J· Biosci., Vol· 22, Number 5, December 1997, pp 555–565. © Printed in India·

Source and target enzyme signature in serine protease inhibitor active site sequences

BALAJI PRAKASH and M R N MURTHY*
Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India

MS received 30 May 1997; revised 3 October 1997

Abstract. Amino acid sequences of proteinaceous proteinase inhibitors have been extensively analysed for deriving information regarding the molecular evolution and functional relationship of these proteins. These sequences have been grouped into several well defined families. It was found that the phylogeny constructed with the sequences corresponding to the exposed loop responsible for inhibition has several branches that resemble those obtained from comparisons using the entire sequence. The major branches of the unrooted tree corresponded to the families to which the inhibitors belonged. Further branching is related to the enzyme specificity of the inhibitor. Examination of the active site loop sequences of trypsin inhibitors revealed that here are strong preferences for specific amino acids at different positions of the loop. These preferences are inhibitor class specific. Inhibitors active against more than one enzyme occur within a class and confirm to class specific sequence in their loops. Hence, only a few positions in the loop seem to determine the specificity. The ability to inhibit the same enzyme by inhibitors that belong to different classes appears to be a result of convergent evolution.

Keywords. Protease inhibitors; active site; sequence signature; evolution; convergence.

1. Introduction

Proteinase inhibitors are ubiquitous in nature and are important for regulating proteolytic activity of their target enzymes. The number of proteinase inhibitors isolated and identified so far is extremely large and hence form a good system to study aspects of molecular evolution and structure function relationships. The majority of inhibitors known and characterized so far are directed towards serine proteinases·

Protein crystallography has played a major role in elucidating some unique features that exist in the structures of proteinase inhibitors and their complexes with proteolytic enzymes that distinguish these inhibitors from ordinary substrates. The major common element in the structures of the inhibitors of serine proteinases is the primary contact region, or the reactive site loops. In all the inhibitors whose structures have been determined, these reactive site loops are exposed from the body of the inhibitors and are accessible to the active sites of proteolytic enzymes. These loops adopt a conformation that is complementary to the surface of the enzymes and resemble the conformation of a substrate bound to the active site. The exposed binding loops of the different families display a range of conformations. In all these inhibitors, the residue towards the amino-terminal side of the scissile bond (P1) determines the specificity of inhibition. Several discussions on the geometry of this loop and the conformation of the residues around the scissile bond are discussed·in the reviews by Bode and Huber (1992) and Read and James (1986).

*Corresponding author (Fax, 334 1683; Email, mrn@mbu.iisc.ernet.in).

Laskowski and Kato (1980) have classified the serine protease inhibitors into ten major classes. The criteria for classification was sequence homology and the topological relationships of the disulphide bridges. Apart from this major classification, there have been studies where the members of a single class have been grouped into different sub classes· For example, Norioka and Ikenaka (1983) have constructed a phylogenetic tree of legume double headed Bowman-Birk inhibitors (BBI) from the amino acid sequences and classified the double headed inhibitors into four groups. Recently we have classified the BBI into two major classes, those from monocotyledonous seeds and dicotyledonous seeds (Prakash *et al* 1996). The further classification of dicot inhibitors is consistent with that of Norioka and Ikenaka (1983).

In this paper, we examine the extent to which the reactive site loops of protease inhibitors reflect the respective branch position in the phylogeny obtained by conventional methods of whole length sequence comparisons. Surprisingly, branching obtained by using only a twelve residue stretch around the inhibitory loop preserves to a large extent the scheme of Laskowski and Kato (1980). Further we discuss the characteristics of the residues in the loop for a few families·

2. Methods

2.1 *Sequences*

The serine protease inhibitor sequences were obtained from the SWISSPROT databank, release June 1994, available at the Bioinformatics Centre of the Indian Institute of Science, Bangalore. There were 226 sequences in all. A twelve residue stretch around the scissile bond from each of these sequences were used in the present analysis· However, the loops that have been reported to inhibit more than one enzyme were omitted. In this way, sets of loop sequences that were specific to trypsin, chymotrypsin, subtilisin and elastase were obtained. The number of sequences are listed in table 1 according to their specificities· In total, there were 125 sequences of length 12 that were analysed.

