J Comput Aided Mol Des (2008) 22:469–478
DOI 10.1007/s10822-008-9175-x

Quantitative structure–activity relationship analysis of canonical inhibitors of serine proteases

Daniele Dell'Orco · Pier Giuseppe De Benedetti

Received: 8 May 2007 / Accepted: 9 January 2008 / Published online: 23 January 2008

Abstract Correlation analysis was carried out between binding affinity data values from the literature and physicochemical molecular descriptors of two series of single point mutated canonical inhibitors of serine proteases, namely bovine pancreatic trypsin inhibitor (BPTI) and turkey ovomucoid third domain (OMTKY3), toward seven enzymes. Simple quantitative structure–activity relationship (QSAR) models based on either single or double linear regressions (SLR or DLR) were obtained, which highlight the role of hydrophobic and bulk/polarizability features of mutated amino acids of the inhibitors in modulating both affinity and specificity. The utility of the QSAR paradigm applied to the analysis of mutagenesis data was underlined, resulting in a simple tool to quantitatively help deciphering structure–function/activity relationships (SFAR) of different protein systems.

Keywords Protein inhibitor · Protein–protein recognition · QSAR · Serine protease · Single point mutation modeling

Introduction

Proteases, or proteolytic enzymes, form one of the largest and most important groups of enzymes, entrusted to selectively catalyze the hydrolysis of peptide bonds. Proteases are involved in numerous important physiological processes. It is well established that uncontrolled, unregulated, or undesired proteolysis can lead to a number of serious diseases [1, 2]. Proteases inhibitors, thus, have considerable potential utility for therapeutic intervention in a variety of disease states.

Serine proteases and their natural protein inhibitors are among the most intensively studied protein–protein complexes [1–3]. A non-covalent protease-inhibitor complex, which is highly similar to substrate–enzyme interaction, is a very common way of inhibition. The most intensively studied example of substrate-like interaction is that of the so called canonical inhibitors of serine proteases [1]. The majority of serine proteases inhibitors is made up of rigid, stable, and purely beta-sheet or mixed alpha/beta proteins. Intriguingly, inhibitor loops show very similar canonical conformations across the families, despite the completely different amino-acid sequences [3–6]. The recognition of both substrate and inhibitor by most serine proteases involves the accommodation of the solvent-exposed primary binding residue (P1, notation of Schechter and Berger) [7] into the specific pocket (S1-site) of the enzyme. During this process, the side chain of the P1 residue buries into the S1 cavity of the enzyme, and the nature of the amino acid at this position greatly affects both the strength and the specificity of the non-covalent association [8]. The inhibitor loop segment P3–P3′ constitutes the principal contact area, which interacts mostly with the corresponding S3–S3′ region of the enzyme (Fig. 1). This binding loop is quite rigid and its conformation is not significantly altered upon complex formation [8]. The side chain of P1 has particular importance for the protease-inhibitor association and specificity. For this reason, P1 has been the main target of extensive single point mutations experiments, which provided insights into both principles of enzyme-inhibitor

D. Dell'Orco · P. G. De Benedetti (✉)
Department of Chemistry, University of Modena and Reggio Emilia, Via Campi 183, 41100 Modena, Italy
e-mail: deben@unimo.it

D. Dell'Orco
Dulbecco Telethon Institute, Via Campi 183, 41100 Modena, Italy

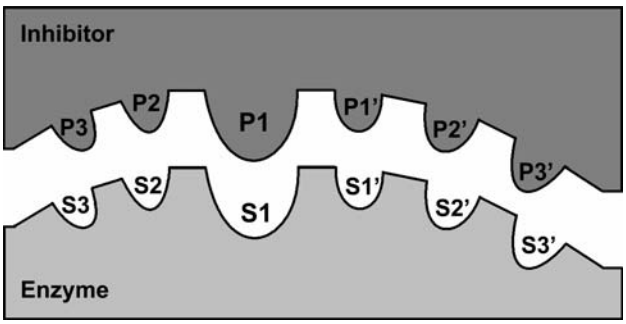

Fig. 1 Schematic diagram of the interaction model between a generic serine protease (light grey) and its proteic inhibitor (dark grey), according to the nomenclature introduced by Schechter and Berger [7]

interaction and general protein–protein recognition mechanisms [8, 9].

Different computational modelling approaches were reported on these bio-systems, which allow to correlate experimental thermodynamic properties of the different enzyme-inhibitor complexes with the change of the structural features of the inhibitors [10]. Particular noteworthy are the elegant works that effectively employed linear interaction energy and free energy perturbation methods [6, 11–13]. However, to the best of our knowledge, no quantitative structure–activity relationship (QSAR) analysis has been done so far on these very interesting bio-systems, in spite of the good and homogeneous experimental data available in the literature. In particular, the systematic single point mutations at the same position on the protein inhibitor, and the consequent binding affinity variations, suggest that the classical chemical situation of the substituent effect on the equilibrium/kinetic constants on a congeneric molecular series is applicable also in this case (i.e. congeneric protein series). In other words, we can assume that the postulates of the well known linear free enthalpy(/energy) relationships (LFERs) [14, 15] are also valid in the case of protein–protein interactions, when the only perturbation introduced in the system is a single point mutation in the same position, substituting for natural and noncoded amino acids. In this situation, the different mutated amino acid side chains with their physicochemical properties can assume a role similar to that of the substituent effect on the properties of congeneric series of small organic molecules, if the LFER (or extrathermodynamic) conditions are satisfied [14, 15]. Under the above assumptions, the simple and fast QSAR approach to mutational data analysis can help deciphering and predicting the molecular/structural mechanisms that influence both the protein intramolecular and intermolecular communication. In the former case, interesting applications and quantitative inferences on protein function were recently done [16, 17]. In the present study, we extend to protein–protein intermolecular interactions the above considerations, by taking into account two inhibitors, namely the bovine pancreatic trypsin inhibitor (BPTI) and the turkey ovomucoid third domain (OMTKY3), which were both systematically mutated at the P1 position to greatly affect the equilibrium constants (K_A) of their association with several serine proteases. The targeted serine proteases are, respectively, bovine α-chymotrypsin (BCHYM), human neutrophil elastase (HNE), bovine β-trypsin (BT) and anionic salmon trypsin (AST), for the BPTI inhibitor (Table 1), and BCHYM, porcine pancreatic elastase (PPE), *Streptomyces griseus* proteinase A (SGPA), and human leukocyte elastase (HLE) for OMTKY3 inhibitor (Table 2). Beside the 20 coded amino acids, systematic mutations at the OMTKY3-P1 position involved as well four noncoded amino acids able to appreciably affect the K_A values (Table 2, Methodology). The analysis has been done to obtain simple QSAR models between the experimental K_A values and selected molecular descriptors of the physicochemical properties of the different amino acids systematically introduced at the P1 position of the two inhibitors. The selected models are based either on single or double linear regressions (SLR or DLR) and show both explicative and predictive capability in all the tested cases. Moreover, they highlight and quantify some features that are observed in the few cases where the X-ray structures of the enzyme-inhibitor complexes are available (Fig. 2).

