. (1993). An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. *Protein Eng*, **6**, 501-12.

Steitz, T. A., and Shulman, R. G. (1982). Crystallographic and NMR studies of the serine proteases. *Annu Rev Biophys Bioeng*, **11**, 419-44.

Stroud, R. M. (1974). A family of protein-cutting proteins. *Sci Am*. **231**, 74-88.

Twining, S. S. (1994). Regulation of proteolytic activity in tissues. *Crit Rev Biochem Mol Biol*, **29**, 315-83.

Suetake, T., Tsuda, S., Kawabata, S., Miura, K., Iwanaga, S., Hikichi, K., Nitta, K., and Kawano, K. (2000). Chitin-binding proteins in invertebrates and plants comprise a common chitin-binding structural motif. *J Biol Chem*, **275**, 17929-32.

Thery, C., Sharpe, M. J., Batley, S. J., Stern, C. D., and Gherardi, E. (1995). Expression of HGF/SF, HGF1/MSP, and c-met suggests new functions during early chick development. *Dev Genet*, **17**, 90-101.

Tsiang, M., Paborsky, L. R., Li, W. X., Jain, A. K., Mao, C. T., Dunn, K. E., Lee, D.W., Matsumura, S.Y., Matteucci, M. D., Coutre, S. E., Leung, L. L., and Gibbs, C. S. (1996). Protein engineering thrombin for optimal specificity and potency of anticoagulant activity in vivo. *Biochemistry*, **35**, 16449-57.

Watorek, W., Farley, D., Salvensen, G., and Travis, J. (1988). Nwutrophil elastase and CatepsinG: structure, function and biological control. *Advn. Expt. Med. Biol.* **240**, 23-31.

Wilke, M. E., Higaki, J. N., Craik, C. S., Fletterick, R. J. (1991). Crystallographic analysis of trypsin-G226A. A specificity pocket mutant of rat trypsin with altered binding and catalysis. *J Mol Biol*, **219**, 525-32.

Wilson, C. A., Kreychman, J., and Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol*, **297**, 233-49.

Yang, F., Gustafson, K. R., Boyd, M. R., and Wlodawer, A. (1998). Crystal structure of Escherichia coli HdeA. *Nat Struct Biol*, **5**, 763-4.

Zarembinski, T. I., Hung, L. W., Mueller-Dieckmann, H. J., Kim, K. K., Yokota, H., Kim, R., and Kim, S. H. (1998). Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc Natl Acad Sci U S A*, **95**, 15189-93.

Zhang, Y. L., Hervio, L., Strandberg, L., Madison, E. L. (1999). Distinct contributions of residue 192 to the specificity of coagulation and fibrinolytic serine proteases. *J Biol Chem*, **274**, 7153-6.

Zimmer, M., Medcalf, R. L., Fink, T. M., Mattmann, C., Lichter, P., and Jenne, D. E. (1992). Three human elastase-like genes coordinately expressed in the myelomonocyte lineage are organized as a single genetic locus on 19pter. *Proc Natl Acad Sci U S A*, **89**, 8215-9.

Figure Legends

Figure 6.1.

The Figure shows typical trypsin fold serine protease domain (shown bovine trypsin; PDB ID 5ptp-). The catalytic residues and loop regions spatially proximate to catalytic residues are marked. The figure was prepared using Setor (Evans, 1993).

Figure 6.2.

The Figure shows the binding site analysis as reported. The binding residues were defined as those having atoms less than 4Å apart. The reasons for considering the redundant structures can be 1) different inhibitors 2) different mutations 3) speciation and 4) Literature 5) deletions etc. Catalytic residues are marked.