2.2 *Multiple sequence alignment*

The program MULTALIN (Corpet 1988) was used to force alignment of active site sequences· In order to avoid all gaps that might be inserted within the active site sequences, the gap penalty was increased to 50, scoring alignments using Dayhoff's (1978) log odds matrix· The similarity scores obtained were then used to determine the hierarchical order of clustering sequences· The alignment of sequences that had the highest score was initially accepted and the aligned pair was treated as a single sequence· Each further step combined either two sequences or clusters or a sequence and a cluster. The similarity measure was reevaluated after each combination. This procedure was continued until all the sequences were merged. After completing multiple alignment, pairwise scores were reevaluated. A binary dendogram was derived on the basis of the order in which the sequences were merged. A simplified representation of the phylogenetic tree was drawn by hand, in order to highlight the branching pattern· The lengths of the arms are arbitrary.

Table 1. List of sequences used.

Code in text	Swiss-Prot Code	Code in text	Swiss-Prot Code
Trypsin inhibitory loops			
BPT7	BPT2_BOVIN	SSI2	SSI2_STRLO
BPT10	HT1A_MANSE	STI8	ID5A_ADEPA
BPT11	HT1B_MANSE	ST19	ID5A_PROJU
BPT12B	IATR_BOVIN	STI12	IDE3_ERYCA
BPT13B	IATR_HORSE	STI13	IDE3_ERYLA
BPT14B	IATR_PIG	STI15	IT1A_PSOTE
BPT15B	IATR_SHEEP	STI16	IT1B_PSOTE
BPT17	IBPC_BOVIN	STI17	IT12_PSOTE
BPT18	IBPS_BOVIN	STI20	ITRA_SOYBN
BPT19	IBP_TURRS	STI22	ITRB_SOYBN
BPT24	ISHP_STOHE	BBI1	IBB1_WHEAT
BPT26	ITR4_RADMA	BBI2	IBB_HORVU (N-TERMINAL 8 K)
BPT30	IVB2_HEMHA	BBI3	IBB_ORYSA (C-TERMINAL 8 K)
BPT31	IVB2_NAJNI	BBI4	IBB2_SETIT
BPT32	IVB2_VIPRU	BBI5	IBB3_SETIT
BPT42	IVBT_NAJNA	BBI6	IBB1_COILA
KZ8A	IOV7_CHICK	BBI7	IBB2_COILA
KZ8B	IOV7_CHICK	BBI8	IBB_HORVU (C-TERMINAL 8 K)
KZ8C	IOV7_CHICK	BBI9	IBB_ORYSA (N-TERMINAL 8 K)
KZ8D	IOV7_CHICK	BBI10	IBB2_WHEAT
KZ24B	IOVO_CHICK	BBI11A	IBB1_ARAHY
KZ30A	IOVO_COTJA	BBI11B	IBB1_ARAHY
KZ54A	IOVO_MELGA	BBI12A	IBB2_ARAHY
KZ73	IPS1_RAT	BBI12B	IBB2_ARAHY
KZ76A	IPSG_FELCA	BBI13A	IBB1_PHAAN
KZ77A	IPSG_MELME	BBI14A	IBB_PHALU
KZ78A	IPSG_PANLE	BBI15A	IBB3_DOLAX
KZ79A	IPSG_VULVU	BBI16A	HGI
KZ81	IPST_BOVIN	BBI17A	IBB4_LONCA
KZ82	IPST_CANFA	BBI18A	IBB4_DOLAX
KZ83	IPST_HUMAN	BBI19A	IBB2_PHAAN
KZ84	IPST_PIG	BBI19B	IBB2_PHAAN
KZ85	IPST_SHEEP	BBI20B	IBB2_PHAVU
KZ86	PSG1_MOUSE	BBI22A	IBB3_SOYBN
SER1	A1AT_HORSE	BBI22B	IBB3_SOYBN
SER10	A1AT_BOMMO	BBI23A	IBB_PHAVU
SER11	A1AT_BOVIN	BBI23B	IBB_PHAVU
SER13	A1AT_CYPCA	BBI24A	IBB1_SOYBN
SER16	A1AT_MACEU	BBI25A	IBB_VICAN
SER43	COTR_CAVPO	BBI26A	IBB_VICFA
SER44	COTR_MOUSE	BBI27A	IBB_MEDSA
SQ2	ITI1_LAGLE	BBI27B	IBB_MEDSA
SQ3	ITR1_CITVU	SQ14	ITR3_CUCPE
SQ6	ITR1_MOMCH	SQ16	ITR4_CUCMA
SQ13	ITR3_CUCMC	SQ18	ITR4_LUFCY

Table 1. (Continued)

