MODELING STUDIES ON THE ANCESTRAL PROTEINASE STEMZYME IDP-B

László SZILÁGYI[1], Beáta ÁBRAHÁM[2], Ildikó MIKLÓSSY[3], Szabolcs SZILVESZTER[4], Judit GÁLICZA[5], Szabolcs LÁNYI[6]

S-a considerat proteinaza ancestrală stemzyme IDP-β ca fiind o proteină-starter adecvată pentru reproiectarea specificității serin-proteinazelor. Din acest motiv, o importană majoră i se acordă analizei secvenței de aminoacizi a acestei proteine şi a elaborării unui model structural in silico*. Deoarece expresia heterologă a acestei enzime necesită atenție specială în ceea ce priveşte reziduurile de cisteină prezente în moleculă, s-a analizat o aliniere multiplă a secvenței proteazei ancestrale cu mai multe proteinaze, în care poziția punților disulfidice este bine documentată. S-au elaborat modele structurale pntru proteina hipotetică pe baza diferitelor structuri de serin-proteinaze cunoscute.*

We considered the ancestral proteinase IDP-β as a good starter-protein to redesign substrate specificity of serine-proteinases. Therefore it is of major importance to analyse its amino acid sequence and to create a good in silico *structural model for this enzyme. As special attention to the cysteine residues present must be paid when taking into account heterologous expression of this enzyme, we analysed a multiple sequence alignment with proteinases where the positions of disulfide bridges are well documented and created structural models for the hypothetical protein based on different known serine-proteinase structures.*

Keywords: serine-proteinases, protein structure modeling

1. Introduction

Stemzyme IDP-β is a synthetic protein predicted theoretically to be the ancestor of the present-day immune defense proteases (IDP); this ancestral proteinase was reported to display an unexceptional tolerance to mutations at the binding site, with different mutations resulting in activities similar to some of the

[1] PhD., Faculty of Sciences, Sapientia University, Cluj Napoca, Romania
[2] PhD student, Faculty of Applied Chemistry and Material Science, University POLITEHNICA of Bucharest, Romania
[3] PhD student, Faculty of Applied Chemistry and Material Science, University POLITEHNICA of Bucharest, Romania, corresponding author, e-mail: miklossyildiko@sapientia.siculorum.ro
[4] PhD student, Faculty of Applied Chemistry and Material Science, University POLITEHNICA of Bucharest, Romania
[5] PhD student, Faculty of Applied Chemistry and Material Science, University POLITEHNICA of Bucharest, Romania
[6] PhD., Faculty of Sciences, Sapientia University, Cluj Napoca, Romania

synthetic enzyme's descendents [1]. The ancestral enzyme also was found to show a remarkably plastic S1 substrate binding site, which can be explained by the fact that the common ancestor of serine proteinases having widely different substrate specificities must tolerate various changes in the substrate binding region during evolution. Considering this ancestral proteinase a good candidate if redesign of substrate specificity of serine-proteinases is targeted, it is of major importance to analyse its amino acid sequence and to create a good *in silico* structural model for this enzyme. As special attention to the cysteine residues present must be paid when taking into account heterologous expression of this enzyme, a multiple sequence alignment with proteinases where the positions of disulfide bridges are well documented was created by Clustal W [2]. The modelling studies, conducted with the MODELLER program developed by A Šali at UCSF [3], using as a template the human mast cell chymase reveal that the 3D structure of stemzyme IDP-β is very similar to that of several proteins of the trypsin superfamily, as well as the presence of a C191-C220 disulfide bond.

2. Experimental

Multiple sequence alignments were constructed using Clustal W.

