Supplementary Information

Supplementary Text T1. Detailed account of molecular modelling

ADDITIONAL MATERIALS AND METHODS

Molecular mechanic simulations were carried out essentially as described by [1] but the energy of interaction between the GHP models (receptors) and the peptides (ligands) were calculated using the Docking module of INSIGHT II and the Discover program. Amber charges were applied to all molecules and the calculations were carried out with an 8.0Å cut off distance for non-bounded interactions. In this study Receptor-Ligand interactions were characterised using explicit van der Waals and electrostatic (Coulombic) energies however the electrostatic desolvation free energy term (ΔG_{solv}) was not calculated. Therefore, the data presented corresponds to interaction energies (iEs) and not binding free energies (ΔG_{bind}). Both the Coulombic and the Lennard Jones interaction energies (iE^{elect}. and iE^{vdw}) were partitioned into interactions terms between each amino acid residue of the ligand and the receptor (Table S2a-c).

Energy minimized protein-ligand docking conformations were also scored using the rescoring application built in of Gold 4.1.2 and the Goldscore scoring function. Molecular structures were saved in the Tripos MOL2 format and imported into Hermes 1.3.1. The centre of the binding site was defined as the hydroxyl oxygen of the active site serine (Ser_{189} in GPH1 Ser_{188} in GPH2 and GPH3) with a 15 Å radius but selection of interacting atoms was restricted to solvent accessible atoms using the cavity detection facility.

RESULTS

Modelling the Eurygaster GHPs. Based on the specificity demonstrated experimentally we used molecular modelling to further characterise the substrate specificity of the GHPs, generating homology models for the three GHP isoforms identified in this study.

The portion of the three *Eurygaster* GHP sequences that was modelled (the mature protein missing 22 amino acids at the *C*-terminus) showed approximately 34% identity and 48% similarity to the crayfish sequence used as template (as this showed the highest similarity in the Protein Structure (PDB) database) (*A. leptodactylus* trypsin, PDB code: 2F91) with 7% of the residues being in gaps located in loops between the main secondary structure elements (Fig. 4). The three catalytic residues (His_{43}, Asp_{96} and $Ser_{188/189}$) were located in conserved core regions which also contain six cysteine residues involved in the formation of three disulfides bonds with a pattern identical to that observed in the crystal structure template. An additional pair of cysteine residues (Cys_{116} and $Cys_{240/241}$) present in the GHP proteins are involved in the formation of another disulphide bond with the C-terminal end of the protein on the opposite side from the active site. This part of the protein could not be modelled in this study because it did not have a counterpart in the template structure. After energy minimisation, comparison of the backbone of the GHP models with that of the crystal structure template reveals a seven residue loop extension corresponding to Gly_{85}-Gly_{91} located just before the catalytic aspartate (as illustrated for GHP3 in Fig.7a and b and Fig. S3a). (The only major difference between the three GHP's modelled is an extended loop formed by residues Glu_{142} to Pro_{145} in GHP1 compared to GHP2 and 3; Fig. S3a and b).

Modelling substrate binding. Mapping the positions of the 30 variant amino acids on the molecular surface of the three GHPs revealed that only four positions: Lys/Leu/Met_{27}, Gln/Lys_{136}; Gly/Thr_{143} and Asn/$Lys_{185/186}$ may have effects on the S1'-S3' binding pocket (Fig.

S5). Pockets S1-S4 are essentially identical suggesting a very similar substrate specificity for the three protease isoforms. In all three models Lys/Leu/Met_{27} and Asn/$Lys_{185/186}$ residues appear to form a narrow trench explaining the requirement for a glycine residue at P1' position of substrate peptides (Fig. S4a and b and FigS5). Another variant amino acid, Gln/Pro/Ala_{90}, is located in the seven residue loop extension specific to the GHP models (Fig. S5). This extension results in the formation of a much deeper S4 binding pocket compared to the template crystal structure. However, in all models the side chain of residue 90 is pointing away from the S4 pocket which, instead, is lined by the conserved Gly_{91} residue. This allows the deeper S4 pocket in our three models to accommodate the side chain of a glutamine residue at the P4 position whereas an alanine residue is bound in the shallower S4 pocket in the crystal structure template (not shown). The peptide PGQGQQGYYP (which was present in the synthetic peptide used to determine the enzyme specificity, and called R6 in the sections that follow, see Fig. S3) appeared to fit very well in the binding pocket of all three models with main chain atoms of the ligand involved in two intrachain H-bonds between the carbonyl oxygens of P5 Gly and P2 Gln and the amide hydrogens of P3 Gly and P1' Gly, respectively (for example, GPH3+R6: Fig. S4a and b). These internal H-bonds are not present in the crystal structure template (not shown). The additional intrachain H-bonds are likely to contribute to stabilisation of the substrate in an optimal low energy conformation favourable for catalytic activity. The presence of Thr_{143} in GHP1, instead of Gly_{143} in GHP2 and GHP3, may result in the formation of one additional H-bond stabilising the P2' Tyrosine in the S2' pocket.