Proteins used for Analysis are as follows:

Trypsin Structures

3tgj	RattusNorvegicus	BPTI
1bra	pig (D189G, G226D)	BENZAMIDINE
1anb	Rattus Rattus (S 214 E)	BENZAMIDINE
1and	Rattus Rattus (R 96 H)	BENZAMIDINE
1bit	Salmon (different crystal)	BENZAMIDINE
1c9p	Pig	Bdellastasin
2sta	Salmon	Squash Seed Inhibitor
1zzz	Bovine	C9H18N4O2 and C5H11N1O2
1ezs	Rattus Norvegicus	Ecotin Mutant
1f0t	Bovine	Rpr131247
1f2s	bovine beta trypsin	Mcti-A
1g3b	bovine beta trypsin	Meta-Amidino Schiff Base
1ldt	pig	Leech-Derived Tryptase Inhibitor
1ntp		C3H8O3P1 (MIP)

1ql9	X99Rt	Factor Xa Specific Inhibitor
1slw	Rat (N143H, E151H)	Ecotin (Nickel bound)
1slx	Rat (N143H, E151H)	Ecotin (Zinc bound)
1taw	Rat	Appi
1tgs		Porcine Pancreatic Secretory Inhibitor
1fy8	trypsinogen delta 16, 17	Bpti
1a0l		Appa
2trm	(D 102 N) at PH 7	BENZAMIDINE

Thrombin

1b7x	Thrombin Y225I	D-Phe-Pro-Arg-Chloromethylketone
1bhx	Human	Sdz (C19H28N6O4S2)
1thp	Human Y225P	D-Phe-Pro-Arg-Chloromethylketone
1bth	Bovine	BPTI
1d6w	Human	Decapeptide Inhibitor
1d9i	Human	Hirugen
1dm4	Human S195A	Fibrinopeptide A
1qhr	Human	TYS (C9H11N1O6S1)
1doj	Human	Rwj-51438
1e0f	Human	Haemadin
1eoj	Human	CPI and TIH
1vr1	Human	Plasminogen Activator Inhibitor-1
2thf	Human Y225F	D-Phe-Pro-Arg-Chloromethylketone

Elastase

1b0f	Human	Mdl 101, 146
1hne	Human	MSACK
1ppf	Human	turkey ovomucoid inhibitor

1ppg	Human	chloromethyl keton inhibitor
1bru	Pig	Gr143783
1qr3	Pig	Fr90127

Coagulation Facotor

1ezq	Human	Rpr128515
1fax	Human	DX9
1fjs	Human	Zk-807834
1kig	Bovine	Anticoagulant Peptide
1xka	Human	Fx-2212A
1pfx	Pig	D-Phe-Pro-Arg

Plasminogen Activators

1a5h	Human	Bis-Benzamidine
1a5i	Vampire Bat	Egr-Cmk
1bda	Human	Dansyl-Egr-Cmk
1bqy	Snake Venom	Chloromethylketone Inhibitor
1c5w	Human	ESI and FLC
1ejn	Human	Phenylguanidine
1lmw	Human	DEOXY-METHYL-ARGININE

Figure 6.3

The alignment of structures was built using STAMP (Russell and Barton, 1992) and annotated using JOY (Overington *et al*., 1990; Mizuguchi *et al*., 1998). Secondary structures and loop regions are as marked. Loop nomenclature was adopted from Peisach *et al*., (1999).

The proteins in the alignments are as follows.
1a0j Trypsin ; 1a0l Beta Tryptase; 1a5h Two-Chain Tissue Plasminogen Activator; 1a5i Saliva Plasminogen Activator; 1a7s Heparin Binding Protein; 1agj Epidermolytic Toxin A; 1ao5 Glandular Kallikrein-13; 1aut Activated Protein C; 1azz Collagenase; 1bio Complement Factor D; 1bqy Plasminogen Activator; 1bru Pancreatic Elastase; 1c1u Alpha Thrombin; 1cgh Cathepsin G; 1ddj Plasminogen Catalytic Domain; 1dpo Anionic Trypsin; 1dva Coagulation Factor VIIa; 1ejn Urokinase Plasminogen Activator; 1ekb Enteropeptidase; 1elt Native Pancreatic Elastase; 1fon Procarboxypeptidase; 1fuj Myeloblastin (PR3); 1fxy Coagulation Factor Xa-Trypsin Chimera; 1hcg Blood Coagulation Factor Xa; 1kig Bovine Factor Xa; 1klt Chymase; 1npm Neuropsin; 1pfx Factor Ixa; 1ppf Leukocyte Elastase; 1pyt Chymotrypsinogen C; 1qnj Pancreatic Elastase; 1qqu Beta Trypsin; 1rfn Coagulation Factor Ixa; 1sgf Nerve Growth Factor; 1sgt Streptomyces griseus trypsin; 1ton Tonin; 1trn Trypsin 1; 1try Fusarium-Oxysporum Trypsin; 2hlc Collagenase; 2pka Pancreatic Kallikrein A; 2sga Streptomyces griseus protease A; 2tbs Trypsin; 3rp2 Mast Cell Protease II; 4cha Alpha Chymotrypsin; 5ptp Beta Trypsin