Table 1 Decimal logarithms of K_A (M^{-1}) for 18 BPTI-P1 coded variants in complex with each serine protease at pH = 8.3, $T = 295$ K

BPTI-P1 variant	BCHYM	BT	HNE	AST
D	4.08	4.80	3.00	4.20
N	6.96	7.34	5.11	7.15
Y	9.88	8.20	4.28	8.18
G	4.90	4.18	5.66	4.63
S	5.43	7.57	5.84	–
A	6.32	5.43	8.67	5.30
L	9.11	6.73	8.76	7.04
M	8.99	7.59	8.11	7.52
V	6.36	4.61	9.79	5.04
I	5.77	5.04	9.20	4.90
T	6.40	5.46	8.08	5.48
W	9.75	6.88	4.04	6.90
F	9.40	8.08	5.04	8.15
R	8.40	–	5.15	–
K	*7.94*	*13.23*	*4.89*	–
E	5.08	6.32	3.48	5.90
Q	7.72	6.36	5.84	6.85
H	7.87	6.79	4.57	6.91

Italic refers to wild type P1. Data are from Ref. 8

Table 2 Decimal logarithms of K_A (M^{-1}) for 24 OMTKY3-P1 coded and non-coded variants in complex with each serine protease at pH = 8.3, T = 295 K

OMTKY3-P1 variant	BCHYM	SGPA	HLE	PPE
D	6.00	6.71	4.61	4.86
N	8.82	8.41	7.36	7.84
Y	12.91	11.00	4.92	3.88
G	6.82	7.65	7.34	8.99
S	7.62	8.20	7.52	8.93
A	7.79	9.28	9.00	10.62
L	*11.28*	*11.48*	*9.79*	*10.62*
M	11.00	11.41	8.86	10.11
V	8.15	9.32	10.15	9.94
I	8.00	8.18	10.32	9.77
T	7.99	9.30	9.11	10.46
W	12.52	9.95	4.26	4.40
F	12.38	11.26	6.00	4.69
R	8.34	7.91	4.52	3.68
K	6.40	7.75	5.58	4.66
E	6.40	6.90	4.20	4.92
Q	9.11	9.08	7.36	7.63
H	9.08	9.41	5.04	4.57
C	9.34	11.52	9.85	10.40
P	4.83	4.66	5.28	5.77
Abu	9.04	10.00	10.15	11.52
Ape	10.34	11.15	10.00	11.38
Ahx	10.90	11.36	9.00	11.28
Hse	9.40	9.81	8.41	9.88

Italic refers to wild type P1. Data are from Ref. 9

Finally, we critically compare the QSAR results with those from other recently developed computational methods to fast assess the effects of point mutations on protein–protein binding affinity.

Methodology

Data sets

The data set concerning the interaction of 18 BPTI variants, holding an amino acid substitution at the P1 position, was selected from the literature [8]. Table 1 reports the binding affinity data values of BPTI variants for BCHYM, BT, HNE and AST serine proteases. Moreover, Table 2 reports the binding affinity data values for the interaction of 24 OMTKY3-P1 variants, also selected from the literature [9]. These include four non-coded synthetic amino acids, namely α-amino-butyric acid (Abu), α-amino-pentanoic acid or norvaline (Ape), α-amino-hexanoic acid or norleucine (Ahx) and α-amino-hydroxy-butyric acid or homoserine (Hse). In all the cases reported in Tables 1 and 2, experimental binding affinity measurements were carried out at pH = 8.3 and temperature T = 295 K [8, 9], providing a homogeneous data set for both the inhibitors series.

Molecular descriptors

A total of 84 molecular descriptors of the physicochemical properties of natural and artificial amino acids were used to build the QSAR models. They include a variety of hydrophobic and hydrophilic parameters [18–22], size descriptors [23, 24], volume and surface area values [25–32], solution and chromatographic properties [25], polarity and polarizability indices [26, 29], and other molecular descriptors developed in the Wold's lab by making use of principal component (PCA) and partial least squares (PLS) analysis on independent families of various hydrophobic, steric and electronic properties of both the 20 coded and noncoded amino acids [29, 33]. Among the employed descriptors, particularly effective for the building of QSAR models were those proposed by Sandberg et al. [29] who characterized 87 amino acids, including artificial ones, by 26 physicochemical descriptors determined from thin-layer chromatography retention values, nuclear magnetic resonance (NMR) chemical shift, semi-empirical molecular orbital calculations, total, polar, and nonpolar surface area, van der Waals side-chain volume, log P, molecular weight, and four variables describing hydrogen bond donor and acceptor properties, and side-chain charge [29]. The resulting principal property scales obtained, i.e. the z-scales, directly reflect, respectively, hydrophobicity (z_1), steric bulk/polarizability (z_2) and electronic (z_3) features of the residue. Table 3 reports a list of the descriptors used in the QSAR models obtained in this study.

