Amino Acids (2012) 43:531–543
DOI 10.1007/s00726-011-1101-1

ORIGINAL ARTICLE

Structure-based prediction of protein–protein binding affinity with consideration of allosteric effect

Feifei Tian · Yonggang Lv · Li Yang

Received: 7 April 2011 / Accepted: 21 September 2011 / Published online: 17 November 2011

Abstract The conformational change upon protein–protein binding is largely ignored for a long time in the affinity prediction community. However, it is widely recognized that allosteric effect does play an important role in biomolecular recognition and association. In this article, we describe a new quantitative structure–activity relationship (QSAR)-based strategy to capture the structural and nonbonding information relating to not only the direct noncovalent interactions between protein binding partners, but also the indirect allosteric effect associated with binding. This method is then employed to quantitatively model and predict the protein–protein binding affinities compiled in a recently published benchmark consisting of 144 functionally diverse protein complexes with their structures available in both bound and unbound states (Kastritis et al. Protein Sci 20:482–491, 2011). With incorporating genetic algorithm and partial least squares regression (GA-PLS) into this method, a significant linear relationship between structural information descriptors and experimentally measured affinities is readily emerged and, on this basis, detailed discussions of physicochemical properties and structural implications underlying protein binding process, as well as the contribution of allosteric effect to the binding are addressed. We also give an empirical estimation of the prediction limit $r^2_{\text{pred}} = 0.80$ for structure-based method used to determine protein–protein binding affinity.

Keywords Protein–protein binding · Noncovalent interaction · Allosteric effect · Quantitative structure–activity relationship · Statistical modeling

Introduction

Protein–protein interactions are essential to many processes within living cells and organisms. The vast majority of proteins bind to other proteins at some time in their existence in order to perform various functions. Processes as varied as cytoskeletal remodeling, vesicle transport and signal transduction are all dependent on physical interactions between proteins (Bogan and Thorn 1998). In addition, many protein–protein interactions are mediated by peptide recognition modular domains that specifically bind a fraction of target proteins, forming typical protein–peptide adducts (Neduva et al. 2005; Petsalaki and Russell 2008). Because of the importance of protein–protein interactions in various physiological and biochemical processes involved in cellular regulatory network, understanding the molecular mechanism of such interactions is crucial step in protein engineering, as well as for designing therapeutic drug and vaccine against diverse diseases, such as cancer and AIDS (Fry 2006). Although numerous studies have been addressed on protein–protein interactions, the principles governing them are not yet fully understood (Jones and Thornton 1996; Vanhee et al. 2009).

F. Tian and Y. Lv contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-011-1101-1) contains supplementary material, which is available to authorized users.

F. Tian · Y. Lv · L. Yang (✉)
Key Laboratory of Biorheological Science and Technology
Under Ministry of Education, '111' Project Laboratory of
Biomechanics and Tissue Repair, and Bioengineering College,
Chongqing University, Chongqing 400044, China
e-mail: liyang@cqu.edu.cn

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

Computational determination of protein–protein binding affinity is very important not only for elucidating molecular mechanism underlying these interactions, but also for developing effective tools to perform accurate protein–protein docking and even to predict complete interactomes involved in a living cell. Previously, quantitative structure–activity relationship (QSAR) studies have been intensively addressed to identify and predict peptides binding to diverse protein receptors (Pripp et al. 2005; Du et al. 2008), such as antigen peptide (Zhao et al. 2007; Tian et al. 2009), antimicrobial peptide (Jenssen et al. 2008), and bitter peptide (Pripp and Ardö 2007), and these works could be regarded as the pioneers of predicting protein affinity to their interacting partners. In past two decades, a number of techniques and methods have been exploited to model and identify the interaction behavior of proteins with their cognate or noncognate partners. In the early stage, limited to the availability of high-quality protein structures a series of sequence-based approaches were used to fulfill this purpose (Kini and Evans 1996; Zhou and Shan 2001; Shen et al. 2007). However, information deriving only from primary sequence of proteins is incapable of effectively capturing the complicated binding profile of protein complexes standing in three-dimensional (3D) spatial configuration. Therefore, most of sequence-based methods are limited to their effectiveness in qualitative classification of protein interaction pattern, but fail to quantitatively predict protein binding affinity. With the number of solved protein–protein complex 3D structures growing up rapidly in recent years, interaction analysis and affinity prediction based on complex structures have received much attention in the structural bioinformatics community. Nowadays, the available methods used for structure-based prediction of protein–protein binding affinity can be roughly categorized into three classes: empirical scoring method (Horton and Lewis 1992; Ma et al. 2002; Audie and Scarlata 2007), knowledge-based method (Jiang et al. 2002; Zhang et al. 2005; Su et al. 2009), and ab initio prediction method (Brandsdal and Smalås 2000; Gandhi and Mancera 2009; Cole et al. 2010). The empirical scoring method defines an energy term-weighed formula on the basis of affinity-known protein complexes (usually used for docking purpose); the knowledge-based method utilizes the frequency of contacts between different residues or atoms in known crystal structures to predict the binding affinity; the ab initio prediction method employs theoretical approaches [such as MM-PB/SA (Gandhi and Mancera 2009), free-energy perturbation (Brandsdal and Smalås 2000) and linear-scaling analysis (Cole et al. 2010)] to directly calculate the interaction energy between two binding partners. Nevertheless, associating a crystal structure to biophysical measurements done in solution is an error-prone process, and the published sets contain many incorrect affinity data (Kastritis and Bonvin 2010). Moreover, the structural data in these sets represent the (bound) complexes, but not their free (unbound) components. Therefore, the models based on them describe the thermodynamics of association reaction by its product only, ignoring the reactants and the structure changes they may undergo.

Very recently, by exhaustively surveying all the published literatures in the field of protein recognition and interaction, Kastritis et al. have first presented a functionally diverse, nonredundant benchmark for protein–protein binding affinity, in which 144 protein complexes that have high-resolution structures available for both the complexes and their unbound components, as well as corresponding dissociation constants measured by biophysical methods, are compiled (Kastritis et al. 2011). With this benchmark, it is now possible to accurately predict protein–protein binding affinity and to effectively evaluate the allosteric effect associated with the binding. In the present study, beyond the three categories of prediction method mentioned above, we herein describe a new quantitative structure–activity relationship (QSAR)-based strategy to characterize the interaction profile of protein complexes, to predict the binding affinity of the interactions, and to assess the contribution of conformation change to the affinity. With built predictive models, the physicochemical properties and structural implications underlying the specific recognition and association between the members of protein complex are analyzed in detail. We also give a discussion of the prediction limit that a structure-based protein–protein affinity model could reach.