To generate 3D models we used the MODELLER 9v7 software, which was released on June 16th, 2009, and developed by A Šali at UCSF. This program generates three-dimensional structures of proteins by satisfaction of spatial restraints. MODELLER is most frequently used for homology or comparative protein structure modeling: In this case the user provides an alignment of a sequence to be modeled with known related structures and MODELLER will automatically calculate a model with all non-hydrogen atoms. First we prepared the input files: the templates were 1klt.pdb (human mast cell chymase), 2cga.pdb (bovine chymotrypsin), 1h4w.pdb (human trypsin 4), 1fq3.pdb (human granzyme B), 1klt.pdb (human mast cell chymase), 1mza.pdb (human pro-granzyme K) - Protein Data Bank atom files containing the coordinates for the template structures (www.ebi.ac.uk); PIR database format type of alignment files of the template structures with the target sequence for each template; script files contanining MODELLER commands for each template that instructs MODELLER what to do. These are ordinary Python scripts (www.python.org). Next we ran the MODELLER with the script files. Molecular graphics images were obtained using the UCSF Chimera package [4].

3. Results and Discussion

3.1. Analysis of the stemzyme-IDP-β sequence

The published amino acid sequence of stemzyme IDP-was firstly analyzed by the ProtParam program of the ExPASy Proteomics Server. The protein is

highly basic, its theoretical isoelectric point being 9.65, and it contains 9 cysteine residues. The serine proteinases are secreted proteins and they contain several disulfide bonds: chymotrypsins have 5 cystines; trypsins of higher vertebrates (except humanoids) have 6 cysteines. Since an unpaired cysteine might induce disulfide rearrangements and therefore can impair the yield of protein expression in heterologous expression systems, it is important to analyze its location in the protein. It can be done by multiple alignments with proteinases where the positions of disulfide bridges are well documented.

Fig. 1 shows the multiple alignment of stemzyme-IDP-β with 5 serine proteases representing different families generated by Clustal W. The alignment was manually adjusted in the region of 216-226. From the alignment one can conclude that 8 out of 9 cysteines present in the stemzyme sequence align with cysteines forming conserved disulfides both in chymotrypsin and trypsin. These disulfide bonds are labeled by differently colored boxes: C42-C58 is red, C136-C201 is green, C168-C182 is blue and C191-C220 is orange.

