Amino Acids (2010) 38:1209–1218
DOI 10.1007/s00726-009-0332-x

ORIGINAL ARTICLE

Toward quantitative characterization of the binding profile between the human amphiphysin-1 SH3 domain and its peptide ligands

Ping He · Wei Wu · Hai-Dong Wang · Kang Yang · Ke-Long Liao · Wei Zhang

Received: 10 April 2009 / Accepted: 22 July 2009 / Published online: 8 August 2009

Abstract A new structure-based approach was proposed to quantitatively characterize the binding profile of human amphiphysin-1 (hAmph1) SH3 domain–peptide complexes. In this protocol, the protein/peptide atoms were classified into 16 types in terms of their physicochemical meaning and biological function, and then a 16 × 16 atom-pair interaction matrix was constructed to describe 256 atom-pair types between the SH3 domain and the peptide ligand, with atoms from peptide and SH3 domain served as the matrix columns and rows, respectively. Three non-covalent effects dominating SH3 domain–peptide binding as electrostatic, van der Waals (steric) and hydrophobic interactions were separately calculated for the 256 atom-pair types. As a result, 768 descriptors coding detailed information about SH3 domain–peptide interactions were yielded for further statistical modeling and analysis. Based on a culled data set consisting of 592 samples with known affinities, we employed this approach, coupled with partial least square (PLS) regression and genetic algorithm (GA), to predict and to interpret the peptide-binding behavior to SH3 domain. In comparison with the previous works, our method is more capable of capturing important factors in the SH3 domain-peptide binding, thus, yielding models with better statistical performance. Furthermore, the optimal GA/PLS model indicates that the electrostatic effect plays a crucial role in SH3 domain–peptide complexes, and steric contact and hydrophobic force also contribute significantly to the binding.

Keywords hAmph1 SH3 domain · Peptide · Non-binding interaction · Atom pair · Quantitative structure–affinity relationship

P. He and W. Wu contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-009-0332-x) contains supplementary material, which is available to authorized users.

P. He · W. Wu · H.-D. Wang (✉) · K. Yang · K.-L. Liao · W. Zhang
Cardiothoracic Surgery Department, Southwest Hospital, Third Military Medical University, 400038 Chongqing, China
e-mail: xxwkwhd@sina.com

Introduction

Many biological processes are regulated through association and dissociation of proteins. These processes include but not restricted to hormone-receptor binding, protease inhibition, antigen–antibody recognition, signal transduction, enzyme-substrate binding, vesicle transport, RNA splicing, and gene activation (Keskin et al. 2004). Often protein–protein interactions are fulfilled by the means of protein recognition modules, i.e. well-conserved domains characterized by a specific function and interaction with short peptides. The Src homology 3 (SH3) domain is a small protein domain of about 50 amino acid residues first identified as a conserved sequence in the non-catalytic part of several cytoplasmic protein tyrosine kinases (e.g. Src) (Stahl et al. 1988). Since then, it has been found in a great variety of other intracellular or membrane-associated proteins. The SH3 module might mediate the assembly of specific protein complexes by binding to proline-rich peptides (Fig. 1). Mutagenesis studies on 3BP1 and 3BP2 proteins have revealed that SH3 ligands are characterized by a PXXP motif and recognize the SH3-binding pocket in a polyproline helix type-II conformation in one of two opposite orientations (Ren et al. 1993). Although the SH3 domain plays essential roles in diverse biological processes

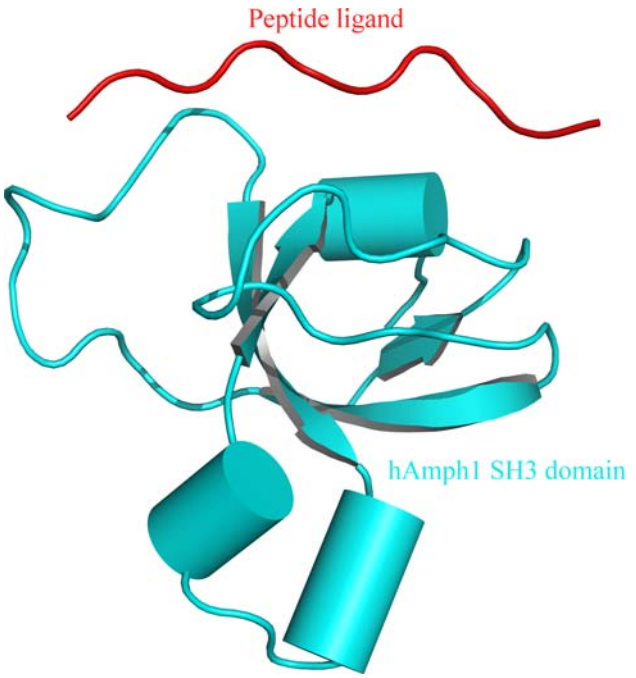

Fig. 1 Stereoview of the hAmph1 SH3 domain–peptide complex. The complex structure was constructed using homology modeling technique (Hou et al. 2006a, b)

such as increasing local concentration of proteins, altering their subcellular location and mediating the assembly of large multi-protein complexes, its function is not well understood today due to the complexity of SH3 domain-participating interaction networks which involve in enormous cellar processes.

