Biomacromolecular quantitative structure–activity relationship (BioQSAR): a proof-of-concept study on the modeling, prediction and interpretation of protein–protein binding affinity

Peng Zhou · Congcong Wang · Feifei Tian · Yanrong Ren · Chao Yang · Jian Huang

Received: 30 July 2012 / Accepted: 12 December 2012 / Published online: 10 January 2013

Abstract Quantitative structure–activity relationship (QSAR), a regression modeling methodology that establishes statistical correlation between structure feature and apparent behavior for a series of congeneric molecules quantitatively, has been widely used to evaluate the activity, toxicity and property of various small-molecule compounds such as drugs, toxicants and surfactants. However, it is surprising to see that such useful technique has only very limited applications to biomacromolecules, albeit the solved 3D atom-resolution structures of proteins, nucleic acids and their complexes have accumulated rapidly in past decades. Here, we present a proof-of-concept paradigm for the modeling, prediction and interpretation of the binding affinity of 144 sequence-nonredundant, structure-available and affinity-known protein complexes (Kastritis et al. Protein Sci 20:482–491, 2011) using a biomacromolecular QSAR (BioQSAR) scheme. We demonstrate that the modeling performance and predictive power of BioQSAR are comparable to or even better than that of traditional knowledge-based strategies, mechanism-type methods and empirical scoring algorithms, while BioQSAR possesses certain additional features compared to the traditional methods, such as adaptability, interpretability, deep-validation and high-efficiency. The BioQSAR scheme could be readily modified to infer the biological behavior and functions of other biomacromolecules, if their X-ray crystal structures, NMR conformation assemblies or computationally modeled structures are available.

Keywords Biomacromolecular quantitative structure–activity relationship · Protein–protein interaction · Regression modeling · Affinity prediction

Introduction

Quantitative structure–activity relationship (QSAR) has long been stood on the arena of drug design and discovery since Hansch and Fujita pioneered a statistical regression method for the correlation of chemical structure and biological activity [1]. During the past half century, QSAR has been widely spread around various fields of, for example, chemistry, pharmacology, toxicology, biology and materials science, and is now used not only to estimate the activity and potency of therapeutic agents, but also to evaluate the property and function of diverse substances such as chemicals (QSPR) [2], toxicants (QSTR) [3], surfactants (QSSR) [4] and nanomaterials (QNAR) [5]. Nowadays, the consideration of additional contributions from cellular environment and interactome networks to the pharmacological profile and pharmacokinetic behavior of drug molecules is underway in the QSAR community (i.e. cell-QSAR [6] and profile-QSAR [7]).

Electronic supplementary material The online version of this article (doi:10.1007/s10822-012-9625-3) contains supplementary material, which is available to authorized users.

P. Zhou (✉) · C. Wang · C. Yang · J. Huang (✉)
Center of Bioinformatics (COBI), School of Life Science and Technology, University of Electronic Science and Technology of China (UESTC), Chengdu 610054, China
e-mail: p_zhou@uestc.edu.cn

J. Huang
e-mail: hj@uestc.edu.cn

F. Tian
School of Life Science and Engineering, Southwest Jiaotong University, Chengdu 610031, China

Y. Ren
Department of Biological and Chemical Engineering, Chongqing University of Education, Chongqing 400067, China

In recent years, the applications of QSAR methodology to tackle biological problems have emerged as a new and promising way to reveal the underlying dependences hidden in complicated biosystems [8]. However, most existing works only attempted to model the relationship between the information derived from the primary sequence of biomacromolecules and their activity, affinity or function, termed as quantitative sequence–activity modeling (QSAM) [9]. Nevertheless, with rapid increase in the quantity of solved atom-resolution structures of proteins, nucleic acids and their complexes with ligands deposited in the Protein Data Bank (PDB) [10], it is now possible to directly correlate structure feature with biological activity for these biomacromolecules by using QSAR strategy, just like that for small-molecule compounds.

Previously, there were already sporadic efforts addressed on the QSAR modeling of biomacromolecular structure–function relationship. For example, González-Díaz and co-workers developed several parameterization methods such as complex network vectors, stochastic molecular descriptors and pseudofolding topological indices to characterize the 3D structure profile of biomolecules, and then successfully applied these methods to identify and infer the interaction behavior, folding mechanism and biological function of diverse biomolecules, including peptides, proteins, RNAs and DNAs [11–16]. In the current work, we introduce the concept biomacromolecular QSAR (BioQSAR) with which the standard QSAR workflow is applied to biomacromolecules instead of traditional small-molecule compounds. Here, a proof-of-concept paradigm for the modeling, prediction and interpretation of the binding affinity of 144 sequence-nonredundant, structure-available and affinity-known protein complexes using the BioQSAR scheme was performed to explore whether the standard QSAR techniques such as structure characterization, variable selection, regression calibration, model validation, and statistical analysis are also applicable for treating protein, the typical biomacromolecule.

Protein–protein interactions (PPIs) play a pivotal role in the organization of life. While some interactions form stable complexes resulting in permanent, multi-protein structures, others are of a transient nature. The latter are abundant in signal transduction, protein–inhibitor complexes, antibody–antigen interactions and others [17]. While it is too time-consuming and expensive to experimentally express all proteins found in a cell and perform protein–protein binding assay to construct the cellular interactome, the fast and reliable prediction of protein–protein binding affinity is fundamentally important but also a challenging task for our understanding of the molecular mechanism and biological implications underlying cell signaling networks, metabolic pathways and the occurrence and development of various diseases. Nowadays, the existing prediction methods include empirical scoring algorithms, knowledge-based strategies and mechanism-type approaches [18]; all of them have been employed to serve as scoring functions in protein–protein docking [19]. However, a recent comparative study revealed a poor correlation between binding affinity and resultant scores for 9 docking algorithms tested on a panel of culled protein complex samples [20]. Therefore, development of new and effective strategies to fulfill the affinity prediction purpose is still needed and would be of great significance to protein docking, function inference and structure design.

