Sacred Heart University
DigitalCommons@SHU

Chemistry & Physics Faculty Publications | Chemistry and Physics

9-2007

A Novel Empirical Free Energy Function That Explains And Predicts Protein–Protein Binding Affinities

Joseph Audie
Sacred Heart University

Suzanne Scarlata
SUNY Stony Brook

Follow this and additional works at: https://digitalcommons.sacredheart.edu/chem_fac

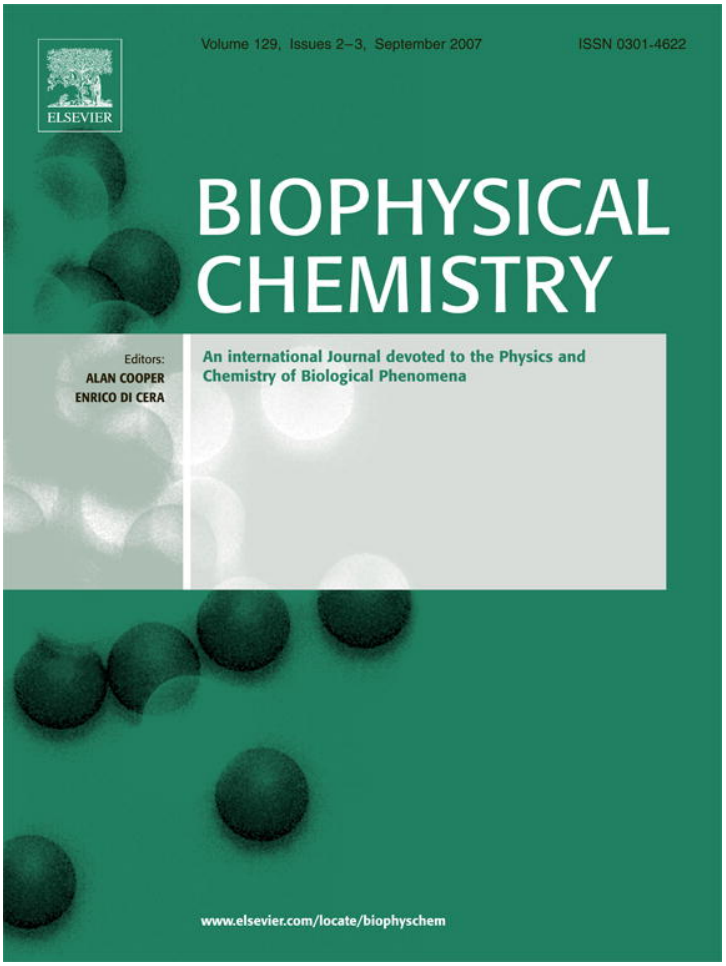
Part of the Analytical Chemistry Commons, and the Biological and Chemical Physics Commons

Recommended Citation

Audie, J., & Scarlata, S. (2007). A novel empirical free energy function that explains and predicts protein–protein binding affinities. *Biophysical Chemistry, 129*(2/3), 198-211. Doi: 10.1016/j.bpc.2007.05.021

This Peer-Reviewed Article is brought to you for free and open access by the Chemistry and Physics at DigitalCommons@SHU. It has been accepted for inclusion in Chemistry & Physics Faculty Publications by an authorized administrator of DigitalCommons@SHU. For more information, please contact ferribyp@sacredheart.edu, lysobeyb@sacredheart.edu.

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article was published in an Elsevier journal. The attached copy is furnished to the author for non-commercial research and education use, including for instruction at the author's institution, sharing with colleagues and providing to institution administration.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

ELSEVIER

Biophysical Chemistry 129 (2007) 198–211

Biophysical Chemistry

http://www.elsevier.com/locate/biophyschem

A novel empirical free energy function that explains and predicts protein–protein binding affinities

Joseph Audie, Suzanne Scarlata *

Department of Physiology and Biophysics, State University of New York at Stony Brook, Stony, Brook, NY 11794, United States

Received 19 April 2007; received in revised form 31 May 2007; accepted 31 May 2007
Available online 7 June 2007

Abstract

A free energy function can be defined as a mathematical expression that relates macroscopic free energy changes to microscopic or molecular properties. Free energy functions can be used to explain and predict the affinity of a ligand for a protein and to score and discriminate between native and non-native binding modes. However, there is a natural tension between developing a function fast enough to solve the scoring problem but rigorous enough to explain and predict binding affinities. Here, we present a novel, physics-based free energy function that is computationally inexpensive, yet explanatory and predictive. The function results from a derivation that assumes the cost of polar desolvation can be ignored and that includes a unique and implicit treatment of interfacial water-bridged interactions. The function was parameterized on an internally consistent, high quality training set giving $R^2=0.97$ and $Q^2=0.91$. We used the function to blindly and successfully predict binding affinities for a diverse test set of 31 wild-type protein–protein and protein–peptide complexes ($R^2=0.79$, rmsd = 1.2 kcal mol^{-1}). The function performed very well in direct comparison with a recently described knowledge-based potential and the function appears to be transferable. Our results indicate that our function is well suited for solving a wide range of protein/peptide design and discovery problems.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Free energy; Protein docking; Desolvation; Computational; Regression; Transfer free energies

1. Introduction

There is a need for computational methods to explain and predict free energy changes for biophysical and biochemical processes [1–6]. An important class of biophysical phenomena is that of non-covalent protein–protein interactions. Such vital activities as cellular growth, self-reproduction, and cellular communication are supported by a byzantine network of signaling cascades and metabolic pathways which rely on the coordinated and tightly regulated activities of interacting proteins, thus making protein–protein interactions attractive targets for therapeutic intervention [7,8]. The ability to estimate the free energy changes that control protein–protein associations will allow us to predict whether these interactions can occur under particular environmental conditions. There is also considerable interest in developing peptide-based drugs [9–12] and thus a concomitant need for accurate methods for estimating protein–peptide binding free energies.

Specifically, free energy functions are needed to solve three problems: (1) predicting and explaining experimentally determinable protein–protein dissociation constants; (2) predicting and explaining how different mutations affect those equilibrium constants; and (3) accurately scoring and ranking the binding poses generated by protein–protein docking algorithms [3,14,15]. Ideally, the function should also be transferable; it should work equally well for a diverse and large number of proteins. Because of the biological and clinical importance of free energy functions and the nature of the scientific challenge, a considerable amount of effort has been devoted to this research [3–6,13,16–23]. However, the development of a function to solve all three problems remains elusive. In part, this is because theoretical validity and physical meaningfulness tend to exclude computational efficiency [1,20]. In this study, we have addressed the first problem of predicting and explaining

* Corresponding author. Tel.: +1 631 444 3071; fax: +1 631 444 3432.
E-mail address: Suzanne.Scarlata@sunysb.edu (S. Scarlata).

