The Table of Interface Forming Residues as the Specificity Indicator for Serine Proteases Bound to Different Inhibitors

R.C. Togawa[1], C. Ribeiro[2,3], I. Mazoni[4], T. Pelligrinelli[4], J.G. Jardine[4] and G. Neshich[4]
[1] Embrapa Genetic Resources and Biotechnology, Brasilia, DF, Brazil;
[2] Department of Biochemistry and Immunology, Univ.Federal de Minas Gerais, Belo Horizonte, MG, Brazil;
[3] Department of Computer Science, Univ. Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[4] Embrapa Information Technologies, Campinas, SP, Brazil;

Abstract - *We propose a novel method for defining the exclusive and exhaustive table of serine proteases specificity determining interface forming residues (IFR). The IFR are obtained by "hard body docking" among 73 structurally aligned, sequence wise non redundant, serine protease structures, with 3 inhibitors: ecotine, ovomucoid third domain inhibitor and basic pancreatic trypsin inhibitor. In silico constructed complexes offered a condition for determining which residues are participating, from both enzyme and inhibitor side, in the ensemble of amino acids that upon binding loose contact with a solvent. Our focus is on offering a thorough study on how the specificity is achieved among serine proteases even though they have very little difference in their tertiary structure (specifically if the position of catalytic triad residues is considered). Presented table of serine protease specificity based on IFR position occupation show clear variations among sub families such as: trypsines, chymotrypsines, elastases and thrombines.*

Keywords: enzyme specificity, hard body docking, interface forming residues, serine proteases.

1 Introduction

Serine proteases play an important role in processes such as blood clotting, digestion and in some pathways of cell development. Serine proteases can hydrolyze either peptide bonds or esters. Proteases digest proteins by hydrolyzing the peptide bonds which are responsible for keeping together amino acids in any polypeptide. The cleavage specificity of elastase, trypsin, chymotrypsin and other serine proteases depends in part on the volume/size, form/shape, and polarity/charge/hydrophobicity of the part of the protein surface where a substrate will be docking – the specificity pocket. There are three amino acid residues responsible for enzymatic activity, present in all serine proteases, and which are denominated as catalytic triad: His 57, Asp 102 and Ser 195. The role of the former two amino acids during trypsin catalysis, for example, is to function as a proton shuttle. Among most known serine proteases are elastases which cleave peptides after small residues. Trypsin cleaves peptides after Lys and Arg residues with co-participation of the Asp 189, which is interacting with the positive charge on peptide. Chymotrypsin on the other hand, cleaves proteins after aromatic (and also large hydrophobic) residues. To achieve such specificity, one can easily identify in chymotrypsin the existence of the hydrophobic pocket, normally shielded by Met 192. Thrombin is the protease which cleaves peptides with more specificity than trypsin: it requires Arg on "P1" position, [1]. Our main objective is to expand the scope of the interest while considering enzyme specificity, broadening it from being focused only on the catalytic triad and binding pocket, to the wider category of amino acids which we have named: the Interface Forming Residues (IFR). Namely, a part of the molecular surface is shielded from a solvent upon complex formation - when one molecule docks to the other. Those residues which lost access to a solvent have an important role in the process of docking and also in defining the specificity [2] [3]. Therefore, we were motivated to first calculate which amino acids are becoming shielded from a solvent upon complex formation among serine proteases and different types of inhibitors, and then create the table (matrix) of all amino acid positions at the interface and their respective occupancy. By mapping those amino acids, we are now able to analyze their characteristics and by doing so we also can make position specific alignment among different sub families of serine proteases. The key step we needed to solve during our procedural approach was to find sufficient number of PDB [4] structures containing complexes of serine proteases with respective inhibitors. It became clear very rapidly that we would need to either produce those by some novel method or abandon our work as there were not enough samples in the PDB. The solution to this challenge is presented in details in materials and methods.

Our work here is providing a unique tool for both structure/function relationship analysis, as well as concrete indications on how specificity of various serine proteases is achieved or could be altered.

Based on results we are reporting here, we are poised to assemble in the near future the serine protease super family interface data resource as an expanding collection of sequence, structural, and functional information about the serine proteases interface forming residues around active site. A combination of graphics images and numerical data will be

used to aid complete analysis of structure/function relationship.

2 Material and Methods

In order to provide sufficient volume of data for analysis of the interfaces around active sites of serine proteases (even if there is no available information on complex formation of one particular protease with any specific substrate and/or inhibitor) we needed to employ the in silico approach for building corresponding complexes. The key feature of our work is mapping the IFR 3D profile into 2D matrix, from known enzyme-inhibitor structure to those with no known structure for such complex. Mapping is done after structurally aligning all serine proteases with non-redundant sequences.