Materials and methods

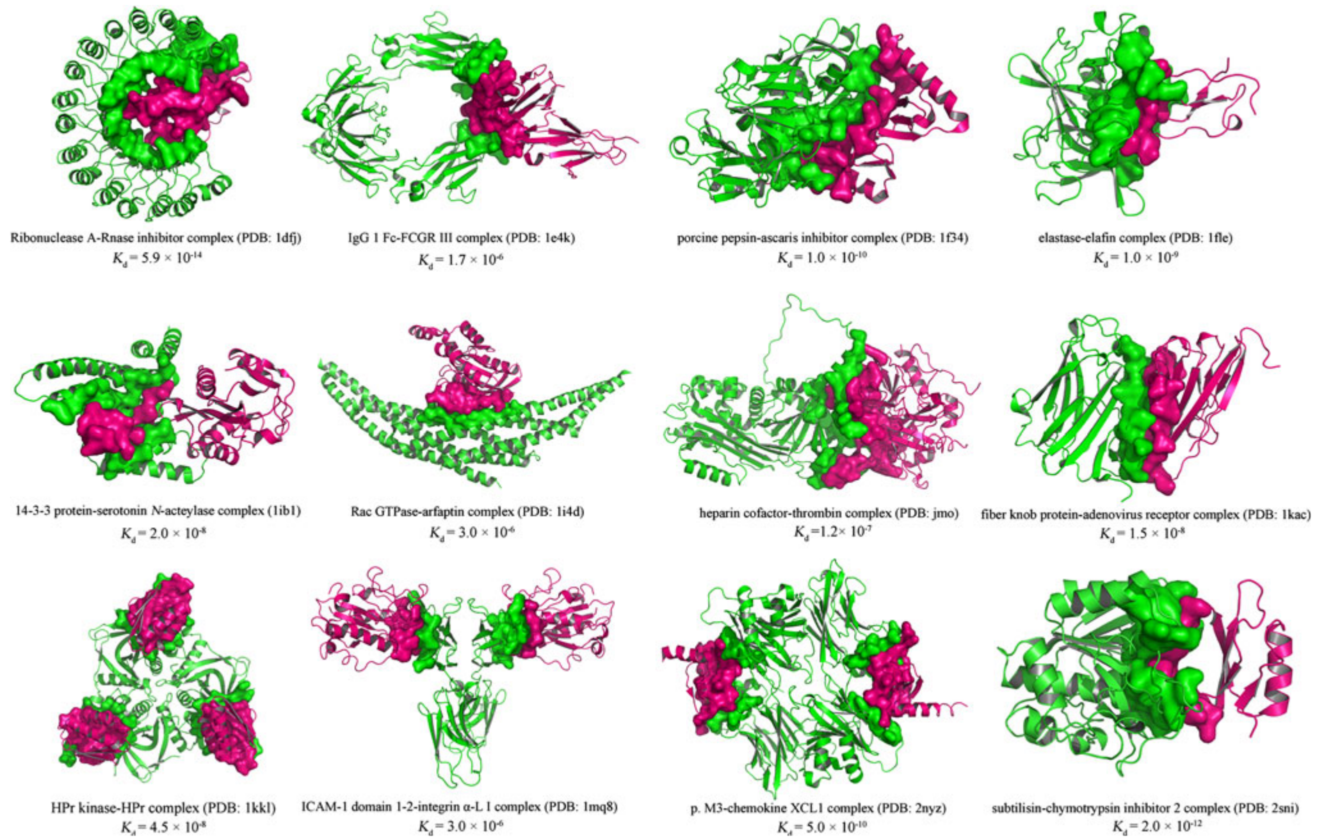
Protein–protein complex dataset

A structure-based benchmark for protein–protein binding affinity compiled by Kastritis et al. [21] was used here to perform BioQSAR modeling. The benchmark consists of 144 protein–protein complexes for which both high-resolution structures and dissociation constants are available. The set is diverse in terms of the biological functions it represents, with complexes that involve G-proteins and receptor extracellular domains, as well as antigen–antibody, enzyme–inhibitor and enzyme–substrate complexes. It is also diverse in terms of the partners' affinity for each other, with K_d ranging between 10^{-5} and 10^{-14} M (Fig. 1). The detailed information about the benchmark is tabulated in Table S1 in Supporting Information. It is worth noting that the binding strength and stability of protein complexes are normally characterized by the change in Gibbs free energy (ΔG) upon formation of the complex architecture, which can be derived from the observable dissociation constant (K_d) of protein–protein binding reaction using the formula $\Delta G = \mathrm{RT}\ln K_d$ [22]. Thus, the K_d is an important physical quantity that quantitatively measures protein–protein binding affinity.

Before performing analyses, the protonation state of titratable residues at pH 7.0 was assigned using PROPKA program [23], and the missing hydrogen atoms and side chains of protein complex structures were added with REDUCE [24] and SCWRL [25] programs, respectively. The SCWRL and REDUCE adopted here are because we have previously demonstrated that these two programs have a good performance in reproducing experimentally determined atomic structures of proteins and protein complexes [26, 27].

Characterization of protein–protein interactions

The place that directly realizes the contacts and interactions of two binding partners is at the interface of formed



Ribonuclease A-Rnase inhibitor complex (PDB: 1dfj) $K_d = 5.9 \times 10^{-14}$

IgG 1 Fc-FCGR III complex (PDB: 1e4k) $K_d = 1.7 \times 10^{-6}$

porcine pepsin-ascaris inhibitor complex (PDB: 1f34) $K_d = 1.0 \times 10^{-10}$

elastase-elafin complex (PDB: 1fle) $K_d = 1.0 \times 10^{-9}$

14-3-3 protein-serotonin *N*-acteylase complex (1ib1) $K_d = 2.0 \times 10^{-8}$

Rac GTPase-arfaptin complex (PDB: 1i4d) $K_d = 3.0 \times 10^{-6}$

heparin cofactor-thrombin complex (PDB: jmo) $K_d =1.2\times 10^{-7}$

fiber knob protein-adenovirus receptor complex (PDB: 1kac) $K_d = 1.5 \times 10^{-8}$

HPr kinase-HPr complex (PDB: 1kkl) $K_d = 4.5 \times 10^{-8}$

ICAM-1 domain 1-2-integrin α-L I complex (PDB: 1mq8) $K_d = 3.0 \times 10^{-6}$

p. M3-chemokine XCL1 complex (PDB: 2nyz) $K_d = 5.0 \times 10^{-10}$

subtilisin-chymotrypsin inhibitor 2 complex (PDB: 2sni) $K_d = 2.0 \times 10^{-12}$

Fig. 1 Some examples in the protein–protein complex set. The interface between two binding partners in a complex is shown as van der Waals surfaces of the residues in contact

protein–protein complex. Thus, the first step of BioQSAR modeling is to characterize protein–protein interface. The interface was defined as the residues from one protein chain that were in contact with other residues from another chain. We considered two residues coming from distinct chains to be in contact if there was a hydrogen bond, a water-mediated hydrogen bond, a van der Waals interaction, or at least one pair of contacting non-hydrogen atoms (< 6 Å) between them [28]. This "contacting" strategy was modified from the approach of Siggers et al. [29] The hydrogen bonds/water-mediated hydrogen bonds and van der Waals interactions were identified with HBPLUS [30] and PROBE [31] programs, respectively.

In fact, the shape and property of protein–protein interfaces in our dataset are extremely diverse, ranging from large and flat plane, to moderate and cupped cave, to small and narrow groove (Fig. 1). Here, we utilized five categories of structural descriptors to extract abundant information from a protein–protein interface: (1) *constitutional descriptors* (such as the numbers of different amino acids, groups and atoms present at the interface, the ratio of the residues at the interface to all residues in the complex, the quantity of hotspot residues, etc.), (2) *contacting descriptors* (such as atomic contact vectors, residue pair numbers, residue interaction indices, empirical contact potentials, etc.), (3) *geometrical descriptors* (such as the accessible surface area, interfacial connectivity indix, interface volume, etc.), (4) *physicochemical descriptors* (such as electrostatic potential, interfacial polarity, molar refractivity, hydrophobicity/hydrophilicity, etc.), and (5) *nonbonded descriptors* (such as the number of hydrogen bonds, hydrophobic forces, salt bridges, water-mediated hydrogen bonds, etc.).

A detailed list of the totally 110 descriptors used in this study to parameterize protein–protein interfaces is summarized in Supporting Information, Table S2.

Machine learning regression

Three sophisticated regression methods, i.e. partial least squares (PLS), support-vector machine (SVM) and Gaussian process (GP), were employed to establish the linear and nonlinear correlations between the interface descriptors and binding affinity of the 144 studied protein complex samples.

PLS is a widely used latent linear multivariate regression technique that relates the information in the response matrix **Y** to the variation over the descriptor matrix **X** [32].

Compared to those traditional linear methods, PLS can analyze the data with many, noisy, collinear, and even incomplete variables in both **X** and **Y**. Nowadays, PLS method has been widely used in the QSAR community [33]. In this research, the determination of significant latent variables in PLS was carried out by tenfold cross-validation.