0301-4622/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2007.05.021

Author's personal copy

experimentally determined standard state binding free energies, specifically, for complexes with a single dominant binding mode that approximates rigid-body association and with an emphasis on computational speed.

Here, we have combined a thermodynamic derivation with a linear combination of physics-based descriptors and applied regression analysis to obtain a function that accurately models the training set binding data in a statistically significant manner. This analysis helps to ensure that all important terms are included in the function and allows us to rule out random chance for any observed correlations. We also show that the magnitudes and signs of each descriptor are consistent with what is known from experiment and theory. We tested the ability of this analysis to predict binding affinities using a leave-one-out cross validation and by blind prediction on test set, with encouraging results. Moreover, our results suggest that the function is transferable. Importantly, the function's fast implementation makes it a viable candidate for addressing the scoring problem.

In summary, we describe a novel, physics-based empirical function that is fast enough to potentially solve the scoring problem yet rigorous enough to accurately predict and satisfactorily explain, within well-defined limits, wild-type binding affinities for a large and heterogeneous set of protein complexes.

2. Materials and methods

2.1. Basic thermodynamic principles and key simplifying assumptions

From the perspective of thermodynamics, protein–protein or receptor–ligand binding reactions can be described in terms of three macroscopic states [1,4],

$$R(N) + L(N) \rightarrow RL \tag{1}$$

where $R(N)$ refers to a 1 M aqueous solution of native-state receptor, with standard state free energy $G_{R(N)}$; $L(N)$ refers to a 1 M aqueous solution of native-state ligand, with standard state free energy $G_{L(N)}$; and RL refers to a 1 M aqueous solution of receptor–ligand complexes, with standard state free energy G_{RL}. The association reaction represented by Eq. (1) can be re-cast in terms of two coupled (standard state) reactions, an isomerization reaction (ΔG_{isomer}) and a binding reaction (ΔG_{bind}),

$$R(N) + L(N) \rightarrow R + L \;\; \Delta G_{\text{isomer}} \tag{1a}$$

$$R + L \rightarrow R + L \;\; \Delta G_{\text{bind}} \tag{1b}$$

where isomerization involves the free ligand and receptor native states adopting their respective "strained" binding conformations. Complex formation is then assumed to proceed between "rigid-bodies" as depicted in the binding reaction (Eq. (1b)). This model allows for the determination of the association constant, K_a, and the standard state free energy of association, $\Delta G_{\text{association}}$, in terms of ΔG_{isomer} and ΔG_{bind},

$$\begin{aligned} \Delta G_{\text{association}} &= G_{RL} - G_{R(N)} - G_{L(N)} \\ &= -RT \ln (K_a) \\ &= \Delta G_{\text{isomer}} + \Delta G_{\text{bind}}. \end{aligned} \tag{2}$$

In all that follows, we make the assumption that $\Delta G_{\text{isomer}} = \Delta H_{\text{isomer}} - T\,\Delta S_{\text{isomer}} \approx 0$ and thus, that $\Delta G_{\text{association}} \approx \Delta G_{\text{bind}}$. This simplification (i.e. rigid-body binding) is often justified for binding reactions that involve only minor conformational changes and allows us to substitute the co-crystallized coordinates of the receptor and ligand (R and L) for their unbound coordinates ($R(N)$ and $L(N)$) [6,21]. Importantly, a relatively large and diverse class of protein–protein and protein–peptide associations entails only minor conformational changes [24–26].

To rigorously predict ΔG_{bind} from the molecular properties of R, L, and RL, one needs to calculate the partition functions (Q_R, Q_L, Q_{RL}) for R, L and RL from an explicit consideration of the molecular ensembles for R, L, and RL and with an explicit solvent model. This approach, however, is too computationally demanding. A simpler method is to assume two-state binding and identify each macroscopic state with a single, experimentally determined molecular structure [1,27] and then derive a mathematical relationship between the molecular properties of each structure and ΔG_{bind} [4,13,18,19]. Finally, because the pressure–volume work contribution is negligible, we can equate the enthalpy change with energy change, $\Delta H \approx \Delta E$ [4].

2.2. Derivation of the master thermodynamic equation

The magnitude and sign for ΔG_{bind} arises from changes in the entropy and energy of binding. We assume that the change in energy can be explicitly estimated from the crystal coordinates of R, L, and RL, while the change in entropy ($\sim$the ensemble sizes for R, L, and RL) can be estimated implicitly from the same coordinate set.

It is helpful to conceptualize the calculation of ΔG_{bind} as the sum of two coupled free changes: (1) a desolvation free energy change (ΔG_{desolv}) that arises from the total elimination of receptor and ligand–water contacts and the formation of new water–water contacts; and (2) a contact free energy change ($\Delta G_{\text{contact}}$) that arises from the formation of new receptor–ligand, receptor–water–ligand, and water–water contacts following association. Thus we have,

$$\Delta G_{\text{bind}} = \Delta G_{\text{desolv}} + \Delta G_{\text{contact}} \tag{3}$$

where

$$\begin{aligned} \Delta G_{\text{desolv}} = {} & \Delta G_{\text{desolv,charge}} + \Delta G_{\text{desolv,polar}} \\ & + \Delta E_{\text{desolv,hydrophobic}} - T\Delta S_{\text{desolv,hydrophobic}} \end{aligned} \tag{4}$$

and

$$\begin{aligned} \Delta G_{\text{contact}} = {} & \Delta G^{\text{r-l}}_{\text{charge–charge}} + \Delta G^{\text{r-l}}_{\text{charge–polar}} + \Delta G^{\text{r-l}}_{\text{polar–polar}} \\ & + \Delta G^{\text{r-l}}_{\text{qm}} + \Delta G^{\text{r-l}}_{\text{vdw}} + \Delta G^{\text{r–w–l}} + \Delta E_{\text{conf}} \\ & - T\Delta S_{\text{conf}} + \Delta G_{\text{trans–rot}}. \end{aligned} \tag{5}$$

Author's personal copy

The first two terms in Eq. (4) refer to the free energy penalties of completely removing charged and polar groups from contact with water. Both terms can be further decomposed into the free energy cost of breaking electrostatic interactions, Van der Waals (vdw) interactions, and hydrogen bonding interactions with the solvent. The last two terms in Eq. (4) give the energetic and entropic changes associated with removing hydrophobic groups from contact with water. The first three terms in Eq. (5) give the electrostatic component to the free energy of forming qm charge–charge, charge–polar and polar–polar contacts at the receptor–ligand interface. $\Delta G^{\mathrm{r-l}}_{\mathrm{qm}}$ is the free energy of forming the quantum mechanical (QM) or covalent component of receptor–vdw ligand hydrogen bonds [28], and $\Delta G^{\mathrm{r-l}}_{\mathrm{vdw}}$ is the free energy of forming receptor–ligand vdw interactions. The final four terms in Eq. (5) refer to the free energy of forming water-bridged receptor–ligand contacts, the change in conformational energy, the change in conformational entropy, and the change in translational–rotational free energy, respectively.