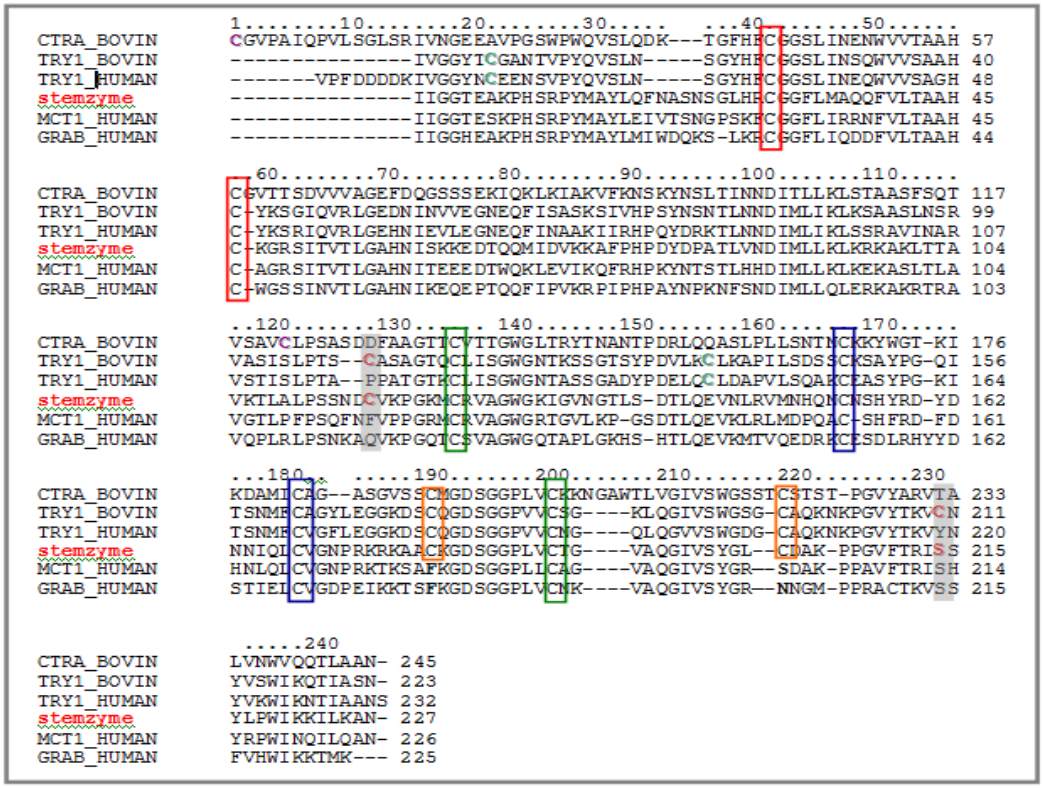

Fig. 1. Multiple sequence alignment of bovine chymotrypsin, bovine cationic trypsin 1, human trypsin 1, stemzyme, human mast cell chymase (MCT1_HUMAN) and human granzyme B; identically colored boxes indicate cysteines belonging to the same a disulfide bond

We found the C191-C220 disulfide bond especially interesting, as it is absent from the proteases presumed to originate form this ancestral enzyme, as illustrated by the last two sequences in the alignment: The reconstruction of this disulfide bond in the hypothetical common ancestor of these enzymes might seem to be erroneous, but it is not the case. This disulfide bond is found in the most primitive trypsin-like proteases found in fungi and insects, and it is conserved through the evolution of the trypsin-like serine protease family. It has definitively been lost in granzymes and chymases, but it is reasonable to suppose, that stemzyme IDP-β, their hypothetical common ancestor did contain it.

The unpaired cysteine in stemzyme is at position 128. There is no cysteine here in chymotrypsins although several other members of the trypsin family (trypsin, thrombin, plasmin, kallikrein) have a C128-C232 bond. The position of this disulfide bond is shaded gray in Fig. 1. It is interesting to note that while this disulfide bond is conserved in most vertebrate trypsins, both of its cysteines are lost in human (and humanoid) trypsins as shown in the sequence of TRY1_HUMAN. It indicates that these segments of the serine protease gene can be regarded as mutation hot spots. On the other hand, in the process of the elimination of a disulfide bond an enzyme with unpaired cysteine must exist because it is highly unlikely that two mutations hitting both cysteines of the same disulfide bond would occur at the same time. With regard to the surprising elimination of disulfide bonds from human trypsins in a previous work we addressed the effect of the unpaired cysteines on the stability and enzymatic activity of trypsin [5]. We concluded that an unpaired cysteine buried in the interior of the protein is highly deleterious while an unpaired cysteine with large accessible surface area is well tolerated. Our analysis showed that in trypsin the C128-C232 disulfide bond lies on the surface of the molecule, and both its mutants containing unpaired displayed high enzymatic activity and sufficient stability for expression in *E. coli.* Since our preliminary modeling studies reveal the 3D structure of stemzyme-IDP-β is very similar to that of trypsins (see Fig. 2B) we are confident that the unpaired cysteine at position 128 will not bring forth unwanted side effects.

3.2 Modeling studies with special respect to the formation of the C192-C220 disulfide bond

As the MODELLER program is most frequently used for homology or comparative protein structure modeling, we used it to construct 3D models for stemzyme-IDP-β on the basis of known related protein structures. In our modeling study first we used the human mast cell chymase (MCT1_HUMAN) as reference sequence and its structure (1klt.pdb), as it was described in the Wouters paper (Fig. 2). It is important to note that human mast cell chymase does not have the C191-C220 disulfide bond. Part A in Fig 2. shows the Cα carbon backbone of

human mast cell chymase (gray, hardly visible) overlaid with stemzyme IDP-β structure obtained with automated modeling (violet). In this case we also obtained an open cysteine containing structure. This structure completely overlaps the chymase structure. However, inclusion of a command line into the modeling script to bind these two cysteines resulted in a structure shown in cyan. In this structure there is a disulfide bond between C192 and C220. Comparing the stemzyme IDP- β structures we see that the position of C192 Cα carbon is practically identical while in the disulfide containing structure the position of C220 Cα carbon moves slightly, (~ 0.35 Å) towards C192 and there is an about 20° rotation around the ψ and φ dihedral angles. Since the movement of the Cα carbon is minimal and the new dihedral angles are within the allowed region of the Ramachandran plot we conclude that there is no evident steric hindrance for the formation of the C191-C220 disulfide bond. Besides our modeling results indicate that the substrate binding pocket (blue circle) is identical in the two forms of stemzyme IDP-β, we did not observe the constriction of the bottom in the open form.