SVM is a machine learning algorithm based on statistical learning theory (SLT), which aims at the structural risk minimization (SRM) rather than the traditional empirical risk minimization (ERM) and is especially suitable for small-sample, high-dimensional and strong collinear problems [34]. SVM uses Lagrange method to transform quadratic convex programming into a dual problem and further employs kernel function to implement the inner product operation of high dimensional Hilbert space in the input space. The grid-searching strategy was used here to determine the parameters of SVM with RBF kernel, i.e. ε-insensitive loss function, penalty C and radial width γ.

As a nonparametric Bayesian regression technique, GP constitutes a good compromise between the comprehensibility and predictive accuracy. It differs from most of the other statistical methods as it does not try to approximate the modeled system by fitting the parameters of the selected basis functions but rather searches for the relationship among measured data [35]. The details of GP algorithm can refer to Refs. [36, 37] Here, a mixed covariance function consisted of constant term, linear term, squared exponential term and noise term was used in GP; initial value of hyperparameter vector $\Theta = [\theta_0, \theta_1, \theta_2, \sigma_v^2, \{r\}_{m=1}^{M}]$ was set according to Obrezanova's rule [38] and further optimized by conjugate gradient method; line search using quadratic and cubic polynomial approximations and the Wolfe-Powell stopping criteria [39] was utilized together with the slope ratio method for guessing initial step sizes.

In this study, the machine learning regressions were performed with the help of in-house Matlab toolbox ZP-explore [40], which implemented the embedded modules ChemoAC, SVMKM and GPML to fulfill PLS, SVM and GP modeling, respectively.

Variable selection

The 110 considered descriptors reflect different aspects of the structural, geometrical and physicochemical properties of protein–protein interface. However, they must have strong collinearity, information overlapping, and background noises. The variable selection is a regular approach in QSAR to solve this problem. Here, we employed parallel genetic algorithm (GA) to perform variable selection for PLS, SVM and GP modeling [41]. During GA-variable selection procedure, the adjustable parameters of PLS and SVM were fixed at that determined by regression modeling with all descriptors as input. The GP hyperparamters were fixed according to the empirical rule suggested by Obrezanova et al. [38], that is, the overall scales θ_0, θ_2 and σ_v^2 (corresponding to constant term, squared exponential term and noise term in GP kernel, respectively) and length scales r_m (corresponding to each variable) were fixed as follows:

$$\theta_0 = \sqrt{n}\tau_{\mathrm{y}} \tag{1}$$

$$\theta_2 = \left(\sqrt{n} + 4\right)\tau_{\mathrm{y}} \tag{2}$$

$$\sigma_v^2 = 0.4 \tag{3}$$

$$r_m = 4\sqrt{M}\tau_{\mathrm{x}}^m \tag{4}$$

where $\tau_{\mathbf{y}}$ is the standard deviation of **y** values, $\tau_{\mathbf{x}}^m$ is the standard deviation of the mth column of descriptor matrix **X**, n is the size of input samples, and M is the number of total descriptors used in GP modeling. For the overall scale θ_1 (corresponding to linear term) that was in absence in literatures, the arithmetic mean of constant term and squared exponential term was used:

$$\theta_1 = \left(\sqrt{n} + 2\right)\tau_{\mathrm{y}} \tag{5}$$

The detailed descriptions of GA-PLS, GA-SVM and GA-GP procedures can be found in our previous publications [42, 43].

Model validation

Internal validation

The tenfold cross-validation was used to test the internal stability of regression models. In a round of the cross-validation, the 100 protein complexes in training set (see below) were randomly divided into 10 groups, each one consists of 10 samples. Subsequently, 10–1 groups were used for training and the prediction was made on the one left-out group. This procedure was repeated for 10 times and the average was computed for obtaining the stability measurement.

External validation

In a highly cited paper, Tropsha and coworkers argued that the good performance of cross-validation appears to be the necessary but not the sufficient condition for the model to have a high predictive power and the external validation is the only way to establish a reliable model [44]. Therefore, based upon D-optimal design technique we split the whole protein complex dataset into a training set consisting of 100 (~2/3) samples for building models and a test set of 44 (~1/3) samples for blindly evaluating the predictive power and generalization ability of the built models. D-optimal design provides an approach for selecting the most dissimilar sample structures and response information in the data set. Therefore, it can guarantee that the training data sets have well balanced structural diversity and are also representative of the entire range of response variable [45]. The splitting result of training and test sets is given in Table S1 in Supporting Information.

Double-cross validation

The double-cross validation randomly partitions the entire sample pool (containing 144 samples) into two size-equal subsets (separately containing 72 samples); each is used as training set for deriving regression models with GA-variable selection and as test set for validating the models exactly once [46].

Statistics

The performance of built BioQSAR models was measured quantitatively by the coefficients of determination deriving from fitting on training set (r^2), tenfold cross-validation on training set (q^2), and prediction on test set (r^2_{pred}), as well as the root-mean-square errors from fitting on training set (RMSF) and prediction on test set (RMSP), respectively [18]:

$$r^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i^{fitting}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_{tr})^2} \tag{6}$$

$$q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i^{cv})^2}{\sum_{i=1}^{n}(y_i - \bar{y}_{tr})^2} \tag{7}$$

$$r^2_{\text{pred}} = 1 - \frac{\sum_{i=1}^{m}\left(y_i - \hat{y}_i^{pred}\right)^2}{\sum_{i=1}^{m}(y_i - \bar{y}_{te})^2} \tag{8}$$

$$\text{RMSF} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i^{fitting}\right)^2} \tag{9}$$

$$\text{RMSCV} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{cv})^2} \tag{10}$$

$$\text{RMSP} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i^{pred}\right)^2} \tag{11}$$

where $n = 100$ and $m = 44$ are the numbers of members in training and test sets, respectively; y_{i} is the experimentally determined affinity of sample i; $\bar{y}_{tr}$ and $\bar{y}_{te}$ are the average values of the y_{i} over all training and test samples, respectively; $\hat{y}_i^{fitting}$, $\hat{y}_i^{cv}$ and $\hat{y}_i^{pred}$ are the estimated affinity for protein complex sample i by fitting, cross-validation and prediction, respectively.

Results and discussion

Regression modeling

The regression modeling procedure consists of two independent steps: Step 1, the 144 protein–protein complex samples were divided into a training set and a test set. The training set was used to perform variable selection and model development, and Step 2, when the step 1 was completed, the built models were then employed to perform prediction on test set. We first performed regression modeling on training set by using PLS, SVM and GP coupled with GA-variable selection. For each modeling, the GA was run for five rounds with parameter settings as follows: population size, 200; convergence criteria, 80 % of population achieving an agreement or genMax = 300; mutation rate, 1 %; hybridization and crossover, 2 points; tenfold cross-validation; data standardization, autoscaling for PLS and GP, or [–1, +1] normalization for SVM. The statistics of obtained optimal GA-PLS, GA-SVM and GA-GP models are listed in Table 1. At a glance, it is evident that the nonlinear SVM and GP performed much well as compared to linear PLS; the SVM and GP exhibited a satisfactory goodness-of-fit profile for the binding affinities of 100 training samples, with coefficients of determination of fitting $r^2 = 0.816$ and 0.802 as well as tenfold cross-validation $q^2 = 0.688$ and 0.691, respectively. In contrast, the PLS model extracted five significant latent variables from GA-selected descriptor pool, which cumulatively account for 76.4 % variance ($r^2 = 0.764$) of dependent variable by fitting and 61.3 % variance ($q^2 = 0.613$) by cross-validation.