Code in text	Swiss-Prot Code	Code in text	Swiss-Prot Code
Chymotrypsin inhibitory loops			
BPT20	ICS3_BOMMO	STI14	IECI_ERYVA
BPT23	ISC1_BOMMO	BBI13B	IBB1_PHAAN
BPT28	IVB1_BUNFA	BBI14B	IBB_PHALU
BPT33	IVB3_VIPAA	BBI15B	IBB3_DOLAX
BPT35	IVBC_NAJNA	BBI16B	HGI_
KZ8E	IOV7_CHICK	BBI17B	IBB4_LONCA
KZ8G	IOV7_CHICK	BBI18B	IBB4_DOLAX
SER22	A2AP_BOVIN	BBI24B	IBB1_SOYBN
SER23B	A2AP_HUMAN	BBI25B	IBB_VICAN
SER25	ACH1_BOMMO	BBI26B	IBB_VICFA
SER26	ACH2_BOMMO		
Elastase inhibitory loops			
KZ75B	1PSG_CANFA	SER58	ILE_HUMAN
KZ76B	1PSG_FELLA	SER59	ILE_PIG
KZ77B	1PSG_MELME	SER79	SERA_MANSE
KZ78B	1PSG_PANLE	BBI20A	IBB2_PHAVU
KZ79B	1PSG_VULVU	BBI21A	IBB2_SOYBN
Subtilisin inhibitory loops			
SSI1	SSI1_STRCI	SSI9	SSI1_STRGI
SSI7	SSI1_STRAO	PCI5	ICT1_PHAAN

3. Results

3·1 *Choice of sequences*

The SWISSPROT databank had a large number of amino acid sequences belonging to the serine protease inhibitor family. As extensively discussed in literature, the importance of the reactive loop in conferring specificity to the inhibitor against a target enzyme relies on the presence of a potential P1 residue and the residues surrounding it. We proposed to investigate if the information present in the loop sequences is sufficient to segregate sequences on the basis of their enzyme specificities. Hence, the inhibitory loops that were directed against more than one enzyme were eliminated from the analysis· This resulted in 125 unique sequences· Among these 90, 21, 10, and 4 were directed against trypsin, chymotrypsin, subtilisin and elastase, respectively.

3·2 *Multiple alignment*

The selected sequences were aligned using MULTALIN.Most trypsin inhibitors had K/R at the P1 site, while the chymotrypsin inhibitors had L/Y/W/F and elastase inhibitors had 'Ala'· The dendogram obtained showed several distinct groups.In order to highlight these groups, a binary tree was drawn manually. Figure 1 shows the various branches that correspond to members belonging to a particular inhibitor family. As seen in the figure, the BBI**,** Kazal inhibitors, BPTI-Kunitz, Serpins,

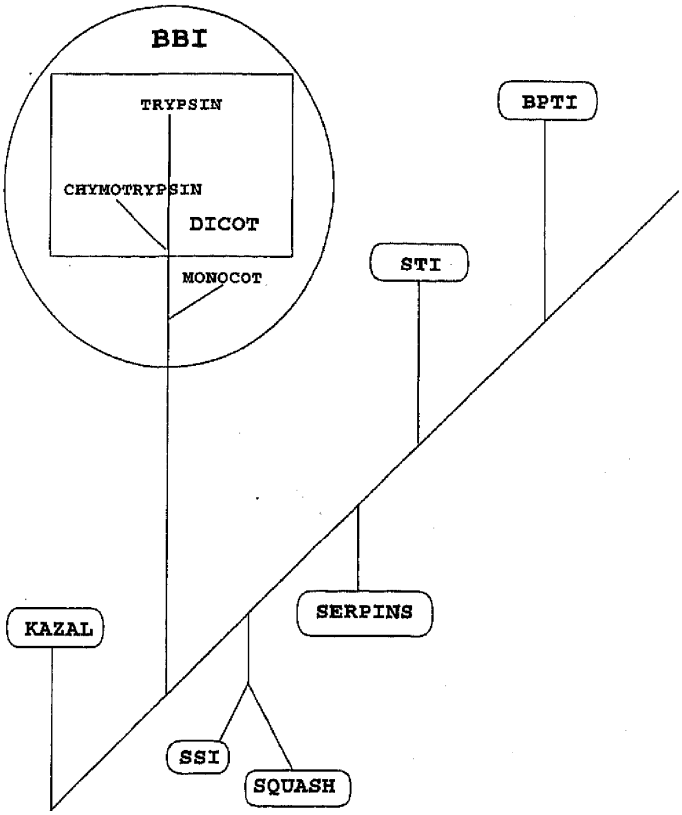

Figure 1. Binary tree constructed for active site loop sequences of serine protease inhibitors. 121 of 125 loop sequences examined confirmed to this pattern. The further branching patterns, in some of the branches with large number of sequences, of this figure are illustrated in figure 2.