To corroborate the results obtained with the R1X5 peptide (Fig. 5) we modelled *in silico* the peptide PGQGQQGHYP (i.e. with histidine at the P2' position) (called X-GHY, see

Fig. S2). After docking and energy minimisation the P2' histidine appeared to fit very well in the S2' pocket of all models which is consistent with the observation that the peptide PQGQGQQGHYPASLQQ is cleaved between the PQGQGQQ and the GHYPASLQQ to liberate the GHYPASLQQ peptide (identified experimentally by N-terminal sequencing of the peptide ladder produced by digestion of R1X5) (Fig. 5).

Calculation of interaction energies. The energies of interaction between the GHP models (receptors) and the peptides (ligands) were computed in order to compare the values for the ligands in various orientations relative to the receptors, and to identify orientations that result in low interaction energies. After energy minimisation the R6 and X-GHY peptides were found to have similar energies of interaction with all three forms of GHP, about -30 kcal/mol (Fig. S6, Table S2a-c and Table S3), confirming the results and observations discussed above. Analysis of the binding of R6 and X-GHY in the three GHP models revealed strong and specific interactions with around 80% of the binding energy concentrated on five residues between the P4 and P2' positions but with distinctively low P3 Glu and high P4 Gln binding energies (Fig. S6, Table S2a-c). The energetic pattern observed for R6 and X-GHY is due to a strong binding of the P4 Gln residue in the unusually deep S4 pocket of the GHPs. All the low energy docking solution obtained after molecular mechanic simulations, were rescored using the Gold scoring function of Gold 4.1. For all structures the fitness scores obtained (Table S3) were consistent with the calculated interaction energies.

Modelling of the cleavage site specificity. To model the cleavage site specificity of the purified GHP demonstrated experimentally (Fig. 6), we docked the peptide LQQPGQGQQG (called the Inverted R6 peptide as, the nonapeptide precedes the hexapeptide, as in Fig. 6) in our GHP models. After docking and energy minimisation the energy of interaction

between this Inverted R6 and the GHP models was about 10 kcal/mol higher than those obtained with the R6 and X-GHY peptides (Table S3). Although, the interaction energy was generally increased for all residues in the Inverted R6 peptide compared to that for the R6 and X-GHY peptides, the difference was more marked for P4, P2 and P2' (Fig. S6, Tables S2a-c and S3). Closer inspection revealed that the conformation of the main chain atoms in docked Inverted R6 is different from that observed for the R6 and X-GHY peptides. The most obvious difference was the absence of the main chain H-bond between the P5 and P3 residues due the absence of an amide hydrogen in the P3 proline residue. Together with the lost of flexibility due to the fixed phi (ϕ) dihedral angle at P3 position, this resulted in a suboptimal docking conformation for the P4 glutamine. The higher binding energy in the S2 and S2' pocket resulted primarily from a less efficient van der Waal interaction (Table S2a-c) rather than loss of hydrogen bonding. In addition mapping the electrostatic potential on the surface of the GHP models showed that the S2' pocket is relatively apolar compared to the S4-S1' region of the proteins (Fig. S4a and b and S5) which is consistent with this pocket being more suitable for interacting with less polar P2' residues.

1. Laskowski, R.A., MacArthur, M.W.; Moss, D.S.; Thornton, J.M. Procheck - a program to check the stereochemical quality of protein structures. *J. Appl. Crystalog.* **1993,** *26,* 283–291.

Table S1.

Sequences of oligonucleotide primers used for PCR and cloning of GHPs from *E. integriceps* salivary glands.

Primer name	Direction	Sequence 5'→3'
degenNterm	Forward	ATHGTNGGNGGNWSNCARGCNYTNGAYAAYGARTAYCCNTGGATGGTNAAR
degenInternal	Reverse	YTTNGCRTANGGNARNARNGCNARRTCRTTNARNGT
InternalFor	Forward	TAAGCATCATCGCAGGCACTTCCG
InternalRev	Reverse	AATCGGAAGTGCCTGCGATGATGC
FL-5'	Forward	ATCATCGTAGCTGGCAAGATG
FL-3'	Reverse	GTCAAGATATAGATTCTATTTATTTATTAG

degenNterm/ degenInternal were the degenerate primers used for the initial amplification reactions.

