Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochimica et Biophysica Acta 1698 (2004) 77–86
BIOCHIMICA ET BIOPHYSICA ACTA
BBA
www.bba-direct.com

Sequence and evolutionary analysis of the human trypsin subfamily of serine peptidases

George M. Yousef[a,b], Marc B. Elliott[a], Ari D. Kopolovic[a], Eman Serry[c], Eleftherios P. Diamandis[a,b,*]

[a] *Department of Pathology and Laboratory Medicine, Division of Clinical Biochemistry, Mount Sinai Hospital, 600 University Avenue, Toronto, ON, Canada M5G 1X5*
[b] *Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada M5G 1L5*
[c] *Faculty of Medicine, Department of Medical Biochemistry, Menoufiya University, Egypt*

Received 3 June 2003; received in revised form 1 October 2003; accepted 27 October 2003

Abstract

Serine peptidases (SP) are peptidases with a uniquely activated serine residue in the substrate-binding site. SP can be classified into clans with distinct evolutionary histories and each clan further subdivided into families. We analyzed 79 proteins representing the S1A subfamily of human SP, obtained from different databases. Multiple alignment identified 87 highly conserved amino acid residues. In most cases of substitution, a residue of similar character was inserted, implying that the overall character of the local region was conserved. We also identified several conserved protein motifs. 7–13 cysteine positions, potentially forming disulfide bridges, were also found to be conserved. Most members are secreted as inactive (pro) forms with a trypsin-like cleavage site for activation. Substrate specificity was predicted to be trypsin-like for most members, with few chymotrypsin-like proteins. Phylogenetic analysis enabled us to classify members of the S1A subfamily into structurally related groups; this might also help to functionally sort members of this subfamily and give an idea about their possible functions.

Keywords: Serine peptidase; Kallikrein; Chymotrypsin-like serine peptidase; Phylogenetic; Gene cluster; Protein domain

1. Introduction

Serine peptidases (SP) are peptidases with an active serine in their catalytic site. Two other residues, a histidine and an aspartate, are associated with the active serine in the catalytic sites of many families of SP including the trypsin (S1), subtilisin (S8), prolyl oligopeptidase (S9), and serine carboxypeptidase (S10) families. These residues form together what is referred to as the ''catalytic triad'' of SP. The positions of these residues are more or less conserved, with the codons for the catalytically essential histidine and serine being almost immediately adjacent to their exon boundary. In the trypsin subfamily, each of the catalytic triad residues is surrounded by a highly conserved motif. The motif ''GDSGGP'' surrounds serine, ''TAAHC'' histidine and DIMLL aspartate [1]. The active serine is situated in an internal pocket with the aspartate and histidine residues closely located in the three-dimensional structure.

Out of the estimated 400–500 peptidases in the human genome, approximately 30% are predicted to be SP [2]. This large family includes the digestive enzymes (e.g., trypsin, chymotrypsin), the kringle domain-containing growth factors (e.g., tissue plasminogen activator), some of the blood clotting factors, and the kallikreins [3–6]. Serine peptidases are involved in many vital functions such as digestion, coagulation and fibrinolysis, tissue remodeling, activation of hormones and growth factors, and extracellular matrix protein degradation.

Peptidases present a challenge for classification and nomenclature, for unlike most enzymes, they cannot easily be defined by activity. In essence they all share a common

Abbreviations: hK, human kallikrein protein; SP, serine peptidases; SCR, structurally conserved region; VR, variable region

* Corresponding author. Department of Pathology and Laboratory Medicine, Division of Clinical Biochemistry, Mount Sinai Hospital, 600 University Avenue, Toronto, ON, Canada M5G 1X5. Tel.: +1-416-586-8443; fax: +1-416-586-8628.
E-mail address: ediamandis@mtsinai.on.ca (E.P. Diamandis).

1570-9639/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbapap.2003.10.008

substrate, a peptide bond, but their specificities vary. The most widely used system for classification of peptidases is the MEROPS Clan System, where enzymes are first sorted into ''clans'' (sometimes referred to as superfamilies) based on evidence of evolutionary relationship [1,7,8]. Evidence for such relationships comes primarily from the linear order of catalytic site residues and the tertiary structure, in addition to distinctive aspects of catalytic activity such as specificity or inhibitor sensitivity. Each clan is given a two-letter identifier, of which the first letter is an abbreviation for the catalytic type, S for serine, C for cysteine, A for aspartic, and so forth (with the letter ''P'' being used for a clan containing families of more than one of the catalytic types serine, threonine and cysteine). Some clans are divided into ''subclans'' because there is evidence of a very ancient divergence within the clan, for example MA(E), the gluzincins, and MA(M), the metzincins. Next, proteins are classified into families (each denoted by a unique number) and subfamilies (denoted by another letter) based on sequence similarity to a chosen 'type example' for that family or to another protein that has already been shown to be homologous to the type example. For example, S1A is the trypsin subfamily; and S8A is the subtilisin subfamily. A protein can also be included in a family if it shows significant homology to another protein in that family which is not the type example.

A number of SP are secreted proteins, produced as inactive ''zymogens'' which require limited proteolysis to release the active enzyme. In many cases, the activator is another serine peptidase. Others are anchored to the cell membrane. Serine peptidases can be divided into two main evolutionary groups, the ''chymotrypsin-like'' SP and the ''subtilisin-like'' pro-protein convertases. The former group is believed to have evolved from a single ancestral gene that duplicated in the course of evolution to give rise to other genes that have gradually mutated and evolved to related peptidases and peptidase subfamilies with new functions, while the subtilisin-like group is believed to be the product of convergent evolution [9,10].

''Non-peptidase homologues'' are proteins that are deemed to be related to family S1 peptidases, but do not display any proteolytic activity. An example is protein Z, which shares significant sequence and structural homology with other blood-clotting factors, but, due to mutation of the catalytic Asp and His, has no peptidase activity.

With the near completion of the Human Genome Project, sequence information became available for almost all SP. In order to better understand the structural, functional and evolutionary aspects of human SP, we analyzed a group of 79 SP proteins, representing all known (confirmed and predicted) members of the S1A subfamily of SP. We provide here information regarding conserved and variable amino acids and protein motifs that might have an impact on function. In addition, we analyzed other structural aspects including the position of conserved cysteine residues, the cleavage site of the zymogen and substrate specificity. We also present a preliminary phylogenetic analysis of selected members of this subfamily.