The first step in our procedure was to select serine proteases from the SCOP database and then subsequently eliminate those which were above the threshold for sequence wise similarity (established to be less than 97%, which means that there should be at least 6 to 8 residues different among selected sequences, usually 240 to 260 amino acids long). We ended up with only 73 structures which covered the following sub families: thrombin, trypsin, chymotrypsin and elastase. In this data set are also included representatives of subfamily of zymogens, chymotrypsinogen, serine proteases from prokaryotes and a category we called "various", containing blood coagulation factor IXa, Xa, D and kallikrein.

The list of all corresponding PDB files for serine proteases used in this work is presented below in the Table I:

Table I: list of all PDB files containing serine protease structures used for structural alignment.

01-1slu	02-1tab	03-1ntp	04-1mct	05-1trn
06-2trm	07-1bra	08-1bit	09-2tbs	10-1try
11-5gch	12-1gcd	13-1acb	14-1chg	15-1eq9
16-2cga	17-1cgi	18-1cgj	19-1pyt	20-1fon
21-1tgc	22-2tgd	23-1tbr	24-1ett	25-1hrt
26-1etr	27-1c5l	28-1ets	29-1ucy	30-1hdt
31-1ppb	32-2hnt	33-1bmm	34-1tmu	35-1aht
36-1bmn	37-1tmt	38-1hxe	39-1lhe	40-1hap
41-1hao	42-1hlt	43-1hbt	44-1dwb	45-1dit
46-1iht	47-1ihs	48-1ppf	49-1hne	50-1elg
51-1elt	52-1hyl	53-1hcg	54-1dst	55-1dsu
56-3rp2	57-1pfx	58-2pka	59-1sgt	60-1fuj
61-1lmw	62-1rtf	63-1ton	64-1arb	65-2sfa
66-3sgb	67-1hpg	68-2sga	69-1p06	70-1p10
71-1gbl	72-1gbe	73-1gbd		

In order to obtain IFR ensemble for those serine proteases for which there was no available structure for the complex with chosen inhibitor (or any inhibitor at all), we first aligned structurally all 73 selected serine proteases using PrISM software [5]. Out of 73 structures, we found that 37 contain the corresponding inhibitors, however, we first decided to test only the 3 most common of them: ovomucoid 3rd domain as in 1ppf.pdb (inhibiting human leukocyte elastases), basic pancreatic trypsin inhibitor – BPTI, as in 1mtn.pdb (inhibiting alpha-chymotrypsin) and ecotin as in 1slu.pdb (inhibiting rat trypsin). By having the structure of the complex formed between a serine protease and one of those three inhibitors, and at the same time, having the multiple structure alignment of 73 serine proteases, we are able to do "hard body docking" of any of the three inhibitors to any of 72 selected serine proteases (the 73-rd complex structure already exists and is available at the PDB). The details of the above described procedure are shown in the figure 1, 2, 3 and 4. The term "hard body docking" is used here to describe what we have done: simple appending of the inhibitor coordinates to each PDB file containing a serine protease structure; Inhibitor coordinates, to be appended to aligned serine proteases PDB file, were taken from the known complex structure and then realigned to the position of that same serine protease in the structure aligned file. Upon completing that procedure, all complex structures could be described as serine proteases spatially oriented in a consistent and unique manner (obeying structure alignment) and also with the inhibitors docked onto binding site in a unique way: following the coordinates of the inhibitor in the PDB entry containing the "real" complex. As the purpose here was to define an interface, a precision as obtained by minimization and dynamics procedures, was not critical and therefore not applied.

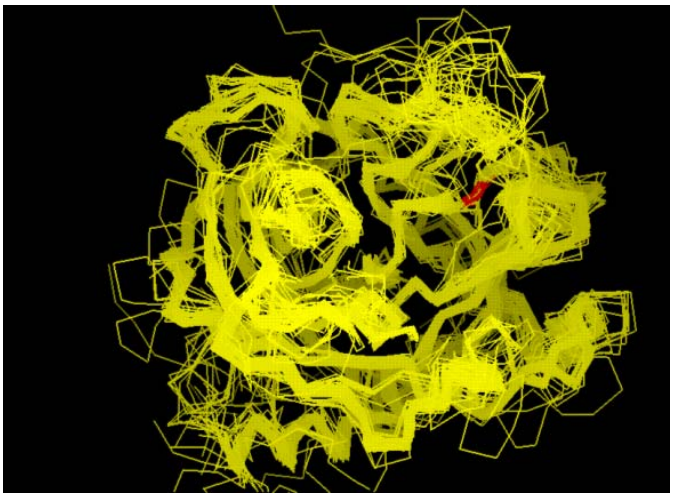

Fig. 1. : Structural alignment of 73 different sequence wise non redundant serine proteases, aligned by PrISM package. The image was produced by INSIGHT II package. Only the main chain is represented. In red, the position of the Ser 195 is depicted.