Rigid-body binding implies that changes in conformational energy can be neglected [2,4,6,13,16,17,21]. Assuming a clash free interface, it is reasonable to assume that the receptor–ligand vdw interactions formed at the interface offsets the binding induced loss of favorable receptor–water and ligand–water vdw interactions [6,29]. Analogously, we assume that the receptor–ligand charge–polar and polar–polar electrostatic interactions formed at the interface help compensate for $\Delta G_{\mathrm{desolv,polar}}$ [30]. Finally, the change in the rotational–translational free energy ($\Delta G_{\mathrm{trans-rot}}$) is assumed to be ≈ 0 [31]. Stated mathematically we make the simplifying assumption that,

$$\begin{aligned} &\Delta E_{\mathrm{conf}} + \Delta G^{\mathrm{r-l}}_{\mathrm{charge-polar}} + \Delta G^{\mathrm{r-l}}_{\mathrm{polar-polar}} + \Delta G^{\mathrm{r-l}}_{\mathrm{vdw}} \\ &\quad + \Delta G_{\mathrm{desolv,polar}} + \Delta G_{\mathrm{desolv,charged,vdw}} \\ &\quad + \Delta E_{\mathrm{desolv,hydrophobic}} + \Delta G_{\mathrm{trans-rot}} \approx 0 \end{aligned} \tag{6}$$

where only the vdw contribution is included in the case of charge group desolvation and the electrostatic and hydrogen bonding (QM) contributions remain uncompensated. Combining Eqs. (3), (4), (5) and (6), we have

$$\begin{aligned} \Delta G_{\mathrm{bind}} = {} & \Delta G^{\mathrm{r-l}}_{\mathrm{charge-charge}} + \Delta G^{\mathrm{r-l}}_{\mathrm{qm}} + \Delta G^{\mathrm{r-w-l}} - T\Delta S_{\mathrm{conf}} \\ & + \Delta G_{\mathrm{desolv,charge}} - T\Delta S_{\mathrm{hydrophobic}}. \end{aligned} \tag{7}$$

Eq. (7) serves as the master thermodynamic equation and implies that the binding affinity is a function of interchain charge–charge contacts, the covalent or quantum-mechanical contribution of hydrogen bonding interactions, the free energy of interfacial water mediated receptor–ligand interactions, the change in conformational entropy and the free energy of charge and hydrophobic group desolvation.

2.3. Identifying each term in the master equation with a physical descriptor

To write ΔG_{bind} in terms of *R*, *L* and *RL*, each of the terms in Eq. (7) must be explicitly associated with a molecular property or physical descriptor characteristic of *R*, *L*, and *RL*. The simplest approach is to assume a linear relationship between each thermodynamic term and its corresponding physical descriptors. The equations describing the relationship between each term in Eq. (7) and its hypothesized *a priori* plausible descriptor are given below:

$$\Delta G^{\mathrm{r-l}}_{\mathrm{charge-charge}} + \Delta G^{\mathrm{r-l}}_{\mathrm{qm}} = \alpha_{\mathrm{sb}} X_{\mathrm{sb}} + \alpha_{\mathrm{hb}} X_{\mathrm{hb}} \tag{8}$$

$$\begin{aligned} &\Delta G_{\mathrm{desolv,\,charge}} - T\Delta S_{\mathrm{hydrophobic}} \\ &\quad = \alpha_{+/-}\left(X_{RL,+/-} - X_{R,+/-} - X_{L,+/-}\right) \\ &\qquad + \alpha_{\mathrm{s/c}}\left(X_{RL,\mathrm{c/s}} - X_{R,\mathrm{c/s}} - X_{L,\mathrm{c/s}}\right) \end{aligned} \tag{9}$$

$$-T\Delta S_{\mathrm{conf}} = \alpha_{\mathrm{tor}}\left(X_{RL,\mathrm{tor}} - X_{R,\mathrm{tor}} - X_{L,\mathrm{tor}}\right) \tag{10}$$

$$\Delta G^{\mathrm{r-w-l}} = \alpha_{\mathrm{gap}} X_{\mathrm{gap}}. \tag{11}$$

In Eq. (8), X_{sb} refers to the total number of salt bridges, defined as the difference between the total number of favorable and non-favorable electrostatic contacts at the interface and X_{hb} refers to the total number of hydrogen bonds. Favorable contacts include oppositely charged atoms separated by 4 or less; unfavorable contacts include like charges separated by 4 or less. The nitrogens of Lys and Arg side chains were assumed to carry positive charges, while the oxygens of Glu and Asp chains were assumed to carry negative charges. Likewise, N-terminal nitrogens, carboxyl oxygens and phosphate oxygens were all assumed to carry positive and negative charges, respectively. Histidine residues were assumed to be neutral. All hydrogen bonds were calculated according to the chemical, angle and distance criteria of Arthur Lesk [32].

In Eq. (9), each $X_{i,+/-}$ refers to the total number of solvent exposed charged groups, where i=*RL*, *R* or *L*. A group is counted as exposed if its solvent accessible surface area (SASA) is >1.0 ^{2}. Similarly, each $X_{i,\ \mathrm{c/s}}$ refers to the total number of exposed (SASA>1.0 ^{2}) hydrophobic groups (carbons or sulfurs).

The $X_{i,\mathrm{tor}}$ terms in Eq. (10) refer to the total number of exposed side-chain torsions including main-chain torsions for peptide ligands. All torsions associated with a side chain (or main chain in the case of peptides) possessing a relative SASA>60% are counted as exposed. All SASA calculations were made using the program NACCESS [33].