InternalFor/ InternalRev internal forward and reverse primers.

FL-5'/ FL-3' primers used to obtained full length clones.

SUPPLEMENTARY Table S2a-c. Interaction energy profiles.

Interaction energy profiles for each residue of ligands peptides used in the docking experiments between GHPs 1, 2 and 3 models. (a) GHP1 with R6, X-GHY and Inverted R6. (b) GHP2 with R6, X-GHY and Inverted R6. (c) GHP3 with R6, X-GHY and Inverted R6.

a	Interaction Energy for each residue in ligands										Total Interaction Energy
		P6	P5	P4	P3	P2	P1	P1'	P2'	P3'	P4'	
GHP1 +R6		P	G	Q	G	Q	Q	G	Y	Y	P	
	VDW	-0.89	-0.31	-4.50	-0.69	-2.92	-7.29	-3.61	-2.82	-1.73	-1.18	-25.94
	Elect	-0.10	-0.03	0.48	-0.04	-0.40	-2.40	-0.83	-1.34	-0.05	-0.16	-4.87
	Total	**-0.99**	**-0.34**	**-4.02**	**-0.73**	**-3.32**	**-9.69**	**-4.44**	**-4.16**	**-1.78**	**-1.34**	**-30.80**
GHP1 +XGHY		P	G	Q	G	Q	Q	G	H	Y	P	
	VDW	-0.69	-0.32	-4.48	-0.55	-3.58	-7.27	-3.43	-2.53	-1.89	-0.94	-25.67
	Elect	-0.03	-0.05	0.30	-0.03	-0.38	-2.61	-0.89	-0.98	-0.03	-0.15	-4.86
	Total	**-0.73**	**-0.37**	**-4.18**	**-0.57**	**-3.96**	**-9.88**	**-4.32**	**-3.51**	**-1.91**	**-1.09**	**-30.53**
GHP1 +Inverted R6		L	Q	Q	P	G	Q	G	Q	Q	G	
	VDW	-0.33	-0.26	-2.85	-1.42	-1.61	-6.64	-3.31	-1.31	-0.31	-0.65	-18.70
	Elect	-0.01	-0.02	0.46	-0.07	0.27	-1.01	-0.78	-0.40	-0.40	0.03	-1.92
	Total	**-0.34**	**-0.28**	**-2.39**	**-1.49**	**-1.34**	**-7.65**	**-4.09**	**-1.71**	**-0.71**	**-0.61**	**-20.62**

b		Interaction Energy for each residue in ligands										Total Interaction Energy
		P6	**P5**	**P4**	**P3**	**P2**	**P1**	**P1'**	**P2'**	**P3'**	**P4'**	
GHP2 +R6		**P**	**G**	**Q**	**G**	**Q**	**Q**	**G**	**Y**	**Y**	**P**	
	VDW	-0.84	-0.52	-4.21	-0.65	-3.05	-6.23	-2.62	-2.60	-1.92	-1.50	-24.13
	Elect	0.06	-0.08	0.13	-1.27	-0.23	-2.86	-1.34	-0.43	-0.03	0.05	-6.00
	Total	**-0.77**	**-0.60**	**-4.08**	**-1.92**	**-3.28**	**-9.09**	**-3.96**	**-3.03**	**-1.95**	**-1.45**	**-30.14**
GHP2 +XGHY		**P**	**G**	**Q**	**G**	**Q**	**Q**	**G**	**H**	**Y**	**P**	
	VDW	-1.04	-0.42	-4.29	-0.60	-2.95	-6.69	-2.83	-2.31	-1.70	-1.55	-24.37
	Elect	0.01	0.04	0.06	-1.36	-0.31	-2.31	-1.42	-0.45	-0.07	-0.01	-5.83
	Total	**-1.03**	**-0.39**	**-4.23**	**-1.96**	**-3.25**	**-9.00**	**-4.25**	**-2.76**	**-1.77**	**-1.56**	**-30.20**
GHP2 +Inverted R6		**L**	**Q**	**Q**	**P**	**G**	**Q**	**G**	**Q**	**Q**	**G**	
	VDW	-0.64	-0.09	-2.83	-0.92	-1.51	-6.11	-3.37	-1.48	-0.61	-0.65	-18.19
	Elect	0.02	-0.04	0.46	-0.63	0.13	-1.18	-0.45	-0.21	-0.10	0.08	-1.94
	Total	**-0.62**	**-0.12**	**-2.38**	**-1.54**	**-1.38**	**-7.30**	**-3.82**	**-1.69**	**-0.71**	**-0.57**	**-20.13**