Accurate identification of the SH3 domain-binding ligands, i.e. short peptide fragments, is the first step to understand the molecular mechanism of diverse biological roles that SH3 domain plays, and several experimental methods such as yeast two hybrid (Ito et al. 2002; Tong et al. 2002), peptide library (Rickles et al. 1994, 1995), and SPOT synthesis (Reineke et al. 2001; Santonico et al. 2005) have been developed for achieving this purpose. Although the experimental approach can give a conclusion on which peptides the SH3 domain binds and how strong the binding is, it is too time consuming and expensive to synthesize all potential peptides in a complete proteome and to perform the SH3 domain–peptide-binding assay to identify the SH3-binding partners (Hou et al. 2006a, b). Alternatively, the computational approach provides a promising way to straightforwardly elucidate the structural basis of SH3 domain–peptide interactions and to rapidly predict the binding affinities. A number of qualitative methods have been proposed to identify SH3 domain-binding peptides in protein primary sequences. For example, Brannetti et al. (2000) constructed a position-specific contact frequency matrix based on the crystal data and applied it to assess the probability that a peptide would bind the given SH3 domain. Zhang et al. (2006) employed machine-learning approaches coupled with structure data to build sequence-sensitive predictors for inferring interaction specificity between the SH3 domain and peptides. Ferraro et al. (2007) presented a web server called SH3-Hunter for the recognition of putative SH3 domain interaction sites on protein sequences, this tool evaluates which peptide–domain pair is a possible interaction pair and produces as output a list of significant interaction sites for each query protein. Recently, researchers were turning their interest to the quantitative prediction of binding affinity of SH3 domain–peptide complexes. In the work of Liang et al. (2008), a set of FASGAI descriptors derived from factor analysis of 335 physicochemical properties were used to quantitative structure–affinity relationship (QSAR) modeling of 2,018 SH3 domain–peptides. Zhou et al. (2008) employed genetic algorithm–Gaussian processes (GA–GP) to mine hidden dependences in the SH3 domain–peptide complex system and found that the binding was co-contributed by diverse properties. More significantly, in a series of publications by Hou et al. (2006a, 2008, 2009) the structure-based approach was used in characterization of SH3 domain–peptide interaction interface, and based on the obtained interaction information they developed several 3D QSAR models such as CoMFA, CoMSIA and MIEC to model the peptide affinity for the binding pocket of SH3 domain.

By comparing sequence-based approaches described by Liang et al. (2008) and Zhou et al. (2008) with structure-based approaches suggested by Hou et al. (2006a, 2008), we found that the structure-based approaches, although not statistically better than sequence-based approaches, can give intuitive insights into the interaction profile and binding behavior between the SH3 domain and its peptide ligands by adding structure information of receptor–ligand complexes to theoretical models, thus benefiting the design of peptidomimetic inhibitors of SH3 domain-containing proteins. In the present study, we propose a new approach for predicting and explaining the interaction behavior of peptides bound to the human amphiphysin-1 (hAmph1) SH3 domain. This method codes the information on non-covalent interaction fields extracted from the protein–peptide-binding interfaces and classifies them into 768 components in terms of 256 atom-pair types and three interaction manners. Subsequently, the partial least squares (PLS) regression, with or without GA-variable selection, was employed to correlate the non-covalent interaction parameters with the binding affinity of peptide ligands. Using this approach, we successfully developed several structure-based models for a set of hAmph1 SH3 domain-binding peptides with known affinities, and we also explored the structural implication of hAmph1 SH3

domain–peptide interfaces at atom level and the physicochemical meaning of the binding processes. We expect that this receptor structure-based method could be applied to other protein–peptide interactions as well.

Methods

Atom classification and atom-pair types

The 20 amino acids in proteins and peptides consists of five elements as carbon (C), nitrogen (N), oxygen (O), sulfur (S) and hydrogen (H). In folded proteins, atoms and groups bear distinct properties that are determined by self-state and local environment. For example, the C of sp3 hybridization conducts an electron-donator, whereas sp2-hybridizated C performs electron-withdrawing behavior. Based on the previous works (Gerstein et al. 1995; Li and Nussinov 1998; Lazaridis and Karplus 1999; Seeliger and de Groot 2007), here we classified the atoms of proteins into 16 types in terms of their physicochemical property and biological function. In this classification, the first emphasis is laid on the non-covalent-forming ability of atoms; for instance, the charged N and O can form salt bridges between them (Kumar and Nussinov 2002), and the polar H can be hydrogen-bonded with N, O and S atoms (McDonald and Thornton 1994). Secondly, the atomic hybridization states are further distinguished. Finally, the influence from neighbors is considered. The classification results are listed in Table 1, and the column 'priority' defines the classification priority of those atoms possessing multiple attributes; for example, the H attached to Nε of Trp is both polar and aromatic, and according to the priority it should be classified as polar H.

From the classification results, it is evident that, in SH3 domain–peptide complex, atom-pair types between the protein receptor and the peptide ligand are at most $16 \times 16 = 256$. The interaction information of these atom pairs accurately codes the binding profile of SH3 domain complexed with its peptide ligand.

Table 1 Classification of the atoms in proteins and peptides into 16 types

Priority	Type	Description	Example
1	Hc	Charged H	The H attached to Nζ of Lys
2	Hp	Polar H	The H attached to backbone N
3	Ha	Aromatic H	The H in benzene ring of Phe
4	H	Aliphatic H	The H attached to Cβ, Cγ, and Cδ of Leu
5	C*	Amide C	The Cγ of Asn
6	Cc	Charged C	The Cζ of Arg
7	Ca	Aromatic C	The C in benzene ring of Phe
8	Cp	Polar C	The Cε of Lys
9	C	Aliphatic C	The Cβ, Cγ, and Cδ of Leu
10	N*	Amide N	The Cδ of Asn
11	Nc	Charged N	The Nη of Lys
12	Na	Aromatic N	The Nε of Trp
13	O*	Amide O	The Oδ of Asn
14	Oc	Charged O	The Oε of Glu
15	Oh	Hydroxyl O	The Oγ of Ser
16	S	All S	The Sγ of Cys and the Sδ of Met