X_{gap}, (Eq. (11)) refers to the gap volume at the interface and was calculated using SURFNET [34]. In this calculation, a trial gap sphere is placed midway between the *vdw* surfaces of two neighboring atoms, and if any neighboring atoms penetrate the trial sphere, the radius is reduced until it just touches the penetrating atom. If the sphere radius falls below some pre-set radius the gap sphere is rejected, otherwise the final gap sphere is saved. The procedure is repeated for all atom pairs and the gap region is filled with spheres. Here, we used the default SURFNET values (minimum radius for a gap sphere=1.0 ; maximum radius=5.0).

Author's personal copy

2.4. Using regression to parameterize the function

Each α_i in Eqs. (8)–(11) is a proportionality constant that was determined by multiple linear regression. Regression assigns optimal α_i's to achieve a least-squares fit between a dependent variable (experimentally determined binding affinties) and a linear combination of independent variables (Eqs. (8)–(11)). Any disagreement between the estimated and experimental binding affinities is attributable to random error (μ) which is assumed to be normally distributed with a constant variance, σ^2. The errors are further assumed to be independent of each other and each estimator X variable. Hence, multiple regression treats the α_i's as estimators of the true population coefficients (β_i's), thus providing a basis for hypothesis testing and statistical inference. It is also assumed that no significant linear correlation exists between the independent variables (multi-colinearity).

2.5. Constructing the training and test sets

Regression was used on a training set of 24 protein–protein and protein–peptide complexes. All 24 complexes were imported from the protein data bank [35] using the molecular modeling package Deep View [36] (http://www.expasy.org/spdbv/). In constructing our training set, we emphasized structural quality and consistency with the underlying assumptions of our method. Hence, we first searched the literature for complexes that passed a screen of 10 explicit physicochemical and structural criteria:

1. structure solved using X-ray crystallography
2. resolution <2.4
3. r-factor <0.25
4. satisfactory What Check report (implemented at PDB)
5. high quality protein–protein interface (assessed visually using DEEP View)
6. ligand free protein–protein interface
7. availability of high quality experimental binding data
8. single, dominant binding mode
9. agreement with our simple charge model
10. rigid-body binding.

As a final quality control, we manually inspected the header files of candidate structures to search for problems not disclosed by the 10 checks described above. Ultimately, we settled on 24 protein–protein complexes.

All 24 training set complexes seem to have a single binding mode and seem to satisfy the rigid-body body approximation. Nine training set structures are known to involve binding associated Cα conformational changes, for both binding partners, of ≤ 0.9 rmsd and thus approximate rigid-body binding [37,38]. Moreover, 19 of the 24 structures have been successfully utilized in past docking and modeling studies that assumed rigid-body association and a single binding mode [6,13,16–19,21,24,26,39]. This set of 19 structures was supplemented by 4 structures that involve binding associated Cα conformational changes of ≤ 0.9 rmsd for at least one binding partner. A final complex (1 DHK) was also included, despite the fact that the bound receptor differs from the unbound receptor by ≈ 1.28 rmsd. The inclusion of 1DHK was deemed a necessary compromise to enhance the diversity of the training set. Agreement with our simple charge model is suggested by the fact that 19 of the 24 complexes have been used in the past modeling work referenced above. Moreover, agreement with our charge model was confirmed through use of the PROPKA pK_a prediction tool (propka.chem.uiowa.edu) [40]. The PDB and CATH [41] codes for all 24 training set complexes are summarized in Table 1.

In constructing the test set the emphasis was on diversifying it with respect to the training set and constructing one large enough to be useful. Thus, we were sometimes forced into relaxing our selection criteria. For example, we included structures solved using NMR and complexes that might undergo binding induced conformational changes that are slightly >0.9 rmsd. However, very large conformational changes and serious structural problems, like missing interface residues or problematic charge assignment, resulted in the elimination of many candidate complexes. These compromises were deemed necessary to construct an adequate test set. The PDB codes for all 35 test set complexes are summarized in Table 3.

Using multiple linear regression, the final equation for predicting and explaining binding affinities was obtained:

$$\begin{aligned}\Delta G_{\text{bind}} = &-0.85\Delta X_{+/-} + 0.067\Delta X_{\text{c/s}} - 0.66X_{\text{sb}} \\ &- 0.90X_{\text{hb}} - 0.00087X_{\text{gap}} - 0.091\Delta X_{\text{tor}} \\ &- 0.54 \end{aligned} \quad (12)$$

where Δ refers to the differences between the unbound and bound states.

2.6. Modifying Eq. (12) to facilitate the analysis of peptide transfer free energies

In the Results and Discussion sections we show the relationship between free energy predictions made using Eq. (12) and free energies obtained from experiment. One test compares side-chain transfer free energies calculated using Eq. (12) with experimentally determined side-chain (X) transfer values from water (W) to octanol (O) obtained from a series of 8 host-guest (ac-Ala-X-Ala-t-butyl) tripeptides and 17 host-guest pentapeptides (Ac-WL-X-LL).

A complicated multi-term equation is probably required to explain the free energy of transfer ($\Delta G_{X,\text{ w–oct}}$) for the peptides. Progress can be made, however, by recognizing that the charge and hydrophobic desolvation term of Eq. (12) could serve as key terms in a more complete expression for estimating $\Delta G_{X,\text{w–oct}}$,

$$\begin{aligned}\Delta G_{X,\text{w–oct}} &= \Delta G_{\text{desolv,charge}} - T\Delta S_{\text{hydophobic}} + \Delta G_{\text{other}} \\ &= -0.85X_{L,+/-} + 0.067X_{L,\text{c/s}} + f(X) \end{aligned} \quad (13)$$

where the first term in Eq. (13) estimates the free energy cost of rupturing contacts between water and charged groups and the second term quantifies the favorable entropy changes that accompany hydrophobic group transfer; $f(X)$ refers to any remaining free energy contributions that are not explicitly accounted for this term is dropped when making actual calculations.