2. Materials and methods

Multiple alignment was performed for 79 protein sequences representing the S1A subfamily, also known as the trypsin subfamily. Non-peptidase homologues and proteins where no complete sequence was available, such as the MASP3 product, were not included in this alignment. Sequences were obtained from the MEROPS [11] (release 5.9), SwissProt [12] (release 40.26), TrEMBL [13] (release 21.7 with daily updates) and GenBank [14] databases.

The amino acid sequence of each protein was scanned using the ProfileScan (http://hits.isb-sib.ch/cgi-bin/PFSCAN?) and ScanProsite algorithms (http://ca.expasy.org/tools/scanprosite/). Domains and secondary structural features were screened by several resources including the PFam database (release 7.5) (http://pfam.wustl.edu) and the PROSITE databases of profiles and verified in some cases with data available from the SwissProt and InterPro databases (http://www.expasy.org). The GenBank database was also searched for recent submissions of potential new serine proteases not yet included in other databases.

Multiple sequence alignment was performed using the ''ClustalW'' software package [15]. Different alignment parameters were tested and manual editing was performed in some cases to bring the sequences into the most biologically relevant alignment. Alignment viewings were done using the ''Boxshade'' (http://www.ch.embnet.org/software/BOX_form.html) and ''Chroma'' (http://www.lg.ndirect.co.uk/chroma/) programs.

Evolutionary analyses were performed using the ''Phylip'' software package (http://evolution.genetics.washington.edu/phylip.html), and the Molecular Evolutionary Genetics Analysis, 'MEGA' program (http://www.megasoftware.net). Different trees were constructed using a range of methods (UPGMA, Neighbor joining, Minimum Evolution and Maximum Parsimony), with different distance option models (Number of Differences, p-Distance, Poisson Correction and Gamma Distance).

3. Results

3.1. Conserved and variable amino acids

Members of the largest family of serine proteases, S1, from clan SA and their related proteins, in addition to related annotated sequence information from other databases (see above), were included in our analysis. Multiple alignment of members of this subgroup is presented in Fig. 1. A list of conserved amino acids is presented in Table 1 with the percentage of conservation and the major substitutions present at each position.

Table 1 shows that seven residues are absolutely conserved in humans: Pro[28], His[57], Asp[102], Cys[168], Cys[182], Ser[195] and Gly[196]. Three of these seven positions, 57, 102 and 195, constitute the catalytic triad of SP, and Gly[196] is next to serine in the "GDSGGP" motif of the active serine residue and is perhaps required for steric reasons, so as not to occlude the serine's hydroxyl group during catalytic cleavage. The remaining residues are two cysteines and a proline and are likely essential for structural reasons. The cysteines have been identified as disulfide bonding partners, a bond which is likely required to maintain the shape of the active site. Due to its special character, a proline can be assumed to be conserved for structural purposes. Of note is that Gly[193], whose backbone amine hydrogen is necessary for the formation of the oxyanion hole [16], is not absolutely conserved.

Although only seven residues showed 100% conservation, an additional 15 showed almost complete conservation (95%+). Eight of these were within close proximity to one of the three catalytic residues. Indeed, all six residues of the GDSGGP motif around Ser[195] showed at least 95% conservation. Of the remaining seven residues, two were a disulfide bonding pair, two were members of a conserved GWG motif, and one was Ser[214], which has been identified as being potentially important in the formation of the S1 binding pocket [8]. The other two, Leu[155] and Trp[237], have yet to have their significance established. In total, 48 residues were found to be more than 80% conserved, and 87 residues were found to display greater than 50% conservation.

Conserved residues tended to group together, likely representing certain necessary structural or functional domain elements. This conclusion is supported by the fact that in most cases of substitution, a residue of similar character (i.e. size, hydrophobicity, polarity) was inserted, implying that the overall character of the local region was conserved for proper function, more so than some of the individual amino acid identities. For example, in cases where the consensus residue is an aromatic amino acid (Trp, Tyr, Phe), an aromatic substitution occurs in 89% of cases. For example, position 29, which has a conserved Trp, has 17 substitutions, 15 aromatic and 2 serines (please note that these are the absolute numbers of the percentages presented in Table 1, raw data is available from the corresponding author). Position 94, a consensus Tyr, has 28 substitutions, 22 of which are aromatic; of the remaining 6 residues, only 2 are hydrophilic (Arg and Ser), so the conservation of hydrophobicity at that position is largely maintained.

Where the consensus residue was aliphatic (Leu, Ile, Val), a non-aliphatic substitution occurred in only 37% (233/630) of the time. In some cases, aliphatic character was completely conserved, such as position 103, which is neighbouring to the catalytic Asp[102]. The consensus Ile at this position was substituted 20 times, 15 times by Leu and 5 times by Val. In other cases, it was not so conserved, such as Val[138], which was substituted 38 times, in 22 cases by either an Ile or a Leu residue.