Non-covalent interactions between SH3 domain and peptides

In a previous report, Hou et al. (2006b) demonstrated that the electrostatic interactions and van der Waals contacts contribute significantly to SH3 domain–peptide association, and this conclusion was further confirmed by Zhou et al. (2008). In addition, Liang et al. (2008) found the hydrophobicity was responsible for the binding. Therefore, here we only considered the three effects of electrostatic, van der Waals and hydrophobic interactions in the coding procedure. Although a number of studies revealed that hydrogen bonds also play an important role in proteins/peptides binding to SH3 domain (Pisabarro and Serrano 1996; Matsuda et al. 1996; Wittekind et al. 1997), we did not explicitly treat this non-covalent type in the atom-pair coding. This is due to (1) hydrogen bond potentials cannot be calculated accurately by available empirical approaches, (2) it is difficult to properly discriminate hydrogen bonds from close dipole–dipole and long-range electrostatic interactions in complex biological environment, and (3) weak hydrogen bond potentials can be reproduced by the combination of Coulomb and Lennard–Jones terms and, for computational convenience, many force fields adopted this scheme to bypass the explicit calculation of hydrogen bonding (Cornell et al. 1995; Jorgensen et al. 1996; MacKerell et al. 1998).

Electrostatic interaction

Electrostatic roles in biomolecular stability and association process have been investigated intensively (Honig and Nicholls 1995), and many studies demonstrated that the electrostatic effect, together with hydrophobic force, contribute major factors to protein folding and protein–protein/nucleic acid recognitions (Perutz 1978; Pace et al. 2000; Vizcarra and Mayo 2005). Electrostatic potential (EP) between two point charges can be accurately described by classical Coulomb's law that the EP is proportional to atomic charges and reciprocal of distance.

$$\mathrm{EP}_{ij} = \kappa_{\mathrm{E}} \frac{q_i q_j}{\varepsilon_0 d_{ij}} \quad (1)$$

where subscripts i and j denote the atoms from protein receptor and peptide ligand, respectively. q_i is the partial charge of atom i, d_{ij} is the distance between atoms, i and j. ε_0 is the dielectric constant and usually defined as a distance-dependant form $\varepsilon_0 = d_{ij}\,\kappa_E$, the constant term of this equation. In this study, the AMBER charge was used as the atomic partial charge (Cornell et al. 1995).

van der Waals interaction

The van der Waals interaction, although transient and weak, can provide an important component of protein structures because of their sheer number. Most atoms of a protein are packed sufficiently close to others to be involved in transient van der Waals attractions (Roth et al. 1996). The formula of this interaction type is commonly expressed as the Lennard–Jones 12–6 potential (LP) that consists of a repulsive term and an attractive term.

$$\mathrm{LP}_{ij} = \ell_{ij}\left[\left(\frac{D^*_{ij}}{d_{ij}}\right)^{12} - 2\left(\frac{D^*_{ij}}{d_{ij}}\right)^{6}\right] \quad (2)$$

where ℓ_{ij} and D^*_{ij} are potential well and contacting distance between atom i and j, respectively, and can be determined using the empirical mixing rules (known as the Lorentz–Berthelot rules): $\ell_{ij} = \sqrt{\ell_{ii}\ell_{jj}}$ and $D^*_{ij} = \frac{D^*_{ii}+D^*_{jj}}{2}$ (Basdevant et al. 2007). The ℓ_{ii} and D^*_{ii} separately indicate potential well coefficient and van der Waals diameter of sole atom i, and in this work their values are taken from the AMBER parm96 [note that the ℓ_{ij} and D^*_{ij} are combined in AMBER (Cornell et al. 1995)].

Hydrophobic interaction

Hydrophobic interaction is very important in the folding and packing of proteins, while it is not easy to be tested experimentally due to the hydrophobicity involves both the enthalpy and entropy effects. Israelachvili and Pashley (1982) demonstrated that the strength of hydrophobic interaction between two molecules was an exponential distance-dependent form, proportional with the inherent hydrophobicity of interacting atoms. An empirical equation was further proposed to quantitatively describe the hydrophobic potential (HP) (Zhou et al. 2007).

$$\mathrm{HP}_{ij} = -\left(S_i\rho_i + S_j\rho_j\right)e^{-d_{ij}} \quad (3)$$

where ρ represents the inherent hydrophobicity of atoms and can be measured using Eisenberg's scale (Eisenberg and McLachlan 1986). S is atomic solvent accessible surface area, and we used the MSMS program (Sanner et al. 1996) to compute it for each atom in SH3 domain–peptide complexes.

Partial least squares regression

Partial least squares regression have been successfully applied in several 3D-QSAR methods such as CoMFA (Cramer et al. 1988), CoMSIA (Klebe et al. 1994) and HASL (Doweyko 1988), and here we employed it to correlate the coded non-covalent information with the binding affinity of SH3 domain–peptide complexes. The detailed description of PLS algorithm can be found in the previous publications (Geladi and Kowalski 1986; Wold et al. 2001).

Overview of the method

The flow of our method is shown in Fig. 2. It starts at the definition of the interaction region in SH3 domain–peptide complexes. This is required to accurately identify every residue contributing to receptor–ligand binding. Common methods used for identifying protein–protein/peptide-binding interfaces include Voronoi polyhedra-based approach, changes in solvent accessible surface area, and various radial cutoffs (Zhou et al. 2009a). Here, we adopted the suggestion by Fischer et al. (2006): two residues respectively from the receptor and the ligand would be considered in contact if there is at least one pair of non-hydrogen atoms in 6 Å between them. Using this protocol, we can readily define the interfacial residues of SH3 domain interacting with peptides.