STI-Kunitz, Streptomyces subtilisin and squash seed inhibitors form separate branches of the tree, although occasionally members from one class are groupted with members of the other. A closer examination of each of the branches shows further interesting branching patterns as follows.

3.2a *Kazal family:* In the Kazal inhibitors, most of the trypsin inhibitors are grouped as members of one branch (figure 2a). The inhibitors against elastase and chymotrypsin are fewer in number and hence the sub grouping of these inhibitors are not as significant as those of trypsin inhibitors. The members Kz75b, Kz79b, Kz78b, Kz77e, Kz76b (elastase inhibitors), Kz8g (chymotrypsin inhibitor), and Kz30a (trypsin inhibitor) cluster together· None of the members of Kazal family fall into the cluster corresponding to other families of inhibitors.

3.2b *Squash seed and Streptomyces subtilisin inhibitor families:* The number of squash inhibitors used for the study after initial elimination were the following seven: Sq2, Sq3, Sq6, Sql3, Sql4, Sql6, Sql8. The Streptomyces subtilisin inhibitors (SSI) were SSI1,

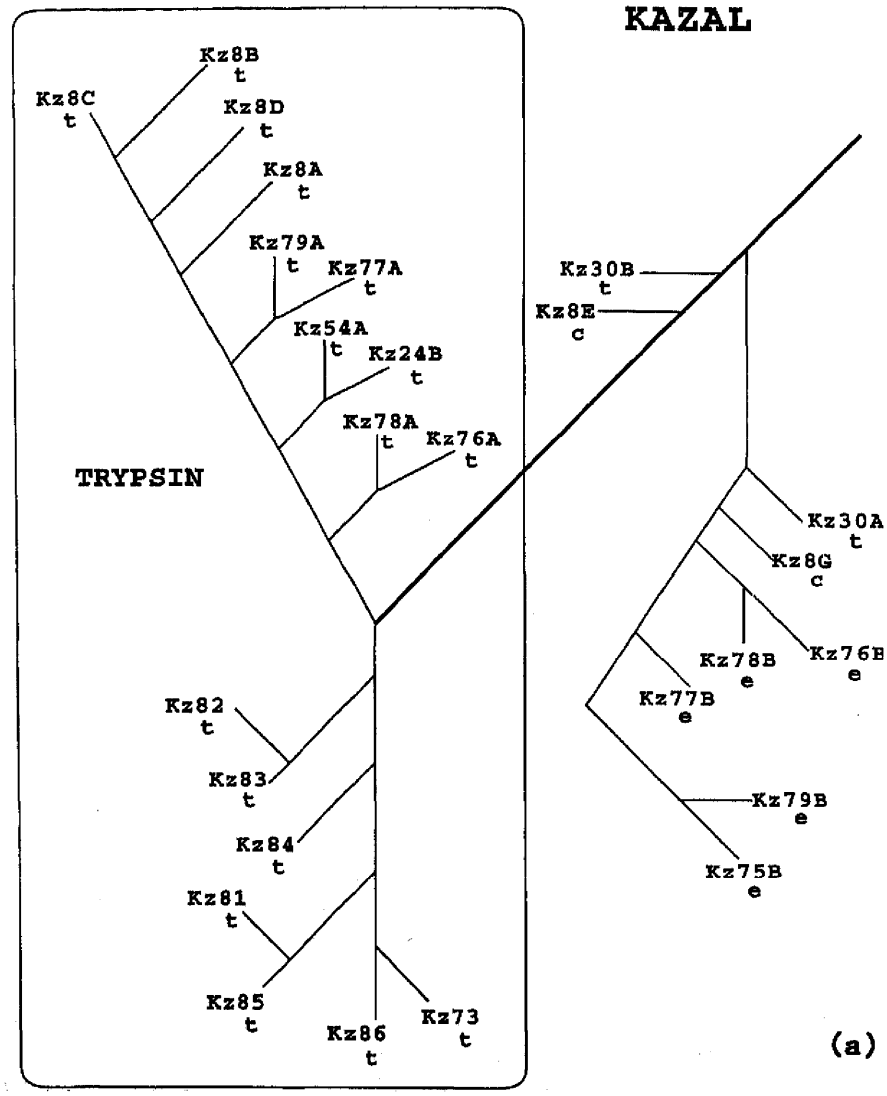

Figure 2a.