It is well known that the high fitting ability of a QSAR model does not guarantee its strong predictive power, and the external validation is the only way to establish a reliable predictor [44]. Therefore, the built GA-PLS, GA-SVM and GA-GP models based on internal training set were further used to carry out prediction on external test set. As might be expected, the SVM and GP also possess a better generalization ability than PLS, as given by the higher predictive correlations $r^2_{\text{pred}} = 0.694$ and 0.714 for SVM and GP versus the lower value $r^2_{\text{pred}} = 0.637$ for PLS. The substantial increase in modeling performance of SVM and GP relative to PLS implies the inclusion of predominant nonlinearity in the protein complex system, which could be attributed to the fact that the binding of protein partners into the huge biomacromolecular architecture is a complicated event that cannot be described accurately using linear approach solely. The findings that nonlinear methods performed better than linear one are consistent with that arising from our previous modeling of biomolecular flexibility profile [47] and protein–peptide behavior [48], albeit

Table 1 The statistics of GA-PLS, GA-SVM and GA-GP regressions

Method	Model parameter	Training set				Test set	
		r^2	RMSF	q^2	RMSCV	r^2_{pred}	RMSP
GA-PLS	$N^a = 5$	0.764	1.050	0.613	1.376	0.637	1.198
GA-SVM	$\varepsilon = 0.2$, $C = 100$, $\gamma = 4$	0.816	0.935	0.688	1.104	0.694	1.004
GA-GP	$\theta_0 = 0.07$, $\theta_1 = 0.16$, $\theta_2 = 0.42$, $\sigma_v^2 = 0.21$	0.802	0.977	0.691	1.013	0.714	0.972

[a] *N*, the number of significant latent variables extracted from the original descriptors by GA-PLS

the linear method appears to be much efficient and fast in carrying out such modeling.

The scatter plots of model-derived affinities versus experimentally measured values are shown in Fig. 2. As can be seen, the resultant scatter points are closely distributed around the slope line fitting through these sample points, no matter linear PLS or nonlinear SVM and GP were used. However, the PLS model exhibited an obvious systemic error in both fitting and prediction, that is, the low- and high-affinity samples were considerably overestimated and underestimated, respectively, giving the fitted lines with small slopes (0.59 and 0.51). The systemic error is a common phenomenon in statistically treating biological system. For example, significant efforts have been previously addressed to predict the binding of peptide ligands to Src homology 3 (SH3) domain, and most obtained models were observed to have noticeable systemic error [49]. Hence, it is suggested that (1) the linear PLS is incapable of effectively capturing complicated nonlinear dependences involved in the protein complex system, and (2) the nonlinear factors such as interactive effects among the descriptors used in modeling are particularly significant for the protein–protein binding with extreme (lowest or highest) affinity.

Furthermore, we performed double-cross validations with GA-PLS, GA-SVM and GA-GP. The statistics generated from the validations are listed in Table 2. As might be expected, the predictive power of different regression models based on double-cross validation appears to have slight or moderate decreases as compared to that based on D-optimal design (r^2_{pred} from 0.637 to 0.565/0.593 for GA-PLS, from 0.694 to 0.650/0.624 for GA-SVM, and from 0.714 to 0.667/0.646 for GA-GP). Even so, the degenerations are not substantial if considering that the statistical correlations implied in the double-cross validation r^2_{pred} are also significant, which satisfy the Tropsha's criteria q^2 (and r^2_{pred}) > 0.5 [44].

Furthermore, it is worth noting that there are many factors that could introduce errors into the set of protein–protein binding affinity data. The experimental errors for both measured affinity and solved structure are inevitable: K_d values are usually reported in publications with standard errors of 20–50 %, equivalent to 0.1–0.25 kcal/mol for ΔG [22]; the resolution level of complex structures deposited in the PDB database commonly ranges from 1 to 3 Å, corresponding to atomic movement of about 0.05–0.2 Å [50]. Besides, the difference between complex structures in crystallized (static) and dissolved (dynamic) states is also an important source of errors. So, we previously estimated that the $r^2_{\text{pred}} = 0.80$ could be regarded as the upper limit of the predictive accuracy of structure-based protein–protein affinity predictors [18].

Comparison of BioQSAR with traditional methods

Traditionally, three kinds of prediction methods are usually used to estimate protein–protein binding affinity, i.e. empirical scoring algorithms, knowledge-based strategies and mechanism-type approaches. The empirical scoring algorithms define an energy term-weighed formula on the basis of affinity known protein complexes (usually used for docking purpose); the knowledge-based strategies utilize the frequency of contacts between different residues or atoms in known crystal structures to predict the binding affinity; the mechanism-type approaches attempt to ab initio compute the interaction energy between two binding partners using sophisticated force fields.

Herein, we gave a brief comparison of the BioQSAR to four widely used methods in blind prediction of the 44 test samples' affinities. The four methods include two empirical scoring algorithms (HADDOCK [51] and ZDOCK [52]), one knowledge-based strategy (DFIRE [53]) and one mechanism-type approach (OPLS-AA [54]). HADDOCK is a high-performing docking approach that uses a search algorithm in which experimental data can be incorporated and drive the docking procedure. ZDOCK implements fast evaluation based on shape complementarity, desolvation energy and electrostatic potential at the binding interface of a protein complex. DFIRE utilizes a potential of mean field (PMF) theory to provide estimation for the potentials between different atom types from distinct protein chains. OPLS-AA is all-atom force field that supplies parameters for nonbonded van der Waals and Coulombic electrostatic interactions of protein–protein binding. Here, the HADDOCK, ZDOCK