Statistical analysis

Linear correlation analysis was done between the association constants (K_A's) values for each variant and the whole set of molecular descriptors by determining the correlation coefficient matrix corresponding to each enzyme-inhibitor system. Only the statistically meaningful correlations were considered for building QSAR models. SLR and DLR were used to build the QSAR models, the intercepts and coefficients of which are reported with their 95% confidence intervals. To test the stability of the models, leave-one-out (LOO) cross validation tests were carried out against each model. The statistical quality was judged by the correlation coefficient (R), the standard deviation from the regression (s), the Fisher's F-test value (F) and, for DLR models,

Table 3 Molecular descriptors employed in QSAR SLR and DLR models (see Eqs. 1–12)

Descriptor	Significance	Reference
α	Residue polarizability	[22]
rh	Estimated free energy difference for the transfer of a residue from a random coil conformation in water to an α-helical conformation in a lipophilic phase	[34]
KD	Estimated scale of natural amino acids hydrophilic and hydrophobic properties	[19]
z_1	First principal property of amino acids: empirically and theoretically derived residue lipophilicity scale	[29]
z_2	Second principal property of amino acids: empirically and theoretically derived residue steric bulk/polarizability scale	[29]
z_3	Third principal property of amino acids: empirically and theoretically derived residue polarity scale	[29]

R_{cv} and s_{cv} for LOO cross validations. All the statistical analyses were performed by means of the free software STATIST v. 1.3.0 (http://statist.wald.intevation.org).

Solvent accessible surface areas calculations

Solvent accessible surface areas (ASAs) were calculated by means of the QUANTA 2005 package (www.accelrys.com) on the available X-ray structures of the enzyme-inhibitor complexes, i.e. BPTI inhibitor in complex with BT and BCHYM enzymes and OMTKY3 inhibitor in complex with HLE and BT enzymes (see Fig. 1). A 1.4 Å probe radius was used for the solvent and the computation was extended both on the exposed and buried hydrophobic and hydrophilic atoms. The ASAs computations concerned both the P3–P3′ and S3–S3′ regions, and the individual P1 side chain (see Fig. 1).

Results

Among the wide set of molecular descriptors considered, those reported in Table 3 provided satisfactory SLR and DLR QSAR models for each tested case of serine protease in complex with either BPTI or OMTKY3 inhibitors. The selected models are reported and discussed in the next paragraphs.

BCHYM–BPTI QSAR modeling

Among the analyzed enzyme-inhibitor sets, the interaction between BCHYM and wild type-BPTI (Table 1) is the one with lower binding affinity, i.e. $K_A = 8.8 \times 10^7\ M^{-1}$ [8]. Details on the binding specificity, in terms of the P3–P3′ inhibitor region interacting with the corresponding S3–S3′ enzyme region are represented in Figs. 1 and 2a. The best SLR model obtained for this system is:

$$\log K_a = 14.48\ (\pm 2.53)\alpha + 4.62\ (\pm 0.52) \quad (1)$$

$$n = 18,\ R = 0.82,\ s = 1.06$$

where α is the polarizability of the substituting residue [22]. Moreover, by performing DLR analysis omitting the outliers D and I, it was possible to obtain two further QSAR models for this system:

$$\log K_a = -0.30\ (\pm 0.07)z_1 + 0.47\ (\pm 0.09)z_2 + 7.54\ (\pm 0.20) \quad (2)$$

$$n = 16,\ R = 0.89,\ F = 24.50,\ s = 0.82$$

where z_1 and z_2 are two principal properties of the amino acids (z-scales) [29] which describe, respectively, the hydrophobic and steric bulk/polarizability effect of the amino acids side chains on the K_a values. The other DLR model obtained omitting the outliers I and W is:

$$\log K_a = 17.12\ (\pm 2.10)\alpha + 0.18\ (\pm 0.05)\mathrm{rh} + 4.31\ (\pm 0.40) \quad (3)$$

$$n = 16,\ R = 0.92,\ F = 37.48,\ s = 0.73$$

where α is the residue polarizability and rh is the free energy (kcal/mol) of transfer for the isolated P1 residue from a random coil/aq state to a α-helix/lipid state [34].

It is worth noting that the linear correlation between α and z_2 ($R = 0.81$) and between rh and z_1 ($R = -0.78$) makes consistency in the three models. In summary, the specificity of the BPTI–BCHYM binding results from a combination of hydrophobic effect and polarizability of the P1 residue. Despite the binding affinity is quite high for this system ($\Delta G^\circ = -11$ kcal/mol), the P3–P3′ BPTI region is buried in the cleft including the BCHYM S3–S3′ region less tightly than in other cases (Fig. 2a) and with lower specificity. Indeed the K15 residue, which corresponds to the wild type BPTI-P1, adopts a less extended side-chain conformation than that adopted when bound to BT (Fig. 2b). In fact, a detailed inspection of the 3D structure of the BCHYM–BPTI complex (PDB entry: 1CBW)