Materials and methods

Dissection of a protein–protein binding

Protein–protein binding is an induced-fit process which can be divided into two independent thermodynamic steps: allostery and association (Fig. 1). In the first allosteric step two interacting partners are changed in their conformations to define geometrically and physicochemically complementary surfaces (state 1 → state 2); in the following association step the complementary surfaces are fitted together to form a functional unit of protein complex (state 2 → state 3). In the binding process, many kinds of noncovalent interaction such as hydrogen bond, salt bridge, and van der Waals contact are formed and broken, and solvent effect also contributes to the binding significantly. As a result, the built complex architectures range from being permanent (Ford 1987; Shi et al. 2006), to being metastable (Lim et al. 2002), to being transient (Tang et al. 2006; Blobel et al. 2009), which can be quantitatively characterized using equilibrium dissociation constant (K_d),

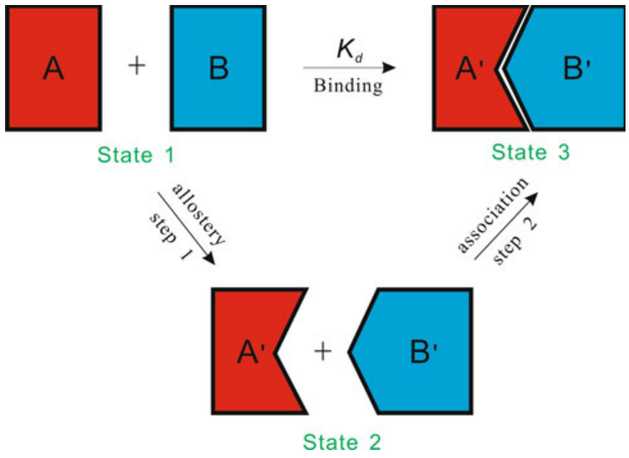

Fig. 1 Schematic representation of protein–protein binding process

which is measured at equilibrium or derived from the reaction kinetics, and the related Gibbs free energy of dissociation ΔG.

Characterization of a protein–protein complex

Previously, we proposed a set of rotation-translation invariants called 3D holographic vector of atomic interaction field (3D-HoVAIF) (Tian et al. 2007) to characterize the 3D structure properties of a small molecule. This method classifies organic atoms into 10 types in terms of their families in the periodic table of elements and hybridization states, and calculates 55 cross-interaction terms between the 10 atomic types to describe the nonbonding profile of an organic molecule. 3D-HoVAIF has been successfully applied to model and predict diverse properties and activities of various molecular entities, such as steroid derivatives (Zhou et al. 2007a), artemisinin antimalarial agents (Ren et al. 2008), and neuraminidase inhibitors (Sun et al. 2010). In this study, we extend the 3D-HoVAIF method to parameterize structural characteristics of biomacromolecules and their complexes.

A biomolecule such as protein could be regarded as an enlarged version of small organic molecule, which is constructed by the basic structural units of amino acid residues. Therefore, amino acid entities can be treated as "pseudo-atoms" of the "protein molecule", just like common atoms in small molecule. Theoretically, there are at most 210 (20 × 21/2) possible combinations between the 20 types of natural amino acid. Here, we used the 210 combination terms to describe inter-interaction profile involved in the static structure of a protein or a protein–protein complex, which can be straightforwardly expressed as a symmetric matrix M (Table 1), in which the matrix element $m(q, p)$ $[q \le p]$ represents the sum of all interaction potentials (vide post) between amino acid types q and p in a protein or a complex.

Characterization of a protein–protein binding

As aforementioned, the inter-interaction profile involved in a protein or a protein–protein complex can be characterized using 210 combination terms between 20 amino acid types. Here, let us look at the Fig. 1; the unbound components and bound complex are assigned to states 1 and 3, respectively, and the allosteric intermediate is the state 2. The changes in structure properties from states a to b (a and b = 1, 2, or 3) can be quantified by the difference of inter-interaction profiles between these two states:

$$\Delta M^{a\to b} = M^b - M^a, \tag{1}$$

where the element $\Delta m^{a\to b}(i, j)$ of profile-changed matrix $\Delta M^{a\to b}$ is obtained by subtracting $m^a(i, j)$ from $m^b(i, j)$, i.e., $\Delta m^{a\to b}(i, j) = m^b(i, j) - m^a(i, j)$. The unbound states 1 and 2 can be regarded as the extreme conditions of protein–protein complex with their two components separating from each other infinitely, and their inter-interaction profiles are thus computed as so. In this way, three profile-changed matrices parameterizing the changes in structure properties between different states could be given as follows:

$\Delta M^{1\to 3}$ parameterizing A–B binding (binding matrix);

$\Delta M^{1\to 2}$ parameterizing the allosteric effect of two isolated binding partners A and B (allosteric matrix);

$\Delta M^{2\to 3}$ parameterizing the association of two allosteric binding partners A′ and B′ (association matrix).

It is evident that $\Delta M^{1\to 3} = \Delta M^{1\to 2} + \Delta M^{2\to 3}$.

Interaction potentials between amino acid residues

It is well known that noncovalent interactions such as hydrogen bonding, hydrophobic interaction, and van der Waals contact are the basic chemical forces dominating biomolecular folding and binding. In the famous 3D-QSAR method comparative molecular field analysis (CoMFA) (Cramer et al. 1988) only electrostatic and steric potentials are considered to describe the nonbonding field distributions around a group of aligned drug compounds. However, it is apparent that other factors such as solvent effect and hydrogen bonding are also essential to biomolecular recognition and association. We recently observed diverse noncovalent interactions present at the interface and in the interior of protein–protein complexes (Zhou et al. 2009a), but most of them can be attributed into four kinds of basic types: electrostatic force, hydrogen bonding, van der Waals contact, and hydrophobic interaction. Therefore, here these four nonbonding types were used to describe the interaction potentials between amino acid residues in a studied protein

Table 1 The 20 types of natural amino acid and the 210 combination terms between them

	AAs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	Ala	1–1	1–2	1–3	1–4	1–5	1–6	1–7	1–8	1–9	1–10	1–11	1–12	1–13	1–14	1–15	1–16	1–17	1–18	1–19	1–20
2	Arg		2–2	2–3	2–4	2–5	2–6	2–7	2–8	2–9	2–10	2–11	2–12	2–13	2–14	2–15	2–16	2–17	2–18	2–19	2–20
3	Asn			3–3	3–4	3–5	3–6	3–7	3–8	3–9	3–10	3–11	3–12	3–13	3–14	3–15	3–16	3–17	3–18	3–19	3–20
4	Asp				4–4	4–5	4–6	4–7	4–8	4–9	4–10	4–11	4–12	4–13	4–14	4–15	4–16	4–17	4–18	4–19	4–20
5	Cys					5–5	5–6	5–7	5–8	5–9	5–10	5–11	5–12	5–13	5–14	5–15	5–16	5–17	5–18	5–19	5–20
6	Gln						6–6	6–7	6–8	6–9	6–10	6–11	6–12	6–13	6–14	6–15	6–16	6–17	6–18	6–19	6–20
7	Glu							7–7	7–8	7–9	7–10	7–11	7–12	7–13	7–14	7–15	7–16	7–17	7–18	7–19	7–20
8	Gly								8–8	8–9	8–10	8–11	8–12	8–13	8–14	8–15	8–16	8–17	8–18	8–19	8–20
9	His									9–9	9–10	9–11	9–12	9–13	9–14	9–15	9–16	9–17	9–18	9–19	9–20
10	Ile										10–10	10–11	10–12	10–13	10–14	10–15	10–16	10–17	10–18	10–19	10–20
11	Leu											11–11	11–12	11–13	11–14	11–15	11–16	11–17	11–18	11–19	11–20
12	Lys												12–12	12–13	12–14	12–15	12–16	12–17	12–18	12–19	12–20
13	Met													13–13	13–14	13–15	13–16	13–17	13–18	13–19	13–20
14	Phe														14–14	14–15	14–16	14–17	14–18	14–19	14–20
15	Pro															15–15	15–16	15–17	15–18	15–19	15–20
16	Ser																16–16	16–17	16–18	16–19	16–20
17	Thr																	17–17	17–18	17–19	17–20
18	Trp																		18–18	18–19	18–20
19	Tyr																			19–19	19–20
20	Val																				20–20