Author's personal copy

Table 1
Training set PDB and CATH codes, complex types, crystallographic resolution, experimental and predicted binding affinities

PDB	Complex type	Resolution ( )	CATH (Rec)	CATH (Lig)	Rigid body[a]	$\Delta G_{bind,exp}$	ΔG_{bind}
1brs	Barnase/barstar	2.0	3.10.450.30	3.30.370.10	1a2p (0.49) 1a19 (0.44)	−17.3	−17.8
1cho	Chymotrypsi/ovomucoid	1.8	2.40.10.10	3.30.60.30	5cha (0.49) 2ovo (0.79)	−14.4	−13.2
1cse	Subtilisin/eglin C	1.2	3.40.50.200	3.30.10.10	1scd (0.36) 1acb (0.64)	−13.1	−14.0
1ppf	Elastase/ovomucoid	1.8	2.40.10.10	3.30.60.30		−13.5	−13.9
1tec	Thermitase/eglin C	2.2	3.40.50.200	3.30.10.10		−14	−14.0
2ptc	Trypsin/PTI	1.9	2.40.10.10	4.10.410.10	2ptn (0.34) 6pti (0.36)	−18.1	−17.4
2sec	Proteinase/inhibitor	1.8	3.40.50.200	3.30.10.10		−14	−13.8
2sic	Subtilisin/SSI	1.8	3.40.50.200	3.30.350.10	1sup (0.25) 3ssi (0.63)	−12.7	−13.1
2sni	Subtilisin/C12	2.1	3.40.50.200	3.30.10.10	1sup (0.26) 2ci2 (0.46)	−15.8	−15.4
3cpa	Carboxypeptidase A/Gly–Tyr	2.0	3.40.630.10	Peptide		−5.3	−4.9
3sgb	Protease B/ovomucoid	1.8	2.40.10.10	3.30.60.30		−12.7	−12.6
3tpi	Trypsinoge/PTI	1.9	2.40.10.10	4.10.410.10		−17.3	−18.2
4sgb	Protease B/PCI-1	2.1	2.40.10.10	3.30.60.30		−11.7	−12.6
4tpi	Trypsinogen/inhibitor	2.2	2.40.10.10	4.10.410.10		−17.7	−17.0
1vfb	Fv D1.3/lysozyme	1.8	2.60.40.10	1.10.530.10	1vfa (0.47) 8lyz (0.51)	−11.5	−12.0
1yqv	Fab HyHel5/lysozyme	1.7	2.60.40.10	1.10.530.10		−14.5	−13.4
2tpi	Trypsinogen/Ile–Val	2.1	2.40.10.10	Peptide		−5.8	−6.2
2tgp	Trypsinogen/BPTI	1.9	2.40.10.10	4.10.410.10		−17.8	−17.5
1tpa	Trypsin/BPTI	1.9	2.40.10.10	4.10.410.10		−17.8	−17.4
2pcc	Cytochrome peroxidase/cyt C	2.3	1.10.520.10*	1.10.760.10	1cca (0.35) 1ycc (0.49)	−7	−6.3
1acb	Chymotrypsin/eglin C	2.0	2.40.10.10	3.30.10.10	5cha (0.69)	−13.1	−13.5
1st f	Papin/stefin B	2.4	3.90.70.10	3.10.450	1cse (0.63) 1ppn (0.32)	−13.5	−13.4
1ycs	p53/53BP2	2.2	2.60.40.720	2.30.30.40*	2ioo (0.69)	−10.3	−11.3
1dhk	Alpha-amylase/inhibitor	1.9	3.20.20.80	2.60.120.200	1pif (1.28)	−14.3	−14.5

[a]Status of the rigid-body approximation: The first 19 structures have been employed successfully in the past modeling work that assumed rigid-body association and are thus assumed to satisfy the rigid-body approximation. By searching the literature and PDB we were sometimes able to locate unbound receptor and ligand coordinates [24,26]. In these cases, we calculated the rmsd between the unbound structures and the bound ones. It is assumed that the rigid-body approximation is satisfied for Cα rmsd ≤ 0.9 . The unbound receptor PDB identifier is listed first and the ligand identifier is listed second; calculated rmsd values are given in parentheses. We were not always able to find unbound receptor and ligand structures. Despite an rmsd of ≈ 1.28, 1DHK was included in the training set and is assumed to satisfy the rigid-body approximation (see text). The CATH codes are for the unbound receptor and ligand, respectively and are from (http://www.cathdb.info/latest/index.html). An * indicates a structure with more than one CATH code. The binding affinity data was collected from multiple sources [6,13,16,17,19,39]. All binding affinities are in kcal mol^{-1}.

It is important to recall that Eq. (13) is only meant to estimate the charge desolvation and hydrophobic contributions to transfer. As such, it is unreasonable to assume that we can use it to fully account for transfer free energies. However, a good correlation between experimentally determined values for $\Delta G_{X,w-oct}$ and the values estimated using Eq. (13) would suggest that Eq. (13) does capture the essential physics of charge–water and hydrophobic–water interactions. Thus, by comparing predictions made using Eq. (13) with experimental side-chain transfer free energies we can test the functional form and coefficients of Eq. (13). This is especially important given the assumptions that lead to Eq. (13), in particular the assumption that polar–polar and charge–polar interactions help offset the free energy penalty of polar group desolvation, that the net polar desolvation penalty is negligible.

In practice we used Eq. (13) to calculate the free energy of transferring a given tripeptide or pentapeptide form water to octanol. In making the calculations we assumed all peptide atoms had an aqueous phase SASA > 1.0 and an octanol phase SASA < 1.0. Because $X_{L,+/-}$ and $X_{L,c/s}$ are calculated as negative (the change in the number of solvent exposed charged or hydrophobic atoms, respectively), the charged term makes an unfavorable contribution to transfer while the hydrophobic term makes a favorable one.

The contribution of individual side chains can also be calculated,

$$\Delta\Delta G_{X,w-oct} = \Delta G_{X,w-oct} + \Delta G_{ref,w-oct} \quad (14)$$

where the reference amino acid (ref) is Gly in the case of the tripeptides and Ala in the case of the pentapeptides. The predicted $\Delta\Delta G_{X,w-oct}$ values calculated using Eqs. (13) and (14) can now be meaningfully compared with experimentally determined side-chain free energies of transfer.

Author's personal copy

2.7. Statistical analysis

All statistical analyses were performed using Microsoft Excel and Sagata Regression Professional (http://www.sagata.com/). Additional details on the statistical methods employed in this study can be found elsewhere [42].

3. Results

3.1. Training set and statistical analysis of the regression model

Eq. (12) was obtained by regression analysis on a training set of 24 protein–protein complexes. Information regarding all 24 complexes is summarized in Table 1. The descriptor values for all 24 complexes are presented in Table 2. The training set consists of very high quality structures and is conspicuous for the large number of protease-inhibitor complexes. There is good reason to believe that each of the 24 complexes approximates rigid-body binding, exhibits a single binding mode and is consistent with our default charge model (see Materials and methods). The data presented in Tables 1 and 2 clearly indicate large variations in descriptor values and binding free energies across all 24 complexes.