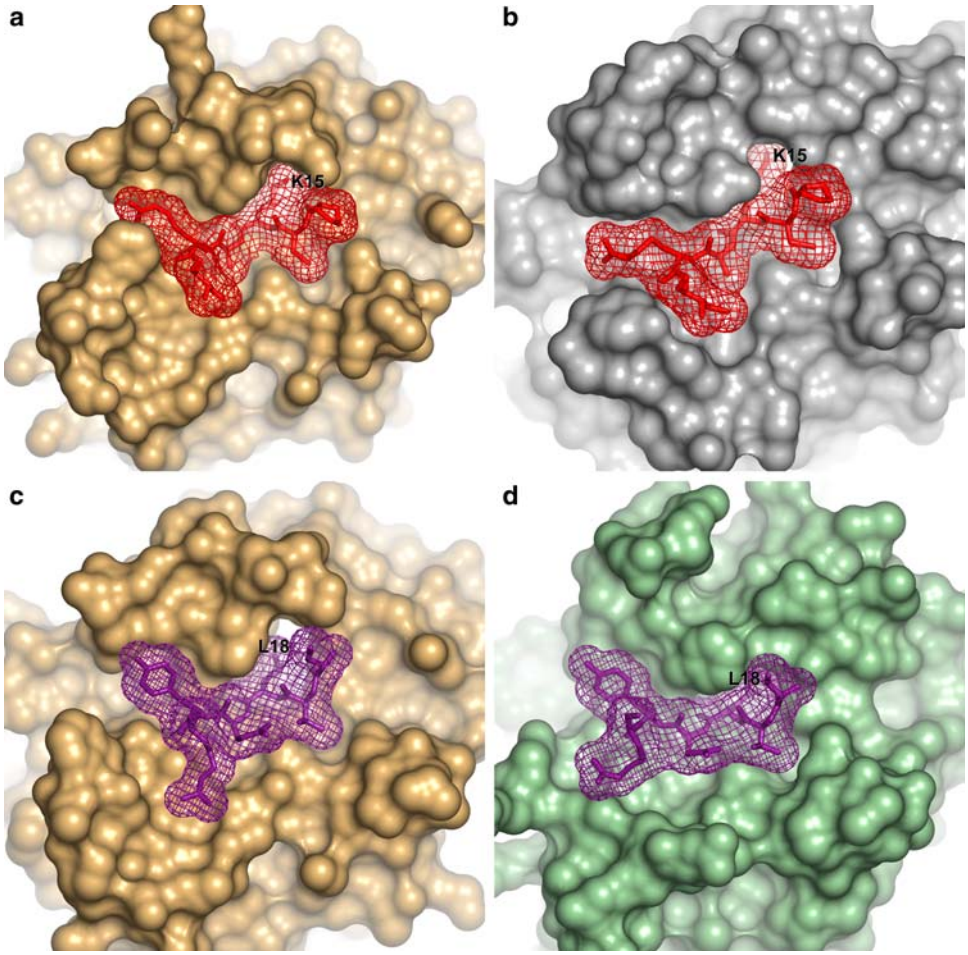

Fig. 2 X-ray structures of the binding regions of four serine proteases-inhibitor complexes analyzed in this study. The P3–P3′ inhibitor regions are represented as sticks within a mesh, representing the molecular surface, whereas the S3–S3′ enzyme cleft and the surrounding regions are represented by the enzyme molecular surface. For clarity, only the inhibitor residues involved in the binding region are represented, i.e. those corresponding to the P3–P3′ sequence. The P1 residue in each inhibitor is labelled. (**a**) BCHYM–BPTI interaction (PDB entry: 1CBW). Here, as in (**b**), BPTI-P3–P3′ is coloured in red, whereas BCHYM is coloured in light orange, as in **c**). (**b**) BT-BPTI interaction (PDB entry: 3BTK). Here BT is coloured in gray. (**c**) BCHYM-OMTKY3 interaction (PDB entry: 1CHO). Here, as in **d**), OMTKY3-P3–P3′ is coloured in purple. (**d**) HLE-OMTKY3 interaction (PDB entry: 1PPF). Here, HLE is coloured in light green. Drawings were done by means of the software PYMOL 0.97 (www.pymol.source-forge.net)

reveals that the pattern of H-bonds (H-bonds) is made up of four main-chain to main-chain H-bonds as well as four side-chain to main-chain H-bonds. This reflects the extremely lower affinity of BCHYM–BPTI compared to the BT–BPTI case, whose H-bond pattern is made up of seven main-chain to main-chain bonds, four side-chain to main-chain bonds and one side-chain to side-chain bond (PDB entry: 3BTK).

The predictive capability of the QSAR model in Eq. 3 is reported in Fig. 3a. The statistical parameters for the LOO cross-validation remain satisfactory ($R^2_{cv} = 0.74$, $s_{cv} = 0.86$).

BT–BPTI QSAR modeling

The interaction between BT and wild type-BPTI (Table 1) is the one with the highest affinity among the studied cases ($K_A = 1.7 \times 10^{13}\ M^{-1}$, i.e. $\Delta G^\circ = -18.0$ kcal/mol) [8]. The details of binding specificity are highlighted by the X-ray structure (Fig. 2b), which clearly shows the P3–P3′ BPTI region tightly packed (at variance with respect to the BCHYM–BPTI complex, Fig. 2a) into the cleft formed by the S3–S3′ BT region. In this case, the P1-K15 side chain protrudes in an extended conformation within a narrow cleft towards the carboxyl of BT-D189 (Fig. 2b). A detailed inspection of the 3D structure (PDB entry: 3BTK) confirms that a network of H-bonds greatly contributes to optimize the interaction, as illustrated above.

However, when building the QSAR models for this system, the extremely high K_A value of the wild type P1 (i.e. a lysine residue) with respect to all other substitutions created a high perturbation in the distribution of the affinity data values (Table 1), suggesting to exclude this point from the correlation analysis. This issue was already discussed in our previous work where we noticed that the data concerning the wild type were measured much earlier and in different conditions than those regarding all the other substitutions [35].

The SLR model obtained for this system is:

$$\log K_a = 0.54\ (\pm 0.10) z_2 + 6.67\ (\pm 0.20) \qquad (4)$$

$$n = 14,\ R = 0.84,\ s = 0.73$$

omitting the D and W outliers.

By considering the same data set, a DLR model was built, that shows improved statistical parameters:

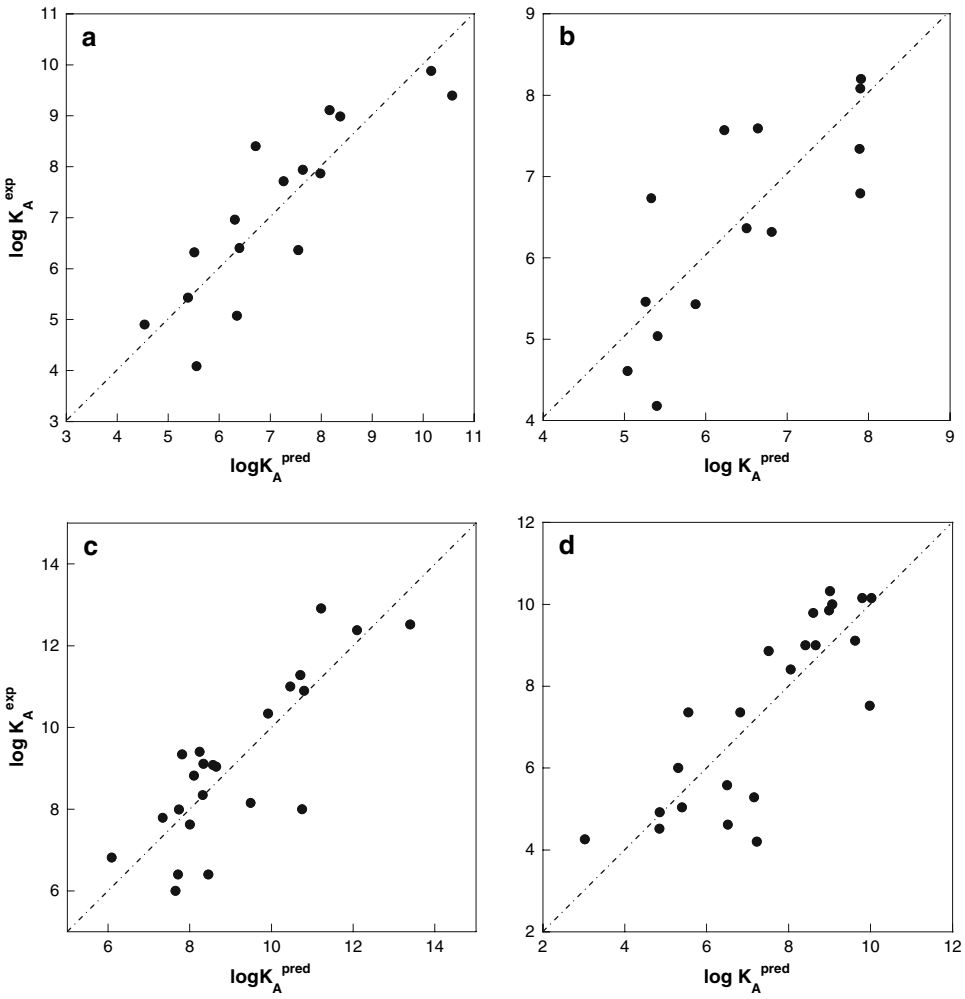

Fig. 3 Cross-validation prediction plots for LOO tests performed on the same four systems reported in Fig. 2. (**a**) BCHYM–BPTI interaction. The fitted correlation equation is: $\log K_A^{exp} = 0.9 \log K_A^{pred} + 0.7$ ($R_{cv}^2 = 0.74$, $s_{cv} = 0.86$). (**b**) BT–BPTI interaction. The fitted correlation equation is: $\log K_A^{exp} = 0.9 \log K_A^{pred} + 0.5$ ($R_{cv}^2 = 0.61$, $s_{cv} = 0.60$). (**c**) BCHYM–OMTKY3 interaction. The fitted correlation equation is: $\log K_A^{exp} = 0.9 \log K_A^{pred} + 0.6$ ($R_{cv}^2 = 0.68$, $s_{cv} = 0.67$). (**d**) HLE–OMTKY3 interaction. The fitted correlation equation is: $\log K_A^{exp} = 1.0 \log K_A^{pred} + 0.4$ ($R_{cv}^2 = 0.67$, $s_{cv} = 0.62$)

$$\log K_A = 0.48\ (\pm 0.10) z_2 + 0.32\ (\pm 0.19) z_3 + 6.68\ (\pm 0.19) \quad (5)$$

$$n = 14,\ R = 0.87,\ F = 17.58,\ s = 0.69$$

where z_2 and z_3 are two principal properties of amino acids (z-scales) [29] describing, respectively, the bulk/polarizability and electronic properties of the amino acids.

In this case, while the bulk effects and the residue polarizability seem to constitute the major requirement for binding affinity and specificity, fine tuning of the electronic properties such as the residue polarity, represented by the z_3 scale, contribute as well to modulate the K_A values. The predictive capability of the Eq. 5 is reported in Fig. 3b. The statistical parameters for the LOO cross-validation are quite satisfactory ($R_{cv}^2 = 0.61$, $s_{cv} = 0.78$), although the point distribution is significantly more scattered than in the previous case.

HNE–BPTI QSAR modeling

Also in the case of the HNE–BPTI complex, we could obtain QSAR models useful to interpret the experimental findings (Table 1). The first SLR model obtained excluding the F outlier is:

$$\log K_a = 0.61\ (\pm 0.09) \mathrm{KD} + 6.63\ (\pm 0.29) \quad (6)$$

$$n = 17,\ R = 0.86,\ s = 1.14$$

where KD is the hydropathy index of amino acids proposed by Kyte and Doolittle [19], which assumes positive values for nonpolar side chains and negative values for polar and charged side chains. When the same outlier is omitted in the DLR analysis, a DLR model is obtained which includes the same descriptor of Eq. 6 and z_3, showing improved statistics:

$$\log K_a = 0.60\ (\pm 0.08) \mathrm{KD} - 0.40\ (\pm 0.17) z_3 + 6.47\ (\pm 0.26) \quad (7)$$

$$n = 17,\ R = 0.90,\ F = 30.00,\ s = 1.01$$

Interestingly, the models reported in Eqs. 6 and 7 indicate that, on one hand, affinity increases with the hydropathy (KD) values of the P1 residue, on the other, it consistently decreases with the polarity (z_3) of the substituting residue.

The statistical parameters for the LOO cross-validation of the model in Eq. 7 are satisfactory ($R^2_{cv} = 0.75$, $s_{cv} = 0.86$).

AST–BPTI QSAR modeling

No X-ray structure is available for the AST–BPTI complex, and less data points, i.e. fifteen P1 substitutions, were achievable from the literature compared to other BPTI-interacting systems (Table 1). While no SLR model could be built, a DLR model was built omitting the D and W points:

$$\log K_a = -0.12\ (\pm 0.06\ )z_1 + 0.52\ (\pm 0.08)z_2 + 6.56\ (\pm 0.17) \quad (8)$$

$$n = 13,\ R = 0.91,\ F = 23.31,\ s = 0.57$$

Interestingly, the same outliers were found for the models reported in Eqs. 4 and 5, which refer to the BT–BPTI interaction. This is consistent with the observation that BT and AST show about 65% amino acid sequence homology and have very similar architecture of the binding pocket [8].

The statistical parameters for the LOO cross-validation of the model in Eq. 8 are satisfactory ($R^2_{cv} = 0.72$, $s_{cv} = 0.84$).

BCHYM–OMTKY3 QSAR modeling

The high affinity of OMTKY3 for BCHYM ($K_A = 1.9 \times 10^{11}\ M^{-1}$, i.e. $\Delta G^\circ = -15.4$ kcal/mol; Table 2) results from the tight packing of the P3–P3′ into the S3–S3′ cleft, characterized by the protrusion of the P1(L18) residue towards the narrow enzyme apolar pocket (Fig. 2c). A detailed inspection of the 3D structure as compared with that of the BPTI inhibitor highlights the rotation of BCHYM-F39's benzyl towards the solvent to better accommodate the P2′–P3′ region in the interaction area surrounding BPTI-R17 and OMTKY3-Y20, respectively. Other conformational changes concern the side chain of BCHYM-N150, which ameliorate the interaction with OMTKY3-Y20 (i.e. the P2′ residue), and other minor slight rearrangements.