SSI2, SSI7 and SSI9. Surprisingly, these inhibitors display similarity in the loop and luster together as shown in figure 2b. All the squash seed inhibitors are specific to trypsin and group together. Three of the SSI's are directed towards subtilisin, while SSI2 is against trypsin. These inhibitors are grouped together.

3.2c *BBI:* Sequences corresponding to 33 active site loops of BBIs are available in the data base· The loops from monocots and dicots clearly segregate (figure 2c). Furthermore, among the dicot inhibitors, there is a clear demarcation of the chymotrypsin inhibitors from the trypsin inhibitors. An inspection of the BBI branch clearly shows that the first separation is based on the plant class and then the specificity of the

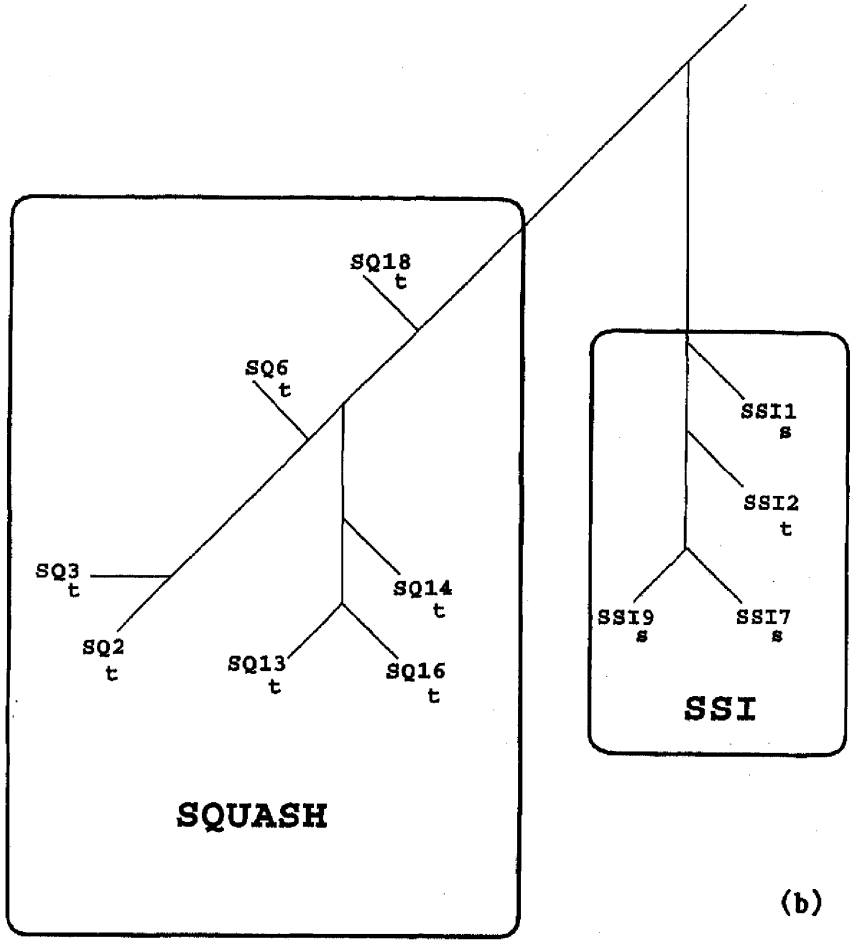

Figure 2b

inhibitor. Of the 33 members that have been used, only BBI9 (N-terminal of the 16 K BBI from rice) cluster with the serpins. We have earlier shown that this loop is different from those of the other BBIs, in terms of the activity and the number of cysteine residues (Prakash *et al* 1996). The members BBI1la, BBI12a and BBI17a which are from dicots do not fall on the corresponding branch although they cluster along with other BBIs.

3.2d *BPTI-Kunitz family:* There are 21 members in this family that have been used for analysis. The inhibitors from snake venom, with the exception of Bpt35, form a separate sub branch (figure 2d). Although a sub branch of this family contains inhibitors specific to trypsin, there is no discernible segregation of the enzyme specificities in the other two sub branches.

3.2e *STI-Kunitz:* The ten STI-Kunitz inhibitors are grouped together. Of these, nine are specific to trypsin and one to chymotrypsin. Three inhibitors belonging to serpin family also occur on this branch.