The three sets of 73 enzyme-inhibitor complex structures were then studied with respect to solvent accessibility of surface amino acids before and upon inhibitor binding. The calculation of the change of solvent accessibility for surface amino acids for both inhibitor and enzyme (in isolation and upon binding) was done by using SurfV software [6]. Residues with identified change in accessibility to a solvent were compiled into the ensemble of IFR.

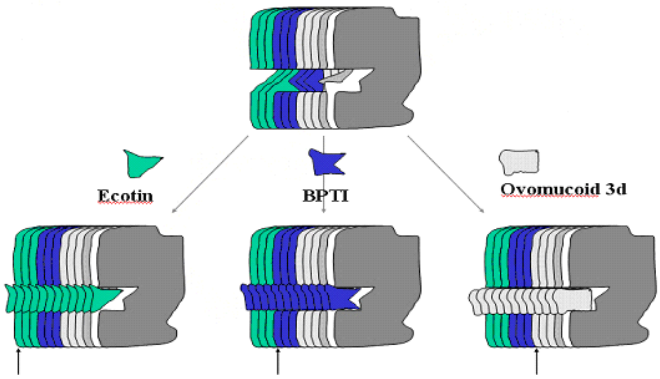

Fig. 2. : Schematic diagram of our experiment for obtaining the structure of three different sets of 73 serine proteases with either one of three selected inhibitors: Ecotine, BPTI and Ovomucoid third domain. From each obtained set the IFR ensemble was extracted and then analyzed. Note that many complexes are in fact "forced" to be formed in silico by "hard body docking" procedure, but due to various factors, including non complementary shape of the inhibitor and the corresponding binding pocket on the enzyme side, they would never be formed naturally. The arrows placed below each of the three sets of 73 structurally aligned serine proteases are indicating that one structure only had coordinates for both serine protease and its corresponding inhibitor (indicated above that particular set). All the other proteases "received" that inhibitor in a position identical to the one found in the known enzyme – inhibitor complex (indicated by the arrow below the structurally aligned complexes).

For each set of 73 complexes, we mapped all the sequence positions that do belong to the IFR ensemble. For the case of ecotine, we compiled 53 residues occupying IFR positions, ranging in the primary sequence from 35 to 224. Respective numbers for BPTI and Ovomucoid third domain (tabular data not shown here due to space limitation) are: 47 from 35 to 226 and 54 from 34 to 227. The aligned residues are listed in rows in the table where the columns are reserved for the each of 73 different serine proteases (Table VI.). Those positions which are not identified as IFR for the particular pair of serine protease and the inhibitor, are presented by "-" sign.

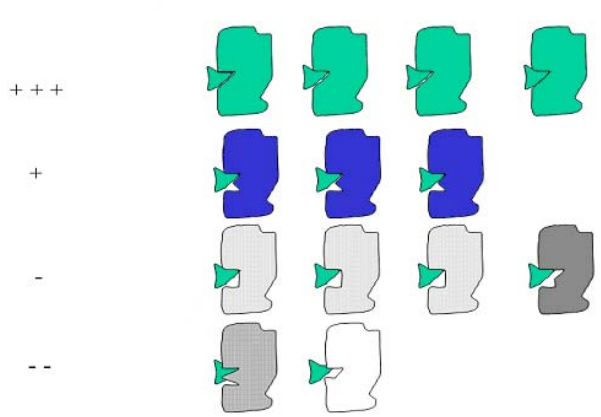

Fig. 3. : The first set of complexes obtained as described in Fig. 1. above, is de-convoluted here in order to demonstrate space compatibility of the Ecotine inhibitor and the binding pocket of 73 different proteases. The "+" and "-" signs are introduced to quantify visual complementarity of the surfaces of the inhibitor and the corresponding binding pocket. Nonetheless, we used those structures not to evaluate binding compatibility but to identify the IFR