By this procedure, residues in the peptide-binding pocket of SH3 domain were determined. Then three atom-pair interaction matrixes with 16 × 16 orders were separately constructed to describe electrostatic, van der Waals and hydrophobic interactions between the peptide-binding pocket of SH3 domain and peptide ligands; atoms from peptides were defined as the columns, and those from SH3 domain as the rows (a column or a row represents an atomic type in peptide or SH3 domain). By the method described in "Non-covalent interactions between SH3 domain and peptides", there were totally 3 (16 × 16) = 768 elements included in the three matrixes, and they corporately encoded the interaction information in a SH3 domain–peptide complex. Taken these information as the independent X, PLS was employed to build linear regression models with dependent Y (binding affinity) and to explore the underlying relationship.

Material preparation

Landgraf et al. (2004) used the phage display coupled with SPOT synthesis to identify peptides in yeast proteome

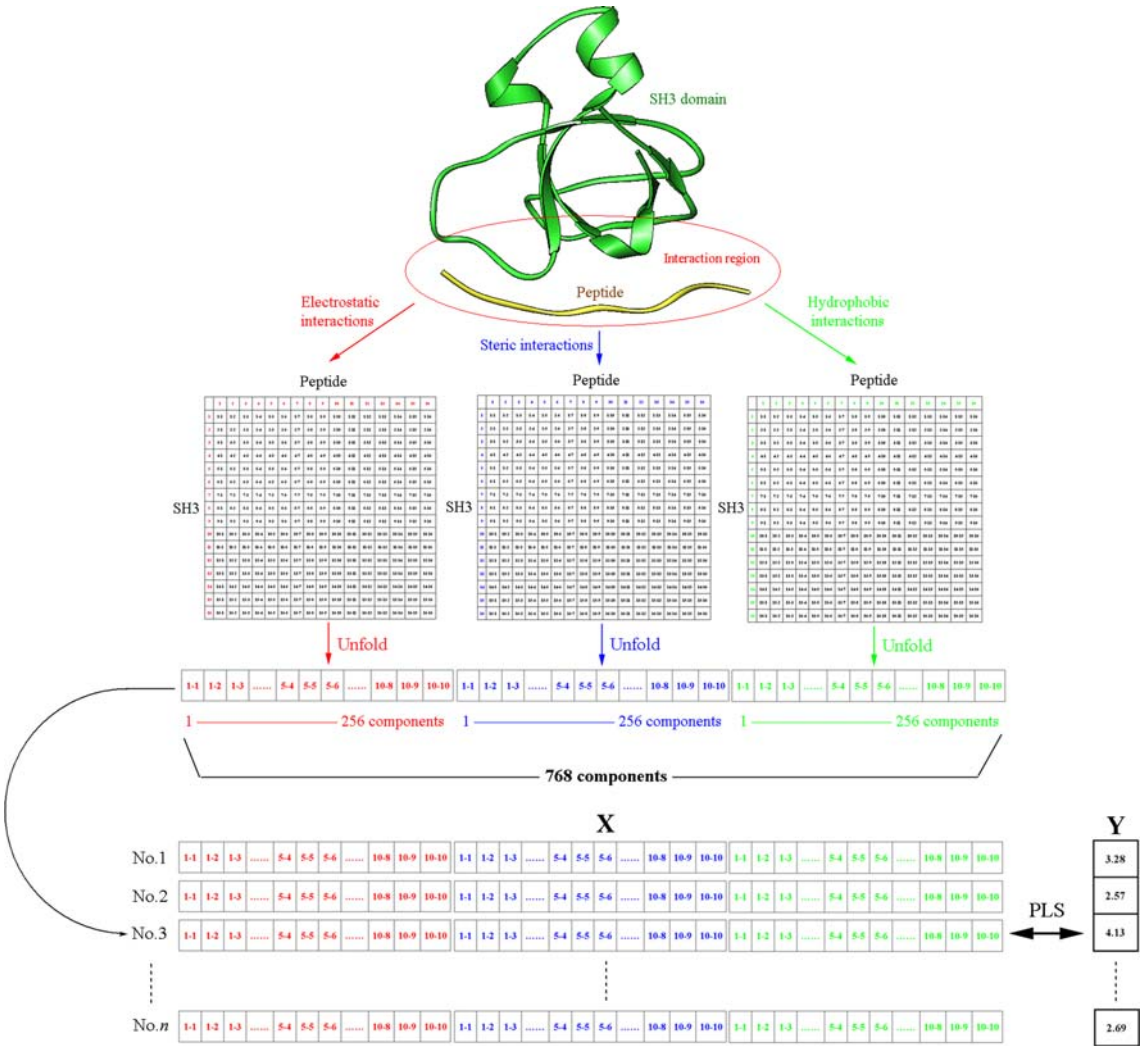

Fig. 2 The flow of the non-binding information-coding algorithm. First, the interaction region in SH3 domain-peptide complex is detected. Secondly, the electrostatic, van der Waals and hydrophobic interactions between the SH3 domain and the peptide are separately calculated for all 256 atom-pair types, thereby yielding three atom-pair interaction matrixes. Finally, 768 independent variables involved in the three atom-pair interaction matrixes are correlated with the binding affinity by the PLS approach

binding to eight yeast SH3 domains and peptides in human proteome binding to two human SH3 domains. The SPOT signal intensity of each peptide was measured in Boehringer light units (BLUs). In their work, 2,031 decapeptides with potential to bind to the hAmph1 SH3 domain were synthesized and assayed. Considering that the BLUs only indirectly reflect the peptide-binding affinity, and that its experimental value also involves strong noise and other undetermined factors, Zhou et al. (2008) have suggested two criteria to filter these samples for the purpose of reliability: (1) there should be at least two measurements for each peptide, and (2) the standard deviation of the repeated measurements should be less than half of the average value. Applying these criteria, 592 peptides were selected for further analysis (provided in Supporting Information).