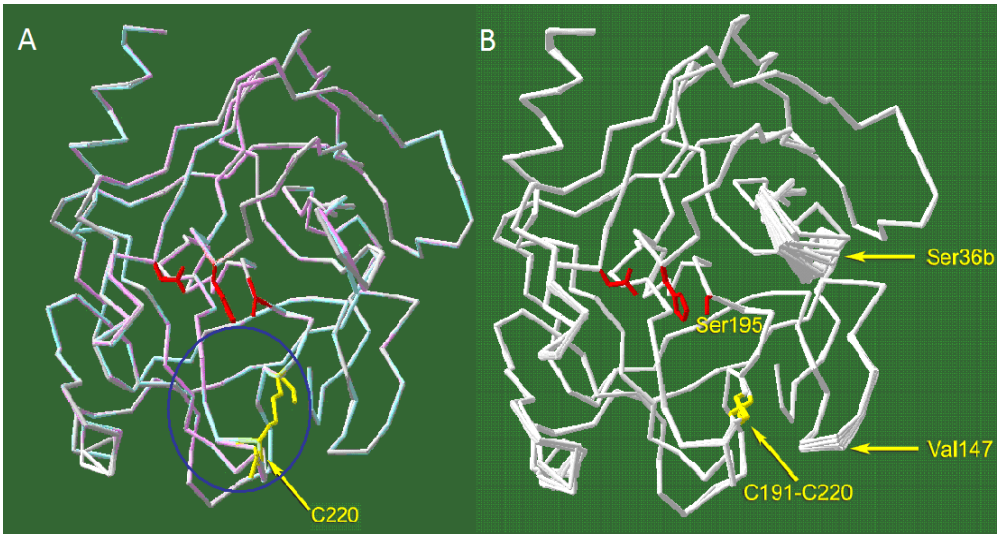

Fig. 2. Part A: Structure of human mast cell chymase (1klt.pdf gray) overlaid with the stemzyme IDP- β structure modeled without specific restrain (violet) and with the restrain to form C192-C220 disulfide bond (cyan). The residues of the catalytic triad (from left to right: Asp102, His57 and Ser195) labeled red. Blue circle indicates the substrate binding pocket;
Part B: 20 independent structures of stemzyme IDP- β structure modeled with the restrain to form C192-C220 disulfide bond

To assess the reliability of the modeling process we performed two sets of experiments. First we generated 20 independent models based on the chymase structure 1ktl.pdb with the restrain to form the disulfide bond. Part B of Fig. 2

shows these structures. (The chymase structure is not shown.) Next we used five different serine protease structures as templates. The results are shown in Fig. 3.