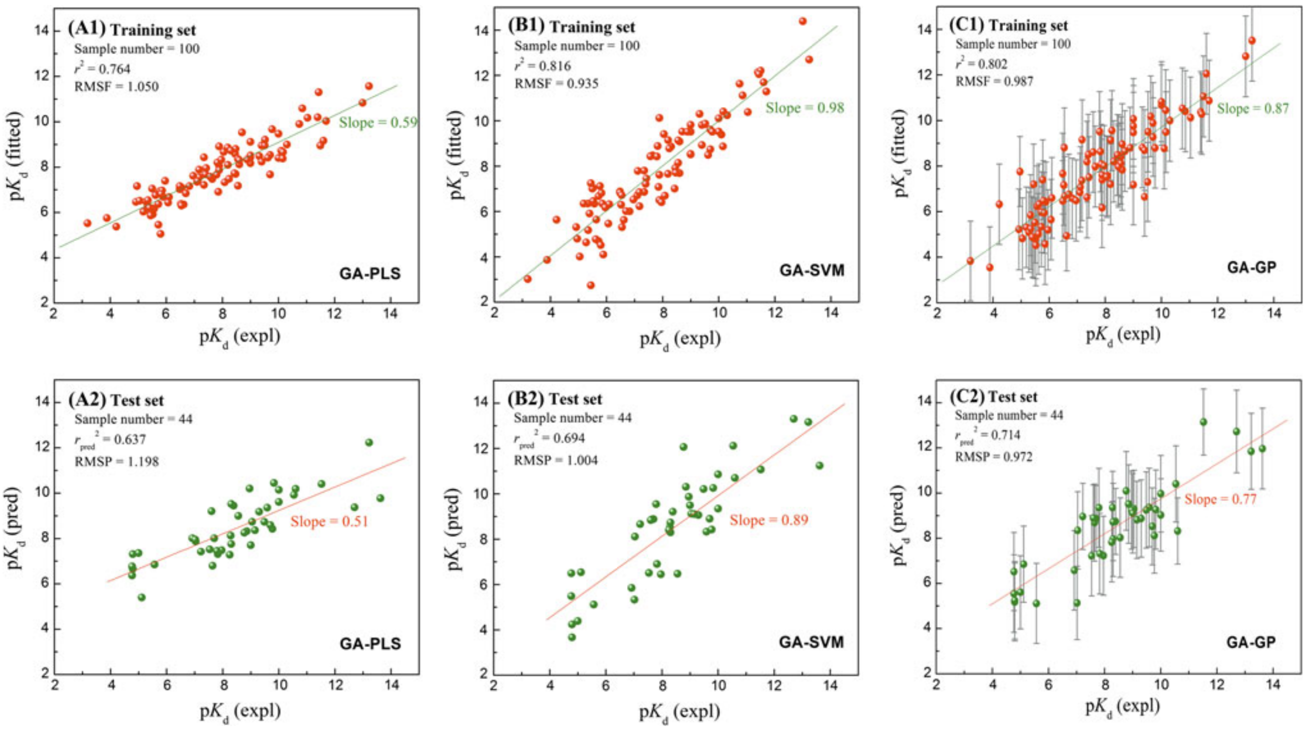


Fig. 2 Scatter plots of fitted affinities versus experimental values for 100 protein complexes in training set (A1, B1 and C1) and predicted affinities versus experimental values for 44 protein complexes in test set (A2, B2 and C2) by using GA-PLS, GA-SVM and GA-GP regressions. The calculated values resulting from GA-GP are shown with *error bars* of 90 % confidence

Table 2 The statistics of double-cross validations

Method	r^2_{pred} based on double-cross validation		r^2_{pred} based on D-optimal design
	Fold 1	Fold 2	
GA-PLS	0.565	0.593	0.637
GA-SVM	0.650	0.624	0.694
GA-GP	0.667	0.646	0.714

and DFIRE scoring were directly available via online servers, and the OPLS-AA can be implemented in standalone program TINKER [55] with consideration of GB/SA solvent effect [56]. The Pearson correlation coefficients R of experimental affinity with the calculated values by using different methods are shown in Fig. 3.

As can be seen, the mechanism-type approach OPLS-AA performed worst among all the adopted methods; its Pearson correlation R = –0.177 (R^2 = 0.031) is very close to zero. That is to say, the OPLS-AA predicted results are, in statistical viewpoint, not appreciably better than those generated from a randomizer. In fact, there were only very few reports for the successful use of modern force fields to treat biomolecular affinity problem. In most of these works, however, the predictions were carried out on either a small panel of samples [57], or a group of congeneric ligands binding to their common receptor [58]. Therefore, mechanism-type approach seems not to be suitable for modeling and predicting the complicated protein–protein binding with diverse interface features.

The two empirical scoring algorithms HADDOCK and ZDOCK gained better, but not much better, predictions as compared to OPLS-AA; their R values are −0.363 and −0.284 (R^2 = 0.131 and 0.081), respectively. In fact, Bonvin and co-workers have recently assessed the performance of nine docking scoring functions in their ability to predicting binding affinities of 81 protein complexes. They also obtained a relatively poor correlation between binding affinity and scores for all algorithms tested [20]. Thus, it is suggested that the empirical scoring algorithms are not the idea predictors for quantitatively evaluating protein–protein affinity, but might be used for the qualitative purpose to, for example, rank the binding modes generated from docking sampling.

The knowledge-based approach DFIRE appears to be acceptable in performing such prediction, albeit its correlation R = −0.475 (R^2 = 0.226) is yet far from perfect. The DFIRE was originally trained by complex structures of proteins with small ligands (M_W < 1000), but, as stated by Zhang et al. [53] it can also give an accurate prediction of the binding affinity of protein–protein and protein–DNA complexes. However, a significant fluctuation of R values (ranging from −0.16 to 0.73 [20, 53, 59, 60]) over different

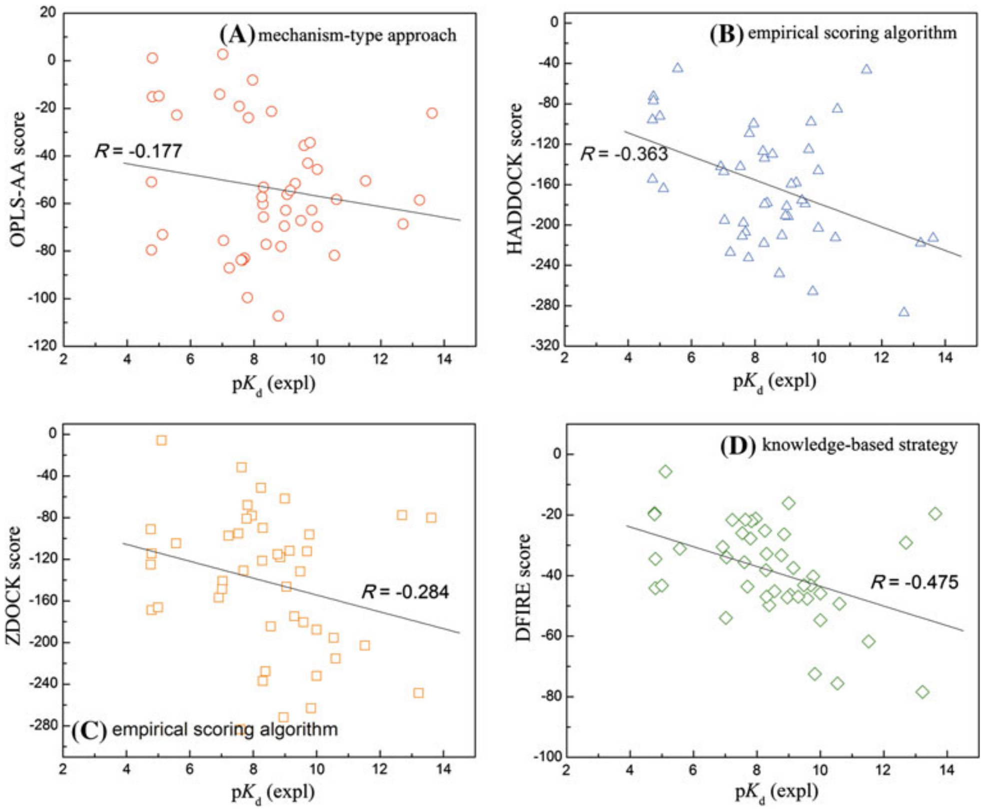


Fig. 3 The Pearson correlations of different predicted scores with experimental affinity for the 44 test protein complexes. **a** OPLS-AA score (mechanism-type approach), **b** HADDOCK score (empirical scoring algorithm), **c** ZDOCK score (empirical scoring algorithm), **d** DFIRE score (knowledge-based strategy)

protein complex sets suggests that the DFIRE method is very sensitive to samples used and the structural quantity of the samples.

