doi:10.1016/j.jmb.2007.04.029 J. Mol. Biol. (2007) 370, 63–79

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

# When the Surface Tells What Lies Beneath: Combinatorial Phage-display Mutagenesis Reveals Complex Networks of Surface–Core Interactions in the Pacifastin Protease Inhibitor Family

**Borbála Szenthe[1], András Patthy[2], Zoltán Gáspári[3] Adrienna Katalin Kékesi[4,5], László Gráf[1,2] and Gábor Pál[1]***

[1]*Department of Biochemistry Eötvös Loránd University Budapest, H-1117, Hungary*

[2]*Biotechnology Research Group of the Hungarian Academy of Sciences, Eötvös Loránd University, Budapest H-1117, Hungary*

[3]*Institute of Chemistry Eötvös Loránd University Budapest, H-1117, Hungary*

[4]*Department of Physiology and Neurobiology, Eötvös Loránd University, Budapest H-1117, Hungary*

[5]*Proteomics Group of the Biology Institute, Eötvös Loránd University, Budapest H-1117, Hungary*

**Corresponding author*

Pacifastin protease inhibitors are small cysteine-rich motifs of ~35 residues that were discovered in arthropods. The family is divided into two related groups on the basis of the composition of their minimalist inner core. In group I, the core is governed by a Lys10–Trp26 interaction, while in group II it is organized around Phe10. Group I inhibitors exhibit intriguing taxon specificity: potent arthropod-trypsin inhibitors from this group are almost inactive against vertebrate enzymes. The group I member SGPI-1 and the group II member SGPI-2 are extensively studied inhibitors. SGPI-1 is taxon-selective, while SGPI-2 is not. Individual mutations failed to explain the causes underlying this difference. We deciphered this phenomenon using comprehensive combinatorial mutagenesis and phage display. We produced a complete chimeric SGPI-1 / SGPI-2 inhibitor-phage library, in which the two sequences were shuffled at the highest possible resolution of individual residues. The library was selected for binding to bovine trypsin and crayfish trypsin. Sequence analysis of the selectants revealed that taxon specificity is due to an intra-molecular functional coupling between a surface loop and the Lys10-Trp26 core. Five SGPI-2 surface residues transplanted into SGPI-1 resulted in a variant that retained the "taxon-specific" core, but potently inhibited both vertebrate and arthropod enzymes. An additional rational point mutation resulted in a picomolar inhibitor of both trypsins. Our results challenge the generally accepted view that surface residues are the exclusive source of selectivity for canonical inhibitors. Moreover, we provide important insights into general principles underlying the structure–function properties of small disulfide-rich polypeptides, molecules that exist at the borderline between peptides and proteins.



*Keywords:* phage-display; protease inhibitor; pacifastin; taxon-specificity; core–surface interaction network

## Introduction

Serine proteases are ubiquitous enzymes that fulfil vital roles in a wide range of biological functions, including food protein digestion, blood coagulation, blood clot removal and immune defense mechanisms. These processes need to be regulated, and one of the many regulatory strategies relies on the use of naturally evolved protease inhibitor proteins.

There is a phylogenetically and structurally extremely diverse group of proteins with members that nevertheless inhibit serine proteases through a common canonical mechanism.[1–4] These are the reversible inhibitors that interact with the protease in a substrate-like manner, forming a non-covalent complex with the enzyme. The inhibitor in the complex can be cleaved, but the cleavage leads to a

Abbreviation used: hGH, human growth hormone.

E-mail address of the corresponding author: palgabor@elte.hu

thermodynamic equilibrium, and the cleaved inhibitor remains active. Although the 18 known families of these inhibitors are structurally unrelated, each possesses a common element, an exposed surface loop. This loop forms an antiparallel β sheet with the enzyme, and one of its side-chains, the P1 residue (for nomenclature see Schechter & Berger[5]) enters the binding pocket of the enzyme. The nature of the P1 residue correlates with the specificity of the inhibited protease; trypsin inhibitors commonly have arginine or lysine, while the most frequent P1 for chymotrypsin inhibitors is leucine. The scissile peptide bond is located at the carbonyl group of P1 between the P1-P1′ residues, and the minimal interacting region spans about six residues between the P3-P3′ residues.

The first reversible inhibitor belonging to the Pacifastin family was discovered about 20 years ago in the haemolymph of the crayfish *Pacifastus leniusculus*.[6] Its amino acid sequence was determined ten years later and was shown to contain nine cysteine-rich domains.[7] Inhibitors with homologous domains were isolated also from the locusts *Schistocerca gregaria* and *Locusta migratoria*, and the wasp *Pimpla hypochondriaca*.[8–11] On the basis of the findings of a recent bioinformatic study, we predict that the large-scale genome projects will soon deliver numerous additional members of this family.[12] A conserved motif emerged from the available sequences having the following characteristic consensus:

$$C^1X_{9-12}C^2N\ C^3XC^4X_{2-3}G\ X_{3-4}C^5T\ X_3C^6$$

The six cysteine residues form the following three-disulfide bridges; $C^1$–$C^4$, $C^2$–$C^6$, and $C^3$–$C^5$, resulting in the pattern *abcacb*. A standard nomenclature has been suggested for these motifs in the form of GSPI (standing for Genus Species Pacifastin-related Inhibitor).[13] We adopt this logical terminology and, for clarity, we also mention the original names of the peptides.

The two major sources for the best characterized pacifastin peptides are the two locust species *L. migratoria* and *S. gregaria*; therefore, the corresponding inhibitors are named as LMPI and SGPI, respectively. The first 3D structures were solved by NMR for LMPI-1 (PMP-D2),[14] and LMPI-2 (PMP-C),[15,16] which revealed that while these peptides exhibit strong structural resemblance, they belong to two separate groups (group I and group II) mostly on the basis of the composition of the inner core.

While globular proteins form a large, densely packed hydrophobic core that establishes a stable native structure, the small pacifastin peptides have only a very small core region and acquire remarkable thermostability through a relatively large number of disulfide bridges. The minimalist group I-type core of LMPI-1 is formed by a pair of residues, in which the aliphatic segment of a lysine side-chain at position 10 interacts with a large hydrophobic tryptophan side-chain at position 26. A quite different core exists in the paralogous group II peptide LMPI-2, where there is phenylalanine at position 10, which is surrounded by small hydrophobic residues, position 26 being occupied by alanine. The same pattern emerged by solving the NMR structures of two homologous peptides from *S. gregaria*, SGPI-1 (SGTI), an ortholog of LMPI-1, and SGPI-2 (SGCI), an ortholog of LMPI-2.[17] The functional importance of W26 in the group I core was proved through a W26A mutation, in an LMPI peptide, which was detrimental for the folding of the molecule.[18] The role of the interactions made by F10 in group II inhibitors was revealed in a study of rationally designed cyclopeptide analogs of SGPI-2.[19]

The P1-P1′ sequence is R-K in SGPI-1 and L-K in SGPI-2. Thus, formally SGPI-1 should be a trypsin inhibitor, while SGPI-2 should be a chymotrypsin inhibitor. When our group first studied the inhibitory properties of SGPI-1 and SGPI-2 on well-established model bovine enzymes, it turned out that SGPI-2 is a strong chymotrypsin inhibitor with a $K_i$ of 6 pM, but SGPI-1 is a rather poor trypsin inhibitor, with $K_i$ in the low micromolar range.[20] Moreover, SGPI-1 with its P1 Arg residue was only a slightly better inhibitor of trypsin; on the other hand, SGPI-2 could be readily transformed to a nanomolar trypsin inhibitor with a simple L30R P1 mutation. An additional K31M mutation based on the known P1′ preference of bovine trypsin[21–23] resulted in a picomolar SGPI-2 variant trypsin inhibitor.[20] It turned out that the analogous P1′ K to M mutation in SGPI-1 failed to increase its trypsin inhibiting activity significantly, suggesting at that time that SGPI-1 might not have the proper native structure to function as a protease inhibitor. However, the simple R to L P1 mutation transformed this peptide into a sub-nanomolar chymotrypsin inhibitor. This finding predicted that wild-type SGPI-1 was a potent inhibitor, but its unknown trypsin-like biological target had some important properties not shared with bovine trypsin.

This hypothesis was clearly verified when our group first demonstrated that wild-type SGPI-1 is a picomolar inhibitor of two arthropod trypsins.[24] Therefore, an unprecedented five orders of magnitude specificity index was found for the arthropod *versus* the vertebrate trypsin inhibition. In subsequent work, the same phenomenon was described for the homologous group I LMPI-1 and LMPI-3 peptides, which are poor inhibitors of vertebrate trypsins but excellent inhibitors of a trypsin isolated from the midgut of *L. migratoria*.[18,25] Moreover, an analogous finding was described for SGPI-5A, another group I inhibitor.[25]

When a simple L to R P1 mutation switched the group II chymotrypsin inhibitor SGPI-2 to a highly potent inhibitor of both bovine and arthropod trypsins,[24] it became evident that taxon specificity is strictly related to the group I fold.

A complete exploration of taxon-specificity elements in SGPI-1 through individual mutations would have been difficult, as there are 16 differences between the SGPI-1 and SGPI-2 sequences besides

the two core positions 10 and 26. Moreover, when completeness cannot be achieved, the decision about which particular positions should be analyzed is unavoidably biased by the researcher's prejudice about what might and what might not be functionally important in the protein.

Therefore, we applied an approach that is able to handle both types of problem. We produced all the possible combinations of the SGPI-1 and SGPI-2 sequences through a combinatorial mutagenesis regime, without judging which one of these could be important. Then, instead of attempting to analyze the $2^{18}$ (131,072) variants individually, we applied phage display and expressed the variants clonally on the surface of M13 phage. Then we carried out *in vitro* selections on immobilized bovine and crayfish trypsin separately.

Sequence analysis of binding-selected clones clearly demonstrated that the group I core of SGPI-1 does not hinder bovine trypsin inhibition; therefore, the core by itself cannot be responsible for taxon specificity. It turned out that it is a particular combination of the 20-22 TPT surface turn and the K10-W26 core that is incompatible with inhibition of bovine trypsin. Two individual SGPI-1 mutants containing five or six mutations were synthesized on the basis of this notion and were found to be able to inhibit both the vertebrate and the arthropod enzyme with high affinity. Thus, the major cause of taxon specificity has been localized, and it has been demonstrated that group I core variants showing no sign of taxon specificity can be readily created.

## Results and Discussion

### Displaying wild-type SGPI inhibitors on phage

Display efficiencies of SGPI-1 and SGPI-2 were assessed through the presence of an N-terminally fused FLAG-tag epitope in ELISA format using an anti-FLAG-tag monoclonal antibody. Then, functionality and binding specificity of the wild-type fusion peptides were assessed on immobilized bovine and crayfish trypsin and on bovine chymotrypsin. As expected, phage-displayed SGPI-1 bound to crayfish trypsin with a high level of efficiency, while showing very low signals on the bovine enzymes. Phage-displayed SGPI-2, as anticipated, bound selectively to bovine chymotrypsin. The FLAG-tagged and phage-displayed wild-type forms were also analyzed in solution using standard enzyme inhibition assays. These further validated the applied format, as the inhibitors fully retained their known activity and specificity. The number of inhibitor molecules expressed per phage particle was calculated on the basis of titration experiments using the tightest inhibitor–enzyme pairs (crayfish trypsin for SGPI-1 and bovine chymotrypsin for SGPI-2). The titrations showed that for both inhibitors the average number of inhibitor per phage is about 1, therefore the display is monovalent (data not shown).

### Producing all the possible chimeras of SGPI-1 and SGPI-2 displayed on phage

There are many ways to produce a diverse phage-displayed protein or peptide library.[26] As the SGPI gene is only 105 bp long, we decided to make a fully synthetic DNA library. We used a DNA cassette technology based on the use of antiparallel pairs of forward and reverse primers that introduced the designed randomizations at 18 predefined positions (Figures 1 and 2). We aimed to introduce only two types of amino acid residues at each randomized position, but the structure of the genetic code necessitated the inclusion of two more residue types at several positions, as shown in Figure 2.