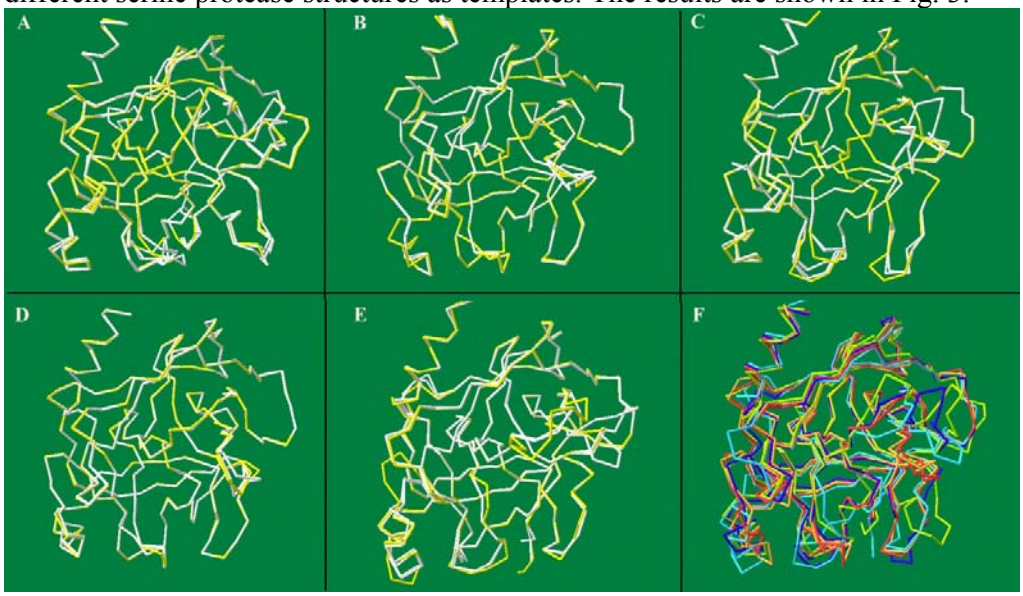

Fig. 3. Modeling of stemzyme IDP- β against different serine proteases.
A-E: stemzyme IDP- β structure gray, template structure yellow. The templates were: A. bovine chymotrypsin A (2cga.pdb); B: human granzyme B (1fq3.pdb); C: human trypsin 4 (1h4w.pdb); D: human mast cell chymase (1klt.pdb); E: human pro-granzyme K (1mza.pdb); F: comparison of the different modeled stemzyme IDP- β structures (templates are not shown)

Part B of Fig. 2 shows a series of models generated according to the human mast cell chymase. Although the vast majority of the Cα carbon positions are well defined there are two loops in where the structures differ significantly. One such loop contains residues 36a-36c; three amino acids insertion compared to the chymotrypsin sequence. Interestingly this insertion is present in the template mast cell chymase sequence too. This loop is very flexible even in the chymase structure corresponding to the high B factors (~50) given in 1klt.pdb (not shown). It seems so that the restrains for this loop are not sufficient for MODELLER to generate a well defined structure. A similar by much lower flexibility can be observed in the so-called autolysis loop (residues 142-152) which is also reflected in the structures around Val147. Since these loops are in the vicinity of the active site they might have some role in the substrate specificity [6] therefore they are potential targets for introducing mutations to optimize substrate specificity.

In Table 1. we summarize the root-mean-square deviations (RMS) of some Cα positions in 20 different structures generated by MODELLER.

Table 1.

RMS deviation of C α positions in different stemzyme IDB- β structures generated by MODELLER compared to the lowest energy structure

code #	Rel. energy	RMS Å					
		Backbone	**Ser36b**	**Val147**	**Cys191**	**Cys220**	**Ser195**
0153.pdb	1307.63904	-	-	-	-	-	0
0140.pdb	1329.67407	0.06	0.089	0.073	0.055	0.117	0.026
0154.pdb	1339.59082	0.15	0.732	0.025	0.051	0.149	0.025
0143.pdb	1340.33264	0.08	0.132	0.082	0.016	0.057	0.036
0147.pdb	1348.12512	0.11	0.542	0.066	0.020	0.047	0.012
0157.pdb	1350.67834	0.17	1.388	0.132	0.041	0.036	0.030
0141.pdb	1356.09753	0.11	0.456	0.137	0.071	0.104	0.026
0155.pdb	1362.99426	0.13	0.802	0.029	0.032	0.067	0.054
0156.pdb	1365.2616	0.16	1.423	0.173	0.063	0.220	0.027
0142.pdb	1369.57104	0.21	1.234	1.341	0.074	0.151	0.043
0144.pdb	1386.87451	0.28	0. 698	0.040	0.026	0.052	0.015
0149.pdb	1392.08398	0.42	3.578	0.195	0.110	0.162	0.041
0148.pdb	1392.87598	0.15	0.464	0.083	0.052	0.133	0.014
0159.pdb	1400.61804	0.11	0.406	0.170	0.027	0.122	0.014
0151.pdb	1420.74744	0.27	2.779	1.454	0.072	0.092	0.020
0158.pdb	1431.06702	0.17	0.624	0.202	0.031	0.132	0.033
0146.pdb	1432.45422	0.27	0.246	0.115	0.048	0.068	0.034
0150.pdb	1448.74377	0.25	0.261	0.041	0.047	0.128	0.032
0152.pdb	1485.96802	0.38	2.745	1.535	0.062	0.092	0.036
0145.pdb	1493.58948	0.24	1.352	1.463	0.042	0.050	0.014
mean		*0.196*	*1.050*	*0.387*	*0.0495*	*0.104*	*0.028*
standard deviation		0.098	*0.986*	*0.566*	*0.0229*	*0.048*	*0.011*