system. In this study, the electrostatic and van der Waal potentials were calculated using the classical Coulomb's law and Lennard-Jones function, respectively, which can be conducted by AMBER03 force field (Duan et al. 2003). The hydrogen bonds involved in a protein complex were identified with the HBPLUS program (McDonald and Thornton 1994), and their potential was characterized using the angle-weighted Lennard-Jones-like 8-6 function (Boobbyer et al. 1989). In addition, an empirical formula developed in our lab was employed to quantitatively describe hydrophobic potential $E_{mn}^{hp} = -(S_m\rho_m + S_n\rho_n)e^{-d_{mn}}$ (Zhou et al. 2007b), where d_{mn} is the distance between two atoms m and n, ρ is the atomic solvation parameters (Eisenberg and McLachlan 1986), and S is the atomic solvent accessible surface area defined in the MSMS program (Sanner et al. 1996) with Bondi radii set (Bondi 1964). For a protein, interaction energy U_{ij} between its ith and jth residues could be calculated in atom-pairwise additive manner: $U_{ij} = \sum_{m=1}^{M}\sum_{n=1}^{N} E_{mn}$, where m represents the mth atom of total M atoms in residue i, n indicates the nth atom of total N atoms in residue j, and E_{mn} is the potential between mth atom of residue i and nth atom of residue j. Once interaction energy U_{ij} between ith and jth residues of the protein is determined using the approach described above, it would be added into one of the 210 terms——the term that meets the combination of residue types for the ith and jth residues.

Each of the three profile-changed matrices, $\Delta M^{1\to3}$, $\Delta M^{1\to2}$ and $\Delta M^{2\to3}$, associated with protein–protein binding can be described separately by the four kinds of nonbinding potentials. In this way, the changes in structure properties between any two states a and b can thus be characterized using four counterparts (each counterpart represents a nonbonding type) of profile-changed matrix $\Delta M^{a\to b}$, or totally $4 \times 210 = 840$ variables. These variables will be further correlated with experimentally measured affinities compiled in a recently published protein–protein binding benchmark (vide post) via the sophisticated partial least squares (PLS) regression (Wold et al. 2001), the most widely used latent regression method for linearly relating two data matrices with many, noisy, collinear and even incomplete variables, just as the case faced in this study. Here, the PLS algorithm was implemented with the help of in-house Matlab program ZP-explore (Zhou et al. 2009b).

Data set of protein–protein complexes

Very recently, Kastritis et al. have published a nonredundant benchmark consisting of 144 protein–protein complexes that have high-resolution structures available for both the complexes and their unbound components (Kastritis et al. 2011). The members of this set are diverse in terms of the biological functions they represent, with complexes that involve G-proteins and receptor extracellular domains, as well as antigen–antibody, receptor–inhibitor, and enzyme–substrate adducts. It is also diverse in terms of the partners' affinity for each other, with dissociation constants K_d ranging from 10^{-5} to 10^{-14} M. In this study, we employed this benchmark in conjunction with the newly proposed method described above to develop predictable QSAR models for accurately predicting the binding affinity between protein and its interacting partners, and for effectively evaluating the contribution of allosteric effect to the affinity. Before performing study, the missing hydrogen atoms and side chains of these protein complexes and their free monomers were added with the REDUCE (Word et al. 1999) and SCWRL (Krivov et al. 2009) programs, respectively. REDUCE was demonstrated in our previous study to be capable of precisely reproducing the hydrogen positions determined by neutron diffraction (Zhou et al. 2009c).

The relevant information about this data set is tabulated in Supporting Information, Table S1.

Model validation

An excellent regression model should be robust and generalized including high internal correlation and cross-validated performance. However, Golbraikh and Tropsha (2002) pointed out that a more reliable result should be obtained by dividing external test set. Hence, 100 (~2/3) out of the 144 samples were randomly selected as training set for building QSAR models, and the remaining 44 (~1/3) were as independent test set for validating the built models.

For a QSAR model, its performance, in statistical viewpoint, could be measured quantitatively by the coefficients of determination of fitting (r^2), leave-one-out (LOO) cross-validation (q^2), and prediction (r^2_{pred}) on training set, training set and test set, respectively, as well as the root-mean-square errors of fitting (RMSF) and prediction (RMSP) on training set and test set, respectively (Tian et al. 2011):

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

$$q^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i^{cv}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_{tr}\right)^2} \tag{3}$$

$$r^2_{\text{pred}} = 1 - \frac{\sum_{i=1}^{m}\left(y_i - \hat{y}_i^{pred}\right)^2}{\sum_{i=1}^{m}\left(y_i - \bar{y}_{te}\right)^2} \tag{4}$$

$$\mathrm{RMSF} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(y_i - \hat{y}_i^{fitting}\right)^2} \quad (5)$$

$$\mathrm{RMSP} = \sqrt{\frac{1}{m}\sum\nolimits_{i=1}^{m}\left(y_i - \hat{y}_i^{pred}\right)^2}, \quad (6)$$

where $n = 100$ and $m = 44$ are the member's numbers of 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 affinities for sample i by fitting, cross-validation and prediction, respectively.

We further performed a particularly stringent Monte Carlo cross-validation (MCCV) (Xu and Liang 2001) to deeply test the stability and reliability of built QSAR models. In MCCV procedure, the whole data set was partitioned randomly into two parts; one part (100 training samples) was used to build model, the remaining part (44 test samples) was used for prediction, and the whole process was repeated thousands of times——we herein adopted $2^{12} = 4096$ repetitions as recommended by Manchester and Czerminski (2008)——to achieve convergent expression of statistics.