Although our library became binary for seven positions and tetranomial at 11 positions, it is im-

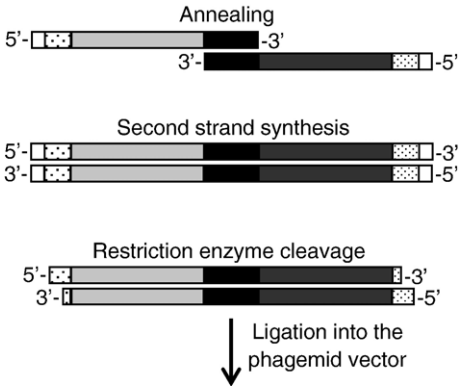


**Figure 1.** The overall scheme of the library production. The library was produced by an elaborate strategy of mixing two types of forward and four types of reverse degenerate mutagenesis oligonucleotides in eight (2×4) separate reactions as explained below. For clarity, only one of the eight reactions is shown using, one forward and one reverse primer. The forward and reverse primers annealed through their 3′ end at a short region (shown as a black box) coding for a Cys-Asn-Thr-Cys segment invariantly present in all SGPI peptides. The light grey segment in the forward primer and the dark grey segment in the reverse primer illustrate regions with randomized positions (for details see Materials and Methods). The reason for using 2×4primers was as follows. At position 10, the mixing of a Lys codon, AAA (or AAG) and Phe codon, TTT (or TTC) would have allowed for the occurrence of a stop codon, TAA (or TAG). To avoid this, we used two forward primers that differed only at position 10, one with an AAA for Lys and another one with TTT for Phe at this position. We used four reverse primers to cover the four possible combinations of insertions at two positions, one at position 24 (in SGPI-2) and another one at position 36 (in SGPI-1). In this way, each position that differed between SGPI-1 and SGPI-2 was randomized to allow for both SGPI-1 and SGPI-2 residue types. An *in vitro* second strand synthesis produced the complete double-stranded DNA cassette with restriction sites (dotted boxes) at the two termini. The cassette was digested and ligated into the phagemid vector.

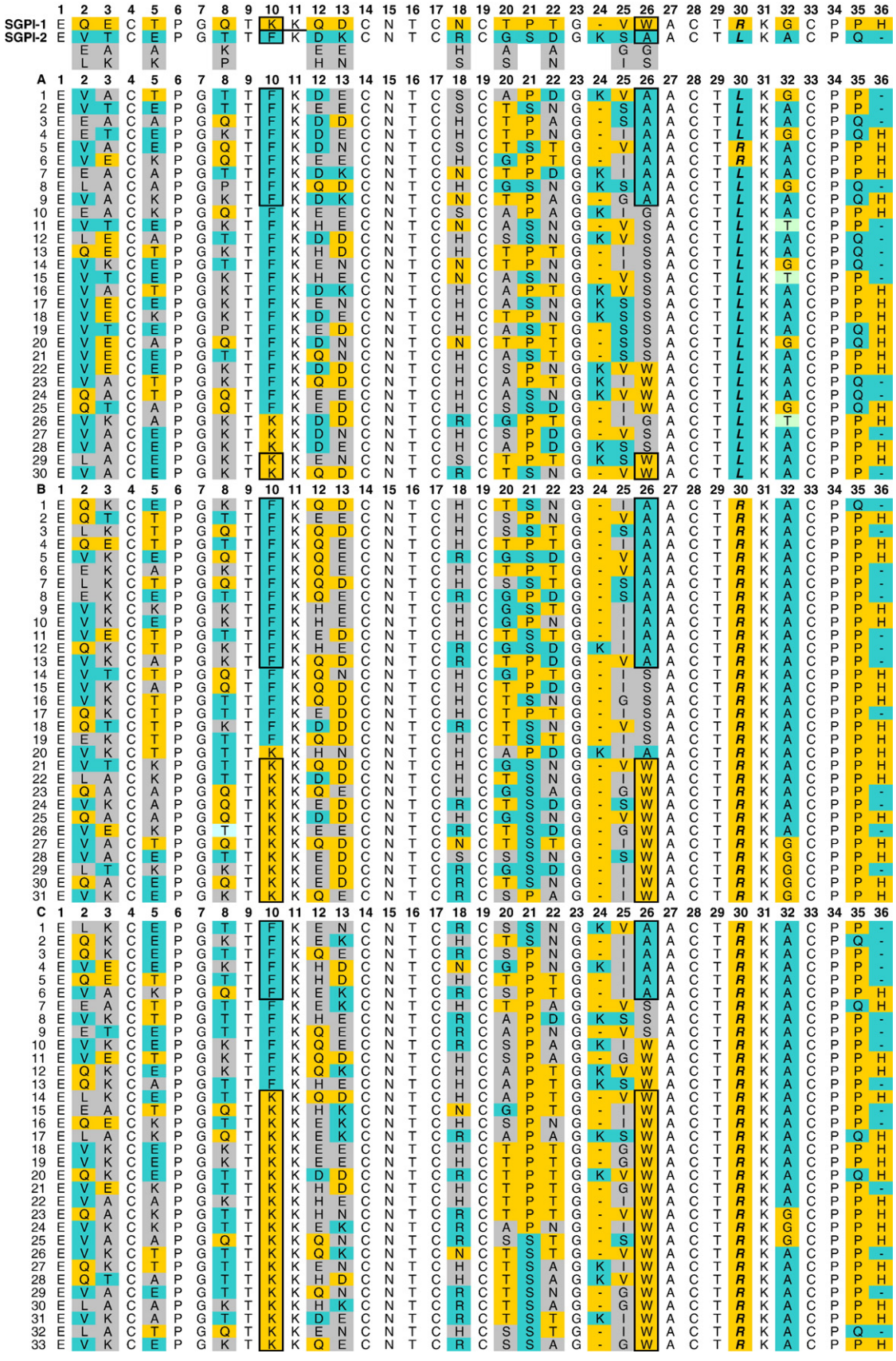


**Figure 2** (*legend on next page*)

portant to note that at each randomized position the initial ratio of the SGPI-1 *versus* SGPI-2 residue types is 1. The theoretical size for this library is $4^{11} \times 2^7$, about $5.3 \times 10^8$ clones. The measured size of our phage library was $5.8 \times 10^9$ individual clones, affording a greater than tenfold over-sampling of the covered sequence space. The composition of the randomized library is shown in Figure 2.

## Selecting functional clones from the chimera library

When protein variants are displayed on phage and selected for binding to an immobilized target, there are two intertwined types of selections. One works inside the *Escherichia coli* cell through proteolytic degradation of variants having poor folding properties. This effect can be amplified by using *in vitro* proteolytic degradation as the selection method.[27,28] This selection promotes structurally viable, stable variants irrespective of their functional properties, resulting in the phenomenon called "display bias".[29] The other type of selection is the binding selection through the immobilized target. This process selects for variants that are both properly folded and functional. We used an invariant N-terminal FLAG-tag epitope on each clone and applied selection through binding to an anti-FLAG-tag antibody to account for display bias in the selected pool. In this way, the effects of display bias on amino acid frequency could be deconvoluted from the combined results of the protease-binding selection at each randomized position, similar to the strategy described earlier.[30]

Binding selections were carried out separately on bovine and crayfish trypsin. The selection cycles were applied on parallel plates containing either the immobilized target, or bovine serum albumin. The relative titer of clones eluted from immobilized target *versus* albumin is the enrichment value. After two or three selection cycles, a thousand-fold enrichment was detected for all three selections. At this stage, 96 individual clones were isolated for clonal phage-ELISA assays for each target, as described.[31] The amino acid sequence for about 30 ELISA-positive SGPI clones was determined through DNA sequencing for each of the three selections. The results are shown in Figure 2.

## Analysis of the display-selected sequences

Figure 2(a) shows the aligned sequences of display-selected SGPI clones. The sequences are ranked so as to highlight any patterns related to the 10-26 core as described in the Figure legend.

### *Dominance of SGPI-2 related cores*

It is immediately apparent that clones with non wild-type core are abundant in the population. Nevertheless, the composition is not random. Among the 30 sequenced clones, there are only two with G26, while the random distribution would dictate a quarter of 30, around seven or eight. The remaining three amino acid types A, S and W, are distributed more evenly, although not equally (nine A, 13 S and six W). For cores with F10, we find nine F10-A26 (wild-type SGPI-2), 11 F10-S26, four F10-W26 and one F10-G26 pair. For cores with K10 we find only two K10-W26 (wild-type SGPI-1), two K10-S26, one K10-G26 and no K10-A26 pair. This suggests that selection for efficient folding favors the SGPI-2 type cores over the SGPI-1 type.

This can be explained as follows: for F10-A26, the wild type SGPI-2 core, the F10-S26 pair should be a proper substitute, as the small side-chains of Ala and Ser are often interchangeable. This was demonstrated in a recent phage-display work that compared the results of Ala-scanning and Ser-scanning of human growth hormone (hGH).[32] On the other hand, in the case of K10-W26, the wild-type SGPI-1 core, the large tryptophan has no alternative homolog in the tetranomial set. Thus, clones with the K10 residue are depleted. Interestingly, four clones with F10-W26 cores were also selected, suggesting that the volume of the core can be increased compared to that of the wild-type.

### *Dominant SGPI-1 residues*

From the two wild-type residue sets, display bias prefers the SGPI-1 version at four positions. At position 13 there is a threefold preference for the negatively charged D over the positively charged K. Note that the frequently selected non-wild-type E is also acidic; therefore, two-third of the clones here carry a negative charge. At position 18 there is a

**Figure 2.** Aligned sequences of binding-selected SGPI-1/SGPI-2 chimeras. The uppermost section shows the aligned sequences of SGPI-1 (yellow background) and SGPI-2 (blue background). Positions shared by both sequences have no background. For complete shuffling of the two wild-type sequences, the structure of the genetic code dictated the inclusion of two additional non-wild-type residue types at several positions. These are shown with grey background. The sequences are grouped in three blocks. In (a) the sequences were selected for efficient display based on binding to a monoclonal antibody through an invariant epitope tag. In (b) and (c), the sequences are from bovine trypsin and crayfish trypsin selection, respectively. The sequences are ranked to highlight any patterns related to the 10-26 core. We used a simple alphabetical arrangement corresponding to the one-letter code of amino acid residues. At position 10, sequences with phenylalanine (F) are listed first, followed by lysine (K). Then, when looking for the F10 sub-library, the residues are again in alphabetical order at position 26. The same is true for the K10 sub-library, and the same logic applies for the enzyme-selected populations. By this logic, the list starts with the F10-A26 wild-type SGPI-2 core, and ends with the K10-W26 wild-type SGPI-1 core. Between these two, we find non-wild-type core compositions. The wild-type cores are highlighted with a black frame in all three blocks. Bold and italic highlight the P1 residue that enters the binding pocket of the inhibited protease.

threefold preference for N over the large R. However, it is the non-wild-type H that dominates here, existing in more than half of the clones. Positions 20 and 22 are located at a surface loop. At position 20 there is a fourfold dominance of T over G, but the small non-wild-type residues A, or S can readily substitute for the slightly larger T. The negative selection against G might again be due to its destabilizing effect on the folded structure. At position 22 there is a twofold preference for T over D, but the non-wild-type N is the most frequent residue. T and N are of similar size and both polar residues can form H-bonds.

### *Dominant SGPI-2 residues*

There are five positions where display selection favored the SGPI-2 residue type. At positions 2 and 30, the non-polar V and L dominate over the polar Q or R, respectively. Both positions are surface-exposed, position 30 being the P1 residue of the inhibitor. Therefore, the dominance of hydrophobic residues might be unexpected. Nevertheless, a similar effect was found in a recent comprehensive saturation scanning of hGH.[33] At position five, half of the clones have a negatively charged E. Position 12 is also acidic due to the presence of about 50% wild-type D and 30% non-wild-type E. At position 32, which is the P2′ site of the inhibitor, there is a more than threefold dominance of A over G. The higher conformational freedom of G might lower the stability of the folded molecule. At position 32 we found three sporadic threonine residues (green background in Figure 2) that were not designed in the library. This is a result of a guanine to adenine mutation in the codon. As this segment of the library was made with primers representing the non-coding strand, the mutation could have occurred in the oligonucleotide in the form of a cytosine to thymine change due to oxidative deamination. Nevertheless, only these three undesired point mutations were detected in the entire pool of about 100 sequences.