The models of hAmph1 SH3 domain–peptide complex structures were constructed according to the strategy proposed by Hou et al. (2006a, 2008). Briefly, the SH3 domain–peptide PLPRRPPRAA complex structure was modeled by homology modeling and molecular mechanics optimization, then the template peptide PLPRRPPRAA was virtually mutated to the other sequences. Subsequently, energy minimization for each complex was carried out using AMBER 9.0 (Cornell et al. 1995); the maximum number of minimization steps was set to 3,000. The first 500 steps were performed with the steepest descent algorithm, whereas the rest of the steps were performed with the conjugate gradient algorithm. In this procedure, the solvent effect was considered using the generalized Born model (IGB = 2). A detailed description of this procedure

Table 2 Statistics of the models constructed using different combinations of non-covalent types and modeling methods

Model	Combination	NC	Training set				Test set	
			r^2	RMSEE	q^{2c}	RMSCV[a]	q^2_{ext}	RMSEP
M_E	E	4	0.614	0.565	0.538	0.619	0.511	0.635
M_S	S	6	0.573	0.589	0.492	0.643	0.477	0.653
M_H	H	3	0.596	0.575	0.529	0.622	0.497	0.648
M_{E+S}	E + S	6	0.687	0.502	0.616	0.563	0.584	0.588
M_{E+H}	E + H	5	0.712	0.477	0.630	0.524	0.607	0.568
M_{S+H}	S + H	5	0.621	0.558	0.559	0.603	0.534	0.610
M_{E+S+H}	E + S + H	6	0.758	0.446	0.665	0.502	0.626	0.554
M^{*b}_{E+S+H}	E + S + H	8	0.798	0.408	0.722	0.463	0.705	0.493

E electrostatic effect, *S* steric effect, *H* hydrophobic effect, *NC* number of significant components in the constructed PLS models

[a] Tenfold cross-validation

[b] Assisted by GA-variable selection

can be found in the works of Hou et al. (2006a, 2008). By this way, all 592 SH3 domain–peptide complex structures were in turn constructed.

Results and discussion

Statistical modeling and analysis

A reliable statistical model should undergo rigorous validations. Previous studies show that the high value of cross-validation q^2 appears to be the necessary but not the sufficient condition for a QSAR model to have high-predictive power, and the external validation is the only way to establish reliable models (Golbraikh and Tropsha 2002). Therefore, the 592 peptides were split into two parts, a training set used for generating models and a test set for validating the models constructed on the training set. The underlying goal at this step is to ensure that both the training and test sets separately span the whole structure space occupied by the entire data set and the chemical domains in the sets are not too dissimilar (Tropsha et al. 2003). We employed a recently published Monte Carlo-based SpScore method to perform data set splitting (Zhou et al. 2009b). This method attempts to find the optimal solution for the splitting, yielding maximum of diversity in training and test sets and a minimum of dissimilarity between the training and test sets. Using this procedure, 92 peptides were selected from the entire data set to serve as test set; the remaining 500 samples thus composed the training set (see Supporting Information). It is worth noting that which method used for splitting data set is not important, since there are a large number of available hAmph1 SH3 domain-binding peptides.

To test the importance of the three non-covalent types of association, and of their interactions, in SH3 domain-peptide binding, we investigated the performances of different combinations of electrostatic, van der Waals, and hydrophobic effects. The statistics of resulting models are listed in Table 2, in which some uncertainties may be imposed by the limited data set. By comparing models constructed separately by electrostatic, van der Waals and hydrophobic effects and their combinations, the model involving electrostatic descriptors was demonstrated to be statistically satisfactory. The model M_E had significantly superior fitting and predictive ability to M_H and M_S, while the model M_{S+H} is inferior to M_{E+S} and M_{E+H} in modeling performance. Therefore, we assumed that electrostatics played important roles in SH3 domain-peptide binding. This assumption is consistent with the previous reports (Hou et al. 2006b). In addition, the model M_{E+S+H} including three kinds of non-covalent effects is the best in fitting ability, stability and predictive ability, suggesting that, in addition to the electrostatic interaction, van der Waals and hydrophobic effects also have contributions to the binding that cannot be neglected, and inclusion of these contributions could further improve the modeling performance.

Further analysis of the model M_{E+S+H} is as follows. From the original 768 variables, six significant principal components were extracted by PLS that explained 75.8% of the variance for dependent variable *Y*, and predicted 66.5% of the variance for *Y* by cross-validation. Then, the model was used as a predictor for test set samples, predicting 62.6% of the variance for test *Y*. Figure 3 shows the calculated and predicted affinities for 500 training samples and 92 test samples by model M_{E+S+H} versus their experimental values. This model better describes the activity of most peptides with low and medium affinities, while underestimating some high-affinity peptides. Similar findings have already been reported (Liang et al. 2008; Zhou et al. 2008). For example, Zhou et al. (2008) constructed predictors for SH3-binding peptides separately by

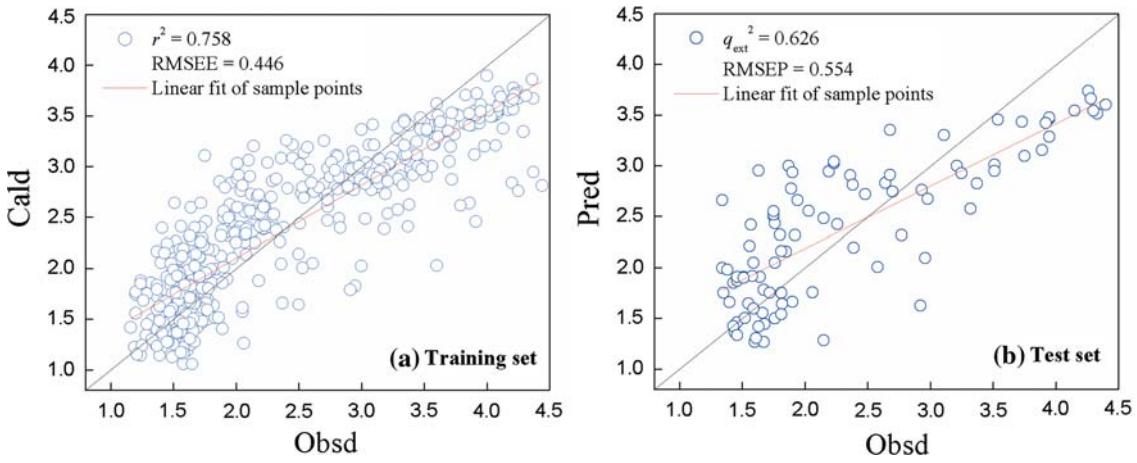

Fig. 3 Calculated and predicted affinities for **a** 500 training samples and **b** 92 test samples by model M_{E+S+H} versus their experimental values