3.2. Conserved protein motifs in human SP

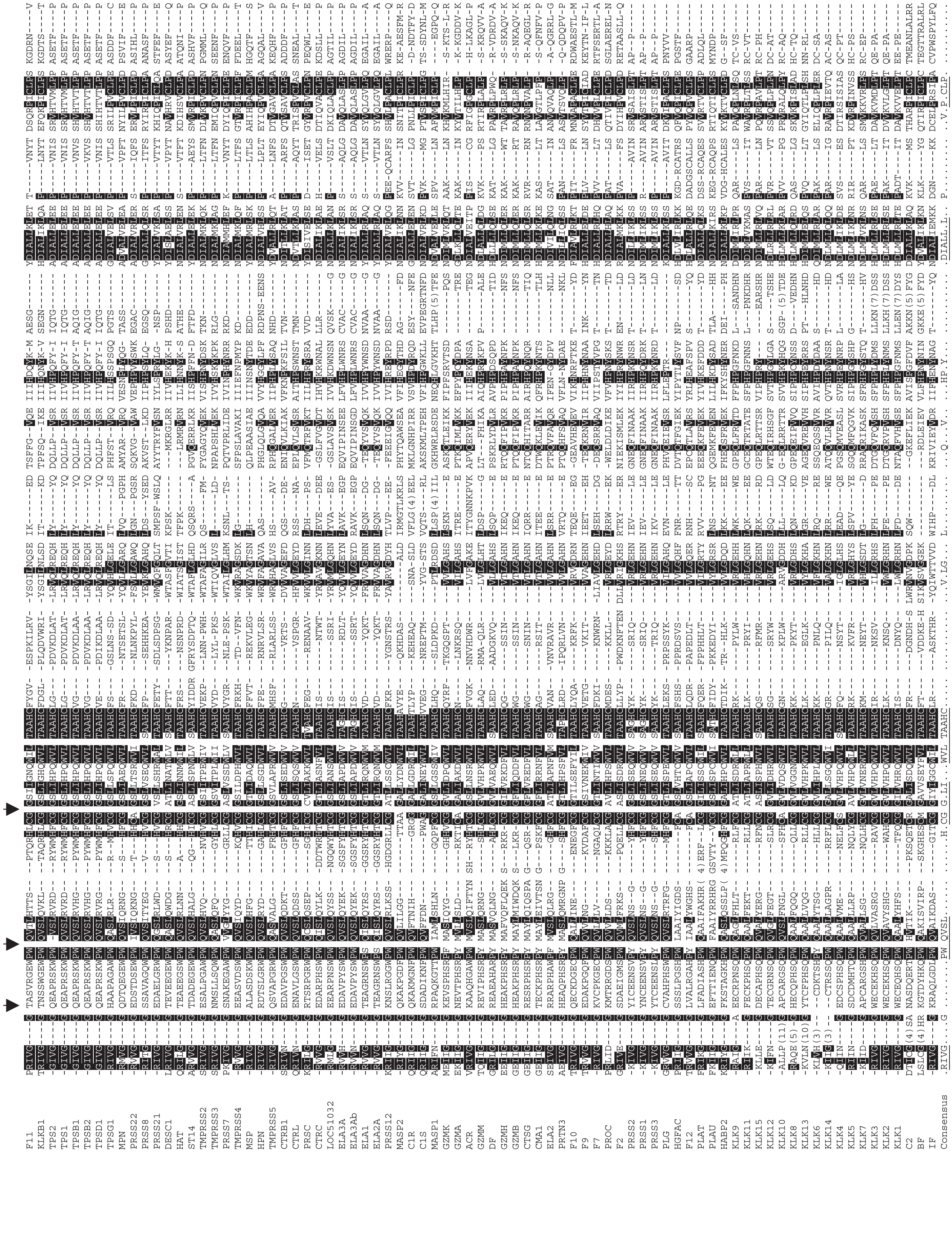
A number of conserved amino acid motifs are shown at the bottom of Fig. 1. All motifs around the catalytic triad, WVL**TAA*H*C** (positions 51–58), ***D*IALL**L (positions 102–108), **GD*S*GGP** (positions 193–198), are highly conserved. In addition, other short motifs, e.g., VxGWG (where x represents Tyr or Ser or Ala) at positions 140–142, **CGG**(S/T)**L**(**I**/L/V) (positions 42–47), SWG that contains the critical Ser[214] (positions 214–216), P(**W**/Y)(**Q**/M)(**V**/A)X(**L**/I/V) (positions 28–33) and the (**R**/K)(**I**/V/L)(**V**/I/L)**GG** trypsin cleavage site at the start of the enzyme (positions 15–19) (characters in bold represent more conserved positions).

3.3. Conserved cysteine residues

Cysteine residues form disulfide bridges that help to keep the molecule intact and to maintain the conformation of elements of the active site. Moreover, in some cases like thrombin, an internal peptide is also excised during activation, and the two resultant peptide chains remain linked by a disulfide bridge [17]. Our global alignment of the 79 proteins showed 13 conserved cysteine residues (positions 22, 42, 58, 122, 127, 136, 157, 168, 182, 191, 201, 220 and 232) (Fig. 2), nine of which were found to be more than 50% conserved, eight showed 70% conservation and six were 90% conserved.

Dayhoff, in 1978, comparing a set of 11 SP from different species, identified 14 conserved or semi-conserved cysteine positions that formed intramolecular disulfide bonds [18]. Comparing our multiple alignment with that of Dayhoff and the information from the literature, bonds were established to exist between positions 42 and 58, 136 and 201, 168 and 182, and 191 and 220. In human SP, the third pair (168 and 182) is absolutely conserved. The three other pairs were also highly conserved, and any gaps were mostly coincident (e.g. the bond between residues 136 and 201 was conserved in 62 of 79 sequences (78%), with 14 of 17 deletions being coincident). A less conserved bond was found to exist in the kallikreins and trypsins between positions 22 and 157 and another was found only in

Fig. 1. (shown on next two pages). Multiple alignment of 79 members of the human S1A family of proteases. Conserved residues are highlighted in black (an arbitrary cut-off of 50% was used for conservation). Dashes represent gaps, introduced for the best alignment. A consensus sequence is shown at the bottom of each column, with conserved amino acid motifs underlined. For full gene (protein) names, see Appendix A. Numbers in brackets represent sparse linker regions that have been excluded for the sake of the compactness of the alignment. Any stretch of residues more than four amino acids in length that was not present in at least 80% of the represented sequences was replaced in this manner.