c		Interaction Energy for each residue in ligands										Total Interaction Energy
		P6	P5	P4	P3	P2	P1	P1'	P2'	P3'	P4'	
GHP3 +R6		P	G	Q	G	Q	Q	G	Y	Y	P	
	VDW	-0.73	-0.32	-4.35	-0.99	-3.29	-5.95	-3.43	-2.77	-1.93	-1.20	-24.95
	Elect	-0.01	-0.05	0.19	-0.52	-0.36	-2.63	-0.66	-0.70	-0.43	-0.06	-5.23
	Total	**-0.74**	**-0.37**	**-4.16**	**-1.51**	**-3.65**	**-8.58**	**-4.09**	**-3.47**	**-2.36**	**-1.26**	**-30.18**
GHP3 +XGHY		P	G	Q	G	Q	Q	G	H	Y	P	
	VDW	-0.73	-0.33	-4.45	-0.98	-3.30	-5.72	-3.50	-2.44	-1.96	-1.01	-24.41
	Elect	-0.03	-0.05	0.21	-0.60	-0.22	-2.64	-0.70	-1.06	-0.49	-0.31	-5.89
	Total	**-0.76**	**-0.38**	**-4.25**	**-1.58**	**-3.52**	**-8.36**	**-4.20**	**-3.50**	**-2.45**	**-1.32**	**-30.31**
GHP3 +Inverted R6		L	Q	Q	P	G	Q	G	Q	Q	G	
	VDW	-0.37	-0.40	-2.82	-1.43	-1.73	-5.16	-3.37	-1.40	-1.10	-0.60	-18.38
	Elect	-0.04	-0.06	0.40	-0.14	0.24	-1.74	-0.39	-0.46	0.42	0.00	-1.77
	Total	**-0.41**	**-0.46**	**-2.42**	**-1.57**	**-1.49**	**-6.90**	**-3.76**	**-1.86**	**-0.68**	**-0.60**	**-20.15**

Table S3. **Binding energy and Gold scores of peptide ligands docked in the three GHP models.** Total Interaction Energies (kcal/mol) calculated using the docking module of INSIGHT II and the Discover program, Gold scores generated using Gold 4.1.2

	R6	**X-GHY**	**Inverted R6**
	Interaction Energy / Gold Score		
GHP1	-30.80 / 78.67	-30.53 / 76.54	-20.62 / 66.43
GHP2	-30.14 / 78.01	-30.20 / 77.19	-20.13 / 67.44
GHP3	-30.18 / 75.95	-30.31 / 76.75	-20.15 / 69.66

Supporting information Figure legends.

Fig. S1. **Detection of glutenin hydrolysing proteinases (GHPs) from salivary glands and whole guts from over-wintering and summer generations of E. integriceps separated by isoelectric focusing (IEF).**

Adults *E. integriceps* overwinter in Russia therefore extracts from salivary glands and whole guts from over-wintering and summer generations were analysed. Activity was detected in salivary glands from the summer generation of two populations collected in the Samara and Saratov regions, respectively, of Russia, the patterns (Tracks a-d, i-l) being similar to those shown in Figure 1. No activity was detected when similar analyses were carried out on the over-wintering populations from these regions (Tracks e-h) unless the exposure was prolonged (not shown). Again, no activity was detected in preparations from whole guts (m-o) unless the exposure was prolonged for many hours (not shown). This demonstrates that production of glutenin-digesting proteinases is largely restricted to the salivary glands in the generation of insects feeding on the wheat grain and absent in the overwintering population feeding on the vegetative parts of wheat.

Fig. S2. **Recombinant, synthetic and *in silico* peptides used for determination of GHPs site specificity.**

Panel a. recombinant peptides used for *in vivo* digests and analysed by SDS PAGE. Panel b. synthetic peptides used for mass spectrometric analysis of site specificity of purified GHP. Panel c. *in silico* peptides used in modelling analysis of site specificity. Repeating motifs are indicated by colour.

Fig. S3. **Molecular modelling of GHPs.**

(a) The three GHP models constructed superimposed upon each other (GHP1: blue; GHP2: green; GHP3: pink) showing the extended loop formed by residues Glu_{142} to Pro_{145} in GHP1 (Arrowed). The R6 peptide is shown as a transparent stick model in atomic colours. (b) Magnification of the previous panel showing the conserved catalytic residues in the three GHP structures.