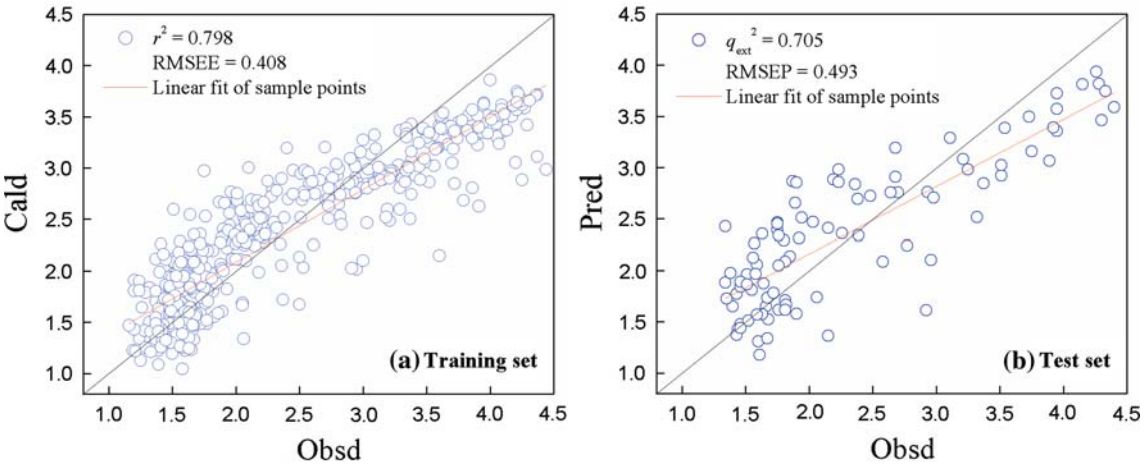

Fig. 4 Calculated and predicted affinities for **a** 500 training samples and **b** 92 test samples by model M^*_{E+S+H} (by GA-variable selection. Population size: 250, mutation rate: 1.0%, fitting function: 10-fold cross-validation, variable preprocessing: autoscaling) versus their experimental values

linear PLS method and nonlinear ANN, SVM and GP methods, and the results showed that nonlinear methods outperformed the linear ones. Therefore, we believed that simple linear models are not suitable for an appropriate modeling of the binding behavior of some peptides (especially the high-affinity peptides) due to the nonlinear complexity of SH3 domain-peptide binding.

GA-variable selection

Genetic algorithm is a widely employed variable selection tool that has been successfully used in many QSAR studies (Cho and Hermsmeier 2002). We combined the PLS with GA-variable selection (GA/PLS) to construct more robust models for predicting and explaining the binding behavior of peptides. GA/PLS was performed using Matlab toolboxes GATBX (Genetic Algorithm Toolbox for MATLAB, University of Sheffield, UK) and ChemoAC (ChemoAC Calibration Toolbox, VICIM, Belgium). As a result, a variable subset consisting of 267 descriptors were extracted from the initial descriptor pool (which consists of 768 descriptors), in which 129 ones are electrostatic terms, 57 ones are steric terms, and 81 ones are hydrophobic terms. The calculated results and statistics of this improved model M^*_{E+S+H} are shown in Fig. 4 and Table 2. By GA-variable selection, modeling performance was advanced noticeably, having its fitting ability r^2, cross-validated q^2 and the predictive q^2_{ext} on external test set increased to 0.798, 0.722 and 0.705, respectively. By comparing Figs. 3 and 4, it is found that the M^*_{E+S+H} was able to correctly describe the samples possessing large calculated errors in the M_{E+S+H}, and therefore the M^*_{E+S+H}, compared with M_{E+S+H}, has few large-error predictions. In the interaction region in SH3 domain–peptide complex, different sites and properties show distinct contributions to the peptide binding, while via GA-variable selection procedure, those descriptors

Table 3 Comparison between the models constructed in this work and previous studies

Method	Modeling tool	References	Training set			Test set	
			Samples	r^2	q^2	Samples	q^2_{ext}
CoMFA[a]	PLS	Hou et al. (2006a, b)	200	0.746	0.633[b]	684	0.578
CoMSIA[a]	PLS	Hou et al. (2006a, b)	200	0.767	0.636[b]	684	0.624
MIEC	GA/PLS	Hou et al. (2008)	442	0.691	0.648[b]	442	0.643
FASGAI	GA/PLS	Liang et al. (2008)	884	0.603	0.562[b]	1,134	0.533
DPPS	GA/GP	Zhou et al. (2008)	296	0.862	0.767[c]	296	0.697
Atom pairs	GA/PLS	This work	500	0.798	0.713[c]	92	0.705
					0.722[d]		

[a] Assisted by region focusing technique

[b] Leave one out cross-validation

[c] Threefold cross-validation

[d] Tenfold cross-validation

unrelated with the binding are eliminated, thus, significantly improving the statistical quality of models.