Results and discussion

Model development

We first used PLS regression to separately correlate the three profile-changed matrices, $\Delta M^{1\rightarrow3}$, $\Delta M^{1\rightarrow2}$ and $\Delta M^{2\rightarrow3}$, with the experimental binding affinities of protein–protein complexes on the basis of 100 training samples. The resulting statistics are listed in Table 2. As can be seen, the coefficient of determination r^2 and cross-validated q^2 are distinct dramatically when different matrices used. As might be expected, the model based on binding matrix $\Delta M^{1\rightarrow3}$, which characterizes the whole binding process of proteins with their partners, performed fairly well with respect to its fitting ability $r^2 = 0.869$ and stability $q^2 = 0.694$ (Fig. 2a), whereas the modeling using allosteric matrix $\Delta M^{1\rightarrow2}$ gave rise to quite modest results, given the relatively low values of $r^2 = 0.661$ and $q^2 = 0.501$. The statistical quality of association matrix $\Delta M^{2\rightarrow3}$-based model was between that of other two, indicating that the structural information involved in the $\Delta M^{2\rightarrow3}$, which does not consider conformational change associated with protein–protein binding, only contribute partially, but not entirely, to binding affinity. These models were further used to predict the affinity values of 44 test samples. The ordering of model's predictive powers on test set was in agreement with that of fitting abilities on training set, i.e., $\Delta M^{1\rightarrow3} > \Delta M^{2\rightarrow3} > \Delta M^{1\rightarrow2}$, but the difference between fitting and prediction were relatively significant, suggesting overfitting phenomenon existed within these models. This is not unexpected if considering that large amount of variables (840 descriptors) used in the modeling would ineluctably lead to strong fitting and weak generalization, since not all these variables contribute significantly to binding. In the QSAR field, one solution for the overfitting problem is to perform variable selection before developing statistical models, and genetic algorithm (GA) is recognized as one of the most effective strategies to do so (Cho and Hermsmeier 2002). Therefore, we rebuilt the three models using GA-variable selection. The parameters setting for GA running was basically consistent with that adopted in our previous study (only slight modification) (Tian et al. 2009). It can be seen from Table 2 that the predictive power (r^2_{pred}) of these rebuilt models received substantial improvement as compared to those based on all the 840 variables, albeit their fitting ability (r^2) seems to be impaired more or less during the variable selection procedure. The decrease in fitting ability and increase in predictive power clearly pronounce that the overfitting attached to these models were largely eliminated through GA-variable selection, giving rise to the best prediction of $r^2_{\mathrm{pred}} = 0.693$ using the model deriving from $\Delta M^{1\rightarrow3}$ (Fig. 2b).

From the Table 2 it is evident that modeling performances based on the three profile-changed matrices increase in the order $\Delta M^{1\rightarrow2} < \Delta M^{2\rightarrow3} < \Delta M^{1\rightarrow3}$, no

Table 2 Statistics of PLS and GA-PLS models

Methods	Descriptor	NL	Training set (100 samples)			Test set (44 samples)	
			r^2	q^2	RMSF	r^2_{pred}	RMSP
PLS	$\Delta M^{1\rightarrow2}$	6	0.661	0.501	1.287	0.478	1.472
PLS	$\Delta M^{2\rightarrow3}$	7	0.736	0.632	1.135	0.617	1.251
PLS	$\Delta M^{1\rightarrow3}$	9	0.869	0.694	0.798	0.630	1.201
GA-PLS	$\Delta M^{1\rightarrow2}$	5	0.624	0.545	1.357	0.522	1.384
GA-PLS	$\Delta M^{2\rightarrow3}$	7	0.773	0.656	1.053	0.675	1.104
GA-PLS	$\Delta M^{1\rightarrow3}$	8	0.815	0.722	0.951	0.693	1.006

NL number of significant latent variables extracted from the original variables by PLS

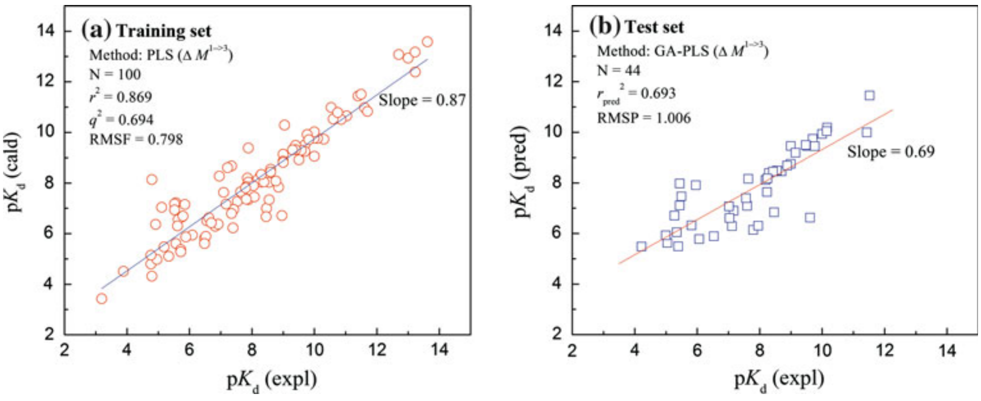

Fig. 2 **a** Scatter plot of calculated versus experimental binding affinities for the 100 training samples using the PLS method in conjunction with binding matrix $\Delta M^{1\to3}$; **b** scatter plot of predicted versus experimental binding affinities for the 44 test samples using the GA-PLS method in conjunction with $\Delta M^{1\to3}$

mater if GA-variable selection was addressed. This is not surprise because the binding matrix $\Delta M^{1\to3}$ contains all information involved in both $\Delta M^{1\to2}$ and $\Delta M^{2\to3}$, and hence can give optimal result for modeling. In almost all of previous works (for instance, see literatures Ma et al. 2002; Zhang et al. 2005; Audie and Scarlata 2007), allosteric effect associated with protein–protein binding was completely ignored and the binding was simply regarded as a process of docking between two rigid bodies. However, our study addressed here demonstrated that the indirect allosteric effect has a solid contribution to binding, albeit this contribution is less than that of direct interactions between two protein partners. This is coming to light from two aspects: (1) the predictive power of GA-PLS model solely based on allosteric matrix $\Delta M^{1\to2}$ is accepted, as its $r^2_{\text{pred}} = 0.522$——this value satisfies the criterion $r^2_{\text{pred}} > 0.5$ recommended by Tropsha et al. (2003) for a predictable QSAR model, and (2) there is a considerable improvement in model's quality if allostery was engaged; for example, the r^2_{pred} value rises from 0.617 to 0.675 (for PLS modeling) or from 0.630 to 0.693 (for GA-PLS modeling) when the allosteric information were introduced into association matrix $\Delta M^{2\to3}$, resulting in binding matrix $\Delta M^{1\to3}$. It is worth noting that, though the indirect allosteric effect is important for developing a reliable model to accurately predict protein–protein binding affinity, direct nonbonding interactions yet dominate the binding, which can be rationalized by the fact that $\Delta M^{2\to3}$-based model performed much well as compared to that deriving from $\Delta M^{1\to2}$. Furthermore, the best GA-PLS model was tested by stringent MCCV, and resulting mean values of r^2 (fitting on training samples) and r^2_{pred} (prediction on test samples) over 4,096 repetitions were 0.831 and 0.676, respectively. As might be expected, these two statistics generating from MCCV are roughly consistent with that from "single splitting" validation (0.815 and 0.693, respectively), indicating that the "single splitting" validation could properly reflect the quality and performance of built models.