## Analysis of the bovine trypsin-selected clones

### *Dominant SGPI-1 residues*

Figure 2(b) lists the clones selected on bovine trypsin. There are several striking differences here, compared to display selection. First of all, at position 30, the P1 position, the dominant L is invariantly replaced with an R from SGPI-1. This is in perfect accordance with the fact that the substrate-binding pocket of bovine trypsin requires a positively charged P1 side-chain for tight binding. Another interesting phenomenon is the apparent dominance of SGPI-1 character at positions 12, 13, 24 and 35. While position 12 was quite acidic in the display bias, two-third of the clones becomes neutral or slightly basic due to the abundance of a wild-type Q or a non-wild-type H. Position 13, where display selection preferred acidic residues, increased the proportion of these residues to 93% due to the further selection of a wild-type D, or a non-wild-type E. At position 24 the SGPI-1 specific gap dominates, as the SGPI-2 specific insertion of a lysine is eliminated selectively. As such an effect was not detected in the display selection, it is possible that this lysine is removed as a potential trypsin cleavage site. The almost exclusive presence of P35 suggests that this proline contributes to binding in a universal, core-independent manner.

### *Dominant SGPI-2 residues*

There are two positions of this type. At position 18 there is a tenfold preference for R over N, while the opposite trend was observed for display selection. The most abundant residue is still the non-wild-type H, as it was detected in the display selection. The other phenomenon is the preference for A over G at position 32, which was already detected in the display selection.

### *The dominance of the wild-type cores*

An important phenomenon not observed in the display selection is the emerging dominance of the wild-type cores. Here, 77% of the clones carry a wild-type core and, although more SGPI-2 cores are present, it is the SGPI-1 core that increased its proportion significantly compared to the display selected pool. About 20% of the clones have an F10-S26 core, which suggests that an A26S replacement is not deleterious for the functionality of the molecule. There is no F10-G26 or F10-W26 clone, and there is only one sporadic K10-A26 clone.

### *Core-specific amino acid preferences*

There are some positions with obvious core-specific amino acid preferences, suggesting that the core and these positions are functionally coupled. Once this phenomenon was observed, we decided to conduct a systematic pairwise covariance analysis for each residue pair and each selection as described in Materials and Methods. The findings of this analysis are described at the end of Results, but the two most striking core-dependent phenomena will be evaluated here and throughout the main text. The two sections presenting well-defined core-specific variations are the 20-22 turn and residue 32. At the 20-22 region, clones with the K10-W26 (SGPI-1) core clearly avoid residues P21 and T22, while those with either the F10-A26 (SGPI-2) core, or with the similar F10-S26 core are completely insensitive to the sequence of this segment.

This finding provides crucial information for the understanding of the cause of species specificity. It suggests that the K10-W26 core interacts with the TPT turn and positions it such that it hinders binding to bovine trypsin. It appears that the SGPI-2-like cores do not participate in a similar interaction.

At position 32, which is the P2′ position of the protease-binding loop, an SGPI-2 type A dominates over the SGPI-1 type G residue. However, the distribution of the G32 residues is clearly core-specific. While clones with SGPI-1 core show equal preference for A32 and G32, G32 is completely missing from clones with the SGPI-2 core. Although it is theoretically possible that the extra methyl group on A contributes to the protease binding at the P2′ position, it is more likely that instead of a positive selection for alanine, there is a negative selection against glycine. The higher conformational freedom of G32 could destabilize the protease-binding loop, while the K10-W26 core might somehow attenuate this effect.

## Analysis of the crayfish trypsin-selected clones

### *Dominant SGPI-1 residues*

Sequences from the crayfish trypsin selection are given in Figure 2(c). There are many similarities but also important differences here compared to the bovine enzyme selection. Similar to the bovine enzyme, the crayfish trypsin selects exclusively for a P1 R at position 30, slightly selects against the display-biased negative charge at position 12, and selects for a P at position 35.

A clear difference is the pronounced selection for T at positions 20 and 22, which will be discussed later. The selection against the SGPI-2 insertion K at position 24 observed for the bovine enzyme does not apply here, as the proportion of this K is the same as in the display-biased pool.

### *Dominant SGPI-2 residues*

A clear similarity to the bovine enzyme selection is the preference for R at position 18 and A at position 32. One difference is the fourfold selection for T over D at position 8, which was not observed in the display or in the bovine trypsin selection.

### *The dominance of the wild-type cores*

In this selection, the proportion of wild-type clones is 78%, practically identical with that found for the bovine enzyme. However, the proportions of various core types are different within both the wild-type and the non-wild-type categories. In the wild-type category there are three times as many SGPI-1 type as SGPI-2 type clones, suggesting that the K10-W26 core provides some advantage over the other wild-type form. In the non-wild-type category it is remarkable that, while bovine trypsin eliminated any F10-W26 clone, here 12% such variant is selected. The same proportion of F10-W26 cores was found in the display selection, suggesting that this core has no particular positive or negative effect on crayfish protease inhibition. This suggests that the crayfish protease might have greater plasticity at the interacting site compared to the bovine enzyme, as this protease can accommodate to the non-wild-type core.

### *Core-specific amino acid preferences*

The two enzymes have overlapping core-specific preferences at the 20-22 turn and at position 32, the two areas with the most obvious correlations. Notably, both enzymes present the core-dependent tolerance of a G at position 32, the P2′ site. Here, only the K10-W26 core is compatible with G32.

Although both enzymes show core-dependent preference at the 20-22 turn, there are important enzyme-specific differences. The bovine enzyme has no preference at position 20, and selectively eliminates the SGPI-1-type P21 and T22 residues from the K10-W26 core. The arthropod enzyme, on the other hand, eliminates the SGPI-2-type G and D from positions 20 and 22, respectively, while it has no preference at position 21. Nevertheless, the SGPI-1-type residues hardly exceed the 50% proportion at these three positions, suggesting that the 20-22 TPT turn is not a significant player in the efficient inhibition of crayfish trypsin.

In the bovine trypsin selection, the SGPI-2-like cores had no preference for any particular TPT sequence. Here, the number of wild-type SGPI-2 cores is so small that no trend could be observed. Nevertheless, it appears that clones with non-wild-type F10-S26 or F10-W26 cores prefer P at position 21.

## Constructing an SGPI variant that has a group I core but lacks taxon specificity

By using the information obtained from the three types of selections, we designed an SGPI-1 variant that was predicted to fold efficiently and to inhibit both the vertebrate and the arthropod enzyme. Starting with the wild-type SGPI-1 sequence, we aimed to minimize the number of mutations while maximizing the functional changes. An important guideline was to omit the observed preferences for non-wild-type residues and use only SGPI-2 residue types.

The following mutations were introduced from SGPI-2 into SGPI-1: T5E; N18R; T20G; P21S and T22D. Position 5 was chosen because it was apparent that display selection preferred E5 over T5, and the same trend was observed for clones with the K10-W26 core in the bovine trypsin selection. Therefore, we believe that this mutation somehow stabilizes the SGPI-1 fold. The N18R mutation was introduced on the basis of the notion that N18 is perfectly eliminated by the bovine enzyme selection and, although a non wild-type H is the most preferred, the second most abundant residue is R18. At the 20-22 turn it was clear that positions 21 and 22 should be switched to the SGPI-2 type based on their dominance over the SGPI-1 type. At position 20, the two types are equally represented,

but the SGPI-2 type G20 increases its frequency compared to the display selection, suggesting that it has a functional role in protease binding. The (T5E; N18R; 20-22 TPT→GSD) variant of SGPI-1 was chemically synthesized and named SGPI-1-PO-1 (standing for Schistocerca Gregaria Protease Inhibitor-1-Phage Optimized 1). The inhibitory constants for this variant were determined on bovine and crayfish trypsin, and the values were compared to those obtained previously for the wild-type SGPI-1 and SGPI-2 forms. The data and the corresponding peptide sequences are shown in Table 1.

On bovine trypsin the SGPI-1-PO-1 shows a 60-fold improvement over the parent SGPI-1 molecule, demonstrating that the predictions were correct. With this improvement, the variant acquired an inhibitory constant in the nanomolar range. What is more important, this SGPI-1 variant has practically the same inhibitory efficiency as the SGPI-2 L30R P1 mutant. Moreover, the picomolar range $K_i$ values of SGPI-1-PO-1 and SGPI-2 L30R are practically identical on crayfish trypsin. Therefore, we produced a group I core variant, which is isofunctional with a group II core variant on both enzymes. As SGPI-1-PO-1 lost its dramatic taxon specificity, we can state that the group I core cannot by itself be responsible for taxon specificity.

### *A P1′ mutation further improves bovine trypsin inhibition*

It was known that bovine trypsin prefers methionine over lysine at the P1′ position; thus, we synthesized a P1′ K to M mutant of SGPI-1-PO-1 and named it SGPI-1-PO-2. The $K_i$ values for this variant were also determined on both trypsins, and the data are shown in Table 1.

It was previously shown that in the SGPI-2 L30R framework the P1′ K31M mutation boosted the bovine trypsin inhibitory efficiency over 1100-fold, while in the SGPI-1 framework it had only a moderate sevenfold effect. In the context of SGPI-1-PO-1, this mutation causes a 175-fold improvement. There is a sixfold difference (1100-fold *versus* 175-fold) between the effects of the same P1′ mutation on the interaction of SGPI-2 L30R and SGPI-1-PO-1 with bovine trypsin. This suggests that, in spite of having the same $K_i$ on bovine trypsin, there might be subtle differences in the binding mechanisms of these two inhibitor variants. The same conclusion might apply for the interaction with crayfish trypsin as, while the P1′ mutation decreased the inhibitory efficiency of SGPI-1-PO-1 about twofold, the same mutation actually increased the efficiency of SGPI-2 L30R with the same magnitude. Nevertheless, as these effects are small, even high-resolution structures might not shed light on the underlying causes.

## Chimera mutagenesis detects complex networks of residue interactions

Once it became evident that residues are not selected independently of one another, we decided to conduct a systematic pairwise covariance analysis for each residue pair and each selection as described in Materials and Methods. The pairwise covariance values and the corresponding $P$ values for statistical significance are shown in Figure 3, while the complex networks based on the statistically significant residue pair co-variations

**Table 1.** Comparative inhibitory data of SGPI variants on bovine and crayfish trypsins and the corresponding peptide sequences

| Inhibitor variant | P1-P1′ | Core | Bovine trypsin | | Crayfish trypsin | |
|---|---|---|---|---|---|---|
| | | | $K_i$ (pM) | $K_{i\ SGPI-1}/K_{i\ variant}$ | $K_i$ (pM) | $K_{i\ SGPI-1}/K_{i\ variant}$ |
| [a]SGPI-1-PO-1 | R-K | K-W | 3500 | 60 | 2.6 | 1.2 |
| [a]SGPI-1-PO-2 | R-M | K-W | 20 | 10,500 | 6.4 | 0.5 |
| SGPI-2-L30R | R-K | F-A | 5500 | 38 | 2.0 | 1.5 |
| SGPI-2-L30R/K31M | R-M | F-A | 5 | 42,000 | 1.2 | 2.5 |
| SGPI-1 | R-K | K-W | 210,000 | 1.0 | 3.0 | 1.0 |
| SGPI-1-K31M | R-M | K-W | 30,000 | 7.0 | 5.0 | 0.6 |

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SGPI-1-PO-1 | E | Q | E | C | E | P | G | Q | T | K | K | Q | D | C | N | T | C | R | C | G | S | D | G | - | V | W | A | C | T | R | K | G | C | P | P | H |
| SGPI-1-PO-2 | E | Q | E | C | E | P | G | Q | T | K | K | Q | D | C | N | T | C | R | C | G | S | D | G | - | V | W | A | C | T | R | M | G | C | P | P | H |
| SGPI-2,L30R | E | V | T | C | E | P | G | T | T | F | K | D | K | C | N | T | C | R | C | G | S | D | G | K | S | A | A | C | T | R | K | A | C | P | Q | - |
| SGPI-2,L30R,K31M | E | V | T | C | E | P | G | T | T | F | K | D | K | C | N | T | C | R | C | G | S | D | G | K | S | A | A | C | T | R | M | A | C | P | Q | - |
| SGPI-1 | E | Q | E | C | T | P | G | Q | T | K | K | Q | D | C | N | T | C | N | C | T | P | T | G | - | V | W | A | C | T | R | K | G | C | P | P | H |
| SGPI-1,K31M | E | Q | E | C | T | P | G | Q | T | K | K | Q | D | C | N | T | C | N | C | T | P | T | G | - | V | W | A | C | T | R | M | G | C | P | P | H |

For clarity, the sequences of the inhibitor variants are color-coded: SGPI-1, orange; SGPI-2, cyan. Residues shared by both types are not specified here but are indicated in Figure 2. A P1′ residue type, M not present in wild-type SGPI-1 or SGPI-2 is highlighted with magenta background. Note that we use consensus numbering.