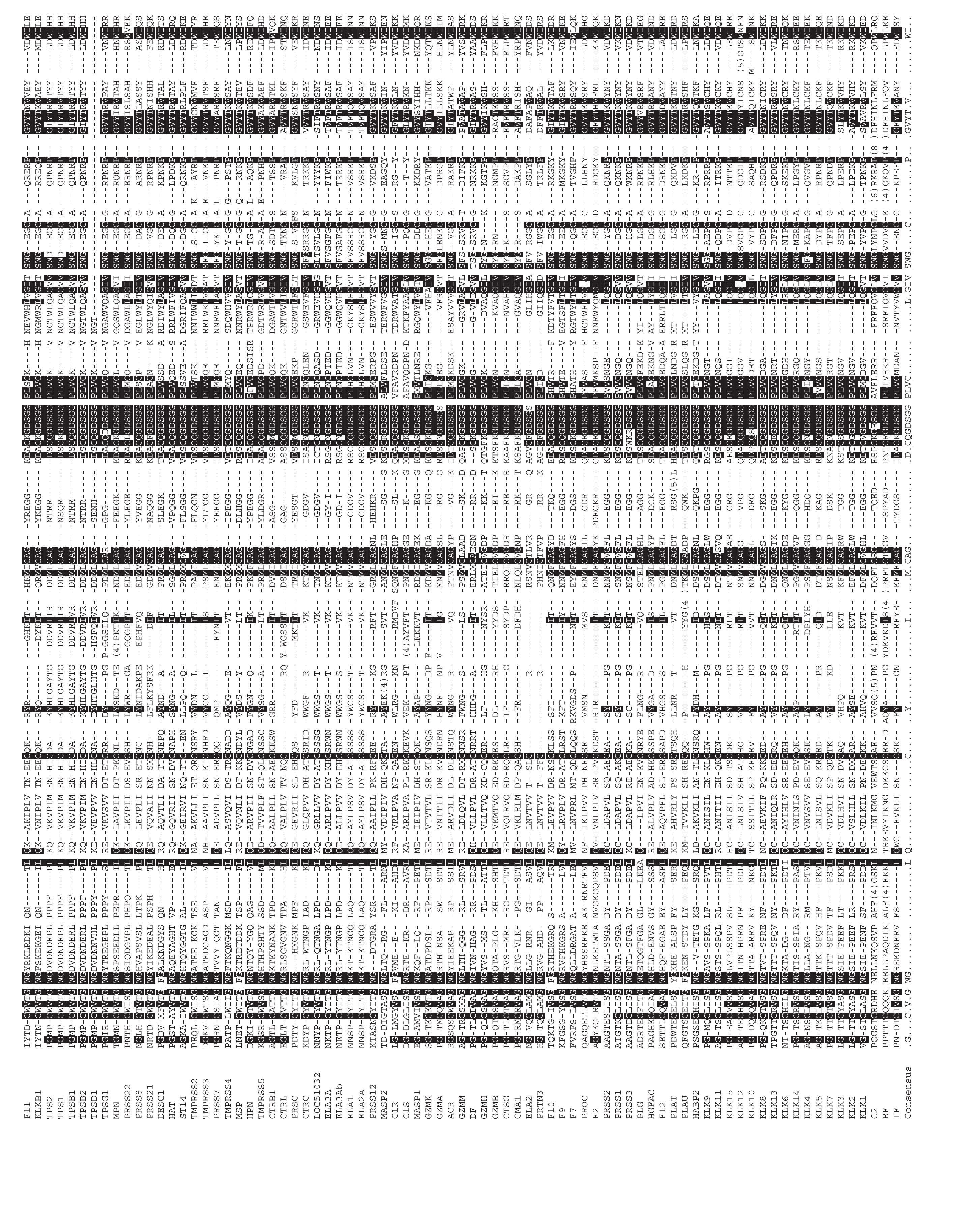


Table 1
Conserved amino acids in members of family S1A of human serine proteases

Pos.[a]	Res[b]	%[c]	Pos.	Res.	%	Pos.	Res.	%	Pos.	Res.	%	Pos.	Res.	%
15	**R**	**63**	**53**	**L**	**72**	**108**	**L**	**95**	**184**	**G**	**91**	**220**	**C**	**84**
	K	15		I/V	24		I	3		A	3		A	6
	E	9		M	4		M	2		S	1		none	6
16	**I**	**73**	**54**	**T**	**81**	**111**	**P**	**53**	**189**	**D**	**72**	**225**	**P**	**87**
	V/L	23		S	19		K/R	19		S/T	15		Y	10
	M	4	**55**	**A**	**99**		S/T	14		G	4		F	1
17	**V**	**56**		V	1	**118**	**V**	**65**	**191**	**C**	**94**	**226**	**G**	**73**
	I/L	35	**56**	**A**	**93**		I/L	30		F/W	6		T/S	9
	Y/F	5		G	5		Q	3					A	6
18	**G**	**80**		T	1	**120**	**P**	**58**	**192**	**Q**	**51**	**227**	**V**	**76**
	N	10	**57**	**H**	**100**		T/S	19		K/R	23		I/L	15
	D/E	6					L/V/I	14		N	9		F	6
19	**G**	**94**	**58**	**C**	**96**	**122**	**C**	**52**	**193**	**G**	**96**	**228**	**Y**	**81**
	A	3		T	1		S/T	15		S	1		F	11
	none	3		A	1		P	6		D	1		H	3
28	**P**	**100**	**66**	**V**	**76**	**123**	**L**	**90**	**194**	**D**	**99**	**229**	**T**	**80**
29	**W**	**78**		I/L	10		V/I	7		E	1		V/I	10
	Y/F	19		A	9		F/W	3	**195**	**S**	**100**		A	5
	S	3	**68**	**L**	**68**	**124**	**P**	**84**	**196**	**G**	**100**	**231**	**V**	**77**
30	**Q**	**76**		A	13		S	5	**197**	**G**	**98**		I/L	19
	M	14		V/I	10		A	4		S	3		A	1
	I/L	5	**69**	**G**	**85**	**133**	**G**	**57**	**198**	**P**	**95**	**237**	**W**	**99**
31	**V**	**57**		R	8		N/Q	10		A	4		none	1
	A	36		none	4		D/E	10		V	1	**238**	**I**	**91**
	I	4	**73**	**L**	**52**	**136**	**C**	**82**	**199**	**L**	**82**		L	4
32	**S**	**53**		I/V	21		none	5		V	6		V	3
	A	20		R/K	8		G	4		F	5			
	V/L	9	**81**	**Q**	**51**	**138**	**V**	**56**	**200**	**V**	**63**			
33	**L**	**75**		I/L/V	19		I/L	28		I/L	9			
	I/V	23		K/R	6		A	8		M	8			
	T	1	**85**	**V**	**59**	**140**	**G**	**98**	**201**	**C**	**78**			
40	**H**	**52**		I/L	19		E	1		S/T	10			
	I/L	18		A	10		K	1		V/L	5			
	F	5	**91**	**H**	**81**	**141**	**W**	**90**	**209**	**L**	**57**			
42	**C**	**96**		N	8		F	6		V/I	18			
	A	3		Y	4		Y	1		Q	14			
	G	1	**92**	**P**	**66**	**142**	**G**	**98**	**211**	**G**	**98**			
43	**G**	**89**		S/T	9		E	3		A	1			
	A	8		D/E	8	**155**	**L**	**98**		none	1			
	S	1	**94**	**Y**	**65**		I	1	**212**	**I**	**51**			
44	**G**	**84**		F/W	28		A	1		V/L	44			
	A	14		V	2	**156**	**Q**	**53**		A	1			
	C	1	**102**	**D**	**100**		K/R	28	**213**	**V**	**57**			
46	**L**	**89**	**103**	**I**	**73**		N	6		T	24			
	I/V	10		V/L	27	**168**	**C**	**100**		I/L	14			
	Y	1	**104**	**A**	**59**				**214**	**S**	**98**			
47	**I**	**61**		M	26	**180**	**M**	**86**		T	1			
	V/L	39		L/V	6		N	4		none	1			
51	**W**	**89**	**105**	**L**	**95**		E	4	**215**	**W**	**73**			
	F	10		V/I	4	**182**	**C**	**100**		F/Y	20			
	Y	1		M	1					G	3			
52	**V**	**89**	**106**	**L**	**54**	**183**	**A**	**82**	**216**	**G**	**86**			
	I/L	19		I/V	36		V	11		V	8			
	A	1		M	8		S/T	5		D	3			

[a] Amino acid position by chymotrypsin numbering.