Model analysis

As aforementioned, GA-variable selection can significantly improve the quality of PLS models. Therefore, we herein gave a further discussion on optimal GA-derived model, the $\Delta M^{1\to3}$-based GA-PLS model. GA algorithm has extracted 378 variables from the crude panel consisted of 840 descriptors, of which 114, 101, 88 and 75 are hydrophobic, steric, hydrogen bonding and electrostatic terms, respectively. At an initial glance, all the four kinds of noncovalent interaction cast effective potency to binding, and the hydrophobic and steric effects appear to be the most important facets that significantly influence the recognition and association between protein partners. This notion is consistent with early investigations on high-resolution crystal structures that protein–protein interface is mainly composed of nonpolar amino acid residues, with the mosaic of polar residues onto it (Tsai and Nussinov 1997; Tsai et al. 1997). According to this theory, hydrophobic force drives binding event and, when the initial complex is formed, large-scale van der Waals contacts as well as specific hydrogen bonding can further shape the exquisite structure of binding interface (Xu et al. 1997; Song and Zhao 2005). We further decomposed the 378 selected variables into two independent reaction steps, i.e., allostery and association, and separately used these two sets of decomposed variables to develop PLS models to make prediction on test samples. As might expected, the predictive power of the model based on association appears to be stronger than that on allostery ($r^2_{\text{pred}} = 0.626$ vs. 0.514), indicating that association is the dominant step in whole protein–protein binding process, while allostery serves as an assistor to the association. This conclusion is in agreement with recent findings by Stein et al. (2011) that protein interactions with many partners only undergo smaller changes upon binding, and are less likely to freely explore larger conformational changes.

Furthermore, variable importance in the projection (VIP) (Wold et al. 2001) of GA-PLS model can, in

Table 3 The most significant 10 terms in the GA-PLS model

No.	Residue-pair	Nonbinding property	VIP
1	Ile-Tyr	Hydrophobic	2.334
2	Arg-Asp	Electrostatic	2.218
3	Gln-Thr	Hydrogen bonding	2.156
4	Met-Val	Steric	2.074
5	Ile-Trp	Hydrophobic	1.955
6	Pro-Phe	Steric	1.930
7	Glu-Lys	Hydrogen bonding	1.809
8	Ser-Asp	Hydrogen bonding	1.767
9	Leu-Gly	Hydrophobic	1.712
10	Val-Cys	Electrostatic	1.695

statistical viewpoint, give a preliminary insight into the important residue-pairs and nonbonding properties exerting to binding. The most significant ten terms and corresponding VIP values are listed in Table 3. It is shown that diverse properties contribute remarkable effects to binding, which agrees to the classical protein–protein interaction model (Jones and Thornton 1996). As seen in Fig. 3, the schematic representation of nonbonding interaction pattern at the binding interface of nuclease A with its cognate inhibitory protein, a high-affinity complex with picomolar dissociation constant ($K_d = 3.2 \times 10^{-12}$), various noncovalent elements present at the interface confer both specificity and stability for complex architecture. This point confirms that co-working of diverse chemical forces at the binding interface is a basic rule for protein–protein recognition and association (Otlewski and Apostoluk 1997), because only in this way the binding process can be modulated in a subtle and exquisite manner to ensue the accurate (and space–time-specific) recognition and interaction between two binding partners.

The bulky Ile, Tyr, Phe and Leu, polar Gln, Thr and Ser, as well as charged Arg, Asp and Lys seem to be crucial for protein–protein interaction. The nonpolar Ile and Leu were found to have a relatively high propensity occurring in the interface of protein complexes and thus thought to exert large amount of non-specific hydrophobic and van der Waals potentials toward binding (Conte et al. 1999), while the charged Lys, Arg, and Asp are the major components of hot spots (Bogan and Thorn 1998), which are key regions that fundamentally contribute to the interaction free energy between two proteins (Clackson and Wells 1995). These charged residues are able to form strong salt bridges and electrostatic attractions across binding interface. In addition, the polar Gln, Thr and Ser are the good acceptors and donors of hydrogen bonds and have been frequently observed in various protein–protein interfaces, such as HIV-1 protease dimer (Wlodawer et al. 1989), MHC–antigen adduct (Madden 1995), and H5N1 hemagglutinin complex (Stevens et al. 2006). It should be noted here that