Fig. S4. **Connolly surface representation of GHP3.**

Electrostatic potential mapped on the surface of the protein. The position of the S6-S4' binding pockets are indicated. Red –ve, Blue +ve, white non-polar. (b) After docking and energy minimisation the backbone of substrate peptides forms two main-chain P5-P3 and P2-P1' hydrogen bonds stabilising the dihedral conformation of P4 and P1 residues, respectively. H-bonds are shown as yellow dotted lines.

Fig S5. **Connolly surface representation of GHP3 mapping the 30 variant amino acids on the surface of the protein.**

Light blue: whole surface; Mid-blue: residues not in contact with the substrate binding pockets; Dark blue: the 5 variant residues involved in the formation of the binding cavity are labelled with the amino acid code and number. The docked R6 peptide (PGQGQQGYYP) is shown as a stick model in atomic colours.

Fig S6. **Interaction energy profiles between GHPs 1, 2 and 3 models and the ligand peptides used in the docking experiments.**

Light grey: residues at P6-P1 positions; mid-grey: residues at P1'-P4' positions. The interaction energy between each amino acid residue of the ligand and the GHPs were calculated as described in the methods and expressed in kcal/mol.

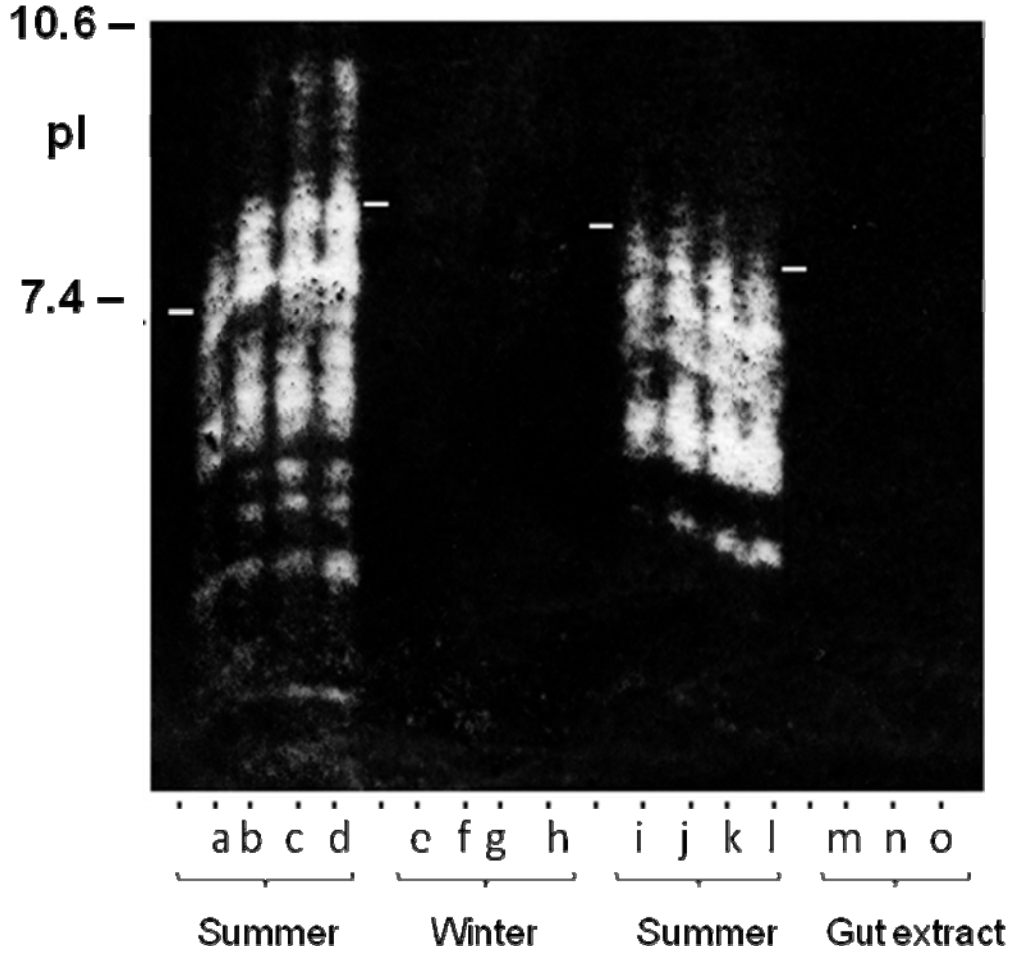
10.6 –
pI
7.4 –
a b c d
e f g h
i j k l
m n o
Summer
Winter
Summer
Gut extract

Figure S1.