The results of the regression analysis are summarized in Fig. 1, Tables 5–7. Fig. 1 displays the fit between the regression

Table 2
Training set descriptor values

PDB[a]	$\Delta X^{b}_{+/-}$	$\Delta X^{c}_{c/s}$	X^{d}_{sb}	X^{e}_{hb}	X^{f}_{gap}	ΔX^{g}_{rot}
1brs	−3	−58	11	10	2662	−29
1cho	0	−53	0	9	3780	−25
1cse	−5	−44	0	15	3801	−22
1ppf	−1	−53	0	9	5219	−21
1tec	−5	−56	0	14	3528	−18
2ptc	−3	−56	2	15	3133	−20
2sec	−5	−49	0	14	4015	−20
2sic	−3	−66	0	11	3580	−25
2sni	0	−60	0	11	3863	−26
3cpa	−4	−25	2	6	168	−8
3sgb	0	−41	0	9	3053	−15
3tpi	−3	−52	2	16	3220	−19
4sgb	0	−44	0	8	3362	−11
4tpi	−5	−56	3	16	3145	−23
1vfb	−4	−30	0	13	3644	−21
1yqv	−8	−39	6	14	3720	−29
2tpi	−2	−26	3	5	62	−10
2tgp	−2	−57	2	14	3231	−20
2pcc	0	−16	2	2	5165	−32
1acb	−1	−46	0	10	4753	−26
1tpa	−1	−56	2	13	3184	−20
1stf	0	−64	0	8	4003	−23
1ycs	−3	−34	6	9	3300	−43
1dhk	−9	−106	−3	15	7149	−34

[a]PDB codes for all 24 training set complexes. [b]Change in the number of solvent exposed charged groups following complex formation. [c]Change in the number of solvent exposed hydrophobic groups (carbons and sulfurs). [d]Total number of interfacial salt bridges and [e] hydrogen bonds. [f]Interface gap volume. [g]Change in the number of solvent exposed side-chain (and main-chain torsions in the case of peptide ligands) torsions following association. A linear combination of all six descriptors, weighted by linear regression on the 24 member training set, gives Eq. (12). See text for additional details.

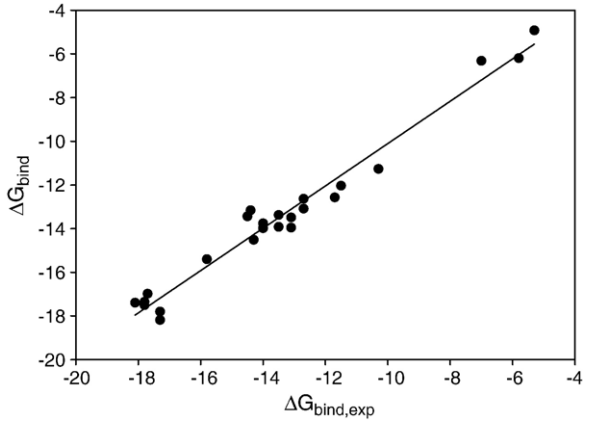


Fig. 1. Training set binding affinities versus predicted affinities. Experimental training set binding affinities ($\Delta G_{bind,exp}$) versus binding affinities (ΔG_{bind}) estimated using Eq. (12) (kcal mol^{-1}); see Tables 1 and 2 for data on all 24 training set complexes; R^2=0.97; rmsd=0.62 kcal mol^{-1}; Q^2=0.91.

estimates of Eq. (12) and the experimental binding affinities of the training set complexes. Table 5 summarizes the statistics (R=0.98, R^2=0.97, R^2-adj=0.96, rmsd=0.62 kcal mol^{-1}) we calculated to assess the goodness-of-fit between Eq. (12) and the training set binding data and the overall statistical significance (f-test, $p<0.05$) of the model. Table 5 also includes the results of the leave-one-out cross validation (Q^2=0.91) we performed as a quantitative first test of the equations goodness-of-prediction. Table 6 summarizes the sensitivity analysis we performed to assess each terms contribution to the observed fit with the experimental data. Table 7 gives the comprehensive regression diagnostics performed to evaluate the statistical legitimacy of each coefficient in Eq. (12).

3.2. Evaluating the descriptors

Each of the physical descriptors in Eq. (12) has been identified with a specific thermodynamic quantity, as

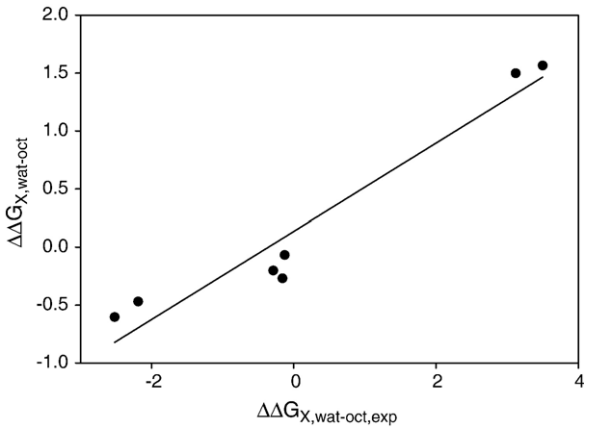


Fig. 2. Predicted versus experimental tripeptide side-chain transfer free energies. Predicted water–octanol ($\Delta\Delta G_{X,w-oct}$) versus experimental water–octanol ($\Delta\Delta G_{X,w-oct,exp}$) side-chain transfer free energies (kcal mol^{-1}) for seven host-guest (^{XO}) tripeptides. Eqs. (12) and (13) were used to make the predictions. R^2=0.93, R=0.97, rmsd=1.36 kcal mol^{-1}. The experimental data is from Kim and Szoka [51].

Author's personal copy

indicated in Eqs. (8)–(11): the formation of salt bridges, hydrogen bonds and water-bridged interactions at the receptor–ligand interface, the burial of hydrophobic and charged groups, and the burial and immobilization of side-chain torsions. Each of these phenomena is thought to be important to complex formation [25,43,44]. Nevertheless, we further evaluated the physics of each a priori association by comparing our results to those obtained from previous modeling and experimental studies.