[b] Consensus residues at given positions are in bold. Residues over 80% conserved are shaded. A threshold value of 50% was set for determining whether a residue was deemed conserved or not.

[c] Percentage of 79 aligned sequences in which the listed residue appears. Percentages are rounded to the nearest number. Only top 3 amino acids are listed.

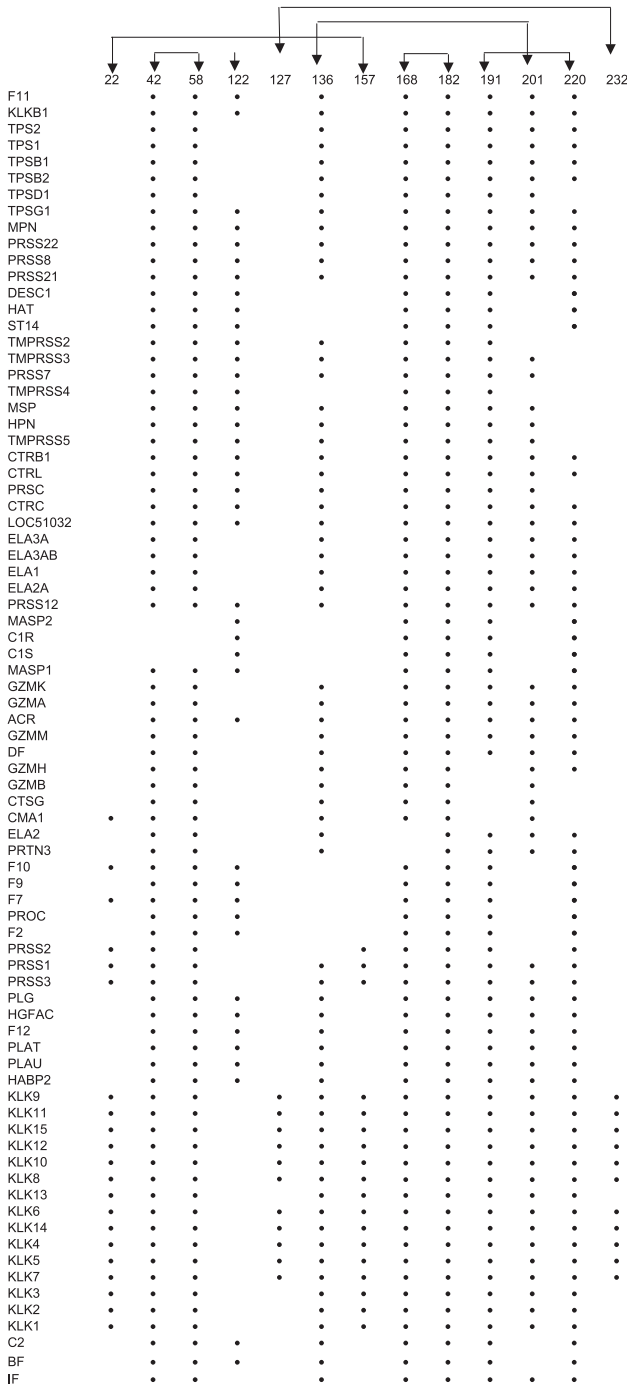


Fig. 2. Schematic presentation of the conserved cysteine residues in 79 proteins of the S1A family of human serine proteases. The numbering system refers to human chymotrypsin. Conserved residues for each protein are shown as dots. For full protein names, see Appendix A. Note that only the mature proteins were used for multiple alignment. For discussion, see text.

kallikreins between positions 127 and 232, except hK1-3 and hK13. This might reflect a distinct structural identity for the kallikrein cluster of SP. It is worth noting that in many cases when one member of a bonding pair is not conserved, it is replaced by an aliphatic or aromatic residue, which may interact with the remaining cysteine and provide weaker structural support to the protein by hydrophobic interactions.

The cysteine residue at position 122 was identified as forming a bond to a residue N-terminal to the trypsin domain. As a result, it is conserved in proteins that possess N-terminal domains necessary for their function, and whose active form is comprised of multiple chains (such as plasminogen/plasmin). Smaller enzymes (e.g. trypsins or kallikreins), whose only functional domain is a trypsin domain, sometimes lose their N-terminal tails upon activation, with Ile^{16} (or equivalent) becoming the new N-terminus. Where conserved and involved in disulfide bonding, it may partner with either a residue in the N-terminal tail of the protein, or to an as-yet unidentified residue elsewhere.

3.4. Protein activation cleavage sites

Many SP are secreted as inactive (pre-pro) enzymes. The signal peptide is cleaved upon secretion of the protein and the inactive (pro-enzyme) is released. This is followed by activation of the molecule by N-terminal cleavage. The conserved domain (R/K)(I/V)(V/I)(G/N) is found at the N-terminal cleavage site of the zymogen (pro-enzyme) end of most SP (Fig. 1). Most enzymes are cleaved after an Arg or Lys, indicating the need for a trypsin-like enzyme for activation. In case of trypsin, cleavage occurs between residues Lys^{15} and Ile^{16} (chymotrypsinogen numbering). After cleavage, Ile^{16} forms the new N-terminus of the protein, and Asp^{194} rotates to interact with it. This rotation and the resulting salt bridge produce a conformational change that completes the formation of the oxyanion hole and the substrate binding pocket, both of which are necessary for proper catalytic activity.