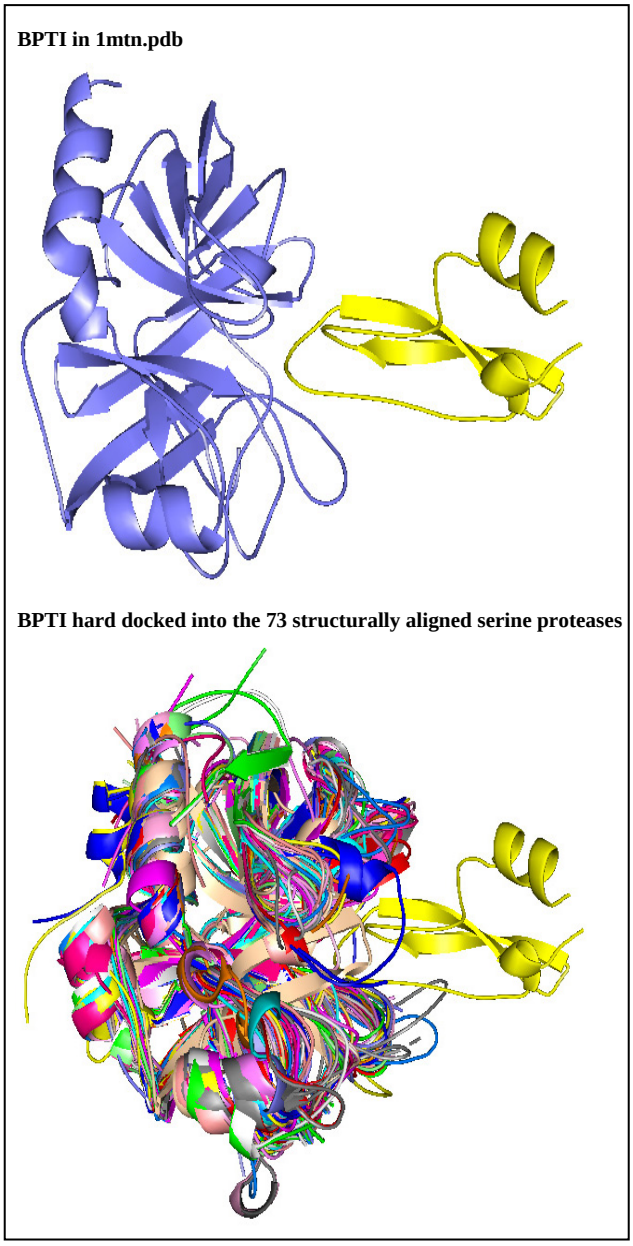

Fig. 4. On the upper side of this figure, the structure of the complex found in 1mtn.pdb is represented and on the lower side is the same BPTI but now "hard body docked" into other 73 structurally aligned serine proteases. One of those 73 enzyme structures is exactly the one presented on the upper side of this figure.

Once we constructed 3 sets of the 73 complexes of proteases with respective inhibitors, we also made some inquiry in relationships among structure characteristics and specificity of enzymes. We calculated the difference in residue type occupancy of the total enzyme (and inhibitor) surface and the respective interfaces. The idea here is to find if the interface is different and if the answer is yes, how it does differ from the rest of the protein surface in terms of type of residues which form them. The results of this inquiry are shown at Figure 5 and are discussed in more details below.

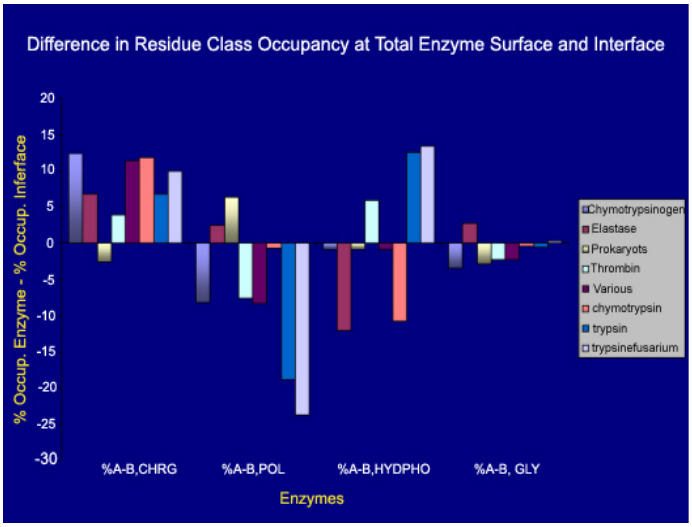

Fig. 5. Using Surfv software, we calculated the total surface area of the enzyme (serine protease) and counted the types of amino acids occupying such surface. By using the same software package, we calculated the area of IFR and counted the type of amino acids occupying this particular surface area. On this figure, we present the difference in percentage occupancy of residues for four different classes (defined as Charged, Polar, Hydrophobic and the special single member sub group – Glycine) between the total enzyme surface and the IFR area only, for all 73 serine proteases bound to inhibitor ecotine. The enzymes are also classified in the following sub-classes: Chymotrypsinogen, Elastase, Prokaryote serine proteases, Thrombin, Chymotrypsin, Trypsin, Trypsin from fusarium and some variety of serine proteases not belonging to any of above defined sub-classes. The average values of percent occupancy are presented for multi member sub-classes of enzyme families.