[a] The names, SGPI-1-PO-1 and SGPI-1-PO-2, stand for the two phage-optimized variants, Schistocerca Gregaria Protease Inhibitor-1-Phage Optimized-1 and 2. Equilibrium inhibition constant ($K_i$) values for these two forms were determined on bovine and crayfish trypsin and compared to those previously published for wild-type and P1-P1′ mutant SGPI variants[20,24]. Comparative data related to the phage-optimized variants are highlighted as bold. As the parent molecule of the phage optimized variants is wild-type SGPI-1, all $K_i$ values are compared to those obtained on this variant.

**(a)**

| | 2 | 3 | 5 | 8 | 10 | 12 | 13 | 18 | 20 | 21 | 22 | 24 | 25 | 26 | 30 | 32 | 35 | 36 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2** | | 0.78 | 0.03 | 0.26 | 0.60 | 0.65 | 0.50 | 0.68 | 0.53 | 0.38 | 0.58 | 0.21 | 0.21 | 0.11 | 1.00 | 0.91 | 0.03 | 0.71 |
| **3** | 0.21 | | 0.44 | 0.99 | 0.11 | 0.37 | 0.51 | 0.24 | 0.90 | 0.13 | 0.15 | 0.03 | 0.69 | 0.09 | 1.00 | 0.06 | 0.94 | 0.62 |
| **5** | 0.79 | 0.46 | | 0.17 | 0.30 | 0.68 | 0.03 | 0.88 | 0.20 | 0.15 | 0.34 | 0.21 | 0.22 | 0.55 | 0.62 | 0.39 | 0.06 | 0.30 |
| **8** | 0.43 | 0.09 | 0.67 | | 0.04 | 0.26 | 0.93 | 0.61 | 0.99 | 0.34 | 0.67 | 0.85 | 0.86 | 0.21 | 0.09 | 0.43 | 0.22 | 0.62 |
| **10** | 0.09 | 0.40 | 0.26 | 0.58 | | 0.83 | 0.77 | 0.18 | 0.82 | 0.62 | 0.46 | 1.00 | 1.00 | 0.32 | 1.00 | 0.58 | 0.13 | 1.00 |
| **12** | 0.26 | 0.49 | 0.33 | 0.55 | 0.07 | | 0.49 | 0.70 | 0.57 | 0.02 | 0.77 | 0.64 | 0.46 | 0.33 | 1.00 | 0.65 | 1.00 | 0.12 |
| **13** | 0.34 | 0.44 | 1.10 | 0.19 | 0.10 | 0.45 | | 0.12 | 0.25 | 0.70 | 0.37 | 0.80 | 0.83 | 0.36 | 0.77 | 0.90 | 0.07 | 0.36 |
| **18** | 0.22 | 0.54 | 0.21 | 0.31 | 0.30 | 0.28 | 0.72 | | 0.19 | 1.00 | 0.97 | 0.23 | 0.89 | 0.17 | 0.77 | 0.10 | 0.10 | 1.00 |
| **20** | 0.32 | 0.20 | 0.66 | 0.11 | 0.07 | 0.38 | 0.63 | 0.60 | | 0.32 | 0.86 | 0.14 | 0.34 | 0.09 | 0.59 | 0.66 | 0.61 | 1.00 |
| **21** | 0.26 | 0.67 | 0.62 | 0.39 | 0.13 | 1.17 | 0.17 | 0.02 | 0.40 | | 0.06 | 0.47 | 0.10 | 0.77 | 1.00 | 0.71 | 1.00 | 0.27 |
| **22** | 0.30 | 0.72 | 0.55 | 0.33 | 0.18 | 0.29 | 0.55 | 0.13 | 0.24 | 0.91 | | 0.55 | 0.52 | 0.82 | 0.14 | 0.58 | 0.91 | 0.07 |
| **24** | 0.39 | 1.09 | 0.51 | 0.10 | 0.00 | 0.18 | 0.14 | 0.45 | 0.67 | 0.19 | 0.21 | | 0.72 | 0.57 | 0.49 | 0.24 | 0.46 | 0.73 |
| **25** | 0.51 | 0.31 | 0.66 | 0.21 | 0.03 | 0.45 | 0.27 | 0.18 | 0.54 | 0.80 | 0.42 | 0.13 | | 0.46 | 0.77 | 0.56 | 0.32 | 0.16 |
| **26** | 0.61 | 0.83 | 0.41 | 0.62 | 0.25 | 0.52 | 0.54 | 0.62 | 0.84 | 0.13 | 0.26 | 0.21 | 0.45 | | 0.13 | 0.76 | 0.90 | 0.89 |
| **30** | 0.04 | 0.03 | 0.06 | 0.24 | 0.02 | 0.02 | 0.06 | 0.04 | 0.09 | 0.01 | 0.17 | 0.10 | 0.05 | 0.19 | | 1.00 | 0.50 | 0.49 |
| **32** | 0.09 | 0.69 | 0.34 | 0.30 | 0.10 | 0.21 | 0.14 | 0.55 | 0.22 | 0.06 | 0.26 | 0.38 | 0.24 | 0.17 | 0.04 | | 0.07 | 0.61 |
| **35** | 0.87 | 0.05 | 0.86 | 0.47 | 0.53 | 0.05 | 0.91 | 0.66 | 0.24 | 0.01 | 0.11 | 0.19 | 0.42 | 0.08 | 0.09 | 0.68 | | 0.26 |
| **36** | 0.13 | 0.21 | 0.43 | 0.22 | 0.00 | 0.64 | 0.40 | 0.02 | 0.02 | 0.43 | 0.89 | 0.05 | 0.66 | 0.07 | 0.10 | 0.12 | 0.43 | |

**(b)**

| | 2 | 3 | 5 | 8 | 10 | 12 | 13 | 18 | 20 | 21 | 22 | 24 | 25 | 26 | 30 | 32 | 35 | 36 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2** | | 0.59 | 0.72 | 0.98 | 0.63 | 0.54 | 0.49 | 0.96 | 0.92 | 0.54 | 0.41 | 1.00 | 0.82 | 0.95 | 1.00 | 0.85 | 0.52 | 0.21 |
| **3** | 0.34 | | 0.42 | 0.17 | 0.00 | 0.07 | 0.55 | 0.87 | 0.70 | 0.14 | 0.10 | 0.49 | 0.68 | 0.00 | 1.00 | 0.04 | 1.00 | 0.21 |
| **5** | 0.32 | 0.46 | | 0.18 | 0.05 | 0.92 | 0.60 | 0.16 | 0.40 | 0.48 | 0.04 | 0.27 | 0.55 | 0.04 | 1.00 | 0.14 | 0.58 | 0.66 |
| **8** | 0.10 | 0.65 | 0.72 | | 0.53 | 0.29 | 0.50 | 0.26 | 0.66 | 0.58 | 0.89 | 0.74 | 0.61 | 0.52 | 1.00 | 0.71 | 0.30 | 0.60 |
| **10** | 0.21 | 1.70 | 0.90 | 0.23 | | 0.11 | 0.48 | 0.14 | 0.93 | 0.13 | 0.11 | 1.00 | 0.89 | 0.00 | 1.00 | 0.01 | 1.00 | 0.49 |
| **12** | 0.37 | 0.78 | 0.20 | 0.53 | 0.61 | | 0.25 | 0.43 | 0.39 | 0.47 | 0.28 | 0.02 | 0.58 | 0.12 | 1.00 | 0.31 | 1.00 | 0.40 |
| **13** | 0.29 | 0.26 | 0.27 | 0.30 | 0.17 | 0.46 | | 0.62 | 0.03 | 0.02 | 0.33 | 0.17 | 0.77 | 0.04 | 1.00 | 0.53 | 1.00 | 0.72 |
| **18** | 0.09 | 0.13 | 0.61 | 0.46 | 0.49 | 0.32 | 0.16 | | 0.96 | 0.56 | 0.00 | 1.00 | 0.70 | 0.20 | 1.00 | 0.01 | 1.00 | 0.34 |
| **20** | 0.15 | 0.26 | 0.48 | 0.27 | 0.07 | 0.43 | 1.05 | 0.09 | | 0.93 | 0.71 | 0.16 | 0.42 | 0.50 | 1.00 | 0.56 | 1.00 | 0.95 |
| **21** | 0.22 | 0.55 | 0.28 | 0.23 | 0.66 | 0.25 | 1.13 | 0.14 | 0.06 | | 0.64 | 1.00 | 0.52 | 0.08 | 1.00 | 0.63 | 1.00 | 0.72 |
| **22** | 0.48 | 0.77 | 1.06 | 0.18 | 0.73 | 0.56 | 0.43 | 1.66 | 0.29 | 0.22 | | 0.09 | 0.54 | 0.27 | 1.00 | 0.74 | 1.00 | 0.04 |
| **24** | 0.02 | 0.07 | 0.12 | 0.04 | 0.01 | 0.28 | 0.14 | 0.01 | 0.13 | 0.01 | 0.19 | | 0.48 | 0.35 | 1.00 | 1.00 | 1.00 | 1.00 |
| **25** | 0.24 | 0.30 | 0.41 | 0.33 | 0.08 | 0.35 | 0.17 | 0.20 | 0.42 | 0.24 | 0.40 | 0.08 | | 0.78 | 1.00 | 0.31 | 1.00 | 0.19 |
| **26** | 0.13 | 1.77 | 1.03 | 0.35 | 4.54 | 0.71 | 0.97 | 0.51 | 0.35 | 0.86 | 0.59 | 0.12 | 0.22 | | 1.00 | 0.00 | 1.00 | 0.26 |
| **30** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 1.00 | 1.00 | 1.00 |
| **32** | 0.06 | 0.48 | 0.37 | 0.09 | 1.21 | 0.25 | 0.12 | 0.82 | 0.13 | 0.08 | 0.08 | 0.01 | 0.23 | 1.06 | 0.00 | | 1.00 | 0.38 |
| **35** | 0.05 | 0.02 | 0.05 | 0.05 | 0.02 | 0.02 | 0.02 | 0.02 | 0.03 | 0.02 | 0.04 | 0.00 | 0.02 | 0.03 | 0.00 | 0.00 | | 0.43 |
| **36** | 0.49 | 0.49 | 0.21 | 0.17 | 0.14 | 0.31 | 0.07 | 0.29 | 0.05 | 0.05 | 1.02 | 0.00 | 0.51 | 0.47 | 0.00 | 0.16 | 0.04 | |