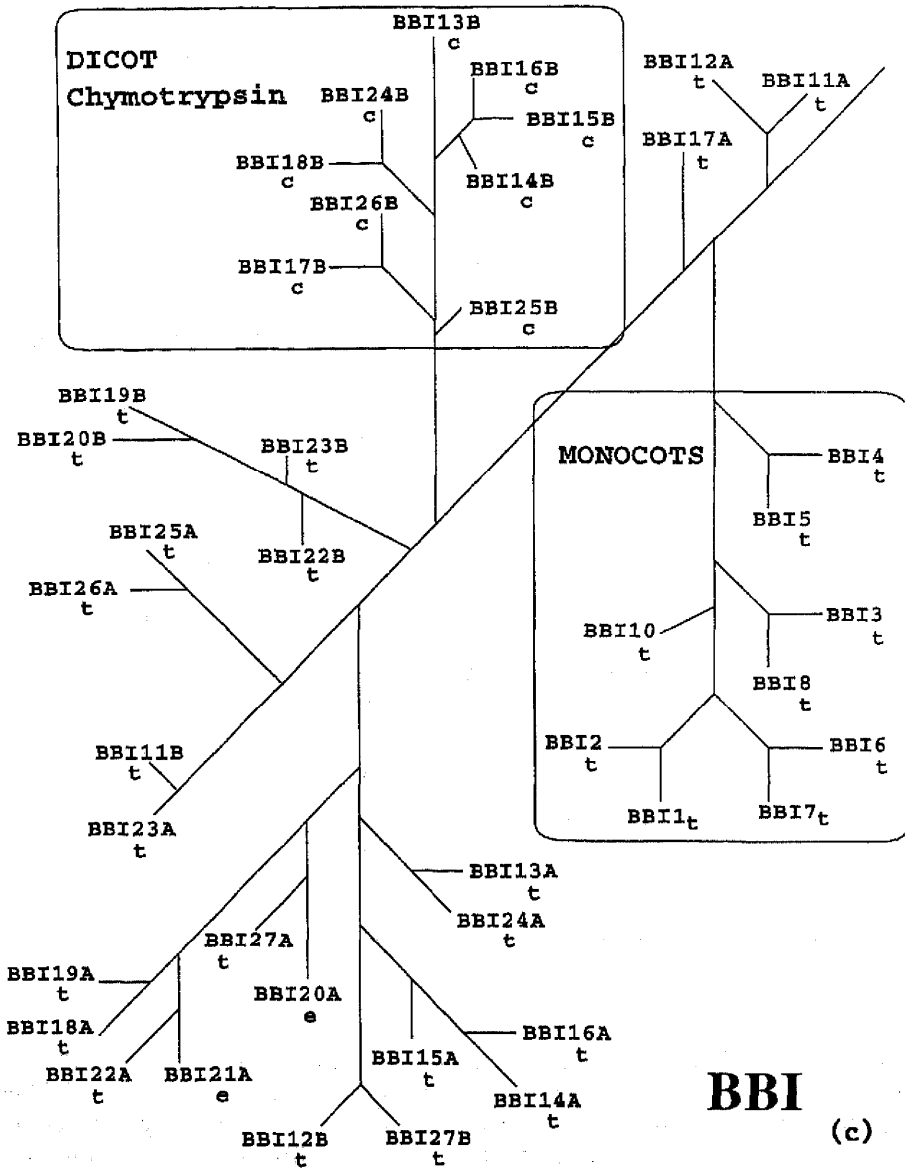

Figure 2c

3.2f *Serpins:* The number of sequences in this class are relatively less. This has 6 trypsin and 3 chymotrypsin and 1 elastase inhibitor loops. However, the sequences of Ser26(CHT), Ser58(Ela), Ser59(Ela) which also belong to this class cluster close to STI-Kunitz inhibitors.

This interesting branching pattern prompted us to analyse all the sequences, without the elimination of sequences active against more than one enzyme. The results obtained from this analysis (not shown as the number of sequences are too large) also reveals