3 Results and discussion

It has been established that the increased structural plasticity of the binding pocket increases the variation of the substrate size that can fit into the critical space pointing toward the catalytic triad and by this, broadens the specificity [7]. Consequently, it is natural to expect that the higher the stiffness around the binding pocket, the higher the selective pressure on the type of substrate that enzyme may operate on. Specificity then is directly proportional to structural limitations imposed first by the size of the docking space and then by the physical chemical characteristics of this space. By focusing our attention on the type of residues occupying the enzyme and inhibitor interface, we purposely set aside the evaluation of how plasticity of binding pocket influences specificity.

Analysis of the data presented in the Table VI clearly indicate subfamily differences with respect to interface engaged in complex formation with the inhibitor(s). We consider that the data compiled here can be considered as an extremely powerful tool for function deciphering, specially for planning of the structure/sequence changes necessary to be introduced in order to alter activity type for example from trypsin to chymotrypsin, (or more generally: from one subfamily type of the serine protease to the other). Namely, in the past, many attempts have been made in order to alter activity with such final goal [8], but focusing on a single or at most, on several residues. Our data shown on Table VI clearly indicate that such function altering could be accomplished by eliminating key differences in IFR occupancy. In case of Trypsin, chymotrypsin, thrombin and elastases, for example, one can see that the most dramatic differences are present at the following 11 positions (out of 53 IFR positions available):

Table II: The IFR positions identified among serine proteases bound to ecotine. Shown positions are among the most different ones in terms of the type of residues occupying them in 4 classes of serine proteases: trypsins; chymotrypsins; thrombins and elastases, respectively.

Pos	Residue	Pos	Residue
35	"-"; D; R; L or Y	95	D or N; N or "-";N or "-"; N or "-"
37	"-"; "-"; P; "-"or S	97	K or N; L or "-"; K or R; "-" or D
39	Y; F; E; "-"or Y	175	K or Q; "-"or K; R; "-"or T
40	H; "-"or H; L; H or "-"	192	Q; "-"or M; E; F or "-" and
59	Y;G or "-"; "-";C or "-"	217	Y or S or D; "-" or S; E; "-"or S
60	K; "-"; L; A or H or D		

A quick analysis of some positions shows that they may be distinct but not so exclusively populated by a particular residue, such as the position 175 (as opposed to 35, 37, 39 and 40 which are much more restricted). Similar analysis is done (corresponding tabular data as in the Table VI, are not shown here) for other two inhibitors - BPTI: 13 (out of 47 total IFR positions available):

Table III: The IFR positions identified among serine proteases bound to BPTI. Shown positions are designated in a same form as in the Table II above.

Pos	Residue	Pos	Residue
35	"-"; D; R; L or Y	94	Y or F; Y; "-"; Y
37	"-"; "-"or T; P; "-"or S	97	K or N; L or "-"; K or R; V or "-"or D
39	Y or S; F; E; "-"or G	174	"-"; "-"; I; "-"
40	H; "-"or H; L; H or "-"	189	D; S or "-"; D; '-'
59	Y; G or "-"; "-"; "-"	192	Q; "-"or M; E; F or N or "-" and
60	K; "-"; L; A or H	217	Y or S or D; "-'or S; E;"-'or S
73	I or "-"; "- "or Q; R; "-"		

and ovomucoid third domain: 9 (out of 54 total IFR positions available):

Table IV: The IFR positions identified among serine proteases bound to Ovomucoid third domain. Shown positions are designated in a same form as in the Table II and Table III above.

Pos	Residue	Pos	Residue
35	"-"; D; R; L or Y	97	K or N; L or "-"; K or R; V or "-"or D
37	S or "-"; T or "-"; P or "-"; "-"	173	"-"; "-"; R; "-"
39	Y; F; E; "-"or Y	192	Q; '-'or M; E; F or "-" or N and
40	H; "-"or H; L; H or "-"	217	Y or S or D; "-"or S; E; "-" or S
60	K; "-"; L; "-" or H		

Continuing the analysis of the interfaces, we explored how different are total surface and the interface in terms of polarity and hydrophobicity. A buried area and total enzyme surface area were compared with respect to the type of residues occupying it. We used the following residue classes: charged, glycine, hydrophobic and polar. Some interesting differences can be observed for trypsin, chymotrypsin and elastase for example:

Table V: The IFR occupancy by four different amino acid classes for the case of the serine proteases bound to ecotine, expressed as a percentage of that amino acid class present in the total enzyme surface. In addition, the area of the interface with respect to the total protein surface is listed in the first row for three serine proteases sub families: trypsin, chymotrypsin and elastase.