In the 267 descriptors obtained by GA-variable selection 129 are electrostatic terms, accounting for almost a half of the total; while steric terms and hydrophobic terms are 57 and 81, respectively. Hence, the electrostatic effect was further confirmed to play a dominant role for the peptide binding, while van der Waals contact and hydrophobic force also play partial roles. In addition, atom pairs aliphatic H–aliphatic C, charged H–charged O, polar H–aminoacyl N, polar C–charged N, aromatic H–aromatic C, polar C–hydroxyl O possess large contributions to the model M^*_{E+S+H}, and all of them have their variable importance in the projection (Wold et al. 2001) of PLS above 1. It can be seen that atom pairs possessing significant contributions are mainly associated with polar and charged atoms, while hydrophobic contacts (e.g. aliphatic H–aliphatic C) and van der Waals stacking (e.g. aromatic H–aromatic C) also have some contributions.

Comparison to previous studies

On the SH3 domain-binding peptide data set, Hou et al. (2006a, b), Liang et al. (2008) and Zhou et al. (2008) had used different modeling methods to conduct QSAR studies at 2D and 3D levels. Here, we made a brief comparison of the current work with their modeling results. Table 3 lists the optimal models by different researches (the optimal model is referred to the best predictor on test set). Hou et al. (2006a, b) used CoMFA and CoMSIA to systematically explore the quantitative correlation between different molecular interaction fields (e.g. electrostatic field, steric field, hydrogen bond donor field, etc.) and peptide-binding affinity. Without variable selection, their models were relatively poor in statistical quality (predictive q^2_{ext} of the optimal CoMSIA model was 0.624). Subsequently, they further taken into account the structure information of receptor and the modeling performance was slightly improved ($q^2_{ext} = 0.643$) (Hou et al. 2008). Liang et al. (2008) used factor analysis and GA-variable selection to perform QSAR modeling on a large-scale data set. Owing to the strong noise included in their unfiltered sample set, resulting models were significantly poor in predictive ability ($q^2_{ext} = 0.533$). Very recently, Zhou et al. (2008) culled a SH3-binding peptide data set and used nonlinear Gaussian process coupled with GA (GA/GP) to construct models. By this procedure, the modeling performance was improved remarkably ($q^2_{ext} = 0.697$). Based on the data set of Zhou et al., we further introduced the information about the non-covalent interaction of SH3 domain with peptides into the modeling procedure, therefore, yielding a more predictable model with $q^2_{ext} = 0.705$.

Conclusions

This study proposed a new method to characterize the binding profile of SH3 domain–peptide complexes. Based on physicochemical properties and biological functions, protein/peptide atoms were divided into 16 kinds, and thereby 256 atom-pair types between ligand and receptor were defined. On this basis, 768 descriptors on atom-pair non-covalent interaction were yielded to describe electrostatic, van der Waals and hydrophobic interactions for 256 atom-pair types. We used this approach, coupled with PLS and GA/PLS, to model, predict and interpret a culled data set consisting of 592 hAmph1 SH3 domain-binding peptides. In comparison with the previous studies, the current study is statistically satisfactory, especially a good predictor on external test set. By analyzing the models

constructed, we confirmed that electrostatic effect played a dominant role in SH3 domain–peptide binding, while hydrophobic force and van der Waals contact an assistant role in this process.

Acknowledgments We gratefully acknowledge the Dr. Tingjun Hou for providing structure model of the hAmph1 SH3 domain–peptide PLPRRPPRAA complex. This work was supported by the National Natural Science Foundation of China (No. 30772145) and the Natural Science Foundation Project of CQ_CSTC (No. CSTC.2006BB5081).

References

Basdevant N, Borgis D, Ha-Duong T (2007) A coarse-grained protein–protein potential derived from an all-atom force field. J Phys Chem B 11:9390–9399

Brannetti B, Via A, Cestra G, Cesareni G, Citterich MH (2000) SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family. J Mol Biol 298:313–328

ChemoAC Calibration Toolbox. Virtual Institute of Chemometrics and Industrial Metrology, Brussels, Belgium

Cho SJ, Hermsmeier MA (2002) Genetic algorithm guided selection: variable selection and subset selection. J Chem Inf Comput Sci 42:927–936

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117:5179–5197

Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110:5959–5967

Doweyko AM (1988) The hypothetical active site lattice: an approach to modeling active sites from data on inhibitor molecules. J Med Chem 31:1396–1406

Eisenberg D, McLachlan AD (1986) Solvation energy in protein folding and binding. Nature 319:199–203

Ferraro E, Peluso D, Via A, Ausiello G, Helmer-Citterich M (2007) SH3-Hunter: discovery of SH3 domain interaction sites in proteins. Nucleic Acids Res 35:W451–W454

Fischer TB, Holmes JB, Miller IR, Parsons JR, Tung L, Hu JC, Tsai J (2006) Assessing methods for identifying pair-wise atomic contacts across binding interfaces. J Struct Biol 153:103–112

Geladi P, Kowalski B (1986) Partial least squares regression: a tutorial. Anal Chim Acta 185:1–17

Genetic Algorithm Toolbox for MATLAB. Department of Automatic Control and Systems Engineering of The University of Sheffield, UK

Gerstein M, Tsai J, Levitt M (1995) The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. J Mol Biol 249:955–966

Golbraikh A, Tropsha A (2002) Beware of q2!. J Mol Graph Model 20:269–276

Honig B, Nicholls A (1995) Classical electrostatics in biology and chemistry. Science 268:1144–1149

Hou T, McLaughlin W, Lu B, Chen K, Wang W (2006a) Prediction of binding affinities between the human amphiphysin-1 SH3 domain and its peptide ligands using homology modeling, molecular dynamics and molecular field analysis. J Proteome Res 5:32–43

Hou T, Chen K, McLaughlin WA, Lu B, Wang W (2006b) Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain. PLoS Comput Biol 2:46–55