**(c)**

| | 2 | 3 | 5 | 8 | 10 | 12 | 13 | 18 | 20 | 21 | 22 | 24 | 25 | 26 | 30 | 32 | 35 | 36 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2** | | 0.38 | 1.00 | 0.10 | 0.55 | 0.83 | 0.45 | 0.29 | 0.72 | 0.69 | 0.98 | 0.41 | 0.09 | 0.45 | 1.00 | 0.86 | 0.07 | 0.87 |
| **3** | 0.46 | | 0.03 | 0.05 | 0.38 | 0.33 | 0.01 | 0.37 | 0.78 | 0.59 | 0.80 | 0.06 | 0.79 | 0.81 | 1.00 | 0.47 | 0.44 | 0.37 |
| **5** | 0.10 | 1.06 | | 0.15 | 0.26 | 0.03 | 0.30 | 0.38 | 0.96 | 0.24 | 0.23 | 0.73 | 0.73 | 0.32 | 1.00 | 0.10 | 0.91 | 0.38 |
| **8** | 0.74 | 0.92 | 0.69 | | 0.64 | 0.84 | 0.28 | 0.55 | 0.20 | 1.00 | 0.40 | 0.78 | 0.04 | 0.69 | 1.00 | 0.56 | 0.53 | 0.31 |
| **10** | 0.22 | 0.32 | 0.48 | 0.13 | | 0.83 | 0.66 | 1.00 | 0.10 | 0.13 | 0.17 | 0.27 | 0.53 | 0.00 | 1.00 | 0.26 | 1.00 | 0.04 |
| **12** | 0.25 | 0.53 | 1.10 | 0.19 | 0.11 | | 0.21 | 0.30 | 0.10 | 0.74 | 0.65 | 0.49 | 0.70 | 0.94 | 1.00 | 1.00 | 0.01 | 0.52 |
| **13** | 0.50 | 1.32 | 0.64 | 0.55 | 0.19 | 0.72 | | 0.40 | 0.83 | 0.09 | 0.99 | 0.76 | 0.84 | 0.94 | 1.00 | 0.15 | 0.23 | 1.00 |
| **18** | 0.48 | 0.42 | 0.47 | 0.24 | 0.00 | 0.50 | 0.48 | | 0.20 | 0.86 | 0.34 | 0.86 | 0.88 | 0.79 | 1.00 | 0.18 | 0.86 | 0.43 |
| **20** | 0.30 | 0.25 | 0.17 | 0.59 | 0.74 | 0.83 | 0.27 | 0.60 | | 0.11 | 0.60 | 0.43 | 0.13 | 0.47 | 1.00 | 1.00 | 0.84 | 0.41 |
| **21** | 0.18 | 0.23 | 0.50 | 0.01 | 0.66 | 0.15 | 0.77 | 0.12 | 0.66 | | 0.58 | 1.00 | 1.00 | 0.83 | 1.00 | 1.00 | 1.00 | 1.00 |
| **22** | 0.11 | 0.22 | 0.64 | 0.39 | 0.55 | 0.33 | 0.12 | 0.44 | 0.34 | 0.23 | | 0.43 | 0.74 | 0.02 | 1.00 | 0.81 | 0.85 | 0.02 |
| **24** | 0.30 | 0.80 | 0.18 | 0.09 | 0.34 | 0.27 | 0.16 | 0.12 | 0.30 | 0.01 | 0.27 | | 0.17 | 1.00 | 1.00 | 0.53 | 1.00 | 1.00 |
| **25** | 0.86 | 0.27 | 0.36 | 0.99 | 0.26 | 0.36 | 0.30 | 0.15 | 0.78 | 0.04 | 0.29 | 0.57 | | 0.10 | 1.00 | 0.85 | 0.64 | 0.33 |
| **26** | 0.28 | 0.13 | 0.37 | 0.15 | 2.81 | 0.12 | 0.11 | 0.11 | 0.27 | 0.12 | 0.85 | 0.00 | 0.58 | | 1.00 | 0.54 | 0.74 | 0.04 |
| **30** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 1.00 | 1.00 | 1.00 |
| **32** | 0.04 | 0.10 | 0.26 | 0.08 | 0.17 | 0.00 | 0.20 | 0.28 | 0.02 | 0.00 | 0.06 | 0.12 | 0.05 | 0.06 | 0.00 | | 1.00 | 0.24 |
| **35** | 0.44 | 0.19 | 0.09 | 0.14 | 0.00 | 0.64 | 0.27 | 0.05 | 0.05 | 0.01 | 0.05 | 0.01 | 0.10 | 0.04 | 0.00 | 0.03 | | 1.00 |
| **36** | 0.08 | 0.33 | 0.40 | 0.35 | 1.16 | 0.25 | 0.04 | 0.24 | 0.33 | 0.00 | 1.22 | 0.00 | 0.42 | 0.85 | 0.00 | 0.23 | 0.01 | |

**Figure 3.** Pairwise covariance of positions detected in binding selected SGPI pools. The covariance for each randomized position pair was analyzed using the formula described in Materials and Methods. The lower quadrant shows the calculated covariance scores. The upper quadrant shows the *P*-values for the null-hypothesis that the two positions vary independently. The *P*-values were calculated on the basis of a bootstrap analysis, as described in Materials and Methods. A *P*-value of 0.02 means,= that there is only 2% probability for the observed covariance scores being due to stochastic error, e.g. random sampling. The following cutoff values were applied: $P < 0.02$, red; $0.02 \le P < 0.06$, orange; $0.06 \le P < 0.1$, green. These colors are used as the background for each residue pair in both quadrants to highlight the level of functional dependence of pairs. Blocks in (a)–(c) correspond to display-selected, bovine trypsin-selected and crayfish trypsin-selected populations, respectively. The networks of residue pairs are illustrated in Figure 4.

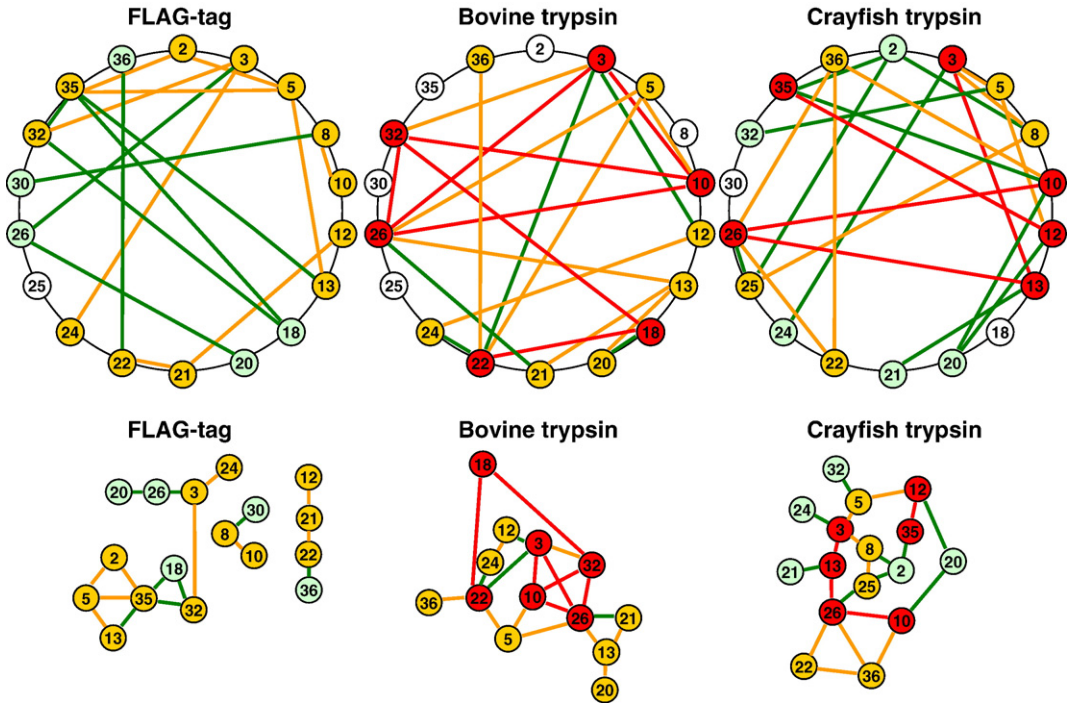


**Figure 4.** Networks of functional dependences based on covariance of residue pairs. Using the covariance values and coloring scheme from Figure 3, the networks of suggested functional dependence were drawn using two schemes. In the upper section, the positions are shown in a circular arrangement so as to easily locate identical connections from the three selections: (a) for folding efficiency, (b) for bovine trypsin binding and (c) for crayfish trypsin binding. In the lower section, the networks are arranged such that the number of crossing lines would be minimized. This arrangement better illustrates the structure of the network. The coloring scheme for the lines is the same as in Figure 3. The positions are colored according to the highest ranked connection they make.

are shown in Figure 4. Note that no information can be obtained for interactions that involve non-randomized positions.

It was readily apparent that results on the protease-selected clones nicely corroborated the structural information about the interaction between residues 10 and 26. Notably, the display selection did not detect a similarly strong correlation between these two positions. While the composition of the 10-26 pair is not fully random, its covariance in display selection is not statistically significant. It suggests that many residue pairs allow for a native-like structure, but proteases select for more stable scaffolds afforded only by a more limited set of residue pairs.

Besides the 10-26 core, there are signs of covariance between numerous other positions. So far we have mentioned only those that are coupled to the core, and we believe these are the most important ones. However, there are many other pairs that require further experimental studies to be interpreted. Nevertheless, we can state, at least at a descriptive level, that display selection provokes a widely distributed and loosely organized network of residue interactions, in which two smaller groups of residues are detached from the rest of the network (Figure 4). Although this network is large, covering all but two randomized positions, none of the interactions is highly significant.

The two enzymes, on the contrary, select for tight networks of residue interactions, in which many interactions have very high significance levels. From the two enzyme selections, the one on the bovine protease appears to be slightly more stringent, as the numbers and significance levels of co-varying pairs are higher for the vertebrate enzyme. Also, the bovine network is more compact, as 13 of the randomized 18 positions participate in the network, while this number is 16in the crayfish trypsin selection. There are only three positions, 3, 10 and 26, that participate in highly significant interactions in both enzyme selections.

The complex nature of the enzyme-selected networks fits nicely to the results of a recent NMR titration experiment, which revealed that local binding provokes long-range structural and dynamic changes in SGPI-2.[34]

## Conclusions

As it was recently reviewed, essentially three types of hypotheses existed to explain the taxon specificity phenomenon.

### The nature of the P1′ residue

The P1′ residue of SGPI-1 is a lysine (K). This K is suboptimal for binding to vertebrate trypsins, because these enzymes usually have a K at a

contacting surface loop (the so-called 60's loop) and the two positively charged lysine residues repulse one another. This hypothesis turned out to be partially right, as a K to M mutation increases the efficiency of pacifastin inhibitors on vertebrate enzymes. However, this effect was small at the group I inhibitors, suggesting that the major factor is located elsewhere.[20,24]

### Rigidity differences between inhibitors of the group I or group II folds

Hydrogen/deuterium exchange and NMR studies have investigated the dynamics of group I and group II inhibitors.[16,17,34,35] In a dynamics study on recombinant $^{15}$N-labeled inhibitors, we found that, compared to SGPI-1, SGPI-2 exhibits enhanced motions on the micro- to millisecond timescale.[17,35] This led to the hypothesis that the relatively high rigidity of SGPI-1 might hinder its steric accommodation to the vertebrate enzyme while not affecting its binding to the arthropod enzyme.[24] For this hypothesis, we assumed that the vertebrate enzymes having larger number of disulfides would be more rigid than the arthropod proteases. The issue is more complicated, as a post-translational fucosylation exists in natural SGPI-2 and LMPI-2, which, at least in the latter inhibitor, provides extra rigidity to the molecule.[16]

### The difference in the P6-P10 loop

The third potential source of taxon specificity was considered to be the difference of sequence and conformation of the P6-P10 loops (residues 20–25) in the group I and group II paralogs. SGPI-1 and its homologous peptides have P at position 21, where there is S in the SGPI-2 homologs. On the basis of the crystal structure of the complex formed between LMPI-2 and bovine chymotrypsin, this P21 was close to G173 in the enzyme.[36] Modeling experiments suggested that in the analogous LMPI-1 vertebrate trypsin complexes, this proline would clash with P173 in these enzymes. Group I LMPI peptides, on the other hand, were good inhibitors of a trypsin from the fungus *Fusarium oxysporum*, which lacked proline at this segment.[18] These data logically pointed towards the steric clash model. However, as the fungal enzyme is a very distant relative of the vertebrate proteases, only cautious conclusions could be drawn about the exact functional differences between these enzymes at selected areas.

Our comprehensive phage display approach pinpointed the cause of taxon specificity in the pacifastin family. We localized it as an incompatibility of the K10-W26 core with the TPT turn for vertebrate trypsin inhibition. The profoundly different positions of the turns in group I and group II inhibitors are observed both in the protease-bound and in the free forms, as illustrated in Figure 5. Apparently, the K10-W26 core positions the TPT turn in an orientation in which a steric clash between P21 and the surface of bovine trypsin hinders the interaction. Therefore, our experimental results corroborate the third hypothesis.