A DLR QSAR model was built for this system, omitting the outlier P:

$$\log K_a = -0.53\ (\pm 0.08)z_1 + 0.43\ (\pm 0.11)z_2 + 9.10\ (\pm 0.22) \quad (9)$$

$$n = 23,\ R = 0.87,\ F = 29.62,\ s = 1.05$$

The predictive capability of the Eq. 9 is reported in Fig. 3c. The statistical parameters for the LOO cross-validation are quite satisfactory ($R^2_{cv} = 0.68$, $s_{cv} = 0.82$).

It is worth noting that substitution of OMTKY3-P1 native leucine with the noncoded amino acids Abu, Ape, Ahx and Hse does not increase the binding affinity (Table 2). The contrary happens with P1-Abu, Ape and Ahx for the PPE enzyme (Table 2).

SGPA–OMTKY3 QSAR modeling

The SGPA–OMTKY3 interaction is characterized by a high binding affinity ($K_A = 3 \times 10^{11}\ M^{-1}$, i.e. $\Delta G^\circ = -15.6$ kcal/mol). We could obtain a SLR model for this system, omitting the outliers I and P:

$$\log K_a = -0.42\ (\pm 0.07)z_1 + 9.45\ (\pm 0.20) \quad (10)$$

$$n = 22,\ R = 0.80,\ s = 0.95$$

The major contribution of P1 to binding affinity arises from its hydrophobicity. The statistical parameters for the LOO cross-validation of the model in Eq. 10 are quite satisfactory ($R^2_{cv} = 0.58$, $s_{cv} = 0.76$).

In this case, P1 substitutions for the four noncoded amino acids exert either little (Ape, Ahx) or destabilizing (Abu, Hse) effect on the complex stability (Table 2).

HLE–OMTKY3 QSAR modeling

The X-ray structure of HLE–OMTKY3 complex suggests neither strong ionic interactions, nor highly specific interactions except for the burying of P1-L18 into a narrow enzyme cleft (Fig. 2d). The presence of several side chain-to-backbone and backbone-to-backbone H-bonds, however, confers on the enzyme-inhibitor complex a high binding affinity ($K_A = 6.1 \times 10^9\ M^{-1}$, i.e. $\Delta G^\circ = -13.3$ kcal/mol).

We obtained a SLR model for this system, omitting the G outlier:

$$\log K_a = -1.02\ (\pm 0.14)z_2 + 7.35\ (\pm\ 0.26) \quad (11)$$

$$n = 23,\ R = 0.85,\ s = 1.2$$

Equation 11 clearly points to the relevant role of the steric/bulk properties of the P1 substituted amino acids. In particular, the negative slope implies that high affinity is achievable only for relatively small side chains.

The predictive capability of Eq. 11 is reported in Fig. 3d. The statistical parameters for the LOO cross-validation are quite satisfactory ($R^2_{cv} = 0.67$, $s_{cv} = 0.82$).

PPE–OMTKY3 QSAR modeling

The PPE–OMTKY3 complex represents an example of enhanced binding affinity following replacement of the wild-type OMTKY3-P1 residue by three noncoded amino

acids, namely Abu, Ape and Ahx (Table 2). In each case, the K_A value is increased of one order of magnitude (Table 2). Similarly to the case of HLE, omitting the G outlier, the model obtained highlights the anti-correlation between affinity and steric bulk/polarizability:

$$\log K_a = -1.34\ (\pm 0.17) z_2 - 7.78\ (\pm\ 0.37) \qquad (12)$$

$$n = 23,\ R = 0.87,\ s = 1.47$$

The predictive capability of the QSAR model in Eq. 12 is confirmed by the LOO cross-validation, which shows quite satisfactory statistical parameters ($R^2_{cv} = 0.70$, $s_{cv} = 0.84$).

Solvent accessible surface areas (ASAs) buried upon enzyme-inhibitor binding

Solvent accessible surface areas (ASAs) buried upon binding were calculated in the S3–S3′/P3–P3′ region for each of the enzyme-inhibitor complexes of known 3D structure (Fig. 2). No significant differences were observed. In particular: (a) a 1,120 Å^2 ASA is buried upon BPTI–BCHYM binding (Fig. 2a), (b) a 1,153 Å^2 ASA is buried upon BPTI–BT binding (Fig. 2b); (c) a 1,099 Å^2 ASA is buried upon OMTKY3–BCHYM binding (Fig. 2c); (d) a 1,006 Å^2 ASA is buried upon OMTKY3–HLE binding (Fig. 2d). In all the tested cases, about 30% of the buried ASA arises from burying polar atoms. Moreover, no significant difference among the test cases was observed when the buried ASA was computed on the P1 residue only.

Discussion

While QSAR analysis has demonstrated extremely powerful in issues related to drug-design/discovery and found a great employment in quantifying structure-related general properties of small molecules (QSPR), it is still in its infancy in the study of bio-macromolecules. Limited applications of QSAR modelling are found in the literature to study, for instance, protein–protein interactions, or the mutational effects on protein structure–function relationships. With respect to the latter case, few QSAR models were obtained for a G protein-coupled receptor (GPCR) by systematically mutating a single residue in a fixed position with all of the 19 coded amino acids [16]. In this way, a series of 20 similar proteins was obtained, which differed only in one amino acid at the same crucial mutation site. The mutated protein set was the object of simple and interesting QSAR models, which helped deciphering the role of different amino acid structures/properties on GPCRs functional modulation [16]. This example showed how QSAR modelling may represent a useful tool for quantitative protein structure–function analysis, i.e. intramolecular interactions. As for protein–protein interactions, i.e. intermolecular interactions, QSAR approaches are not yet a common practice, despite the good results already obtained for antibody-lysozyme interaction [36], endonuclease-inhibitor interaction [37], RAS proteins-effectors interaction [38] and several peptide–protein interactions [39]. In this respect, the canonical protein inhibitors of serine proteases, being representative of the nearly lock and key binding model, constitute an excellent biological system to challenge the QSAR modelling paradigm on protein–protein interactions. Moreover, by considering (a) the impressive development of biochemistry and molecular biology (and the associated technologies), (b) the extensive and systematic approaches to mutational analysis aiming to capture molecular/structural mechanistic features of biosystems structure–function and (c) the explosive production of protein–protein interactions data, overall these points suggest that QSAR modelling of mutational data on protein systems should be a natural consequence of the above considerations and may constitute a simple and fast tool to extract and estimate further information from quantitative experimental data.