Hou T, Zhang W, Case DA, Wang W (2008) Characterization of domain–peptide interaction interface: a case study on the amphiphysin-1 SH3 domain. J Mol Biol 376:1201–1214

Hou T, Xu Z, Zhang W, McLaughlina WA, Casec DA, Xu Y, Wang W (2009) Characterization of domain-peptide interaction interface: a generic structure-based model to decipher the binding specificity of SH3 domains. Mol Cell Proteomics 8:639–649

Israelachvili J, Pashley R (1982) The hydrophobic interaction is long range, decaying exponentially with distance. Nature 300:341–342

Ito T, Ota K, Kubota H, Yamaguchi Y, Chiba T, Sakuraba K, Yoshida M (2002) Roles for the two hybrid system in exploration of the yeast protein interactome. Mol Cell Proteomics 1:561–566

Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118:11225–11236

Keskin O, Tsai C-J, Wolfson H, Nussinov R (2004) A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. Protein Sci 13:1043–1055

Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146

Kumar S, Nussinov R (2002) Close-range electrostatic interactions in proteins. Chem Bio Chem 3:604–617

Landgraf C, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, Volkmer-Engert R, Cesareni G (2004) Protein interaction networks by proteome peptide scanning. PLoS Biol 2:94–103

Lazaridis T, Karplus M (1999) Effective energy function for proteins in solution. Protein 35:133–152

Li AJ, Nussinov R (1998) A set of van der Waals and Coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. Protein 32:111–127

Liang G, Chen G, Niu W, Li Z (2008) Factor analysis scales of generalized amino acid information as applied in predicting interactions between the human amphiphysin-1 SH3 domains and their peptide ligands. Chem Biol Drug Des 71:345–351

MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616

Matsuda M, Ota S, Tanimurai R, Nakamurai H, Matuoka K, Takenawa T, Nagashima K, Kurata T (1996) Interaction between the amino-terminal SH3 domain of CRK and its natural target proteins. J Biol Chem 271:14468–14472

McDonald IK, Thornton JM (1994) Satisfying hydrogen bonding potential in proteins. J Mol Biol 238:777–793

Pace CN, Alston RW, Shaw KL (2000) Charge–charge interactions influence the denatured state ensemble and contribute to protein stability. Protein Sci 9:1395–1398

Perutz MF (1978) Electrostatic effects in proteins. Science 201:1187–1191

Pisabarro MT, Serrano L (1996) Rational design of specific high-affinity peptide ligands for the Abl-SH3 domain. Biochemistry 35:10634–10640

Reineke U, Volkmer-Engert R, Schneider-Mergener J (2001) Applications of peptide arrays prepared by the SPOT-technology. Curr Opin Biotech 12:59–64

Ren RB, Mayer BJ, Cicchetti P, Baltimore D (1993) Identification of a 10-amino acid proline-rich SH3 binding-site. Science 259:1157–1161

Rickles RJ, Botfield MC, Weng Z, Taylor JA, Green OM, Brugge JS, Zoller MJ (1994) Identification of Src, Fyn, Lyn, PI3 K and Abl SH3 domain ligands by screening a random phage display library. EMBO J 13:5598–5604

Rickles RJ, Botfield MC, Zhou XM, Henry PA, Brugge JS, Zoller MJ (1995) Phage display selection of ligand residues important for Src homology 3 domain binding specificity. Proc Natl Acad Sci USA 92:10909–10913

Roth CM, Neal BL, Lenhoff AM (1996) Van der Waals interactions involving proteins. Biophys J 70:977–987

Sanner MF, Olson AJ, Spehner JC (1996) Reduced surface: an efficient way to compute molecular surfaces. Biopolymers 38:305–320

Santonico E, Castagnoli L, Cesareni G (2005) Methods to reveal domain networks. Drug Discovery Today 10:1111–1117

Seeliger D, de Groot BL (2007) Atomic contacts in protein structures: a detailed analysis of atomic radii, packing, and overlaps. Proteins 68:595–601

Stahl ML, Ferenz CR, Kelleher KL, Kriz RW, Knopf JL (1988) Sequence similarity of phospholipase C with the noncatalytic region of Src. Nature 332:269–272

Tong AHY, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295:321–324

Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 22:69–77

Vizcarra CL, Mayo SL (2005) Electrostatics in computational protein design. Curr Opin Chem Biol 9:622–626

Wittekind M, Lee V, Goldfarb V, Friedrichs MS, Meyers CA, Mueller L, Mapelli C (1997) Solution structure of the Grb2 N-terminal SH3 domain complexed with a ten-residue peptide derived from SOS: direct refinement against NOEs, J-couplings and ^{1}H and ^{13}C chemical shifts. J Mol Biol 267:933–952

Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemometr Intel Lab Syst 58:109–130

Zhang L, Shao C, Zheng D, Gao Y (2006) An integrated machine learning system to computationally screen protein databases for protein binding peptide ligands. Mol Cell Proteomics 5:1224–1232

Zhou P, Tian F, Li Z (2007) A structure-based, quantitative structure-activity relationship approach for predicting HLA-A*0201-restricted cytotoxic T lymphocyte epitopes. Chem Biol Drug Des 69:56–67

Zhou P, Tian F, Chen X, Shang Z (2008) Modeling and prediction of binding affinities between the human amphiphysin SH3 domain and its peptide ligands using genetic algorithm–Gaussian processes. Biopolymers (Pept Sci) 90:792–802

Zhou P, Tian F, Shang Z (2009a) 2D depiction of nonbonding interactions for protein complexes. J Comput Chem 30:940–951

Zhou P, Chen X, Wu Y, Shang Z (2009b) Gaussian process: an alternative approach for QSAM modeling of peptides. Amino Acids (in press). doi:10.1007/s00726-008-0228-1)