%	Trypsi	Chymotrypsin	Elastase
Total area buried	7	6	7
Charged area buried	16	5	41
Glycine area buried	9	44	3
Hydrophobic area buried	12	6	38
Polar area buried	63	44	19

Although all three serine protease subfamilies show virtually identical percentage of the enzyme total surface area being buried upon in silico complex formation with the ecotine, the percentage of the residue classes being buried at the interface varies significantly. A composite of buried charged and polar area in trypsin is 80%, in chymotrypsin is ~50% while for elastase it amounts to 60%. High percentage of glycine occupancy at the chymotrypsin interface is not observed in case of trypsin and elastase interface. Details for the difference in occupancy of the total enzyme surface and the interface with the ecotine is presented in Figure 5.Those differences clearly corroborate with the function that each protease performs; trypsin needs to handle the positive charges of the polypeptide, where it will cleave it, by counterbalancing polypeptide charge/polarity with the charge/polarity of its own at the IFR surface. Chymotrypsin counterbalances the space occupancy of the large hydrophobic residues at the polypeptide to be cleaved, by accommodating many glycines into its own IFR surface. Elastases on the other hand have a large area occupied on the IFR surface by the hydrophobic amino acids, managing the space complementarity issues against small hydrophobic amino acids at the cleavage site of a polypeptide. The general trends of the characteristics for the interfaces described above are to be complemented with the specific occupancy of the IFR positions, discussed earlier in this section, in order to obtain the complete portrayal of the specificity for serine protease subfamilies.

4 Conclusions

The analysis of the data produced during this work is presented as a very useful tool for understanding of protease specificity based on Interface Forming Residues (IFR) profile alignment.

A predictive power of presented procedure is very much focalized in its capability to indicate differences between various proteases with respect to specificity, as well as to open a possibility to account for difference in how strong might be the binding between an enzyme and an inhibitor (an issue we need to explore in the future with more details).

Work presented here, although taxonomic in its nature, has offered very interesting insight in structure/function relationship for chosen family of proteins: serine proteases. Choice of focusing our attention on interface is a simple consequence of the fact that it is there where all functionally critical parameters must be concentrated. Superposition of structurally aligned backbones for serine proteases clearly corroborate with this idea: namely, nature is using essentially identical scaffold and yet it achieves variety of very specific activities by simply varying the SURFACE. This might be described as an application of a particular "finishing touch" on a common skeleton. A reader may find in literature a compilation of data where the authors tried to modify the serine protease functionality by treating a functional specificity in a local and limited scope and consequently mostly succeeding to abolish activity of the initial enzyme but not to establish the activity of the planned type. Our approach explains why and corroborates with the results presented by Ma *et al.* [9] and Novozymes Biotech, Inc. (Davis, CA, US) which even patented a technology (Microbial trypsin mutants having chymotrypsin activity - United States Patent 20050037368).

5 Acknowledgment

This work was supported in part by the following grants: FAPESP 01/08895-0, FINEP 1945/01 and CNPq 401695/2003-4 and by NSF grant #DBI-9904841 to Barry Honig. The authors wish to express special thanks to Barry Honig for his valuable insights and continuous interest in this project.

6 References

[1] Schechter I. and Berger A., “On the size of the active site in proteases. *I. Papain”*. *Biochem. Biophys. Res. Commun.*, vol. 27 pp. 157, 1967.

[2] Neshich G., Togawa R. C., Mancini A. L., Kuser P. R., Yamagishi M. E. B., Pappas Jr. G., *et al.*, "STING Millennium: a Web based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence." *Nucleic Acids Res,*. Vol. 31, no. 13, pp. 3386 - 3392, 2003.

[3] Da Silva M. C. M., Grossi de Sá M. F., Chrispeels M. J., Togawa R. C. and Neshich G., "Analysis of Structural and Physico-chemical Parameters Involved in the Specificity of Binding between Alpha-Amylases and their Inhibitors." *Protein Eng.*, vol. 13, no. 3, pp. 167-177, Mar 2000.

[4] Berman H. M., Bourne P. E. and Westbrook J., “The PDB: a case study in management of community data”. *Curr. Proteomics,* vol. 1 pp. 49-57, 2004.

[5] Yang, A.-S. and Honig, B., “Sequence to structure alignment in comparative modeling using PrISM.” *Proteins: Struc., Func. and Genet.*, vol. 3, pp. 66-72, 1999

[6] Sridharan, S., Nicholls, A. and Honig, B. “A new vertex algorithm to calculate solvent accessible surface areas”. *Biophys. J.*, vol 61, pp. A174, 1992.