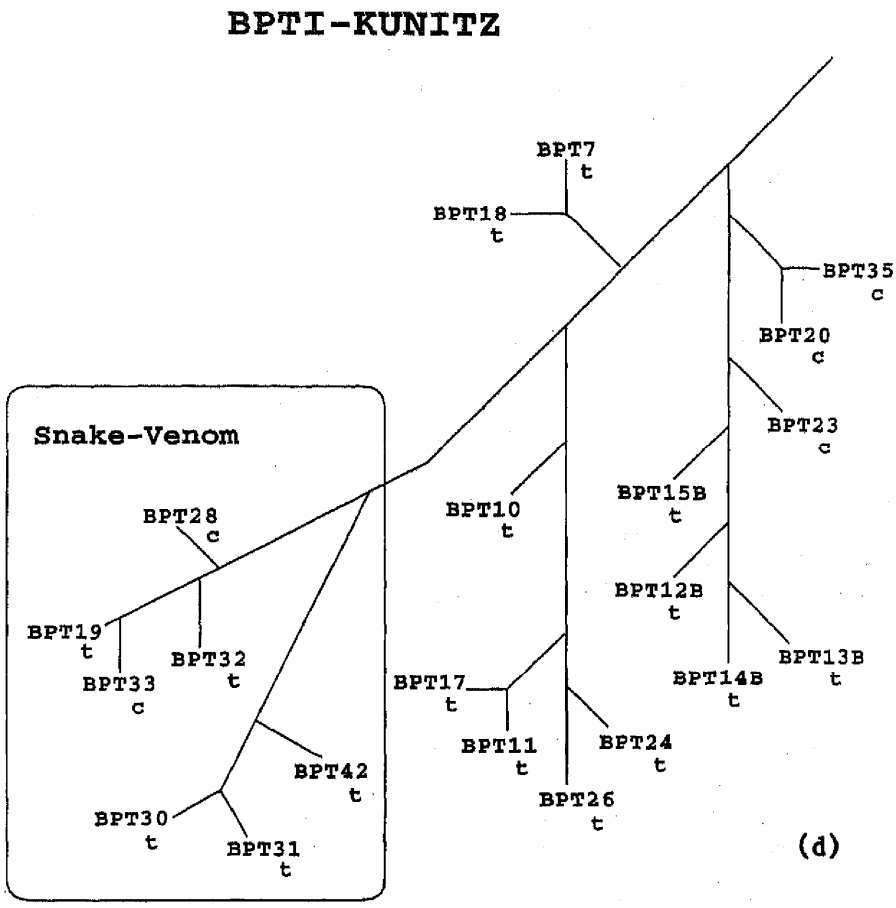

Figure 2. Binary tree representing the loop sequence relationship of inhibitors occurring in **(a)** KAZAL, **(b)** SSI-SQUASH, **(c)** BBI, and **(d)** BPTI branches of figure 1. The inhibitors are referred by their codes as defined in table 1. Small letters associated with the code represents enzyme specificity, t, Trypsin; c, chymotrypsin; e, elastase; s, subtilisin.

branching very similar to that presented here. Some of the interesting features include clear segregation of the wound induced inhibitors (PCI family), snake venom proteins (BPTI family), and ovomucoids (Kazal). A closer examination shows further grouping of the inhibitors based on enzyme specificities with only occasional violation.

3.3 *Analysis of positional preference of amino acids*

In order to understand the chemical characteristics of the inhibitory loops, the following analysis was performed. To begin with, the branches of the Bowman-Birk family were examined. There were three branches. One was that of the trypsin specific monocot inhibitors (BBIMT), second was that of the chymotrypsin specific dicot inhibitors (BBIDC) and the third was that of the trypsin specific dicot inhibitors

(BBIDT). In all the three branches, a Cys residue at P3 and P6' positions, a Pro at P3', a Ser at P 1', are found. In BBIDT and BBIDC, both of which are dicot inhibitors, the P5' residue is polar, while BBIMT has a hydrophobic residue at this position. Similarly, conserved Cys at P5 of dicot inhibitors is usually substituted by an Ala in the monocot inhibitors BBIMT. In the trypsin specific groups, BBIDT and BBIMT, a Pro is preferred at the P4' position. This residue is mostly Ala in the chymotrypsin specific BBIDC class. Similarly, P6 is conserved (Ser) in BBIDC but variable in BBIDT and BBIMT.These observations suggested that there could be a pattern in the preference of amino acids at various positions along the loop. We chose to analyse trypsin specific inhibitors from various families. These were dicot Bowman-Birk trypsin inhibitors (BBIDT), monocot Bowman-Birk trypsin inhibitors (BBIMT), Kazal trypsin inhibittors (KZT), squash seed trypsin inhibitors (SQT), STI-Kunitz trypsin inhibitors (STIT). The number of sequences in each of these classes were 24, 9, 17, 12 and 11, respectively. At each position, the percentage of polar, hydrophobic and charged amino acid residues were calculated. These varied from class to class suggesting inhibitor class specific residue occurrences. These results are summarized in table 2, where capital letter stands for strong preference (occurrence > 70%), small letter stands for weak preference (occurrence between 50 and 70%) and H, P, C +, C — represent hydrophobic, polar, positively charged and negatively charged residues, respectively·