The point mutations in our SGPI variant are restricted to surface residues; therefore, one might think that the dynamic properties of this variant should be similar to those measured for SGPI-1. If this was the case, rigidity would not play a major role in taxon specificity. Nevertheless, our results do not rule out dynamics as a possible factor. The dynamics of our mutants need to be investigated experimentally to answer this question.

Recently, our group presented an exceptionally high-resolution structure for the complex between SGPI-1 and crayfish trypsin.[37] The structure showed that the interaction extends significantly beyond the typical P3-P3′ region, as additional contact sites were found at the P12-P4 and P4′-P5′ segments. On the basis of *in silico* binding energy calculations, we suggested that this extended binding site, including the 20-22 TPT turn, contributes to the higher affinity of SGPI-1 to the crayfish *versus* the bovine enzyme.

While high-resolution structures contribute immensely to our understanding of molecular interactions, it remains valid that structural epitopes and functional epitopes are not identical.[38] The X-ray structure between SGPI-1 and crayfish trypsin portrayed an impressive extended binding site. However, it turned out that a large portion of these interactions, the one through N18 and those through the TPT turn, can be readily replaced with very different interactions (an R18 residue and a GSD segment) without affecting the binding affinity. One important conclusion is that only well-designed mutagenesis experiments can identify which structural interactions contribute to binding, i.e. which parts of the structural epitope constitute the functional epitope.

An important issue that requires further study is the effect of the minimalist core on the functional properties of the inhibitor. On the basis of an extremely thorough and large-scale investigation, universal rules have been proposed to exist for the canonical serine protease inhibitors.[39] It was suggested that in naturally occurring canonical inhibitors the surface residues, and especially those located at the binding loop, have an almost exclusive impact on the affinity of the inhibitor to the proteases. A sequence to reactivity algorithm was established on the basis of an immense number of measurements. An important model emerged, in which individual binding loop positions contribute to the overall binding energy in an additive fashion. While most measurements were done on Kazal-type inhibitors, it was demonstrated that findings on a Kazal-type scaffold are about 70% valid on the unrelated eglin C scaffold.[40] Based on similar studies that included BPTI, a Kunitz-type inhibitor, the authors concluded that interscaffolding additivity may not be perfectly universal, as it might depend on the sequential and conformational similarities of the two inhibitory regions being compared.

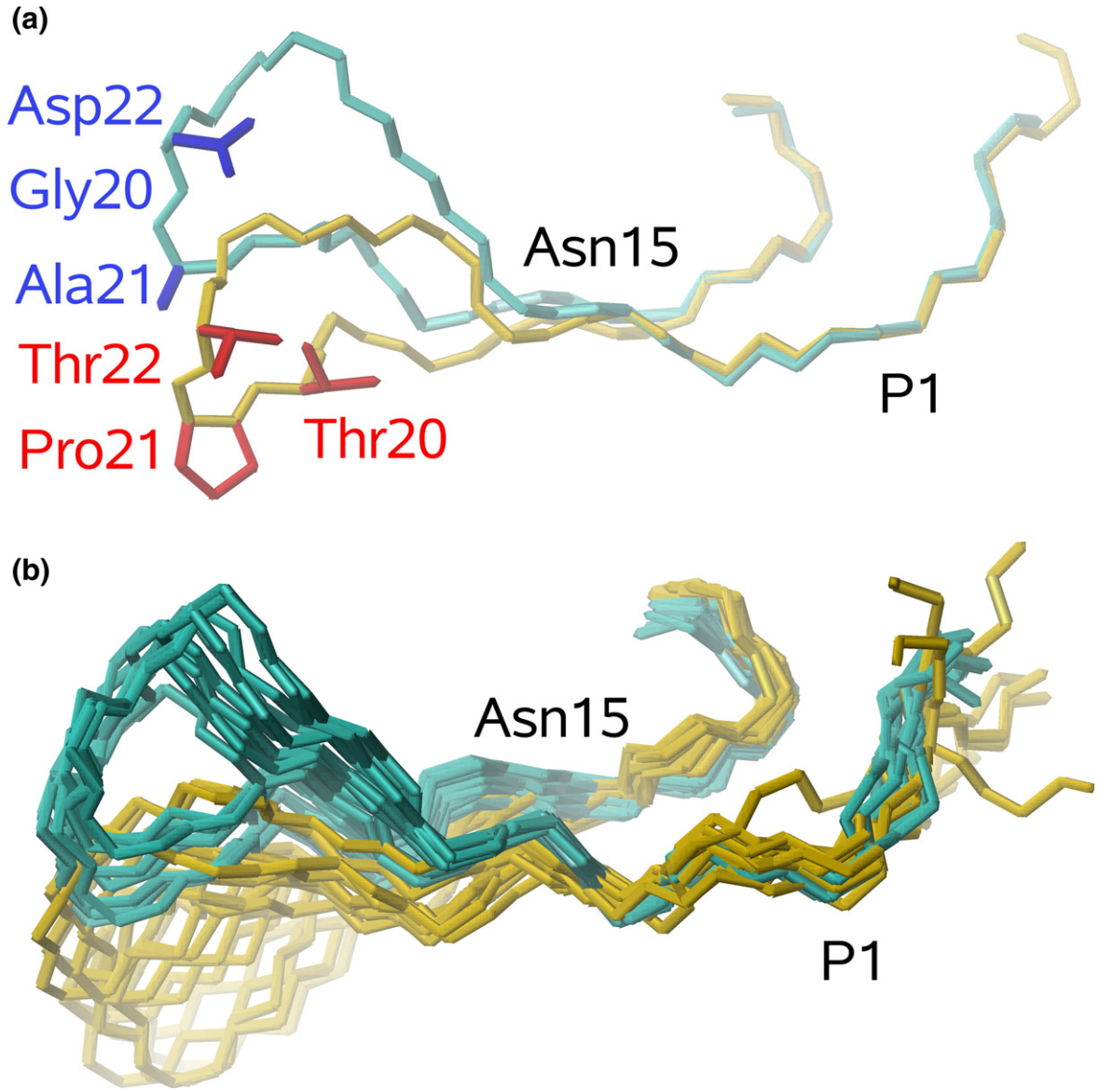


**Figure 5.** Structural comparisons of the 20-22 turns from group I and group II Pacifastin inhibitors. Structures of inhibitors in (a) are from protease–inhibitor complexes and were solved by X-ray crystallography. The free structures of inhibitors in (b) were solved by NMR. For clarity, only the 13-36 segment is shown. (a) Structural comparison of the group I type SGPI-1 (SGTI, colored orange, PDB ID 2F91, chain B) and the group II type LMPI-2 (PMP-C, colored blue, PDB ID 1GL1, chain I). Structures were superimposed at the backbone atoms of the strictly conserved segment 13–16 and residues of the protease binding loop 28-33 (consensus numbering). Positions of the strictly conserved Asn15 and the P1 residue (position 30) on the binding loop, as well as side-chains on the 20-22 turn are indicated. Note that LMPI-2 has alanine at position 21. This residue is serine in SGPI-2. (b) Structural comparison of the free forms of SGPI-1 (SGTI, colored orange, PDB ID 1KGM) and SGPI-2 (SGCI, colored blue, PDB ID 1KJ0) using the 10-conformer ensembles deposited in the PDB. Structures are superimposed as described above. Positions of the conserved Asn15 and the P1 residue are indicated. The structures were superimposed and the Figures were generated with MOLMOL.[49]

We found an important example for this refined model as, through measuring the effects of the P1′ K31M mutation, we detected non-additive effects even between two closely related inhibitors that belong to the same inhibitory family. The network of co-varying residues also supports this notion. The effects found in our system originate mainly from differences between the two types of minimalist cores of these peptides. On the basis of these results and those mentioned above for the larger Kazal and Kunitz-type inhibitors, we argue that the smaller the core is, the larger its effect might be on the overall structural–functional behavior of the inhibitor molecule.

The pacifastin family motifs represent a typical size range that exists at the borderline between marginally stable peptides and small, but stable globular proteins. While the native structure of a

globular protein is organized around and stabilized by a densely packed non-polar inner core, a 35 residue peptide cannot afford a similar core simply because of topology reasons. These peptides have a minimalist core, but the stable native structure requires strategically positioned disulfides. Moreover, while the distinction of "core" and "surface" is a well-established paradigm for globular proteins, it might not be as feasible for these motifs.

It is the first time that inhibitors from the pacifastin family have been functionally displayed on phage. We believe that this achievement opens new dimensions for structure–function studies within this interesting family of inhibitors. Combinatorial mutagenesis approaches combined with structure–function studies on individual mutants should soon build a rich knowledge base for better understanding the properties of these molecules. Small size, high stability and high potency of these inhibitors render them an excellent system for macromolecular interaction analysis. Moreover, these molecules could serve as starting points for protease-inhibiting drug design.

## Materials and Methods

### Materials and bacterial strains

The DNA-modifying enzymes were from Fermentas and New England Biolabs. Oligonucleotides were from Integrated DNA Technology. All common laboratory reagents were from Sigma Aldrich. DNA isolation was done using the appropriate kits from QIAgen. DNA sequencing reactions were done by using the ABI PRISM BigDye v3.0 Kit (Applied Biosystems), and the reactions were run on an ABI PRISM 3700 equipment (Applied Biosystems). Bovine trypsin (TPCK-treated) and bovine chymotrypsin (TLCK-treated) were from Sigma Aldrich, while the crayfish enzyme was isolated as described.[37,41] The Anti FLAG-tag antibody was from Affinity BioReagents. MaxiSorp plates from Nunc International were used for target immobilization. The bacterial strains XL1-Blue, CJ236 and SS320 were from Stratagene, New England Biolabs and Genentech, respectively. The M13-KO7 helper phage was from New England Biolabs.

### The Tag-pGP8 phagemid construction

In order to express all possible SGPI-1 and SGPI-2 chimeras on phage, first we had to construct a phagemid vector that would allow for the monovalent display of properly folded, functional wild-type SGPI-1 and SGPI-2 on the surface of M13 phage. We started with the pS1607 phagemid vector developed at Genentech for the monovalent display of hGH fused to p8, the major coat protein of the M13 phage.[42] The vector had to be modified slightly for our purpose. We deleted a unique XhoI restriction site by opening the vector with XhoI at position 1864 and with SalI at position 1625. The two enzymes create identical sticky ends but, after vector religation, neither of them cleaves the ligated sequence. Then we introduced an adaptor containing a new unique XhoI site between existing unique NsiI and BglII sites using the following oligonucleotides: 5′ TCGGATTATAAAGACGATGATGACAAACTCGAGA3′ and 5′GATCTCTCGAGTTTGTCATCATCGTCTTTATAATCCGATGCA3′ the XhoI site is underlined. The adaptor replaced the first half of the hGH gene with a DNA segment coding for a small linear epitope, (DYKDDDDK) a FLAG-tag. The N terminus of the tag became fused to the C terminus of the PhoA periplasmic signal peptide. We also introduced an Acc65I (KpnI) site upstream from a Ser-Gly coding linker region that connects the recombinant peptides to the N terminus of the p8 coat protein. This site (underlined) was introduced through Kunkel mutagenesis with the mutagenesis primer:[43] 5′ GGCAGCTGTGGCTTCGGTACCGGTGGAGGATCCGG3′ from IDT. Then, the XhoI site of the linker and the Acc65I site were used for directional cloning of the synthetic wild-type SGPI-1 and SGPI-2 genes into the vector.