[7] Bone, R., Silen, J. L., and Agard, D. A. “Structural plasticity broadens the specificity of an engineered protease”. *Nature,* vol. 339 pp. 191-195, 1989.

[8] Page M. J., Di Cera E. ”Is it possible to transform trypsin to thrombin?” *IUBMB Life*, vol. 59, no. 6, pp. 413-414, 2007.

[9] Ma W., Tang C., and Lai L. “Specificity of Trypsin and Chymotrypsin: Loop-Motion-Controlled Dynamic Correlation as a Determinant”; *Biophys J.*, vol. 89, no. 2, pp. 1183–1193. Aug 2005

Table VI: IFR positions arranged so that the rows represent the primary sequence position of IFR is presented in the leftmost column and the columns thereafter are populated by AA from 73 different serine proteases, all (but 1mtn.pdb) with the hard docked inhibitor: ecotine. The color code for amino acids is as follows: Residues: AVLIMFP are colored grey! [small and hydrophobic]; Residues: STYNQWG are colored green [polar]; Residues: D E are colored red! [negatively charged]; Residues: R K are colored blue! [positively charged] ; Residues: C is colored yellow! [disulphide bridge forming].The columns are occupied by following serine protease sub-families: Trypsin(1-10), Fusarium trypsin(11), Chymotrypsin(12-18), Chymotrypsinogen(19-20), Zymogen(21-23), Thrombin(24-48), Elastase(49-53), Compilation of various serine proteases including Kallikreins and blood coagulation factor IXa, Xa, D (54-64), serine proteases from Prokaryotes (65-73), in respective left to right order.

```
         1       10        20        30        40        50        60        70
         ---------|---------|---------|---------|---------|---------|---------|
E0035:   ---------RDD-D-DDD-----RRR-RRRRRRR-RRRRRRRRRRRRLLYY-NLLI-HMMR--AH--------
E0036:   ------------------------------------------------K-E---K------CC--------
E0037:   ---------NTT-----------PPP-PPPP-PP-PPPPPPPPPPPP--S-LE--T----HH----------
E0038:   ------------------------------------------G---R----I-------E---------
E0039:   YYYS-YYYYGFF-F-FFF-----EEE-EEEEEEE-EEEEEEEEEEEE---YREA-KDS--TRE-ITGGAAAAA
E0040:   HHHHHHHHHPHH-H-HHH-----LLL-LLLLLLL-LLLLLLLLLLLLHH-HVGHHVAF-HYFY--GSSSSSSS
E0041:   FFFFFFFFFWFF-F-FFF-----LLL-LLLLLLL-LLLLLLLLLLLFF-TWFLLIFQGFVLL--RRRLLLLL
E0042:   CCCCCCCCCCCC-C-CCC-----CCC-CCCCCCC-CCCCCCCCCCCCCC-CC-CCCCCCCCCC--CCCCCCC
E0057:   HHHHHHHHHHHH-H-HHH-----HHH-HHHHHHH-HHHHHHHHHHHHHH-HHHHHHHHHHHHHH-HHHHHHH
E0058:   CCCCCCCCCCCC-C-CCC---------C---C---C-----------CC-CC---CCCCCCCCC-C--CCCCC
E0059:   YYYYYYYYYV-G-----G-------------------------------V---K-K----------------
E0060:   KKKKKKKKK-------------LLL-LLLLLL-LLLLLLLLLLLAAHDDYEE-E--RIQ-----------
E0073:   ----------------------R-R--RRR-RR--RRR-R--RRRR----Q---------------------
         1       10        20        30        40        50        60        70
         ---------|---------|---------|---------|---------|---------|---------|
E0094:   FYYFYFFYYYYY-Y-YYY---------------------------YY-WF---YYFYYY-Y--FFFFFFF
E0095:   DNDNDDD---NN-N-NNN-----NNN-NN-NN-----NN-NNN-NNN---NNT--NNNN--DI----------
E0096:   RSSGRRRSSGSS-S-SSS-----WWW-WWWWWWW-WWWWWWWWWWWWP--SPKP-SVGGAADV----------
E0097:   KNNNKKKYY-LL-L-LLL-----KKK-KKRRRRR-RRRRRRRRRRRV--DDEDDVNK-EDDH----------