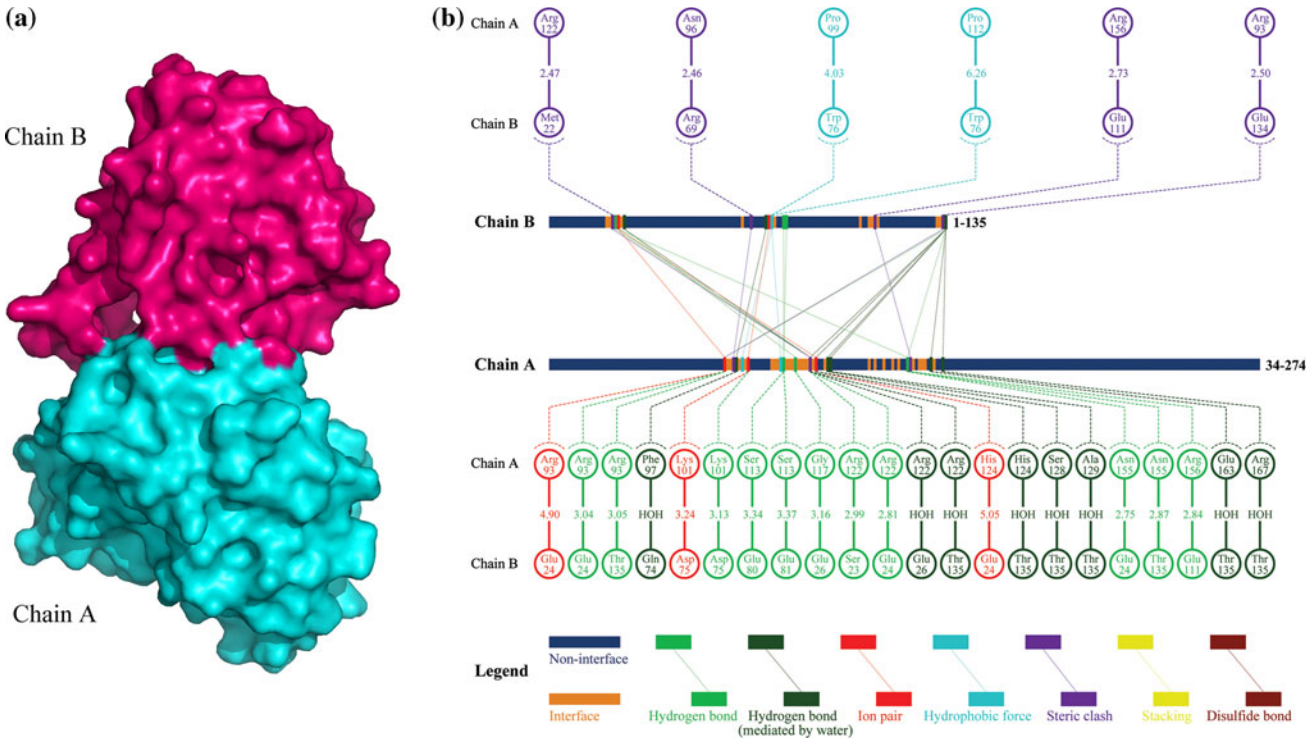

Fig. 3 **a** Stereoview of the nuclease A (chain A)–cognate inhibiter (chain B) complex (PDB entry: 2o3b). **b** Schematic representation of noncovalent interaction pattern across the binding interface of the complex [prepared using the in-house program 2D-GraLab (Zhou et al. 2009a)]

these significant residue-pairs and nonbonding properties throwing their potencies to protein–protein binding perhaps may be not only in direct manner contributing to binding affinity, but also using indirect approach to influence the allosteric effect and conformational change associated with the binding.

Comparison with previous works

Over two past decades, much effort has been addressed in predicting binding affinity of protein complexes on the basis of known structures, and the first attempt to associate binding affinities with a set of structures was due to Horton and Lewis (1992), who collected 15 experimental binding energies from the literatures and showed that these energy values could be fitted by summing contributions of the interface polar and nonpolar groups. After then, a number of new methods and tools were exploited to model and predict the binding affinities of various protein–protein complexes. The empirical scoring method was first developed to evaluate sampling accuracy of protein–protein docking algorithms (Strynadka et al. 1996); even if this strategy is highly promising for the high-throughput screening of numerous candidates within a interactome context, which still remains largely unreliable to do so (Kastritis and Bonvin 2010). With the rapid increase in high-resolution structures of proteins and their complexes with diverse ligands in recent years, knowledge-based protocol was proposed to assess the binding potency of two protein subunits using quasichemical potentials deriving from the frequency of contacts between different residues or atoms in known crystal structures (Jiang et al. 2002; Zhang et al. 2005; Su et al. 2009). However, this method is highly depended on the quality of training structures and may introduce bias for different targets. On the other hand, sophisticated ab initio approach for directly estimating free-energy magnitudes has been reported (Brandsdal and Smalås 2000; Gandhi and Mancera 2009; Cole et al. 2010). This approach is as yet computationally demanding and cannot be used for free-energy screening and binding affinity prediction of large-scale candidates.

Some of representative works are summarized in Table 4, in which only the studies relating to empirical scoring method and knowledge-based protocol, as well as the QSAR-based method present in the present work, are provided. This is because the ab initio prediction approaches can only give very limited predictions at a time and hence are difficult to derive statistically significant correlations (such as statistics r^2 and r^2_{pred}) from the predictions. As shown in Table 4, the empirical scoring methods appear to be pretty well in associating physicochemical parameters with binding affinities of protein–protein complexes. However, just as noted by Kastritis et al. (2011) that the samples adopted for training models were very limited (<30); most of the data concerned protease–inhibitor complexes, and some of them was spurious, which would largely impair the generalization ability of theoretical models. In addition, the high fitting correlations ($r^2 > 0.9$) deriving from internal training samples were not tested on an independent set, leading to potential uncertainty of these models when used to conduct external predictions. In contrast, statistics arising from knowledge-based protocols could be more reliable due to abundant structures used and independent set tested, albeit their predictive coefficients of determination seem to be quite modest ($r^2_{\mathrm{pred}} < 0.6$). In this study, we first proposed the use of QSAR approach to characterize not only the binding profile of protein partners, but also the allosteric effect associated with the binding. As can be seen, the QSAR-based models performed pretty well on both internal training set and external test set; the resulting fitting and predictive correlations using sophisticated PLS regression were $r^2 = 0.869$ and $r^2_{\mathrm{pred}} = 0.630$, respectively. Note that the predictive power of the model could be further enhanced to 0.693 if GA-variable selection was implemented. It is all coming together to show that the performance of QSAR-based strategy is comparable with or even better than that of previously engaged methods in the context of structure-based prediction of protein–protein binding affinity, even more that conformational change associated with the binding was first taken into account in prediction algorithm could give another significant advantage for the newly proposed strategy.

The prediction limit of structure-based methods

Accurate estimation of prediction limit for structure-based method is dramatically difficult, since there are many factors that could directly or indirectly influence the final predicted results of a statistical model. Generally speaking, these factors could be roughly divided into two aspects: experiment and modeling. In experimental aspect, affinity value of a complex system is always associated with measure methods used and experimental conditions adopted. For example, there are a number of techniques such as isothermal calorimetry (ITC), surface plasmon resonance (SPR) and other spectroscopic methods that could be used to assay binding affinity (Leavitt and Freire 2001; Hartmann-Petersen and Gordon 2005), and more significantly, the measure performed under different temperature, ionic strength and pH could give rise to an observable variation over assay results. The experimental errors for both measured affinity and solved structure are also ineluctable: K_{d} values are usually reported in publications with standard errors of 20–50%, equivalent to 0.1–0.25 kcal/mol for ΔG; the resolution level of complex structures deposited in the PDB

Table 4 Comparison of the modeling statistics obtained in different works for predicting protein–protein binding affinities

Authors	Method type	R^2 on training set (N)	R^2_{pred} on test set (M)
Horton and Lewis (1992)	Empirical scoring	0.92 (15)	–
Ma et al. (2002)	Empirical scoring	0.90 (20)	–
Audie and Scarlata (2007)	Empirical scoring	0.97 (24)	–
Jiang et al. (2002)	Knowledge-based	–	0.56 (28)
Zhang et al. (2005)	Knowledge-based	–	0.53 (82)
Su et al. (2009)	Knowledge-based	–	0.58 (86)
This work (PLS)	QSAR-based	0.869 (100)	0.630 (44)
This work (GA-PLS)	QSAR-based	0.815 (100)	0.693 (44)

N number of samples in training set, *M* number of samples in test set

database (Berman et al. 2000) commonly ranges from 1 to 3 Å, corresponding to atomic movement of about 0.05–0.2 Å (Fields et al. 1994; Acharya and Lloyd 2005). Besides, the difference between complex structures in crystallized (static) and dissolved (dynamic) states is an important source of errors. For modeling facet, all of the data set collected, characterization method used, and statistical tool employed can fundamentally affect final results of the modeling remarkably; redundant and limited sample set could introduce bias to built models, improper characterization method would lead to significant noises involved in independent variables, and inappropriate statistical tool is incapable of sufficiently capturing complicated dependences hidden in the investigated system.