Figure 6.4

Solvent accessibilities of 43 proteases structures listed in legend of Figure 6.3 were calculated using PSA (Lee and Richards, 1971) after removing all nonprotease entries from the PDB files. The mean accessibility (Y-axis) for each alignment position (X-axis) is shown as a solid bar and the root mean square deviation is shown as error bar.

Secondary structures are marked. The gap regions were assigned an arbitary accessibility of 100. The analysis was done using loop regions spatially proximate to catalytic triad.

Figure 6.5

Figure5 displays a tree calculated by CLUSTALX8.1 (Jeanmougin and Thompson, 1998) using Neighbour-Joining method (Saitou and Nei, 1987) from alignment of serine proteases shown in Figure3. The protein codes are as described in legend of Figure 6.3. Branch lengths are proportional to sequence divergence and can be measured relative to bar shown (top right). Branch labels record the stability of the branches over 1000 bootstrap replicates.

Figure6 6

The multiple alignment was prepared using profile mode of CLUSTALX8.1 (Jeanmougin and Thompson, 1998), where annotated sequences from the Swissprot database (Bairoch and Apweiler, 2000) are added to the structure based alignment prepared using STAMP (Russell and Barton, 1992). The residues identified as "binding" residues (see results) are marked with boxes. It should be noted that the marked residues are either absolutely conserved or sub-family specific, hence assumed to render specificity to the proteases.

The swissprot accession numbers (for structures please see legend of Figure 6.3) and short description for the sequences are as follows.
TRY2_MOUSE sp|P07146| Trypsin II Anionic Precursor; TRY1_CANFA sp|P06871| Trypsinogen Cationic Precursor; TRY1_CHICK sp|Q90627| Trypsin I-P1 Precursor; TRY1_XENLA sp|P19799| Trypsin Precursor; TRY1_GADMO sp|P16049| Trypsin I Precursor; TRY1_SALSA sp|P35031| TRypsin I Precursor; TRYP_SQUAC sp|P00764| Trypsin Precursor; 1TRY-_FUSOX Trypsin; 1SGT-_STRGR Streptomyces Griseus Trypsin; TRYA_DROER sp|P54624| Trypsin Alpha Precursor; TRYA_DROME sp|P04814| Trypsin Alpha Precursor; TRY4_LUCCU sp|P35044| Trypsin Alpha-4 Precursor; TRYE_DROER sp|P54627| Trypsin Epsilon Precursor; TRYE_DROME