E0098:   TTTTTTTNN-TT-T-TTT-----NNN-NNNNNNN-NNNNNNNNNNNNNPWDTTT-PK-TN-TD----------
E0099:   LLLLLLLIINII-I-III-----LLL-LLLLLLL-LLLLLLLLLLLLLV-VYYIINYY-KHYH----------
E0141:   WW-WWWWWWW---W---------W-W--WWWWWW--WWWWW----WFW-----W--------RRRRR
E0142:   G----GGGG--G-G-G-------GGG-GGGGGGG--GGGGG--GGGG----GGG----------ATTTTTTTT
E0143:   H----NNNNALL----LL-----NNN-NNNN-NN-NNNNNNNNNNNLL-KQR-IKR--RKK--E--------
E0147:   ---------G-----T-------TT--T-TT-TT-TT--TTT-T----R------R-P---L--S--------
E0148:   ---------G-------------WWW-WW-W-WW-WW--WW---------------RDG--------------
E0149:   -T---VV--S------AA-----T-T-TT----T-TT----T---TT--L--------FG-YPM----------
E0151:   HYYYYEE--TTT----TT-----QQQ-QQQQ-QQ-QQQQQQQQQQQQII-L--RR-SFQPYYVN---------
E0168:   ------------------------------------------------------------------TTTTT-TT
E0169:   ------------------------------------------------------------------IW-VVVAAAAA
E0170:   ----------------------------------------------------R-------EG-NNNNNNNN
E0171:   ------------------------------------------------Y---Y---------YYYYYYYY
E0172:   YYY-Y-Y--YWW-W-WWW-------------------------------WP-HH--HY--L---GGGAA---
E0173:   PP--PP----------------RRR-RRRRRRR-RRRRRRRRRRR----PSDD-KP---NK--GDS--AAA
E0174:   GGGG-GG--STT-T-TTT-----III-IIIIIII-IIIIIIIIIII----GF---FDN--RD-TG-SEEEEE
E0175:   KQQQKKKMMAKK-K-KKK-----RRR-RRRRRRR-RRRRRRRRRRR---TII-A-TKE-ET---DG-GGGGG
E0176:   ------------------------------------------------------------VPIAAAAA
         1       10        20        30        40        50        60        70
         ---------|---------|---------|---------|---------|---------|---------|
E0190:   SSSSSSSSSSS-S-SSS-----AAA-AAAAAAA-AAAAAAAAAAAAVV---A--ASTTISATE------A-A
E0191:   CCCCCCCCCCCC-C-CCC-----CCC-CCCCCCC-CCCCCCCCCCCCCC-CCC--FCCCCCC-P------G-G
E0192:   QQQQQQQQQQMM----MM-----EEE-EEEEEEE-EEEEEEEEEEEFF-NFQKKMQMQFQQAG-ASAMMRRR
E0193:   GGGGGGGGGGGG-G-GGG-----GGG-GGGGGGG-GGGGGGGGGGGGGG-GGGGGGGGGGGGG--GGGGGGGG
E0194:   ----------D-D-------------------------D------D--------D------S-D-D-D---
E0195:   SSSSSSSSSSS-S-SSS-----SSS-SSSSSSS-SSSSSSSSSSSSSS-SSSSSSSSSSSS--SSSSSSSS
E0213:   VVVVVVVVVVVV-V-VVV-G---VVV-VVVVVVV-VVVVVVVVVVVVAA-TVV--VITVDVITP-THT--M-M
E0214:   SSSSSSSSSSS-S-SSS-C---SSS-SSSS-SS-SSSSSSSSSSSSSS-SSS--SSSSSS-S-SSSSSSS
E0215:   WWWWWWWWWWW-W-WWW-----WWW-WWWWWW-WWWWWWWWWWWFF-FFWWSYWWWFWWG--GGGGGGGG
E0216:   GGGGGGGGGGG-G-GGG-----GGG-GGGGGGG-GGGGGGGGGGGVV-VVGGGGGGGVGGG--GSGGGGLA
E0217:   YSSYDYYYYNSS-S-SSS-----EEE-EEEEEEE-EEEEEEEEEEER--SSESS-EHYIRLA--SSSNNNNN
E0218:   ----------SS-S-SSS-------------------------GG-SAGRR--T-WG-T--G-G--VVV
E0219:   GGGGGGGGGG---------K---GGG-GGGGGGG-GGGGGGGGGGGGG-----V-E-G--GP---G------
E0220:   CCCCCCCCCCCC-C--CC-----CCC-CCCCCC--CCCCCC-CCCC---C-C-C-CCC-CCC----------
E0221:   ---------A-------------RDR-RD-R----D-DDDD-DDDDD--T-----H----------------
E0222:   ------------------------------------------------S--R--------------------
E0223:   ------------------------------------------------F---R-A----------------
E0224:   -KKKK----Y-------------KKK-KK-KKKK-KKKKKKKKKKKKYYV---KK--KYFK----------
```