Specifically, the QSAR models obtained in this study and expressed in Eqs. 1–12 allow for some general interpretation concerning the BPTI and OMTKY3 inhibitory potency on several serine proteases, as exerted through tight binding complexes (Tables 1 and 2; Fig. 2). In detail, for BCHYM–BPTI, a high affinity corresponds to a high polarizability of the P1 residue (Eq. 1), and is enhanced by highly hydrophobic residues (Eqs. 2 and 3). Similarly, high polarizability is required for the achievement of a high affinity in the BT–BPTI complex, which, however, is significantly modulated also by the residue polarity arising from the electronic properties of the P1 side chain (Eqs. 4 and 5). Differently from the previous case, the major responsible for high affinity in the HNE–BPTI complex is still the high hydrophobicity of the P1 residue (Eq. 6), but a high polarity exerts, consistently with high hydrophobicity, a negative influence on the binding affinity (Eq. 7). Finally, and similarly to the BCHYM–BPTI case, for the AST–BPTI complexes, higher affinities are correlated with high bulk/polarizability and hydrophobicity of the substituted residues (Eq. 8). It is noteworthy to observe that, excluding Eq. 7, in Eqs. 2–5 and 8 the common outliers are D and W residues. A simple explanation for these outliers could be associated to insufficient description of their physicochemical features and/or to a conformational perturbation due to D and W residues on the BPTI loop interacting with the inhibited enzymes. As for the OMTKY3 inhibitor, the BCHYM inhibition is achieved with high bulk/polarizability and hydrophobicity of the P1 residue (Eq. 9), which

is consistent with the BPTI case. A high P1 hydrophobicity is also required for specificity in the SGPA–OMTKY3 binding (Eq. 10), whereas for HLE–OMTKY3 and PPE–OMTKY3 it is essential that the P1 substitute has low size/polarizability of the side chain in order to achieve the required high binding affinities (Eqs. 11 and 12). This trend is opposite with respect to that observed for the BPTI complexes. Also in this case, a possible explanation for the omitted outliers in Eqs. 9–12 is similar to that proposed for BPTI inhibitor.

Collectively, our QSAR models of serine protease-inhibitor interactions allow for quantitative interpretations of the many inferences previously done by combining data arising from high resolution X-ray determinations with the plenty of mutagenesis experiments, and lead to consistent conclusions concerning the major determinants of high affinity for the bio-systems considered [4, 5, 8, 9, 40]. In this respect, the principal properties of amino acids (z-scales, Table 3) demonstrated particularly effective as molecular descriptors.

Recently, we discussed the performance of a broadly used empirical method developed by Murphy et al. [41, 42] for the thermodynamic analysis of protein–protein interactions, starting from high resolution structures, as for the case of BT–BPTI interaction [43]. This method has been widely used in the literature to empirically calculate the free energy change from structural information. The basic assumption is that enthalpy and entropy changes upon binding can be both linearly related to changes in the solvent accessible surface area (ΔASA) of polar and nonpolar protein atoms through empirical relationships [42]. Although the method demonstrated effective in predicting the absolute ΔG° values, i.e. K_A, for the OMTKY3–PPE interaction [41], when applied to a set of eight P1 variants of BT–BPTI interactions it led to incorrect predictions [43]. Similarly, the method was successful in predicting ΔG° values, i.e. K_A, for the reconstitution of the calcium binding protein calbindin D9k from wild type fragments, whereas it resulted in wrong predictions when the reconstitution occurred starting from mutated fragments [44]. The results presented in this study clearly show that the ASA buried upon binding may not be a proper descriptor to discriminate between different protein–protein interactions when interfaces are somewhat similar. Indeed, all the four cases considered in this study for which high resolution structures are available (Fig. 2), present only small differences in ASAs buried upon binding, but, on the other hand, binding affinity ranges over six orders of magnitude (Tables 1 and 2). Hence, in order to achieve sound explanations and predictions of binding affinity, a more extensive search for molecular descriptors such as the one presented in this study is required in most cases.

Among the computational techniques aimed at predicting structural features of protein–protein interactions such as enzyme-inhibitor complexes, protein docking is gaining an ever increasing importance. One of the main limits of such an approach is that the static prediction of structural features, though accurate, generally offers very restricted information about the dynamic process of protein–protein binding and dissociation, thus requiring further computational demanding simulations such as molecular or Brownian dynamics to get some kinetic and/or thermodynamic insight. We recently demonstrated that rigid-body docking simulations of protein–protein association alone may lead to quantitative models able to describe and predict the thermodynamic and kinetic features of the association process [43, 44]. The docking algorithm chosen in those studies is a successful Fast Fourier Transform algorithm that effectively performs a search for optimization of intermolecular complementarity and provides a score accounting for the shape and the electrostatic complementarities, as well as for desolvation of protein interfaces upon binding [45]. Based upon this algorithm, we set up a computational protocol leading to a convincing correlation between the empirical docking score and ΔG° values for a number of protein–protein complexes including BCHYM–OMTKY3 and HLE–OMTKY3 [44]. The protocol was successfully employed for prediction of thermodynamic and kinetic properties of wild type and mutated protein reconstitutions [44], protein–DNA interactions [46], a number of enzyme-inhibitor interactions [35, 43] and very recently a case of homodimerization of protein domains within a biological membrane [47]. Interestingly, when comparing the results obtained from docking simulations on BT–BPTI complexes, i.e. a serine protease-inhibitor association, with those concerning endonuclease-inhibitor interactions, we found a relevant role of combined desolvation/electrostatics upon binding for the former [35]. Although the docking study was limited to eight BPTI-P1 mutations, the results are in line with the overall findings in the present study that, depending on the serine-protease system, either P1's hydrophobicity, size, polarity, or combinations between them, are important for determining the binding affinity and specificity. Thus, one may conclude that while in the presence of high resolution structures the docking based method may allow for an accurate structural and thermodynamic characterization of the protein–protein complex, independently of the type and the number of mutations, the QSAR analysis presented in this study provides consistent results in the case of systematic amino acid replacements in the crucial mutation site, and allows for very fast predictions even in the absence of a high resolution structure of the complex.