sp|P35005| Trypsin Epsilon Precursor; TRYT_DROER sp|P54628| Trypsin Theta Precursor; TRYT_DROME sp|P42278| Trypsin Theta Precursor; TRYP_SARBU sp|P51588| Trypsin Precursor; TRYI_DROME sp|P52905| Trypsin Iota Precursor; TRYU_DROER sp|P54629| Trypsin Eta Precursor; TRYU_DROME sp|P42279| Trypsin Eta Precursor; TRYZ_DROER sp|P54630| Trypsin Zeta Precursor; TRYZ_DROME sp|P42280| Trypsin Zeta Precursor; TRY1_ANOGA sp|P35035| Trypsin 1 Precursor; TRY3_AEDAE sp|P29786| Trypsin 3A1 Precursor; TRYP_SIMVI sp|P35048| Trypsin Precursor; TRYA_MANSE sp|P35045| Trypsin Alkaline A Precursor; TRYP_CHOFU sp|P35042| Trypsin CFT-1 Precursor; THRB_MOUSE sp|P19221| ProThrombin Precursor; THRB_RAT sp|P18292| ProThrombin Precursor; THRB_BOVIN sp|P00735| ProThrombin Precursor; PLMN_BOVIN sp|P06868| Plasminogen Precursor; PLMN_SHEEP sp|P81286| Plasminogen; PLMN_PIG sp|P06867| Plasminogen; PLMN_MACMU sp|P12545| Plasminogen Precursor; PLMN_CANFA sp|P80009| Plasminogen; PLMN_MOUSE sp|P20918| Plasminogen Precursor; PLMN_ERIEU sp|Q29485| Plasminogen Precursor; PLMN_HORSE sp|P80010| Plasminogen; TPA_BOVIN sp|Q28198| Tissue-Type Plasminogen Activator Precursor; TPA_MOUSE sp|P11214| Tissue-Type Plasminogen Activator Precursor; TPA_RAT sp|P19637| Tissue-Type Plasminogen Activator Precursor; UROK_PAPCY sp|P16227| Urokinase-Type Plasminogen Activator Precursor; UROK_PIG sp|P04185| Urokinase-Type Plasminogen Activator Precursor; UROK_BOVIN sp|Q05589| Urokinase-Type Plasminogen Activator Precursor; UROK_MOUSE sp|P06869| Urokinase-Type Plasminogen Activator Precursor; UROK_RAT sp|P29598| Urokinase-Type Plasminogen Activator Precursor; UROK_CHICK sp|P15120| Urokinase-Type Plasminogen Activator Precursor; HCGA_HUMAN Blood Coagulation Factor Xa; KIGH_BOVINE Factor Xa; A10_RABIT sp|O19045| Coagulation Factor X Precursor; A10_CHICK sp|P25155| Coagulation Factor X Precursor; A10_TROCA sp|P81428| Coagulation Factor X; DVAH_HUMAN Coagulation Factor VIIa; A7_MOUSE sp|P70375| Coagulation Factor VII Precursor; FA7_RABIT sp|P98139| Coagulation Factor VII Precursor; FA7_BOVIN sp|P22457| Coagulation Factor VII; FA9_BOVIN sp|P00741| Coagulation Factor IX; FA9_SHEEP sp|P16291| Coagulation Factor IX; FA9_RAT sp|P16296| Coagulation Factor IX; EL1_BOVIN sp|Q28153| Elastase 1 Precursor; EL1_RAT sp|P00773| Elastase

1 Precursor; EL2_MOUSE sp|P05208| Elastase 2 Precursor; EL2_RAT sp|P00774| Elastase 2 Precursor; EL2_BOVIN sp|Q29461| Elastase 2 Precursor;

Figure 6.7

The figure shows alignment of serine protease like domain of masquerade like sequences with other proteins involved in patterning in Drosophila and bovine trypsin (5ptp-) and human thrombin (1c1uh). The residues conserved in masquerade like sequences are boxed. The functional residues identified are marked as star. Secondary structures shown according to bovine trypsin (5ptp-). The loop regions are as marked. Loop nomenclature was adopted from Peisach *et al*., (1999).

Figure 6.8

The chitin-binding motifs identified from masquerade and GRAAL gene product of Drosophila and Sp22D protein of Anophilis are shown with those identified from plants and other invertebrates. The residue numbers are shown and repeats are indicated by english uppercase letters (A to E). Proposed chitin-binding residues are boxed and conserved cysteines are marked with star.

Invertebrates are as follows: T. tridentatus tachycitin (Tachycitin), Anopheles gambiae chitinase (Ag-chit), Penaeus japonica chitinase 1 (Pj-chit1), Chelonus sp. chitinase (Ch-chit), 44-kDa glycoprotein from Lucilia cuprina (Peritrophin-44), Trichoplusia ni intestinal mucin (Tn-IM), five repeats of Drsophila masquerade (masquerade A, -B, -C, -D and -E), two repeats of Drosophila Graal gene product (Graal A, -B) and two repeats in Anophilis Sp22D protein (Sp22D A, and -B).

Plants are as follows: hevein from rubber tree (Hevein), Amaranthu caudatus antimicrobial protein, 2 (Ac-AMP2) and four homologous domains of wheat germ agglutinin (WGA A, -B, -C, and -D). Alignments of Proteins other than Masquerade, Graal and Sp22D are taken from Sutake et al., (2000). NMR structures of tachycitin (Sutake et al., 2000) and hevein (Anderson et al., 1993) are known.