The position of the amino acids of the catalytic triad is extremely well defined as demonstrated by the data for Ser195. Although the values are significantly larger than that of for Ser195, the positions for C191 and C220 are also well defined in spite of the absence of this disulfide bond in the chymase structure, which we consider as another proof for the formation of this disulfide bond in stemzyme IDP- β.

4. Conclusions

The first conclusion driven from the results presented in the previous chapter is presence of the C191-C220 disulfide bond, based on analysis and adjustment of the multiple alignment of the stemzyme-IDP-β sequence with representative serine-proteinases from different families. The original paper of Wouters et. al, 2003 [1] suggested the absence of this bond which they concluded to be responsible for contraction of the substrate-binding pocket of the protein.

Our modeling results indicate that the substrate binding pocket is identical in the two forms of stemzyme IDP-β (with and without the C191-C220 disulphide), we did not observe the constriction of the bottom in the open form.

Another interesting observation concluded from our modeling studies reveals flexibility of the region called autolysis loop (residues 142-152), around Val 147, as well as of the secondary structural elements formed around Ser36. Being situated close to the active site of the enzyme, these structures have been reported previously to have influence on substrate specificity of serine-proteinases. This is of major importance considering our future research plans targeting optimization or even complete redesign of substrate specificity of this ancestral proteinase, these regions being of special interest when site-directed mutations are planned to modify to some extenct the substrate-binding pocket.

The results shown in the previous chapter are very instructive, on one hand by providing useful information on choosing the adequate expression system for this protein, on the other hand on the regions suitable for introduction of mutations.

The results conclude also in a demonstration of the known fact that the result of homology based protein modeling depends on the choice of template. This observation supports our proposal to determine the 3D structure of stemzyme IDP-β.

R E F E R E N C E S

[1] *M.Wouters, K.Liu K, P.Riek, A.Husain,* "A despecialization step underlying evolution of a family of serine proteases, in Mol. Cell., **vol. 12,** 2003, pp. 343-354

[2] *R.Chenna, H.Sugawara, T.Koike, R.Lopez, T.Gibson, D.Higgins, J.Thompson,* "Multiple sequence alignment with the Clustal series of programs", in Nucleic Acids Res*.,* **vol. 31,** no.13, pp. 3497-500

[3] *A.Šali, T.Blundell*, "Comparative protein modelling by satisfaction of spatial restraints", in J. Mol. Biol., **vol. 234**, 1993, pp. 779-815

[4] *E.Pettersen, T.Goddard, C.Huang, G.Couch, D.Greenblatt, E.Meng, T.Ferrin,* "UCSF Chimera - A Visualization System for Exploratory Research and Analysis", in J. Comput. Chem*,* **vol. 25**, no.13, 2004, pp. 1605-1612

[5] *E.Kénesi, G.Katona, L.Szilágyi*, "Structural and evolutionary consequences of unpaired cysteines in trypsinogen", in Biochem. Biophys. Res. Commun., **vol 309,** 2003, pp. 749-754

[6] *L.Hedstrom, L.Szilágyi, W.Rutter*, "Converting trypsin to chymotrypsin: the role of surface loops", in *Science,* **vol. 255**, no.5049, 1992, pp. 1249-53.