To farthest improve the quality of a statistical model that investigators can only do is to avoid unfavorable factors relating to the modeling aspect as possible as they can. Therefore, the prediction limit for a structure-based method could be regarded as the idea condition that all unfavorable factors of modeling aspect are completely eliminated. Usually, the largest source of variation over experimental results arises from the inconsistency in experimental conditions. For example, according to a previous survey change in temperature (18–35°C) or pH (5.5–8.5) can change K_{d} by a factor of 2 or 10, respectively, corresponding to 0.3–1 logarithmic scales (Kastritis et al. 2011). And the deviations stemming from structural inaccuracy and the difference between measured states are assumed to introduce one-scale bias for pK_{d}. In this respect, experimental errors could result in at least two-scale uncertainty on the predicted pK_{d}, which corresponds to ~20% variance that cannot be explained properly for a sufficiently large data set with the member's K_{d} values varying in 10 orders of magnitude. In other words, the structure-based method is roughly estimated to be capable of explaining at most 80% variance (namely $r^2_{\mathrm{pred}} = 0.80$) of experimentally measured affinities. It should be pointed out that the 0.80 limit is an empirical estimation and hence can only be used for qualitative purpose.

Conclusions

Understanding principles of protein recognition that are pertinent to biological process is one of the long-term goals in the area of protein science. Theoretical approach to estimate the binding affinity of protein–protein interactions is an indispensable tool for protein function and design studies (Jiang et al. 2002). In this article, we report the successfully use of a new QSAR-based approach to accurately characterize the structural and noncovalent profile relating to not only the direct binding behavior between protein partners, but also the indirect allosteric effect associated with the binding, and to quantitatively predict the binding affinity of protein–protein complexes, on the basis of a recently published benchmark consisting of 144 functionally diverse, nonredundant complex samples. Here, we conclude with following remarks to close this work:

1. The QSAR-based approach is a promising alternative for the structure-based prediction of protein–protein binding affinity. This is because not only this method have shown a good performance in structural characterization, statistical modeling, and external generalization as compared to those of traditional prediction protocols, but also it allows us to carry out high-throughput screening over thousands of complex candidates within an accepted time-scale.
2. Incorporation of allosteric effect into modeling method does improve the predictive power and interpretability of the method considerably. It is suggested that the structural information regarding conformational change upon binding should not be ignored if you desire to develop a high-quality model for reliably and accurately predict the binding affinity of protein–protein complexes.
3. Although indirect allosteric effect is an important aspect of protein recognition and association, direct nonbonding interactions between protein partners seem to play a dominant role in protein binding.
4. Diverse properties contribute appreciable potencies to protein recognition. In particular, hydrophobic and

steric effects appear to be the critical facet that dominates binding process, and hydrogen bonding and electrostatic interaction confer specific judgement that further refines the exquisite structural architecture of protein complexes.

5. The bulky Ile and Leu, charged Lys, Arg and Asp, as well as polar Gln, Thr and Ser are statistically determined as the key residues since they are considered as the major contributors of, respectively, hydrophobic force, electrostatic attraction (salt bridge) and hydrogen bonding involved in the binding.
6. Owing to the ineluctable errors and heterogeneities existed in experimentally obtained data for both the structure and affinity of protein complexes, the prediction limit for a structure-based method is estimated no more than $r^2_{\text{pred}} = 0.80$.

Acknowledgments The authors would like to express their gratitude to the anonymous reviewers and the editors for their professional, intensive, and useful comments. This work was supported by the Innovation and Attracting Talents Program for College and University (‘111’ Project) (Grant No. B06023), the National Natural Science Foundation of China (Grant Nos. 11032012 and 30870608), the Key Science and Technology Program of Chongqing CSTC (Grant No. 2009AA5045), and the Sharing Fund of Chongqing university’s Large-Scale Equipment and the Visiting Scholar Foundation of Key Laboratory of Biorheological Science and Technology Under Ministry of Education in Chongqing University.

Conflict of interest The authors declare that they have no conflict of interest.

References

Acharya KR, Lloyd MD (2005) The advantages and limitations of protein crystal structures. Trends Pharm Sci 26:10–14

Audie J, Scarlata S (2007) A novel empirical free energy function that explains and predicts protein–protein binding affinities. Biophys Chem 129:198–211

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

Blobel J, Bernadó P, Svergun DI, Tauler R, Pons M (2009) Low-resolution structures of transient protein–protein complexes using small-angle X-ray scattering. J Am Chem Soc 131:4378–4386

Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. J Mol Biol 1998(280):1–9

Bondi A (1964) van der Waals volumes and radii. J Phys Chem 68:441–451

Boobbyer DNA, Goodford PJ, McWhinnie PM, Wade RC (1989) New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. J Med Chem 32:1083–1094

Brandsdal BO, Smalås AO (2000) Evaluation of protein–protein association energies by free energy perturbation calculations. Protein Eng 13:239–245

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

Clackson T, Wells JA (1995) A hot spot of binding energy in a hormone–receptor interface. Science 267:383–386

Cole DJ, Skylaris CK, Rajendra E, Venkitaraman AR, Payne MC (2010) Protein–protein interactions from linear-scaling first-principles quantum-mechanical calculations. Europhys Lett 91:37004

Conte LL, Chothia C, Janin L (1999) The atomic structure of protein–protein recognition sites. J Mol Biol 285:2177–2198

Cramer RD III, 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

Du QS, Huang RB, Chou KC (2008) Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design. Curr Protein Pept Sci 9:248–259

Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman P (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 24:1999–2012

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

Fields BA, Bartsch HH, Bartunik HD, Cordes F, Guss JM, Freeman HC (1994) Accuracy and precision in protein crystal structure analysis: two independent refinements of the structure of poplar plastocyanin at 173 K. Acta Crystallogr D 50:709–730

Ford RC (1987) Investigation of highly stable Photosystem I chlorophyll–protein complexes from the thermophilic cyanobacterium *Phormidium laminosum*. Biochim Biophys Acta 893:115–125

Fry DC (2006) Fry protein–protein interactions as targets for small molecule drug discovery. Biopolymers (Pep Sci) 84:535–552

Gandhi NS, Mancera RL (2009) Free energy calculations of glycosaminoglycan–protein interactions. Glycobiology 19:1103–1115

Golbraikh A, Tropsha A (2002) Beware of q^2! J Mol Graph Model 20:269–276

Hartmann-Petersen R, Gordon C (2005) Quantifying protein–protein interactions in the ubiquitin pathway by surface plasmon resonance. Methods Enzymol 399:164–177

Horton N, Lewis M (1992) Calculation of the free energy of association for protein complexes. Protein Sci 1:169–181