4. Discussion

The sequence analysis using only a twelve residue stretch around the reactive site loop of the serine protease inhibitors is presented in this paper. Although there are statistically robust algorithms to evaluate a phylogenetic tree, due to the large number of inhibitor sequences to be analysed, we have limited ourselves to the solutions obtained from the MULTALIN program, where the dendogram obtained clearly gave indications of branching that were significant in terms of inhibitor class and specificity (figure 1). Surprisingly, the dendogram thus obtained preserves the classification of Laskowski and Kato (1980) to a large extent. This loop not only distinguishes between the various families (table 2) but also leads to further branches based on enzyme specificity (figure 2). This feature becomes clear in the Bowman-Birk family, where the number of trypsin and chymotrypsin specific inhibitors are reasonably large. This branching pattern is also consistent with our earlier results (Prakash *et al* 1996).

Table 2. Patterns of residue conservation in trypsin inhibitors: Most frequently occurring residue at positions on either side of scissile bond in different inhibitor families·

	P6	P5	P4	P3	P2	P1	P1′	P2′	P3′	P4′	P5′	P6′
BBIDT	p	P	H	P	P		P	H	H	H	P	P
BBIMT		H	h	P	P		P	H	H	H	H	P
KZT	H	H	H	P	h		H	h		H	H	P
SQT		C+	h	P	H		H	H	H		P	C+
STIT	H	p	P		H		p		H	H	H	C−

H, Hydrophobic; C +, positively charged; P, polar; C —, negatively charged.

The branching of BPTI-Kunitz, STI-Kunitz, Serpins, Bowman-Birk, Kazal, SSI, and squash seed inhibitor families was distinct, except for only four sequences, out of a total of 125 sequences. Trypsin specific inhibitors are further clearly segregated in the Kazal, squash seed, BBI, and STI-Kunitz families.In BPTI-Kunitz family, one branch contains inhibitors from venoms, while the rest have other inhibitors. In BBI's the separation between the inhibitors of monocot and dicots is the primary branch, which is followed by the branching of chymotrypsin inhibitors from the trypsin inhibitors in dicots. It is interesting that this kind of branching can be achieved using just a twelve residue stretch. In a separate analysis, it was also found that this branching pattern is valid for all the sequences including those that inhibit more than one enzyme.

The analysis of the sequences of trypsin specific inhibitors from different families, to understand the basis for this segregation, is shown in table 2. The table suggests distinct signature for each family, which is a combination of its preference for either a polar, charged or hydrophobic residue at each position along the loop. It is likely that only a subset of these 12 residues determine the enzyme specificity while the other residues characterize the inhibitor family and hence related to the evolutionary origin of the inhibitor species. Further, the analysis suggests that the enzyme specificities have arisen independently in each inhibitor species by convergent evolution.

Acknowledgements

We thank the Department of Science and Technology, New Delhi for financial support, the staff of Bioinformatics, Indian Institute of Science, Bangalore for help in obtaining the sequences and Dr H S Savithri, Dr S Selvaraj, Dr Lalitha R Gowda, and Mr Y N Sreerama for useful discussions.

References

Bode W and Huber R 1992 Natural protein proteinase inhibitors and their interactions with proteinases; *Eur· J· Biochem·* **204** 433-451

Corpet F 1988 Multiple sequence alignment with hierarchical clustering; *Nucleic Acids Res*. **16** 10881–10890

Dayhoff D 1878 A model of evolutionary change in proteins: Matrices for detecting distant relationships; in *Atlas of protein sequence and structure,* 5 Suppl 3 (Washington DC: National Biomed. Res. Foun.)

Laskowski M Jr and Kato 11980 Protein inhibitors of proteinases; *Annu. Rev. Biochem·* **49** 593-626

Norioka S and Ikenaka T 1983 Amino acid sequences of trypsin chymotrypsin inhibitors (A-I, A-II, B-I and B-II) from peanut *(Arachis hypogaea):* A discussion on molecular evolution of legume Bowman-Birk type inhibitors; *J· Biochem·* **94** 589-599

Prakash B, Selvaraj S, Murthy M R N, Sreerama Y N, Rama Sarma P R, Rajagopal Rao D and Gowda L R1996 Analysis of the amino acid sequences of plant Bowman-Birk inhibitors; *J. Mol Evol*. **42** 560-569

Read R J and James M N G 1986 Introduction to proteinase inhibitors- X-ray crystallography; in *Proteinase inhibitors* (eds) A J Barret and G Salvensen (Amsterdam: Elsevier) pp 301-336

Corresponding editor: M S SHAILA