### Synthesis of wild-typeSGPI-1 and SGPI-2 genes and cloning into the pGP8 phagemid

Wild-type SGPI-1 and SGPI-2 genes were constructed through pairs of synthetic primers that had overlapping complementary 3′ termini (Figure 1). The following oligonucleotides were ordered from IDT:

*WT-SGPI-1-forward*: 5′GTG TGT *CTC GAG* CAG GAA TGC ACC CCG GGT CAG ACC AAA AAA CAG GAT TGC AAC ACC TGC AAC TGC ACC C3′
*WT-SGPI-1-reverse*: 5′GTG TGT *GGT ACC* ATG CGG CGG GCA GCC TTT GCG GGT GCA CGC CCA CAC GCC GGT CGG GGT GCA GTT GCA GG3′
*WT-SGPI-2-forward*: 5′GTG TGT *CTC GAG* GTG ACC TGC GAA CCG GGT ACG ACC TTT AAA GAT AAA TGC AAC ACC TGC CGT TGC GG3′
*WT-SGPI-2-reverse*: 5′GTG TGT *GGT ACC* CTG CGG GCA CGC TTT CAG GGT GCA CGC CGC GCT TTT GCC ATC GCT ACC GCA ACG GCA GG3′

The forward and reverse primer pairs having complementary 3′ termini were annealed to each other and extended to generate a double-stranded DNA product using a single PCR cycle. The two products were digested with XhoI and Acc65I, purified on a GenElute PCR Clean Up column (QIAgen) and ligated into the Tag-pGP8 vector opened with the same enzymes. The right clones were identified through restriction site analysis and DNA sequencing.

### ELISA and activity assays of Tag-wtSGPI-1 and Tag-wtSGPI-2 phages

XL1-blue cells harboring Tag-wtSGPI-1-pGP8 or Tag-wtSGPI-2-pGP8 phagemids were infected with M13-KO7 helper phage, and the inhibitor-phage particles were isolated as described.[31] The presence of SGPI-1 or SGPI-2 on the phage surface was assessed by using as targets bovine chymotrypsin (1 μg/well), crayfish trypsin (1 μg/well) or Anti FLAG-tag antibody (250 ng/well) immobilized on MaxiSorp plates. The phage ELISA was performed as described.[31] Solution inhibitory activity of the inhibitor-phage particles were measured as follows. In a 1 ml final volume $10^{-10}$ M active-site titrated bovine

chymotrypsin or crayfish trypsin was incubated with increasing concentrations of inhibitor-phage. The assay buffer was 50 mM Tris-HCl (pH 8.0), 10 mM $CaCl_2$, 0,001% (v/v) Triton-X100. For trypsin, 50 μM Succ-Ala-Ala-Pro-Lys-AMC, and for chymotrypsin 50 μM Succ-Ala-Ala-Pro-Tyr-AMC were used as fluorogenic substrate. The excitation and emission wavelengths were 380 nm and 460 nm, respectively. The activity measurements were done on a Spex FluoroMAX spectrofluorimeter. The same amount of M13 KO7 helper phage was used as a negative control. The background-normalized residual activity was plotted as a function of phage particle number. At this concentration of enzyme it yielded a linear plot for the SGPI-1 crayfish trypsin and SGPI-2 chymotrypsin pairs, which bind to each other with picomolar binding affinities.

### Synthesis of the chimera SGPI-1/SGPI-2 phage library

Nucleotide degeneracy is indicated using the IUPAC code as follows: A, C, G, T, R (A or G), Y (C or T), M (A or C), K (G or T), B (C, G or T), D (A, G or T), H (A, C or T), V (A, C or G) or N (A, C, G or T). Six degenerated library oligonucleotides were used in the arrangement of two forward and four reverse primers. The construction of the DNA cassette was done essentially as described for the wild-type genes. The following primers were used:

*Chimera-forward-1*: 5′GTG TGT CTC GAG SWG RMG TGC RMG CCG GGT MMG ACC AAA AAA SAK RAW TGC AAC ACC TGC3′
*Chimera-forward-2*: 5′GTG TGT CTC GAG SWG RMG TGC RMG CCG GGT MMG ACC TTT AAA SAK RAW TGC AAC ACC TGC3′
*Chimera-reverse-1-K24;H35*: 5′GTG TGT GGT ACC ATG CKG CGG GCA GSC TTT AMG GGT GCA CGC CSM GMY TTT GCC GKY CGR GSY GCA AYK GCA GGT GTT GCA3′
*Chimera-reverse-2-K24;H35del*: 5′GTG TGT GGT ACC CKG CGG GCA GSC TTT AMG GGT GCA CGC CSM GMY TTT GCC GKY CGR GSY GCA AYK GCA GGT GTT GCA3′
*Chimera-reverse-3-K24del;H35*: 5′GTG TGT GGT ACC ATG CKG CGG GCA GSC TTT AMG GGT GCA CGC CSM GMY GCC GKY CGR GSY GCA AYK GCA GGT GTT GCA3′
*Chimera-reverse-4-K24del;H35del*: 5′GTG TGT GGT ACC CKG CGG GCA GSC TTT AMG GGT GCA CGC CSM GMY GCC GKY CGR GSY GCA AYK GCA GGT GTT GCA3′.

A total of eight library cassettes were generated through the 2 ×4 PCR reactions. The cassettes were ligated into the pGP8 vector and, after a desalting step, were electroporated separately into SS320 cells to generate phage libraries as described.[31]

### Selection of inhibitor phages on FLAG-tag antibody, bovine trypsin and crayfish trypsin

Bovine chymotrypsin (1 μg/well), bovine trypsin (1 μg/well), crayfish trypsin (1 μg/well) or Anti FLAG-tag antibody (250 ng/well) were immobilized on MaxiSorp plates and blocked with bovine serum albumin. The control plates did not contain target, but were blocked by serum albumin. Four cycles of selection and amplification were conducted exactly as described.[31] The inhibitor-phage titers eluted from target and control plates were compared to see how the selection proceeded.

### Phage ELISA of selected library members

The phage ELISA of individual clones was performed as described.[31] Clones for this assay were from the second round of the FLAG-antibody selection and the third round of the protease selection. Clones producing an ELISA signal threefold higher than the background were collected for DNA sequencing.

### DNA sequencing

The sequencing reactions were done using a strategy as described.[30] The gene of the library member was amplified with the following PCR primers annealing to invariant vector sequences. Forward primer, pTacUp35T7: 5′CGAAATTAATACGACTCACTATAGGGCTATAGGGTCTGGATAATGTTTTTTGCGCC3′ and reverse primer, pVIII-rev: 5′GTTATGCTAGTTATTGCTCAGCGGCTTGCTTTCGAGGTGAATTTC3′

The forward PCR primer was designed to contain the sequence of the T7-pro sequencing primer: 5′CGAAATTAATACGACTCACTATAGGG3′, which was then used for the sequencing reaction.

### Pairwise covariance analysis of selected sequences

We applied the observed minus expected squared (OMES) covariance algorithm.[44] The sequences were aligned and for every pair of columns (column *i versus* column *j*), we generated a list *L* of all distinct pairs of amino acids. There were two gaps in the alignment that were also included in the analysis. The score (*S*) for each column pair *i*, *j* is calculated by the following equation:

$$S = \sum_{1}^{L} \frac{(N_{\text{obs}} - N_{\text{exp}})^2}{N}$$

where $N_{\text{obs}}$ is the observed number of the residue pairs, $N_{\text{exp}}$ is the expected number of residue pairs, and $N$ is the number of sequences in the alignment. For an *X*-*Y* residue pair at positions *i* and *j*, the $N_{\text{exp}}$ is calculated simply through the number of occurrences of the two residue types, *X* and *Y* at positions *i* and *j* ($C_{X,i}$ and $C_{Y,j}$) and the number of sequences in the alignment as follows:

$$N_{\text{exp}} = (C_{X,i}\, C_{Y,j})/N$$

The statistical significance of the observed scores was assessed by a bootstrap analysis as follows.[45] For each binding selected pool, the same number of random sequences was generated such that at each residue position the frequencies of the individual amino acid types were kept exactly the same as in the pool of the real selected sequences. With this shuffling of sequences, any co-selected residue pairs were uncoupled. Then, *S* was calculated for the shuffled dataset. This shuffling procedure was repeated 100,000 times, generating 100,000 statistical *S′* scores for each position pair. A significance value was determined for each real *S* score as follows. The number of *S′* scores having values equal to or higher than the observed *S* score was counted and divided by 100,000. Therefore, a *P* value of ≤0.05 means that only 5% of shuffled pairs had the same or higher covariance scores

than those observed, meaning that there is only a 5% chance that the observed value is generated randomly.

### Peptide synthesis

Solid-phase peptide synthesis was performed using the standard Fmoc (*N*-(9-fluorenyl)methoxycarbonyl) chemistry. Cleavage from the resin and simultaneous de-protection were carried out by the trifluoroacetic acid (TFA) method (using 1,2-ethanedithiol, thioanisole, water and phenol as scavengers). After concentrating the solvent to near-dryness and adding cold diethyl ether, the precipitate was dissolved in water and lyophilized. The crude peptides (0.1 mg/ml) were air-oxidized in water (pH adjusted to 8–9 with *N*,*N*-diisopropyl ethylamine) without preliminary purification. The completeness of oxidation was checked by HPLC and mass spectrometry. As the synthesis of peptides with C-terminal P*X* (where *X* is any amino acid) is often problematic, we extended the sequence with an extra C-terminal alanine. This boosted the yield but, as comparative inhibitory assays on SGPI-1 revealed, did not affect the functional properties of the inhibitor (data not shown).

### Mass spectrometry

Mass spectrometry analysis was performed on a HP1100 series HPLC-ESI-MS system using the flow-injection method with the following buffer: 10 mM ammonium formiate in 9:1 (v/v) distilled water/ (pH 3.5). The flow rate was 0.2 ml/min. The parameter settings for MS were as follows: nitrogen was used as drying and nebulizing gas. The drying gas flow rate was 10 l/min, the drying gas temperature was 300 °C, the nebulizing gas pressure was 30 PSI (1 PSI ≈ 6.9 kPa) and the capillary voltage was 3500 V. The total ion current (TIC) chromatogram was obtained in positive ion mode by scanning in the 100–1500 mass/charge range. The mass information was evaluated with Agilent ChemStation software.

### Determination of the inhibitory constant ($K_i$) values

Bovine and crayfish trypsin stock solutions were made by dissolving the proteins in 10 mM HCl, 10 mM $CaCl_2$ or in distilled water, respectively. Active enzyme concentration was determined by active-site titration[46] using fluorescent burst titrant 4-methylumbelliferyl *p*-guanidinobenzoate. The Spex FluoroMAX spectrofluorimeter was calibrated with methylumbelliferone. For the determination of active inhibitor concentration, incremental amounts of inhibitor were incubated with trypsin in 50 mM Tris–HCl (pH 8.0), 10 mM $CaCl_2$, 0.005% Triton X-100 for 10 min at ambient temperature. The final concentration of the enzymes was 1 μM. Residual enzyme activities were measured with 1 mM Nα-Benzoyl-DL-arginine p-nitroanilide hydrochloride. The active inhibitor concentration was determined by linear regression analysis. The equilibrium inhibitory constants ($K_i$) were determined as described.[47,48] The protease was incubated at a concentration close to, but higher than the estimated $K_i$ value with incremental amounts of inhibitor up to about a two- to threefold excess of the inhibitor. After reaching equilibrium, incubation mixtures were assayed by the addition of a substrate most appropriate for determination of the concentration of the free enzyme. The buffer was the same as above. For enzyme–inhibitor pairs with $K_i$ in the $10^{-9}$ M range, the enzyme concentration was 5 nM, and 0.25 mM of the photometric substrate CBZ-Gly-Pro-Arg-pNA was used on a Shimadzu spectrophotometer. For enzyme–inhibitor pairs with $K_i$ in the $10^{-12}$ M range, the enzyme concentration was 0.1 nM, and 50 μM Succ-Ala-Ala-Pro-Lys-AMC was used on a Spex FluoroMAX spectrofluorimeter. The excitation and emission wavelengths were 380 nm and 460 nm, respectively. Numerical $K_i$ values were determined from two parallel measurements through non-linear regression analysis using the LabFit software† and the following equation:

$$\frac{[E]}{[E]_0} = 1 - \frac{[E]_0 + [I]_0 + K_i - \sqrt{([E]_0 + [I]_0 + K_i)^2 - 4[E]_0[I]_0}}{2[E]_0}$$

In the equation [E], $[E]_0$ and $[I]_0$ represent the molar concentrations of the free enzyme, total enzyme and total inhibitor, respectively.