In contrast, three BioQSAR models presented satisfactory profile when engaged to predict the 44 test samples ($R = 0.821$, 0.862 and 0.876 [$R^2 = 0.674$, 0.743 and 0.767] for GA-PLS, GA-SVM and GA-GP, respectively). However, we should be cautious for the conclusion if recalling that the BioQSAR models were built on the basis of supervised machine learning which require to be calibrated with a set of training samples that share a consistency with the test samples. Even so, we could consider that the performance of BioQSAR is at least comparable with or even better than that of most traditional methods in the context of structure-based prediction of protein–protein binding affinity, even more that nonlinear dependences associated with the binding are taken into account in prediction algorithm could gives another advantage for the BioQSAR approach.

Important descriptors in BioQSAR models

In this study, GA algorithm was employed to perform variable selection for PLS, SVM and GP regressions. The number of selected variables in final GA-PLS, GA-SVM and GA-GP models were, respectively, 32, 58 and 44. By a further examination we found that there were 9 common descriptors shared by the three models, which were also indicated as the important contributors in terms of PLS variable importance in projection (VIP) [32] and GP length scales r_m [36]. The 9 important descriptors as well as corresponding VIP and r_m values are listed in Table 3, which include 2 constitutional descriptors, 1 geometrical descriptor, and 2 physicochemical descriptors, 4 non-bonded descriptors—none is the contacting descriptors.

From the listed important descriptors we argue that the entropic effects might play a critical role in protein–protein binding, since there are several properties such as hydrophobicity, flexibility and desolvation energy related to (or governed by) entropic feature of system. However, it is a great challenge to accurately describe the relative contributions of entropy and enthalpy to the binding [61, 62], since the separation of the two components from binding affinity is a hard task if only based on statistical information—which are unable to clarify the details of thermodynamic mechanism.

The number of hotspot residues

The hotspot residues are defined as those that contribute critically ($\Delta\Delta G \geq 2$ kcal/mol) to protein–protein interactions [63]. In this respect, it is reasonable that the quantity of hotspot residues involved in a protein complex directly

Table 3 The 9 important descriptors shared by GA-PLS, GA-SVM and GA-GP models

Descriptor	Type	VIP[a]	r_m[a]
The number of hotspot residues	Constitutional descriptor	1.42	0.76
The number of oxygen atoms	Constitutional descriptor	1.37	1.13
Accessible surface area of interface	Geometrical descriptor	1.55	0.58
Average hydrophobicity	Physicochemical descriptor	1.87	0.82
Average flexibility	Physicochemical descriptor	0.96	1.40
The number of hydrogen bonds	Nonbonded descriptor	2.11	0.88
The number of water-mediated hydrogen bonds	Nonbonded descriptor	1.07	1.02
Desolvation energy	Nonbonded descriptor	0.60	1.97
The electrostatic potential of ion pairs	Nonbonded descriptor	0.89	1.60

[a] Larger VIP and smaller r_m values indicate the more significance of corresponding descriptors in GA-PLS and GA-GP models, respectively; vice versa

determines the binding strength and stability of the complex. Kortemme and Baker [64] pointed out that hotspot residues are responsible for ~80 % of total interaction energy, and Ofran et al. [65] further derived a significant correlation between the observed affinity and energy contribution of hotspot residues to protein–protein binding. Therefore, it is no surprise to see that the number of hotspot residues was selected as one of the most important descriptors in these regression models.

The number of oxygen atoms

Oxygen is a strong electronegative element which defines the polarity and electrostatic profile for a protein. In addition, oxygen atoms play a central role in a variety of noncovalent forces such as hydrogen bonds and salt bridges that confer both stability and specificity for biomolecular folding and binding [66]. These come together to suggest the importance of oxygen atoms in protein–protein interactions.

Accessible surface area of interface

The accessible surface area (ASA) of interface is computed as the change in ASA of protein members upon binding, which characterizes the interface size of protein complexes. Normally, a larger interface means more chance to form effective chemical contacts such as hydrophobic and van der Waals forces at the interface, and hence could provide stronger stabilization energy for binding. For example, the strong and persistent protein–protein interactions are mostly those with large and flat interface, such as HIV-1 protease dimer (1,600 Å) and elongation factor complex (3,660 Å) [67], whereas the weak and transient ones are commonly mediated by protein segments where a globular domain in one protein binds to a short peptide stretch in another, such as peptides bound to PDZ, SH3 and WW domains (<1000 Å) [68].

Average hydrophobicity and desolvation energy

Protein–protein interface is moderately hydrophobic, just like the core of folded protein monomer [69]. Young et al. [70] found that hydrophobic potentials provide a large proportion of total energetic contribution to the binding stability of protein complex, and remaining other nonbonded forces such as hydrogen bonds and salt bridges define specificity for the binding. The desolvation energy is a thermodynamic measure for the contribution of hydrophobic potentials to diverse biomolecular behavior, which was broadly used in the computational biology community to account for hydrophobic effects associated with protein–protein binding.

Average flexibility

Flexibility is an important factor to affect the biological function of proteins through exerting substantial influences on diverse properties of protein–protein interactions, such interface complementarity, entropic penalty and allosteric effect. However, computational analysis of protein flexibility is a challenge task, and most traditional methods therefore ignore the role of flexibility in protein–protein binding [71]. Alternatively, the BioQSAR approach does not directly calculate flexibility contribution to binding. In stead, this method indirectly correlates conformational features with binding affinity, thus avoiding the difficulty in directly calculating protein flexibility.

The number of hydrogen bonds and the number of water-mediated hydrogen bonds

It is surprising to see that two hydrogen bond descriptors were simultaneously selected as the important variables in the three BioQSAR models. Hydrogen bonds do play an invaluable role in governing the structure, stability, dynamics and function of biomolecules. It also confers both

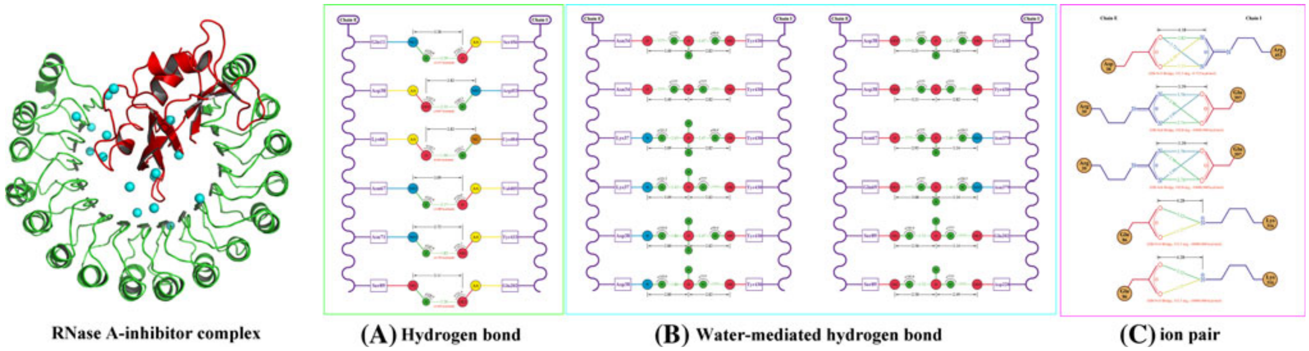


Fig. 4 Schematic representation of hydrogen bonds (**a**), water-mediated hydrogen bonds (**b**) and ion pairs (**c**) at the interface of RNase–inhibitor complex (PDB: 1dfj). This complex possesses a strong binding affinity $K_d = 5.9 \times 10^{-14}$. The figure was produced with in-house program 2D-GraLab [78]

stability and specificity for multiprotein architecture. In fact, protein–protein complex interfaces are typically characterized by complicated hydrogen bond networks, which refine binding specificity and eliminate interface roughness for the complexes (Fig. 4a). The water-mediated hydrogen bond is a hybrid of hydrogen bond and solvent effect. It is commonly used to fill the volume gaps contained in protein–protein interfaces (Fig. 4b). In recent years, the importance of water-mediated hydrogen bonds in biomolecular interactions has received increasing recognition by structural biologists. For example, side-chain truncation experiments showed that such water-mediated hydrogen bond networks can contribute significantly to the free energy of interaction of two proteins: the removal of one of the partners in the network often leads to substantial destabilization [72]. In addition, accounting for the loss of water-mediated hydrogen bonds observed in X-ray crystal structures can improves the accuracy of prediction of the effects of side chain truncations on the binding free-energy [73].