Finally, an attempt to generate a QSAR model which include all the four enzyme systems for the two types of inhibitors considered was unsuccessful.

Conclusions

An exhaustive QSAR analysis of various serine proteases in complex with two engineered protein inhibitors has been performed employing molecular descriptors of the physicochemical properties of the mutated amino acids. The simple SLR and DLR QSAR models obtained allowed for a clear interpretation of many experimental mutagenesis data of the bio-systems considered. Moreover, satisfactory binding affinity data values validations/predictions of the enzymes for the inhibitors were obtained also when the replaced amino acid was noncoded, i.e. artificial. This latter aspect is particularly interesting for protein design purposes and encourages an extensive use of QSAR modelling for handling issues such as the molecular engineering problems of protein–protein interactions. Finally, the high-throughput mutagenesis approaches now available generate large pools of interesting experimental data that can be modelled by many methods, including the well consolidated and simple QSAR paradigm. In this respect, it is now relatively easy to shift from structure–function/activity relationships (SFAR) towards quantitative SFAR.

Acknowledgment We gratefully acknowledge Prof. Francesca Fanelli for helpful discussions.

References

1. Otlewski J, Jelen F, Zakrzewska M, Oleksy A (2005) EMBO J 24:1303
2. Otlewski J, Krowarsch D, Apostoluk W (1999) Acta Biochim Pol 46:531
3. Bode W, Huber R (1992) Eur J Biochem 204:433
4. Helland R, Berglund GI, Otlewski J, Apostoluk W, Andersen OA, Willassen NP, Smalas AO (1999) Acta Crystallogr D Biol Crystallogr 55:139
5. Helland R, Otlewski J, Sundheim O, Dadlez M, Smalas AO (1999) J Mol Biol 287:923
6. Mekonnen SM, Olufsen M, Smalas AO, Brandsdal BO (2006) J Mol Graph Model 25:176
7. Schechter I, Berger A (1967) Biochem Biophys Res Commun 27:157
8. Krowarsch D, Dadlez M, Buczek O, Krokoszynska I, Smalas AO, Otlewski J (1999) J Mol Biol 289:175
9. Lu W, Apostol I, Qasim MA, Warne N, Wynn R, Zhang WL, Anderson S, Chiang YW, Ogin E, Rothberg I, Ryan K, Laskowski M Jr (1997) J Mol Biol 266:441
10. Laskowski M Jr, Qasim MA, Yi Z (2003) Curr Opin Struct Biol 13:130
11. Almlof M, Aqvist J, Smalas AO, Brandsdal BO (2006) Biophys J 90:433
12. Brandsdal BO, Aqvist J, Smalas AO (2001) Protein Sci 10:1584
13. Brandsdal BO, Smalas AO, Aqvist J (2006) Proteins 64:740
14. Franke R (1984) Theoretical drug design methods. Elsevier
15. Selassie CD (2003) In: Abraham DJ (ed) Burger's medicinal chemistry and drug discovery. Wiley
16. Scheer A, Fanelli F, Costa T, De Benedetti PG, Cotecchia S (1997) Proc Natl Acad Sci USA 94:808
17. Bahia DS, Wise A, Fanelli F, Lee M, Rees S, Milligan G (1998) Biochemistry 37:11555
18. Hopp TP, Woods KR (1981) Proc Natl Acad Sci USA 78:3824
19. Kyte J, Doolittle RF (1982) J Mol Biol 157:105
20. Nicholls A, Sharp KA, Honig B (1991) Proteins 11:281
21. Sharp KA, Nicholls A, Friedman R, Honig B (1991) Biochemistry 30:9686
22. Charton M, Charton BI (1982) J Theor Biol 99:629
23. Frommel C (1984) J Theor Biol 111:247
24. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH (1985) Science 229:834
25. Eriksson L, Jonsson J, Sjostrom M, Wold S (1989) Prog Clin Biol Res 291:131
26. Grantham R (1974) Science 185:862
27. Krigbaum WR, Komoriya A (1979) Biochim Biophys Acta 576:229
28. Krigbaum WR, Komoriya A (1979) Biochim Biophys Acta 576:204
29. Sandberg M, Eriksson L, Jonsson J, Sjostrom M, Wold S (1998) J Med Chem 41:2481
30. Sjostrom M, Eriksson L, Hellberg S, Jonsson J, Skagerberg B, Wold S (1989) Prog Clin Biol Res 291:313
31. Skagerberg B, Clementi S, Sjostrom M, Tosato ML, Wold S (1989) Prog Clin Biol Res 291:127
32. Stahle L, Wold S (1988) Prog Med Chem 25:291
33. Mei H, Liao ZH, Zhou Y, Li SZ (2005) Biopolymers 80:775
34. von Heijne G, Blomberg C (1979) Eur J Biochem 97:175
35. Dell'Orco D, De Benedetti PG, Fanelli F (2007) BMC Struct Biol 7:37
36. Freyhult EK, Andersson K, Gustafsson MG (2003) Biophys J 84:2264
37. Wang T, Tomic S, Gabdoulline RR, Wade RC (2004) Biophys J 87:1618
38. Tomic S, Bertosa B, Wang T, Wade RC (2007) Proteins 67:435
39. Wang T, Wade RC (2002) J Med Chem 45:4828
40. Scheidig AJ, Hynes TR, Pelletier LA, Wells JA, Kossiakoff AA (1997) Protein Sci 6:1806
41. Baker BM, Murphy KP (1997) J Mol Biol 268:557
42. Baker BM, Murphy KP (1998) Methods Enzymol 295:294
43. Dell'Orco D, De Benedetti PG, Fanelli F (2006) From computational biophysics to systems biology workshop, vol 34. NIC series, p. 67
44. Dell'Orco D,Seeber M, De Benedetti PG, Fanelli F (2005) J Chem Inf Model 45:1429
45. Chen R, Li L, Weng Z (2003) Proteins 52:80
46. Fanelli F, Ferrari S (2006) J Struct Biol 153:278
47. Dell'Orco D, De Benedetti PG, Fanelli F (2007) J Phys Chem B 111:9114