Table 8 summarizes and quantitatively compares the regression-optimized coefficient magnitudes obtained for Eq. (12) with those from theory and experiment, indicating good agreement between the two. To test the energetic and entropic contributions of hydrophobic and charged group desolvation implied by Eq. (12), we used Eq. (12) to predict water–octanol side-chain transfer free energies for 8 host-guest tripeptides and 17 pentapeptides and compared the predictions with experimental water–octanol transfer free energies. The results of this analysis are presented in Figs. 2 and 3. Our predictions are in good agreement with experiment, further suggesting the physical accuracy of our regression equation.

3.3. Testing the predictive power of Eq. (12)

The high Q^2 value (0.91) presented in Table 5 suggests that Eq. (12) can be used to accurately predict binding affinities for complexes excluded from the training set. To directly evaluate this possibility a test set of 35 diverse complexes was constructed and Eq. (12) was used to make blind binding affinity predictions for the entire test set. Information regarding all 35 protein–protein structures is provided in Tables 3 and 4. In stark contrast with the training set, the test set is conspicuous for its lack of protease-inhibitors, the large number of peptide ligands, the presence of non-proteineous atoms at several complex interfaces (9 phosphorylated-tyrosine residues, pTyr), and the use of lower resolution and NMR structures. It is interesting to note that our test set is considerably larger and more heterogeneous than our training set. The receptor–ligand interactions implied by the CATH codes indicate minimal overlap between the training and test sets. As with the training set, there are large variations in descriptor values and binding free energies across the 35 member test set. With a few exceptions there is good reason to think that the test set complexes satisfy the main assumptions of our method. Experimental and predicted binding affinities for the 35 complexes are summarized in Table 3. A best-fit-regression line between predicted and experimentally determined binding affinities (excluding 4 outliers, $R^2=0.79$, $R=0.89$, rmsd=1.2 kcal mol^{-1}) is given in Fig. 4.

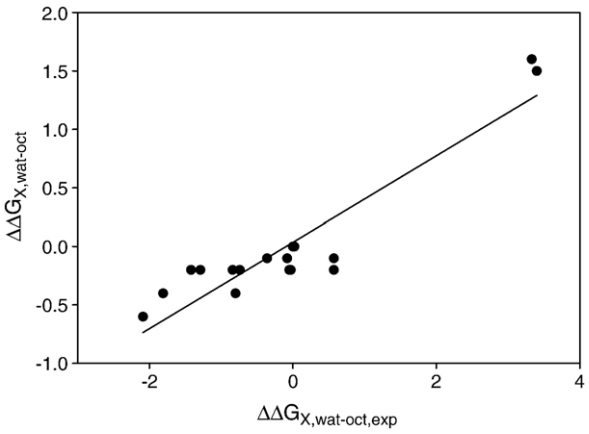


Fig. 3. Predicted versus experimental pentapeptide side-chain transfer free energies. Predicted water–octanol ($\Delta\Delta G_{X,\mathrm{w-oct}}$) versus experimental water–octanol ($\Delta\Delta G_{X,\mathrm{w-oct,exp}}$) side-chain transfer free energies (kcal mol^{-1}) for 17 host-guest (X) tripeptides. Eqs. (12) and (13) were used to make the predictions. $R^2=0.86$, $R=0.93$, rmsd=0.97 kcal mol^{-1}. The experimental data is from Wimley et al. [52]. Arg and Lys were omitted from analysis because salt bridge formation accompanies transfer; Gly was also excluded, as its transfer free energy is problematic [52].

3.4. Direct comparisons with previous work

Ma et al. [19] derived a regression equation to estimate binding affinities for protein–protein and protein–peptide complexes from static complex structures. Their empirical equation was derived for use as a scoring function for the protein–protein docking problem. The direct relevance of the Ma et al. method to our work is obvious and thus a direct comparison was deemed important. A comparison with Eq. (12) in terms of training set size (20/24), number of descriptors (3/6), R^2-adj (0.96/0.90), rmsd (0.66/1.08 kcal mol^{-1}), and Q^2 (0.91/0.88) is provided in Table 9. Ma and co-workers did not include blind prediction on a test set in their study.

Dcomplex is a web-accessible, state-of-the-art, knowledge-based function for estimating binding affinities that was trained and systematically tested on a large number of protein–ligand, protein–protein and protein–nucleic acid complexes [5]. We decided to compare the predicted binding affinities we obtained using Eq. (12) with predictions made using Dcomplex. The 35 member test set described above, minus the 4 outliers we failed to accurately predict, served as the comparison test set. The results of the 31 member test comparison are summarized in Fig. 5 (Eq. (12): $R^2=0.79$, $R=0.89$, rmsd=0.6 kcal·mol^{-1}/Dcomplex: $R^2=0.53$, $R=0.73$, rmsd=2.6 kcal mol^{-1}).

4. Discussion

Our goal was to derive and validate an empirical free energy function to explain and predict protein–protein binding affinities, subject to certain constraints. To accomplish this, we formulated a master thermodynamic equation (Eq. (7)), associated each thermodynamic term in the equation with a linear combination of *a priori* plausible physical descriptors, and used multiple linear regression on a 24 member training set to optimize the proportionality constants. The major result of this paper is Eq. (12); a regression equation that estimates binding affinities in terms of receptor–ligand salt bridges, the quantum-mechanical contribution of hydrogen bonds, and water mediated interactions, respectively, and charge and hydrophobic group desolvation and interfacial torsion immobilization.

Author's personal copy

Table 3
Test set PDB codes, complex types, experimental and predicted binding affinities