Jenssen H, Fjell CD, Cherkasov A, Hancock RE (2008) QSAR modeling and computer-aided design of antimicrobial peptides. J Pept Sci 14:110–114

Jiang L, Gao Y, Mao F, Liu Z, Lai L (2002) Potential of mean force for protein–protein interaction studies. Proteins 46:190–196

Jones S, Thornton JM (1996) Principles of protein–protein interactions. Proc Natl Acad Sci USA 93:13–20

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

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

Kini RM, Evans HJ (1996) Prediction of potential protein–protein interaction sites from amino acid sequence. Identification of a fibrin polymerization site. FEBS Lett 385:81–86

Krivov GG, Shapovalov MV, Dunbrack RL Jr (2009) Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77:778–795

Leavitt S, Freire E (2001) Direct measurement of protein binding energetics by isothermal titration calorimetry. Curr Opin Struct Biol 11:560–566

Lim JH, Bustin M, Ogryzko VV, Postnikov YV (2002) Metastable macromolecular complexes containing high mobility group nucleosome-binding chromosomal proteins in HeLa nuclei. J Biol Chem 277:20774–20782

Ma XH, Wang CX, Li CH, Chen WZ (2002) A fast empirical approach to binding free energy calculations based on protein interface information. Protein Eng 15:677–681

Madden DR (1995) The three-dimensional structure of peptide–MHC complexes. Annu Rev Immunol 13:587–622

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

Neduva V, Linding R, Su Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol 3:e405

Otlewski J, Apostoluk W (1997) Structural and energetic aspects of protein–protein recognition. Acta Biochim Pol 44:367–387

Petsalaki E, Russell RB (2008) Peptide mediated interactions in biological systems: new discoveries and applications. Curr Opin Biotechnol 19:344–350

Pripp AH, Ardö Y (2007) QSAR modeling and computer-aided design of antimicrobial peptides. 102:880–888

Pripp AH, Isaksson T, Stepaniak L, Sørhaug T, Ardo Y (2005) Quantitative structure activity relationship modelling of peptides and proteins as a tool in food science. Trends Food Sci Tech 16:484–494

Ren Y, Chen G, Hu Z, Chen X, Yan B (2008) Applying novel three-dimensional holographic vector of atomic interaction field to QSAR studies of artemisinin derivatives. QSAR Comb Sci 27:198–207

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

Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H (2007) Predicting protein–protein interactions based only on sequences information. Proc Natl Acad Sci USA 104:4337–4341

Shi L, Chen B, Wang Z, Elias DA, Mayer MU, Gorby YA, Ni S, Lower BH, Kennedy DW, Wunschel DS, Mottaz HM, Marshall MJ, Hill EA, Beliaev AS, Zachara JM, Fredrickson JK, Squier TC (2006) Isolation of a high-affinity functional protein complex between OmcA and MtrC: two outer membrane decaheme c-type cytochromes of *Shewanella oneidensis* MR-1. J Bacteriol 18:4705–4714

Song X, Zhao X (2005) The van der Waals interaction between protein molecules in an electrolyte solution. J Chem Phys 120:2005–2009

Stein A, Rueda M, Panjkovich A, Orozco M, Aloy P (2011) A systematic study of the energetics involved in structural changes upon association and connectivity in protein interaction networks. Structure 19:881–889

Stevens J, Blixt O, Tumpey TM, Taubenberger JK, Paulson JC, Wilson IA (2006) Structure and receptor specificity of the hemagglutinin from an H5N1 influenza 312:404–410

Strynadka NC, Eisenstein M, Katchalski-Katzir E, Shoichet BK, Kuntz ID, Abagyan R, Totrov M, Janin J, Cherfils J, Zimmerman F, Olson A, Duncan B, Rao M, Jackson R, Sternberg M, James MN (1996) Molecular docking programs successfully predict the binding of a β-lactamase inhibitory protein to TEM-1 β-lactamase. Nat Struct Biol 3:233–239

Su Y, Zhou A, Xia X, Li W, Sun Z (2009) Quantitative prediction of protein–protein binding affinity with a potential of mean force considering volume correction. Protein Sci 18:2550–2558

Sun J, Cai S, Yan N, Mei H (2010) Docking and 3D-QSAR studies of influenza neuraminidase inhibitors using three-dimensional holographic vector of atomic interaction field analysis. Eur J Med Chem 45:1008–1014

Tang C, Iwahara J, Clore GM (2006) Visualization of transient encounter complexes in protein–protein association. Nature 444:383–386

Tian F, Zhou P, Lv F, Song R, Li Z (2007) Three-dimensional holograph vector of atomic interaction field (3D-HoVAIF): a novel rotation-translation invariant 3D structure descriptor and its applications to peptides. J Pept Sci 13:549–566

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

Tian F, Yang L, Lv F, Luo X, Pan Y (2011) Why OppA protein can bind sequence-independent peptides? A combination of QM/MM, PB/SA, and structure-based QSAR analyses. Amino Acids 40:493–503

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

Tsai CJ, Nussinov R (1997) Hydrophobic folding units at protein–protein interfaces: Implication to protein folding and to protein–protein association. Protein Sci 6:1426–1437

Tsai CJ, Lin SL, Wolfson HJ, Nussinov R (1997) Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. Protein Sci 6:53–64

Vanhee P, Stricher F, Baeten L, Verschueren E, Lenaerts T, Serrano L, Rousseau F, Schymkowitz J (2009) Protein–peptide interactions adopt the same structural motifs as monomeric protein folds. Structure 17:1128–1136

Wlodawer A, Miller M, Jaskolski M, Sathyanarayana BK, Baldwin E, Weber IT, Selk LM, Clawson L, Schneider J, Kent SB (1989) Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. Science 245:616–621

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

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

Xu QS, Liang YZ (2001) Monte Carlo cross validation. Chemometr Intel Lab Syst 56:1–11

Xu D, Tsai CJ, Nussinov R (1997) Hydrogen bonds and salt bridges across protein–protein interfaces. Protein Eng 10:999–1012

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

Zhao C, Zhang H, Luan F, Zhang R, Liu M, Hu Z, Fan B (2007) QSAR method for prediction of protein–peptide binding affinity: application to MHC class I molecule HLA-A*0201. J Mol Graph Model 26:246–254

Zhou HX, Shan Y (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 44:336–343

Zhou P, Tian F, Li Z (2007a) Three dimensional holographic vector of atomic interaction field (3D-HoVAIF). Chemometr Intel Lab Syst 87:114–120

Zhou P, Tian F, Li Z (2007b) 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, Shang Z (2009a) 2D depiction of nonbonding interactions for protein complexes. J Comput Chem 30:940–951

Zhou P, Tian F, Lv F, Shang Z (2009b) 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

Zhou P, Zou J, Tian F, Shang Z (2009c) Fluorine bonding—how does it work in protein–ligand interactions? J Chem Inf Model 49:2344–2355