Recombinant peptides

R6

APGQGQC GYYPTSLQQ PGQGQQ GYYPTSLQQ PGQGQQ GYYPTSLQQ PGQGQQ GYYPTSLQQ
PGQGQQ GYYPTSLQQ PGQGQQ GYYPTSLQQ PGQGQQ GYYPTSLQQ PGQGQQ GYYPTSLQQ
PGQGQQ GYYPTSLQQ PGQGQQ GYYPTSLQQ PGQGQQ GYYPTSLQQ PGQGQQ GYYPTSLQQ
PGQGQQ GYYPTSLQC PGQGQQ

R1X5

APGQGQC GYYPTSLQQ PGQGQQ GYYPTSLQQ PGQGQQ GHYPASLQQ PGQGQQ GHYPASLQQ
PGQGQQ GHYPASLQQ PGQGQQ GHYPASLQQ PGQGQQ GHYPASLQQ PGQGQQ GHYPASLQQ
PGQGQQ GHYPASLQQ PGQGQQ GHYPASLQQ PGQGQQ GHYPASLQQ PGQGQQ GHYPASLQQ
PGQGQQ GYYPTSLQC PGQGQQ

Panel a

Synthetic peptides

(hexapeptide-nonapeptide)

PGQGQQGYYPTSLQQ

(nonapeptide-hexapeptide)

GYYPTSLQQPGQGQQ

Panel b

in silico peptides

R6

PGQGQQGYYP

X-GHY

PGQGQQGHYP

Inverted R6

LQQPGQGQQG

Panel c

Figure S2.

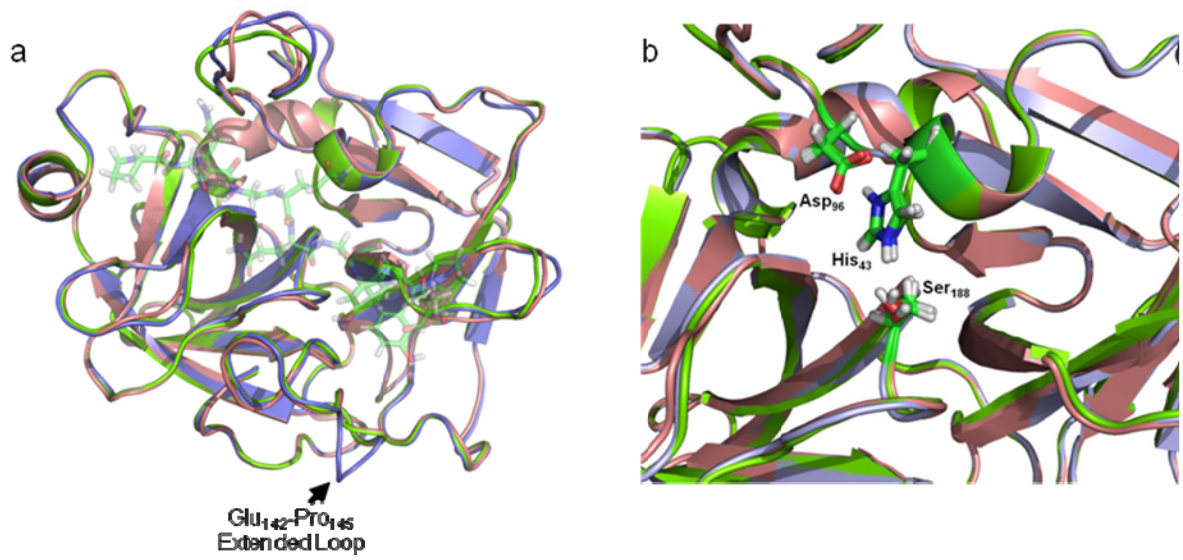

Figure S3a and b.