Certain sequences in our alignment did not display conservation of this trypsin cleavage site, with substitutions at either the 15th or 16th positions (e.g., the granzymes and hK10) (Fig. 1). These substitutions likely result in either cleavage by a peptidase with different specificity, or no cleavage. For instance, granzyme H was shown to be cleaved by dipeptidyl peptidase I, which cleaves between Glu^{15} and Ile^{16} [19]. In all cases, it is probable that even if cleavage occurs, subsequent interactions may not, resulting in an enzyme with no proteolytic activity. As well, in some sequences cleavage at this site has not been definitively established, as it may not be necessary for activation. Changes in other residues or association with a cofactor may serve to stabilize the active conformation of the protein, making cleavage unnecessary.

3.5. Substrate specificity

Serine peptidases exhibit preference for hydrolysis of peptide bonds adjacent to a particular class of amino acids. In the trypsin-like group, the peptidase cleaves peptide bonds following positively charged amino acids such as

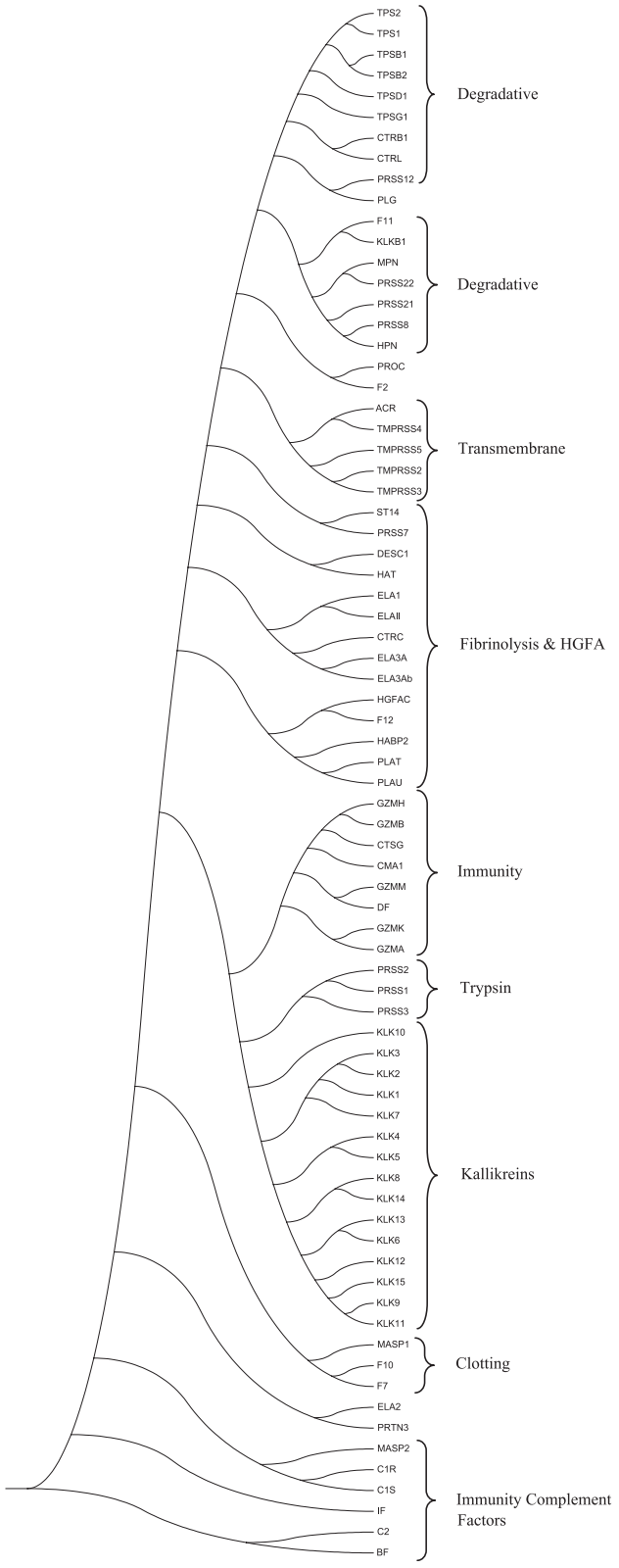


Fig. 3. An example of an evolutionary tree for selected members of the S1A family of SP. This tree was constructed by the UPGMA method with the Poisson correction distance model. In addition to classifying the proteins into structurally related subgroups, it might also help in ''functional'' classification (see text for discussion). For full protein names, see Appendix A.

arginine or lysine, since it has an aspartate (or glutamate) in the substrate-binding pocket that can form a strong electrostatic bond with these residues. The chymotrypsin-like peptidases have a non-polar substrate-binding pocket, and thus require an aromatic or bulky non-polar amino acid such as tryptophan, phenylalanine, tyrosine or leucine. The elastase-like enzymes, on the other hand, have bulky amino acids (valine or threonine) in their binding pockets, thus requiring small hydrophobic residues, such as alanine [20].

The presence of Asp in position 189 of our multiple alignment indicates that most members of the S1A subfamily will have a trypsin-like specificity. In chymotrypsin and chymotrypsin-like peptidases (e.g. hK3, PSA), there is a Ser in this position (Fig. 1). Some few enzymes have other residues at this position, possibly indicating a distinct pattern of substrate specificity.

3.6. Evolutionary analysis

A representative evolutionary tree of members of the S1A subfamily of SP is presented in Fig. 3. Our analysis utilizing the full protein sequence allowed separation of members of the S1A subfamily into different structurally related groups, e.g. kallikreins, transmembrane peptidases and the tryptases. In addition, it might also be helpful in clustering the proteins with similar functional domains in the same sister group (functional categories). The possible functional categories are shown in Fig. 3 beside each group. This functional classification is highly similar to the functional tree constructed by Krem et al. [21], with the exception that ours included more members for the analysis and included only the ''human'' S1A peptidases.

This preliminary functional classification is hard to verify because of the lack of information about the exact function for some members, and although it provides only approximate preliminary evidence, it opens the door for further more detailed studies, ands it might be also of help in predicting the function of some newly identified members. The main purpose of this tree was to give a graphical depiction of the distances between the structures of different members of this subfamily and not to represent an accurate construction of the evolutionary history of the family; statistical analysis has not therefore been performed.

Several trees have been previously constructed for various lineages within the S1 peptidase family. Some of these suggested that non-peptidase domains could have significant evolutionary influence [22]. Others, however, suggested that the peptidase domain, especially its C-terminal part, accounts fully for the functional diversity of SP and is an important element in shaping their evolution [23]. Trees based on this domain were previously published for some families [3]. Krem et al. [21] recently published a dendogram based on the peptidase domain sequences and used it

to classify them into distinct functional groups. The apparent driving force behind this phenomenon was substrate recognition. Another recent evolutionary analysis was based on ‘‘evolutionary markers’’ [8].