The electrostatic potential of ion pairs

An ionic pair is characterized by electrostatic interactions between two oppositely charged residues such as Arg-Glu and Lys-Asp (Fig. 4c); the short-rang, effective ion pairs are conventionally called salt bridges [74], which have been found to render substantial thermostability for protein complex architecture [75]. In fact, most ion pairs impose their influences on protein–protein binding by coupling with other structural properties of proteins, such as flexibility [76] and hydrogen bond [77]. In this way, the energetic effects of ion pairs and salt bridges on binding could be amplified considerably.

Conclusions

In this article, we presented a novel BioQSAR scheme for the fast and reliability prediction of protein–protein binding affinity. In this procedure, various structural descriptors were utilized to characterize the diverse properties of protein–protein interface and then correlated with the experimentally measured affinity of 144 sequence-nonredundant, structure-available and affinity-known protein complexes, by using linear and nonlinear machine learning protocols.

Although our approach was herein used to predict protein–protein binding affinity, the implementation and performing of this approach do not specialize in protein complexes—it could be readily modified to predict the binding affinity as well as biological activity and function of other biological systems, such as protein–nucleic acid complexes and monomeric proteins. We therefore expect that this QSAR-based method can be applied to other biomacromolecules as well.

Acknowledgments This work was supported by the National Natural Science Foundation of China (No. 31200993), the Fundamental Research Funds for the Central Universities (No. ZYGX2012J111), the Young Teacher Doctoral Discipline Fund of Ministry of Education of China (No. 20120185120025) and the scientific research funds of UESTC.

References

1. Hansch C, Fujita T (1964) ρ-σ-π analysis. A method for correlation of biological activity and chemical structure. J Am Chem Soc 86:1616–1626
2. Katritzky AR, Lobanov VS, Karelson M (1995) QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. Chem Soc Rev 24:279–287
3. Siraki AG, Chevaldina T, Moridani MY, O'Brien PJ (2004) Quantitative structure–toxicity relationships by accelerated cytotoxicity mechanism screening. Curr Opin Drug Discov Devel 7:118–125
4. Mozrzymas A, Rózycka-Roszak B (2010) Prediction of critical micelle concentration of nonionic surfactants by a quantitative structure–property relationship. Comb Chem High Throughput Screen 13:39–44
5. Fourches D, Pu D, Tassa C, Weissleder R, Shaw SY, Mumper RJ, Tropsha A (2010) Quantitative nanostructure–activity relationship modeling. ACS Nano 4:5703–5712

6. Natesan S, Wang T, Lukacova V, Bartus V, Khandelwal A, Subramaniam R, Balaz S (2012) Cellular quantitative structure–activity relationship (Cell-QSAR): conceptual dissection of receptor binding and intracellular disposition in antifilarial activities of Selwood antimycins. J Med Chem 55:3699–3712
7. Martin E, Mukherjee P, Sullivan D, Jansen J (2011) Profile-QSAR: a novel meta-QSAR method that combines activities across the kinase family to accurately predict affinity, selectivity, and cellular activity. J Chem Inf Model 51:1942–1956
8. Winkler DA (2002) The role of quantitative structure–activity relationships (QSAR) in biomolecular discovery. Brief. Bioinform. 3:73–86
9. Zhou P, Tian F, Wu Y, Li Z, Shang Z (2008) Quantitative sequence–activity model (QSAM): applying QSAR strategy to model and predict bioactivity and function of peptides, proteins and nucleic acids. Curr Comput Aided Drug Des 4:311–321
10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242
11. Concu R, Podda G, González-Díaz H (2009) In quantitative structure-property relationships from bio-molecular to social networks. Nova Science Publisher, New York
12. González-Díaz H, Vilar S, Santana L, Uriarte E (2007) Medicinal chemistry and bioinformatics — current trends in drugs discovery with networks topological indices. Curr Top Med Chem 7:1025–1039
13. González-Díaz H, Prado–Prado F, Perez-Montoto LG, Duardo-Sanchez A, Lopez-Diaz A (2009) QSAR models for proteins of parasitic organisms, plants and human guests: theory, applications, legal protection, taxes, and regulatory issues. Curr Proteomics 6:214–227
14. Munteanu CR, González-Díaz H, Magalhaes AL (2008) Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. J Theor Biol 254:476–482
15. González-Díaz H, Agüero-Chapin G, Varona J, Molina R, Delogu G, Santana L, Uriarte E, Gianni P (2007) 2D-RNAcoupling numbers: a new computational chemistry approach to link secondary structure topology with biological function. J Comput Chem 28:1049–1056
16. Munteanu CR, Vázquez JM, Dorado J, Pazos-Sierra A, Sánchez-González A, Prado–Prado FJ, González-Díaz H (2009) Complex network spectral moments for ATCUN motif DNA cleavage: first predictive study on proteins of human pathogen parasites. J Proteome Res 8:5219–5228
17. Neuvirth H, Raz R, Schreiber G (2004) ProMate: a structure based prediction program to identify the location of protein–protein binding sites. J Mol Biol 338:181–199
18. Tian F, Lv Y, Yang L (2012) Structure-based prediction of protein–protein binding affinity with consideration of allosteric effect. Amino Acids 43:531–543
19. Heuser P, Schomburg D (2007) Combination of scoring schemes for protein docking. BMC Bioinformatics 8:279
20. Kastritis PL, Bonvin AM (2010) Are scoring functions in protein–protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J Proteome Res 9: 2216–2225
21. Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AM, Janin J (2011) A structure-based benchmark for protein–protein binding affinity. Protein Sci 20:482–491
22. Park C, Marqusee S (2004) Analysis of the stability of multimeric proteins by effective ΔG and effective m-values. Protein Sci 13:2553–2558
23. Li H, Robertson AD, Jensen JH (2005) Very fast empirical prediction and rationalization of protein pK_a values. Proteins 61: 704–721
24. Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 285: 1735–1747
25. Krivov GG, Shapovalov MV, Dunbrack RL (2009) Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77:778–795
26. Zhou P, Zou J, Tian F, Shang Z (2009) Fluorine bonding: how does it work in protein–ligand interactions? J Chem Inf Model 49:2344–2355
27. Tian F, Lv Y, Zhou P, Yang L (2011) Characterization of PDZ domain–peptide interactions using an integrated protocol of QM/MM, PB/SA, and CFEA analyses. J Comput Aided Mol Des 25:947–958
28. Zhou P, Tian F, Ren Y, Shang Z (2010) Systematic classification and analysis of themes in protein–DNA recognition. J Chem Inf Model 50:1476–1488
29. Siggers TW, Silkov A, Honig B (2005) Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. J Mol Biol 345:1027–1045
30. McDonald IK, Thornton JM (1994) Satisfying hydrogen bonding potential in proteins. J Mol Biol 238:777–793
31. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC (1999) Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol 285:1711–1733
32. Wold S, Sjostrom M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemometr Intel Lab Syst 58:109–130
33. Stanton DT (2012) QSAR and QSPR model interpretation using partial least squares (PLS) analysis. Curr Comput Aided Drug Des 8:107–127
34. Cortes C, Vapnik V (1995) Support vector networks. Mach Learn 20:273–293
35. Zhou P, Xiang C, Wu Y, Shang Z (2010) Gaussian process: an alternative approach for QSAM modeling of peptides. Amino Acids 38:199–212
36. Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. MIT Press, Cambridge
37. Ren Y, Chen X, Feng M, Wang Q, Zhou P (2011) Gaussian process: a promising approach for the modeling and prediction of peptide binding affinity to MHC proteins. Protein Pept Lett 18:670–678
38. Obrezanova O, Csanyi G, Gola JMR, Segall MD (2007) Gaussian processes: a method for automatic QSAR modeling of ADME properties. J Chem Inf Model 47:1847–1857
39. Wolfe P (1969) Convergence conditions for ascent methods. SIAM Rev 11:226–235
40. Zhou P, Tian F, Lv F, Shang Z (2009) Comprehensive comparison of eight statistical modelling methods used in quantitative structure–retention relationship studies for liquid chromatographic retention times of peptides generated by protease digestion of the Escherichia coli proteome. J Chromatogr A 1216: 3107–3116
41. Cho SJ, Hermsmeier MA (2002) Genetic algorithm guided selection: variable selection and subset selection. J Chem Inf Comput Sci 42:927–936
42. 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
43. Tian F, Yang L, Lv F, Yang Q, Zhou P (2009) In silico quantitative prediction of peptides binding affinity to human MHC molecule: an intuitive quantitative structure–activity relationship approach. Amino Acids 36:535–554
44. Golbraikh A, Tropsha A (2002) Beware of q^2! J Mol Graph Model 20:269–276