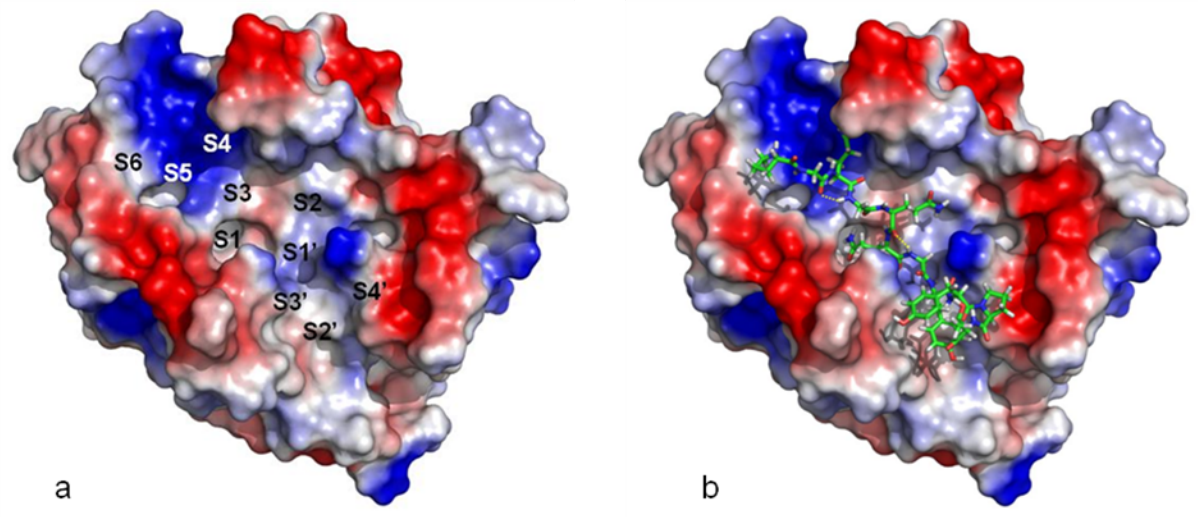

Figure S4a and b.