4. Discussion

In the present study, we performed sequence analysis of 79 members of the S1A subfamily of SP. Our results are consistent with previously published alignment that utilize smaller sets of human SP [4,17,24–27]. In most studies, however, only few human members were included along with other mammalian SP. Multiple alignment of all members of the S1A subfamily was also recently published [3]. This alignment should allow for better detection of conservation and evolutionary changes in the human lineage.

Structurally conserved regions (SCR) usually remain conserved in all members of the subfamily and are usually composed of secondary structure elements, the immediate active site and other essential structural residues of the molecule. For instance, Ser^{214} in chymotrypsin-like peptidases contributes to the S1 binding pocket and appears to be a fourth member of a catalytic tetrad [9]. Between these conserved elements are highly variable stretches (also called variable regions ‘‘VR’’). These are almost always loops that lie on the external surface of the protein and contain all additions and deletions between different protein sequences. The former regions (SCR) have been successfully utilized as the bases for predicting the 3D structure of newly identified SP based on information from existing members [26]. The latter (VR) are important for studying the evolutionary history of SP.

Certain residues with variable degree of conservation can be investigated for their usefulness as ‘‘evolutionary markers,’’ which can give an idea about the history of each enzyme family or clan and allow comparative analysis with other families or clans. Krem and Di Cera [8] identified several such evolutionary markers with proven evolutionary usefulness. In addition to the use of these markers for rooting the phylogenetic trees, an attempt was made to classify SP into functional groups based on these markers and/or their coding sequences. Absolutely conserved non-serine residues will likely yield little evolutionary information but other less conserved residues might be more useful.

Protein alignment of SP can rely on either the amino acid sequence similarity or the overlap of the 3D structure [17,26]. It is always important to be careful in interpreting the ‘‘functional’’ and ‘‘structural’’ aspects of multiple alignment. Specific changes in certain residues may signal a novel functional identity of the protein. For instance, haptoglobin is no longer a ‘‘functional’’ SP although its sequence homology clearly places it among the ‘‘structural’’ family of SP.

Certain sequences in our alignment displayed missing regions that were preserved in most or all of the remaining proteins. These regions might reflect an incomplete identification of the full structure of the gene (especially when located at the C-terminal end, as in case of the TPSD1), or the fact that these proteins were missing a given functional region, which likely would have reduced or modified their enzymatic activity. Another possibility is that these sequences might represent pseudogenes, and might not be expressed.

Appendix A. Proteins used for multiple alignment

Protein name	Symbol	Merops ID	UniGene ID
Acrosin	ACR	S01.223	Hs.183088
Complement factor B	BF	S01.196	Hs.69771
Complement component C1r (activated)	C1R	S01.192	Hs.1279
Complement component C1s (activated)	C1S	S01.193	Hs.284609
Complement component 2	C2	S01.194	Hs.2253
Chymase	CMA1	S01.140	Hs.135626
Chymotrypsin B	CTRB1	S01.152	Hs.74502
Chymotrypsin C	CTRC	S01.157	Hs.8709
Chymopasin	CTRL	S01.256	Hs.405774
Cathepsin G	CTSG	S01.133	Hs.74502
DESC1 peptidase	DESC1	S01.021	Hs.201877
Complement factor D	DF	S01.191	Hs.155597
Pancreatic elastase	ELA1	S01.153	Hs.348395
Neutrophil elastase	ELA2	S01.131	Hs.99863
Pancreatic elastase 2A	ELA2A	S01.155	Hs.2121
Pancreatic endopeptidase E	ELA3A	S01.154	Hs.181289
Pancreatic endopeptidase E form B	ELA3B	S01.205	Hs.425790
Thrombin	F2	S01.217	Hs.76350
Coagulation factor VIIa	F7	S01.215	Hs.36989
Coagulation factor IXa	F9	S01.214	Hs.1330
Coagulation factor Xa	F10	S01.216	Hs.47913
Coagulation factor XIa	F11	S01.213	Hs.1430
Coagulation factor XIIa	F12	S01.211	Hs.1321
Granzyme A	GZMA	S01.135	Hs.90798
Granzyme B	GZMB	S01.010	Hs.1051
Granzyme H	GZMH	S01.147	Hs.348264
Granzyme K	GZMK	S01.146	Hs.3066
Granzyme M	GZMM	S01.139	Hs.268531
Plasma hyaluronan-binding serine protease	HABP2	S01.033	Hs.241363
Human airway trypsin-like enzyme	HAT	S01.301	Hs.132195
Hepatocyte growth factor activator	HGFAC	S01.228	Hs.104
Hepsin	HPN	S01.224	Hs.823
Complement factor I	IF	S01.199	Hs.36602
Human kallikrein 1	KLK1	S01.160	Hs.123107
Human kallikrein 2	KLK2	S01.161	Hs.181350
Human kallikrein 3	KLK3	S01.162	Hs.171995
Human kallikrein 4	KLK4	S01.251	Hs.218366
Human kallikrein 5	KLK5	S01.017	Hs.50915
Human kallikrein 6	KLK6	S01.236	Hs.79361
Human kallikrein 7	KLK7	S01.300	Hs.151254
Human kallikrein 8	KLK8	S01.244	Hs.104570

(*continued on next page*)

Appendix A (*continued*)