45. Baroni M, Clementi S, Cruciani G, Kettaneh-Wold N, Wold S (1993) D-optimal designs in QSAR. Quant Struct Act Relat 12: 225–231
46. Filzmoser P, Liebmann B, Varmuza K (2009) Repeated double cross validation. J Chemometr 23:160–171
47. Tian F, Zhang C, Fan X, Yang X, Wang X, Liang H (2010) Predicting the flexibility profile of ribosomal RNAs. Mol Inf 29:707–715
48. Ren Y, Wu B, Pan Y, Lv F, Kong X, Luo X, Li Y, Yang Q (2011) Characterization of the binding profile of peptide to transporter associated with antigen processing (TAP) using Gaussian process regression. Comput Biol Med 41:865–870
49. He P, Wu W, Wang HD, Yang K, Liao KL, Zhang W (2010) Toward quantitative characterization of the binding profile between the human amphiphysin-1 SH3 domain and its peptide ligands. Amino Acids 38:1209–1218
50. Acharya KR, Lloyd MD (2005) The advantages and limitations of protein crystal structures. Trends Pharm Sci 26:10–14
51. Dominguez C, Boelens R, Bonvin AM (2003) HADDOCK: a proteinprotein docking approach based on biochemical or biophysical information. J Am Chem Soc 125:1731–1737
52. Chen R, Li L, Weng Z (2003) ZDOCK: an initial-stage protein-docking algorithm. Proteins 52:80–87
53. Zhang C, Liu S, Zhu Q, Zhou Y (2005) A knowledge-based energy function for protein–ligand, protein–protein, and protein–DNA complexes. J Med Chem 48:2325–2335
54. 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
55. Ponder JW, Richards FM (1987) An efficient newton-like method for molecular mechanics energy minimization of large molecules. J Comput Chem 8:1016–1024
56. Qiu D, Shenkin PS, Hollinger FP, Still WC (1997) The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. J Phys Chem 101: 3005–3014
57. Almlöf M, Brandsdal BO, Aqvist J (2004) Binding affinity prediction with different force fields: examination of the linear interaction energy method. J Comput Chem 25:1242–1254
58. Khoruzhii O, Donchev AG, Galkin N, Illarionov A, Olevanov M, Ozrin V, Queen C, Tarasov V (2008) Application of a polarizable force field to calculations of relative protein–ligand binding affinities. Proc Natl Acad Sci USA 105:10378–10383
59. Liu S, Zhang C, Zhou H, Zhou Y (2004) A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. Proteins 56:93–101
60. Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11: 2714–2726
61. Biela A, Sielaff F, Terwesten F, Heine A, Steinmetzer T, Klebe G (2006) Ligand binding stepwise disrupts water network in thrombin: enthalpic and entropic changes reveal classical hydrophobic effect. J Med Chem 55:6094–6110
62. Freire E (2009) ITC: affinity is not everything. Eur Pharm Rev 14:44–47
63. Moreira IS, Fernandes PA, Ramos MJ (2007) Hot spots: a review of the protein–protein interface determinant amino-acid residues. Proteins 68:803–812
64. Kortemme T, Baker D (2002) A simple physical model for binding energy hot spots in protein–protein complexes. Proc Natl Acad Sci USA 99:14116–14121
65. Ofran Y, Rost B (2007) Protein–protein interaction hotspots carved into sequences. PLoS Comput Biol 3:e119
66. Xu D, Tsai CJ, Nussinov R (1997) Hydrogen bonds and salt bridges across protein–protein interfaces. Protein Eng 10:999–1012
67. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein–protein recognition sites. J Mol Biol 285:2177–2198
68. Petsalaki E, Russell RB (2008) Peptide-mediated interactions in biological systems: new discoveries and applications. Curr. Opin. Biotech. 19:344–350
69. Tsai CJ, Nussinov R (1997) Hydrophobic folding units at protein–protein interfaces: implications to protein folding and to protein–protein association. Protein Sci 6:1426–1437
70. Young L, Jernigan RL, Covell DG (1994) A role for surface hydrophobicity in protein–protein recognition. Protein Sci 3:717–729
71. Tuffery P, Derreumaux P (2012) Flexibility and binding affinity in protein–ligand, protein–protein and multi-component protein interactions: limitations of current computational approaches. J R Soc Interface 9:20–33
72. Burnett JC, Kellogg GE, Abraham DJ (2000) Computational methodology for estimating changes in free energies of biomolecular association upon mutation. The importance of bound water in dimer-tetramer assembly for beta 37 mutant hemoglobins. Biochemistry 39:1622–1633
73. Jiang L, Kuhlman B, Kortemme T, Baker D (2005) A “solvated rotamer” approach to modeling water-mediated hydrogen bonds at protein–protein interfaces. Proteins 58:893–904
74. Kumar S, Nussinov R (2002) Close-range electrostatic interactions in proteins. ChemBioChem 3:604–617
75. Missimer JH, Steinmetz MO, Baron R, Winkler FK, Kammerer RA, Daura X, van Gunsteren WF (2007) Configurational entropy elucidates the role of salt-bridge networks in protein thermostability. Protein Sci 16:1349–1359
76. Kumar S, Wolfson HJ, Nussinov R (2001) Protein flexibility and electrostatic interactions. IBM J Res Dev 45:499–512
77. Marqusee S, Sauer RT (1994) Contributions of a hydrogen bond/salt bridge network to the stability of secondary and tertiary structure in lambda repressor. Protein Sci 3:2217–2225
78. Zhou P, Tian F, Shang Z (2009) 2D depiction of nonbonding interactions for protein complexes. J Comput Chem 30:940–951