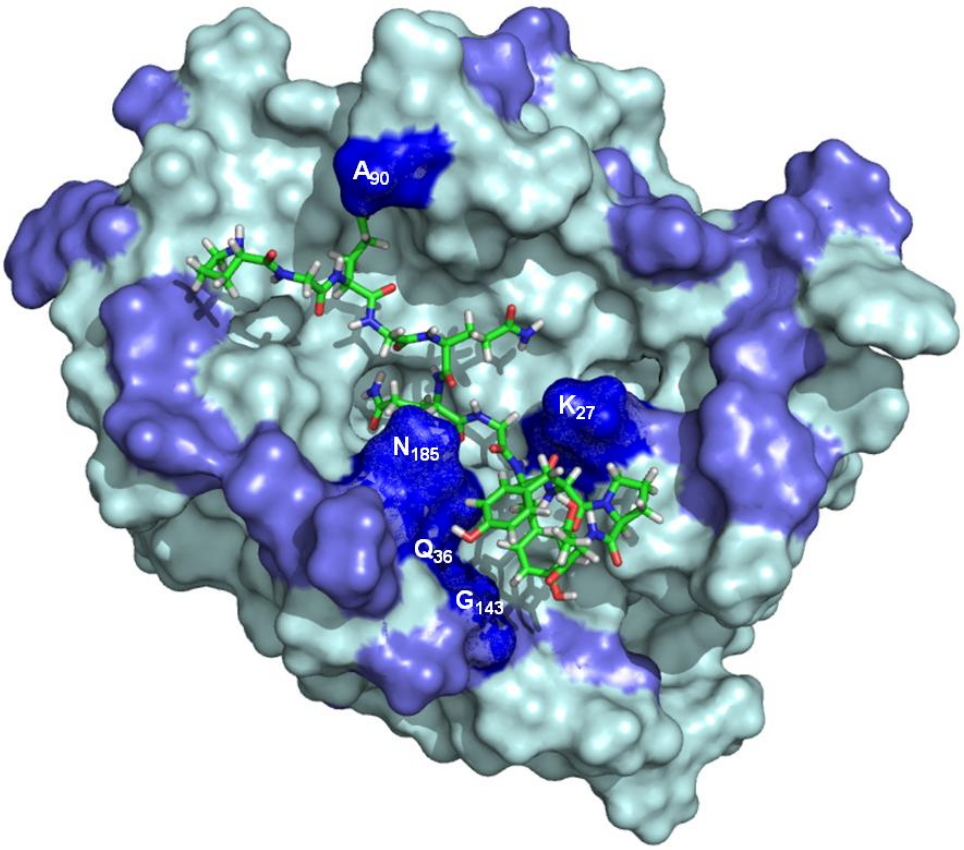

Figure S5

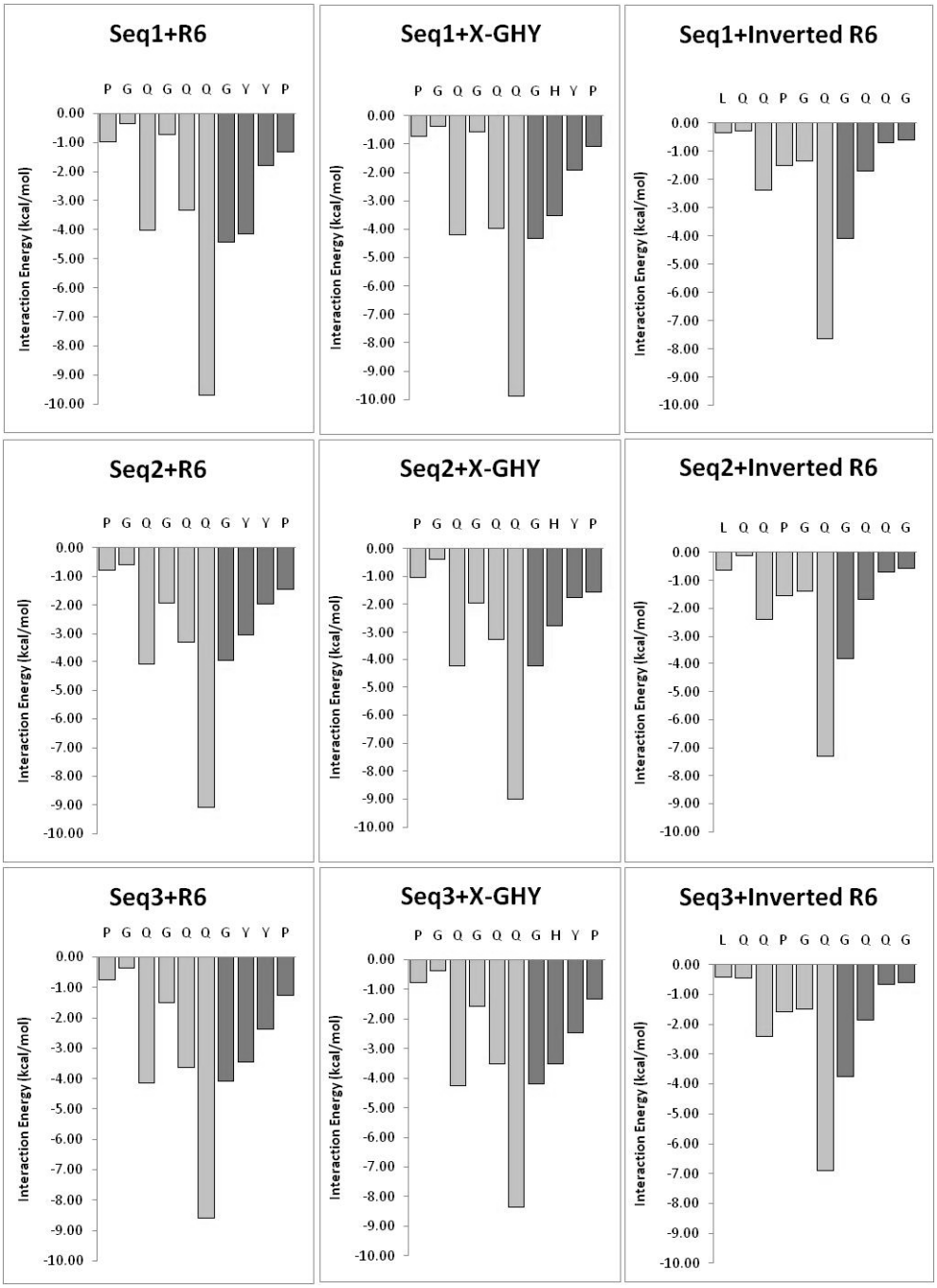
Seq1+R6
P G Q G Q Q G Y Y P
Seq1+X-GHY
P G Q G Q Q G H Y P
Seq1+Inverted R6
L Q Q P G Q G Q Q G
Seq2+R6
P G Q G Q Q G Y Y P
Seq2+X-GHY
P G Q G Q Q G H Y P
Seq2+Inverted R6
L Q Q P G Q G Q Q G
Seq3+R6
P G Q G Q Q G Y Y P
Seq3+X-GHY
P G Q G Q Q G H Y P
Seq3+Inverted R6
L Q Q P G Q G Q Q G
Interaction Energy (kcal/mol)
0.00
-1.00
-2.00
-3.00
-4.00
-5.00
-6.00
-7.00
-8.00
-9.00
-10.00

Figure S6