## Acknowledgements

The authors express their thanks to Dr András Perczel for his important insights regarding the interpretation of the results. We express our thanks for the SS320 bacterial strain and the pS1607 phagemid vector that were developed at Genentech and were a kind gift from Dr Sachdev S. Sidhu. This work was supported by the Hungarian National Science Foundation (OTKA TS049812, K068408), the National Office for Research and Technology (RET 14/2005), the National Development Plan (GVOP-3.1.1-2004-0235) and ICGEB (CRP Hun04-03).

## References

1. Krowarsch, D., Cierpicki, T., Jelen, F. & Otlewski, J. (2003). Canonical protein inhibitors of serine proteases. *Cell. Mol. Life Sci.* **60**, 2427–2444.
2. Laskowski, M. & Kato, I. (1980). Protein inhibitors of proteinases. *Annu. Rev. Biochem.* **533**, 593–626.
3. Bode, W. & Huber, R. (1992). Natural protein proteinase inhibitors and their interaction with proteinases. *Eur. J. Biochem.* **204**, 433–451.
4. Otlewski, J., Jelen, F., Zakrzewska, M. & Oleksy, A. (2005). The many faces of protease-protein inhibitor interaction. *EMBO J.* **24**, 1303–1310.
5. Schechter, I. & Berger, A. (1967). On the size of the active size of proteases. I. Papain. *Biophys. Res. Commun.* **27**, 157–162.
6. Hergenhahn, H. G., Aspan, A. & Soderhall, K. (1987). Purification and characterization of a high-Mr proteinase-inhibitor of pro-phenol oxidase activation from crayfish plasma. *Biochem. J.* **248**, 223–228.
7. Liang, Z., Sottrup-Jensen, L., Aspan, A., Hall, M. & Soderhall, K. (1997). Pacifastin, a novel 155-kDa heterodimeric proteinase inhibitor containing a unique transferrin chain. *Proc. Natl Acad. Sci. USA*, **94**, 6682–6687.

† www.labfit.net

8. Hamdaoui, A., Wataleb, S., Devreese, B., Chiou, S. J., Vanden Broeck, J., Van Beeumen, J. *et al.* (1998). Purification and characterization of a group of five novel peptide serine protease inhibitors from ovaries of the desert locust, *Schistocerca gregaria*. *FEBS Letters*, **422**, 74–78.
9. Kellenberger, C., Boudier, C., Bermudez, I., Bieth, J. G., Luu, B. & Hietter, H. (1995). Serine protease inhibition by insect peptide containing a cysteine-knot and a triple-stranded β-sheet. *J. Biol. Chem.* **270**, 25514–25519.
10. Nakakura, N., Hietter, H., Vandorsselaer, A. & Luu, B. (1992). Isolation and structural determination of 3 peptides from the insect *Locusta migratoria* - identification of a deoxyhexose-linked peptide. *Eur. J. Biochem.* **204**, 147–153.
11. Parkinson, N. M., Conyers, C., Keen, J., MacNicoll, A., Smith, I., Audsley, N. & Weaver, R. (2004). Towards a comprehensive view of the primary structure of venom proteins from the parasitoid wasp *Pimpla hypochondriaca*. *Insect Biochem. Mol. Biol.* **34**, 565–571.
12. Gáspári, Z., Ortutay, C. & Perczel, A. (2004). A simple fold with variations: the pacifastin inhibitor family. *Bioinformatics*, **20**, 448–451.
13. Simonet, G., Claeys, I., Vanderperren, H., November, T., De Loof, A. & Vanden Broeck, J. (2002). cDNA cloning of two different serine protease inhibitor precursors in the migratory locust, *Locusta migratoria*. *Insect Mol. Biol.* **11**, 249–256.
14. Mer, G., Kellenberger, C., Koehl, P., Stote, R., Sorokine, O., Van Dorsselaer, A. *et al.* (1994). Solution structure of PMP-D2, a 35-residue peptide isolated from the insect *Locusta migratoria*. *Biochemistry*, **33**, 15397–15407.
15. Mer, G., Kellenberger, C., Renatus, M., Luu, B., Hietter, H. & Lefevre, J.-F. (1996). Solution structure of PMP-C: a new fold in the group of small proteinase inhibitors. *J. Mol. Biol.* **258**, 158–171.
16. Mer, G., Hietter, H. & Lefevre, J. F. (1996). Stabilization of proteins by glycosylation examined by NMR analysis of a fucosylated proteinase inhibitor. *Nature Struct. Biol.* **3**, 45–53.
17. Gáspári, Z., Patthy, A., Gráf, L. & Perczel, A. (2002). Comparative structure analysis of proteinase inhibitors from the desert locust, *Schistocerca gregaria*. *Eur. J. Biochem.* **269**, 527–537.
18. Kellenberger, C., Ferrat, G., Leone, P., Darbon, H. & Roussel, A. (2003). Selective inhibition of trypsins by insect peptides: role of P6-P10 loop. *Biochemistry*, **42**, 13605–13612.
19. Mucsi, Z., Gáspári, Z., Orosz, G. & Perczel, A. (2003). Structure-oriented rational design of chymotrypsin inhibitor models. *Protein Eng.* **16**, 673–681.
20. Malik, Z., Amir, S., Pál, G., Buzás, Z., Várallyay, E., Antal, J. *et al.* (1999). Proteinase inhibitors from desert locust, *Schistocerca gregaria*: engineering of both P1 and P1′ residues converts a potent chymotrypsin inhibitor to a potent trypsin inhibitor. *Biochim. Biophys. Acta*, **1434**, 143–150.
21. Grahn, S., Kurth, T., Ullmann, D. & Jakubke, H. D. (1999). S′ subsite mapping of serine proteases based on fluorescence resonance energy transfer. *Biochim. Biophys. Acta*, **1431**, 329–337.
22. Schellenberger, V., Turck, C. W., Hedstrom, L. & Rutter, W. J. (1993). Mapping the S′ subsites of serine proteases using acyl transfer to mixtures of peptide nucleophiles. *Biochemistry*, **32**, 4349–4353.
23. Schellenberger, V., Turck, C. W. & Rutter, W. J. (1994). Role of the S′ subsites in serine protease catalysis. Active-site mapping of rat chymotrypsin, rat trypsin?-lytic protease, and cercarial protease from *Schistosoma mansoni*. *Biochemistry*, **33**, 4251–4257.
24. Patthy, A., Amir, S., Malik, Z., Bódi, A., Kardos, J., Asbóth, B. & Gráf, L. (2002). Remarkable phylum selectivity of a *Schistocerca gregaria* trypsin inhibitor: The possible role of enzyme-inhibitor flexibility. *Arch. Biochem. Biophys.* **398**, 179–187.
25. Simonet, G., Breugelmans, B., Proost, P., Claeys, I., Van Damme, J., De Loof, A. & Vanden Broeck, J. (2005). Characterization of two novel pacifastin-like peptide precursor isoforms in the desert locust (*Schistocerca gregaria*): cDNA cloning, functional analysis and real-time RT-PCR gene expression studies. *Biochem. J.* **388**, 281–289.
26. Fellouse, F. & Pál, G. (2005). Methods for the construction of phage-displayed libraries. In *Phage Display in Biotechnology and Drug Discovery* (Sidhu, S. S., ed), pp. 111–142, CRC, BocaRaton, FL.
27. Bai, Y. & Feng, H. (2004). Selection of stably folded proteins by phage-display with proteolysis. *Eur. J. Biochem.* **271**, 1609–1614.
28. Sieber, V., Pluckthun, A. & Schmid, F. X. (1998). Selecting proteins with improved stability by a phage-based method. *Nature Biotechnol.* **16**, 955–960.
29. Kotz, J. D., Bond, C. J. & Cochran, A. G. (2004). Phage-display as a tool for quantifying protein stability determinants. *Eur. J. Biochem.* **271**, 1623–1629.
30. Pál, G., Kossiakoff, A. A. & Sidhu, S. S. (2003). The functional binding epitope of a high affinity variant of human growth hormone mapped by shotgun alanine-scanning mutagenesis: insights into the mechanisms responsible for improved affinity. *J. Mol. Biol.* **332**, 195–204.
31. Sidhu, S. S., Lowman, H. B., Cunningham, B. C. & Wells, J. A. (2000). Phage display for selection of novel binding peptides. *Methods Enzymol.* **328**, 333–363.
32. Pál, G., Fong, S. Y., Kossiakoff, A. A. & Sidhu, S. S. (2005). Alternative views of functional protein binding epitopes obtained by combinatorial shotgun scanning mutagenesis. *Protein Sci.* **14**, 2405–2413.
33. Pál, G., Kouadio, J. L., Artis, D. R., Kossiakoff, A. A. & Sidhu, S. S. (2006). Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. *J. Biol. Chem.* **281**, 22378–22385.
34. Gáspári, Z., Szenthe, B., Patthy, A., Westler, W. M., Gráf, L. & Perczel, A. (2006). Local binding with globally distributed changes in a small protease inhibitor upon enzyme binding. *FEBS J.* **273**, 1831–1842.
35. Szenthe, B., Gáspári, Z., Nagy, A., Perczel, A. & Gráf, L. (2004). Same fold with different mobility: backbone dynamics of small protease inhibitors from the desert locust, *Schistocerca gregaria*. *Biochemistry*, **43**, 3376–3384.
36. Roussel, A., Mathieu, M., Dobbs, A., Luu, B., Cambillau, C. & Kellenberger, C. (2001). Complexation of two proteic insect inhibitors to the active site of chymotrypsin suggests decoupled roles for binding and selectivity. *J. Biol. Chem.* **276**, 38893–38898.
37. Fodor, K., Harmat, V., Hetényi, C., Kardos, J., Antal, J., Perczel, A. *et al.* (2005). Extended intermolecular interactions in a serine protease-canonical inhibitor complex account for strong and highly specific inhibition. *J. Mol. Biol.* **350**, 156–169.
38. Cunningham, B. C. & Wells, J. A. (1993). Comparison of a structural and a functional epitope. *J. Mol. Biol.* **234**, 554–563.

39. Lu, S. M., Lu, W., Qasim, M. A., Anderson, S., Apostol, I., Ardelt, W. *et al.* (2001). Predicting the reactivity of proteins from their sequence alone: Kazal family of protein inhibitors of serine proteinases. *Proc. Natl Acad. Sci. USA*, **98**, 1410–1415.
40. Qasim, M. A., Ganz, P. J., Saunders, C. W., Bateman, K. S., James, M. N. G. & Laskowski, M. (1997). Interscaffolding additivity. Association of P-1 variants of eglin C and of turkey ovomucoid third domain with serine proteinases. *Biochemistry*, **36**, 1598–1607.
41. Fodor, K., Harmat, V., Neutze, R., Szilagyi, L., Graf, L. & Katona, G. (2006). Enzyme:substrate hydrogen bond shortening during the acylation phase of serine protease catalysis. *Biochemistry*, **45**, 2114–2121.
42. Sidhu, S. S., Weiss, G. A. & Wells, J. A. (2000). High copy display of large proteins on phage for functional selections. *J. Mol. Biol.* **296**, 487–495.
43. Kunkel, T. A., Roberts, J. D. & Zakour, R. A. (1987). Rapid and efficient site-specific mutagenesis without phenotypic selection. *Methods Enzymol.* **154**, 367–382.
44. Fodor, A. A. & Aldrich, R. W. (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Struct. Funct. Genet.* **56**, 211–221.
45. Manly, B. J. F. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd edit., Chapman & Hall/CRC, BocaRaton, FL.
46. Jameson, G. W., Roberts, D. V., Adams, R. W., Kyle, W. S. & Elmore, D. T. (1973). Determination of the operational molarity of solutions of bovine alpha-chymotrypsin, trypsin, thrombin and factor Xa by spectrofluorimetric titration. *Biochem. J.* **131**, 107–117.
47. Green, N. M. & Work, E. (1953). Pancreatic trypsin inhibitor. II. Reaction with trypsin. *Biochem. J.* **54**, 347–352.
48. Empie, M. W. & Laskowski, M., Jr (1982). Thermodynamics and kinetics of single residue replacements in avian ovomucoid third domains: effect on inhibitor interactions with serine proteinases. *Biochemistry*, **21**, 2274–2284.
49. Koradi, R., Billeter, M. & Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.

***Edited by F. Schmid***

*(Received 19 January 2007; received in revised form 5 April 2007; accepted 10 April 2007)*
Available online 19 April 2007