Protein name	Symbol	Merops ID	UniGene ID
Human kallikrein 9	KLK9	S01.307	Hs.447142
Human kallikrein 10	KLK10	S01.246	Hs.69423
Human kallikrein 11	KLK11	S01.257	Hs.57771
Human kallikrein 12	KLK12	S01.020	Hs.159679
Human kallikrein 13	KLK13	S01.306	Hs.165296
Human kallikrein 14	KLK14	S01.029	Hs.283925
Human kallikrein 15	KLK15	S01.081	Hs.250770
Plasma kallikrein	KLKB1	S01.212	Hs.1901
Pancreatic elastase II form B	LOC5 1032	S01.206	Hs.169234
Mannose-binding-protein-associated serine peptidase 1	MASP1	S01.198	Hs.356082
Mannose-binding-protein-associated serine peptidase 2	MASP2	S01.229	Hs.119983
Marapsin	MPN	S01.074	Hs.332878
Membrane-type mosaic serine peptidase	MSP	S01.087	Hs.266309
t-Plasminogen activator	PLAT	S01.232	Hs.274404
u-Plasminogen activator	PLAU	S01.231	Hs.77274
Plasmin	PLG	S01.233	Hs.75576
Protein C (activated)	PROC	S01.218	Hs.2351
Corin	PRSC	S01.019	Hs.340634
Cationic Trypsin	PRSS1	S01.127	Hs.419094
Anionic Trypsin	PRSS2	S01.258	Hs.241561
Mesotrypsin	PRSS3	S01.174	Hs.58247
Enteropeptidase	PRSS7	S01.156	Hs.158333
Prostasin	PRSS8	S01.159	Hs.75799
Neurotrypsin	PRSS12	S01.237	Hs.22404
Testisin	PRSS21	S01.011	Hs.72026
Brain serine protease 2	PRSS22	S01.252	Hs.125532
Myeloblastin	PRTN3	S01.134	Hs.928
Matriptase	ST14	S01.302	Hs.56937
Transmembrane serine protease 2	TMPRSS2	S01.247	Hs.318545
Transmembrane serine peptidase 3	TMPRSS3	S01.079	Hs.298241
Transmembrane serine peptidase 4	TMPRSS4	S01.034	Hs.63325
Transmembrane serine protease 5	TMPRSS5	S01.323	Hs.46720
Tryptase alpha 1	TPS1	S01.143	Hs.334455
Tryptase alpha 2	TPS2	S01.015	N/A[‡]
Tryptase beta 1	TPSB1	S01.027	Hs.406479
Tryptase beta 2	TPSB2	S01.242	Hs.294158
Tryptase delta 1	TPSD1	S01.054	Hs.241387
Tryptase gamma 1	TPSG1	S01.028	Hs.278275

[‡] No unigene cluster ID available at time of writing.

References

[1] N.D. Rawlings, A.J. Barrett, Evolutionary families of peptidases, Biochem. J. 290 (1993) 205–218.

[2] C. Southan, A genomic perspective on human proteases as drug targets, Drug Discov. Today 6 (2001) 681–688.

[3] A.J. Barrett, N.D. Rawlings, Families and clans of serine peptidases, Arch. Biochem. Biophys. 318 (1995) 247–250.

[4] G.M. Yousef, E.P. Diamandis, The new human tissue kallikrein gene family: structure, function, and association to disease, Endocr. Rev. 22 (2001) 184–204.

[5] E.P. Diamandis, G.M. Yousef, Human tissue kallikreins: a family of new cancer biomarkers, Clin. Chem. 48 (2002) 1198–1205.

[6] R. Asakai, E.W. Davie, D.W. Chung, Organization of the gene for human factor XI, Biochemistry 26 (1987) 7221–7228.

[7] A.J. Barrett, N.D. Rawlings, Families and clans of serine peptidases, Arch. Biochem. Biophys. 318 (1995) 247–250.

[8] M.M. Krem, E. Di Cera, Molecular markers of serine protease evolution, EMBO J. 20 (2001) 3036–3045.

[9] J.J. Perona, C.S. Craik, Structural basis of substrate specificity in the serine proteases, Protein Sci. 4 (1995) 337–360.

[10] D.I. Liao, K. Breddam, R.M. Sweet, T. Bullock, S.J. Remington, Refined atomic model of wheat serine carboxypeptidase II at 2.2—A resolution, Biochemistry 31 (1992) 9796–9812.

[11] N.D. Rawlings, E. O'Brien, A.J. Barrett, MEROPS: the protease database, Nucleic Acids Res. 30 (2002) 343–346.

[12] A. Bairoch, R. Apweiler, The SwissProt protein sequence database and its supplement TrEMBL in 2000, Nucleic Acids Res. 28 (2000) 45–48.

[13] C. O'Donovan, M.J. Martin, E. Glemet, J.J. Codani, R. Apweiler, Removing redundancy in SwissProt and TrEMBL, Bioinformatics 15 (1999) 258–259.

[14] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, B.A. Rapp, D.L. Wheeler, GenBank, Nucleic Acids Res. 30 (2002) 17–20.

[15] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.

[16] S.P. Bajaj, S.G. Spitzer, W.J. Welsh, B.J. Warn-Cramer, C.K. Kasper, J.J. Birktoft, Experimental and theoretical evidence supporting the role of Gly363 in blood coagulation factor IXa (Gly193 in chymotrypsin) for proper activation of the pro-enzyme, J. Biol. Chem. 265 (1990) 2956–2961.

[17] A.M. Lesk, W.D. Fordham, Conservation and variability in the structures of serine proteinases of the chymotrypsin family, J. Mol. Biol. 258 (1996) 501–537.

[18] M.O. Dayhoff, Atlas of protein sequence and structure, Natl. Biomed. Res. Found. 5 (1978) 79–81.

[19] T.V. Tran, K.A. Ellis, C.M. Kam, D. Hudig, J.C. Powers, Dipeptidyl peptidase I: importance of progranzyme activation sequences, other dipeptide sequences, and the N-terminal amino group of synthetic substrates for enzyme activity, Arch. Biochem. Biophys. 403 (2002) 160–170.

[20] L. Stryer, Biochemistry, 4th ed., W.H. Freeman and Company, New York, 1995.

[21] M.M. Krem, T. Rose, E. Di Cera, Sequence determinants of function and evolution in serine proteases, Trends Cardiovasc. Med. 10 (2000) 171–176.

[22] R.M. Lawn, K. Schwartz, L. Patthy, Convergent evolution of apolipoprotein(a) in primates and hedgehog, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 11992–11997.

[23] M.M. Krem, T. Rose, E. Di Cera, The C-terminal sequence encodes function in serine proteases, J. Biol. Chem. 274 (1999) 28063–28066.

[24] G.M. Yousef, E.P. Diamandis, Human tissue kallikreins: a new enzymatic cascade pathway? Biol. Chem. 383 (2002) 1045–1057.

[25] G.M. Yousef, E.P. Diamandis, Human kallikreins: common structural features, sequence analysis and evolution, Curr. Genom. 4 (2003) 147–165.

[26] J. Greer, Comparative modeling methods: application to the family of the mammalian serine proteases, Proteins 7 (1990) 317–334.

[27] J.F. Bazan, R.J. Fletterick, Viral cysteine proteases are homologous to the trypsin-like family of serine proteases: structural and functional implications, Proc. Natl. Acad. Sci. U. S. A. 85 (1988) 7872–7876.