PDB[a]	Complex type	Resolution (Å)	CATH (Rec)	CATH (Lig)	Rigid body[a]	$\Delta G_{bind,exp}$	ΔG_{bind}
1a0o	Che Y/ChA	3.0	3.40.50.230	3.30.70.400	1ehc (0.72) 1fwp (1.8)	−8.1	−7.6
1hbs	Deoxyhemoglobin	3.0	1.10.490.10	1.10.490.10	[18]	−4.8	−6.5
2iff	Antibody/lysozyme	2.7	2.60.40.10	1.10.530.10	[16]	−11.1	−9.6
1avz	V-1 Nef/Fyn Sh3	3.0	3.30.62.10	2.30.30.40	1avv (0.68) 1shf (0.42)	−6.4	−8.7
1hhg	MHC/peptide	2.6	3.30.500.10	Peptide	[20]	−8.9	−10.4
1hhh	MHC/peptide	3.0	3.30.500.10	Peptide	[20]	−11.6	−10.5
1hhi	MHC/peptide	2.5	3.30.500.10	Peptide	[20]	−11.2	−9.5
1hhk	MHC/peptide	2.5	3.30.500.10	Peptide	[20]	−10.9	−9.4
1lcj	Sh2/pTYR	1.8	3.30.500.10	Peptide	[22]	−8.0	−9.4
1lck	Sh3–Sh2/pTYR	2.5	3.30.500.10	Peptide	[22]	−7.0	−7.4
1lkl	Sh2/pTYR	1.8	3.30.500.10	Peptide	[22]	−8.0	−7.7
1sps	Sh2/pTYR	2.7	3.30.500.10	Peptide	[22]	−9.1	−9.6
1tce	Sh2/pTYR	nmr	3.30.500.10	Peptide	[22]	−5.9	−5.4
2pld	Sh2/pTYR	nmr	3.30.500.10	Peptide	[22]	−9.1	−9.4
1irs	IRS-1 PTB/pTYR	nmr	2.30.29.30	Peptide	[22]	−7.2	−7.6
1shc	PTB/pTYR	nmr	2.30.29.30	Peptide	[22]	−10.0	−7.9
1sha	Sh2/pTYR	1.5	3.30.505.10	Peptide	[22]	−7.6	−7.1
1dkz	DNAK/peptide	2.0	2.60.34.10*	Peptide	[53]	−9.1	−9.5
2er6	Endothiapepsin/peptide	2.0	2.40.70.10	Peptide	1oew (0.6)	−9.8	−10.2
1bbz	Sh3/peptide	1.7	2.30.30.40	Peptide	[54]	−7.7	−7.0
1ebp[c]	EPO receptor/peptide	2.8	2.60.40.30	Peptide	[25] 1ern (2.14)	−11.7	−9.9
1nsn	Antibody/nuclease	2.8	2.60.40.10	2.40.50.90	[25] 1kdc (0.74)	−11.8	−13.5
1jhl	Antigen/antibody	2.4	2.60.40.10	1.10.530.10	[25] 1ghl (0.51)	−11.8	−9.6
1mda	Dehydrogenase/amicyacin	2.5	2.130.10.10*	2.60.40.420	?	−7.3	−6.1
3hfm	Hy/HEL-10/FAB-lysozyme	3.0	2.60.40.10	1.10.530.10	[25]	−13.3	−14.9
1bth	Thrombin/BPTI	2.3	2.40.10.10	4.10.410.10	2htn (poor fit), 6pti (0.48)?	−16.5	−16.8
2kai	Kallikrein A/BPTI	2.5	2.40.10.10	4.10.410.10	2pka (0.53) 6pti (0.46)	−12.5	−13.1
2jel	Jel42 FAB/Hpr	2.5	2.60.40.10	3.30.1340.10	[24] 1poh (0.58)	−11.5	−12.5
1gla	IIIGLC/glycerol kinase	2.6	3.30.420.40	2.70.70.10	[25] 1f3g (0.42)	−7.1	−8.4
1mel	VH antibody/lysozyme	2.5	2.60.40.10	1.10.530.10	[25] 1lza (0.67)	−10.5	−10.2
1bql	Anti-HEL FAB/lysozyme	2.6	2.60.40.10	1.10.530.10	1dkj (0.84)	−14.5	−13.4
1nmb[b]	NC10/neuraminade	2.2	2.60.40.10	2.120.10.10	[25] 7nn9 (0.28)	−10.0	−14.7
1avw[b]	Trypsin/STI	1.8	2.40.10.10	2.80.10.50	2ptn (0.39) 1ba7 (0.47)	−12.3	−21.1
1wej[b]	E8 antibody/cytochrome *C*	1.8	2.60.40.10	1.10.760.10	1qbl (0.91) 1hrc (0.36)	−9.5	−15.0
1hhj[b]	MHC/peptide	2.5	3.30.500.13	Peptide	[20]	−9.0	−13.3

[a]Status of the rigid-body approximation: a reference indicates that the structure was used in the past work that assumed or argued for rigid-body binding and is thus assumed to satisfy the approximation. Where possible, we calculated the Cα rmsd between bound and unbound receptor and ligand structures [24,26]. It is assumed that the rigid-body approximation is satisfied for Cα rmsd ≤ 0.9 Å. The unbound receptor PDB code is listed first and the ligand code below it; calculated rmsd values are given in parentheses. We were not always able to find unbound receptors and ligands. Despite rmsd's of ≈ 1.8 and 2.4, 1a0o and 1ebp were included in the test set and are assumed to satisfy the rigid-body approximation. The "?" indicates that we lack clear independent evidence for the truth or falsity of the rigid-body approximation. [b]Failed predictions (see text). [c]We also retained 1ebp because it includes a large peptide ligand and seems to mark the transition when main-chain torsions should be excluded from peptide calculations. The CATH codes are for the unbound receptor and ligand, respectively and are from (http://www.cathdb.info/latest/index.html). An * indicates a structure with more than one CATH code. The binding data is from multiple sources ([5,16,20–22,39,54]. All binding affinities are in kcal mol^{-1}.

4.1. Explaining binding affinities

To justify the claim that Eq. (12) explains or causally accounts for protein–protein binding affinities, at least four conditions must be satisfied: (1) the equation must have the correct form and include all relevant terms; (2) the equation must accurately model the binding energies of the training set; (3) the correlations implied by the equation cannot be attributed to random chance; (4) the contribution of each term to the binding affinity is consistent with experiment and fundamental theory.

Eq. (7) appears to satisfy condition (1) in that important free energy contributions have been included. The key to this derivation are the assumptions that the cost of polar group desolvation can be ignored and, that $\Delta G_{trans-rot} \approx 0$ and the inclusion of a term for water-bridged receptor–ligand contacts.

The high correlation coefficient ($R = 0.98$) demonstrates near optimal agreement between the binding energies estimated by Eq. (12) and the experimental values. The calculated coefficient of determination ($R^2 = 0.97$) implies that Eq. (12) accounts for roughly 97% of the variation in the experimental binding data, even when adjusted for the number of model descriptors (R^2-

Author's personal copy

Table 4
Test set descriptor values

PDB[a]	$\Delta X^{b}_{+/-}$	$\Delta X^{c}_{c/s}$	X^{d}_{sb}	X^{e}_{hb}	X^{f}_{gap}	ΔX^{g}_{rot}
1a0o	−1	−33	3	4	2494	−24
1hbs	0	−8	0	0	8343	−20
2iff	−7	−42	2	11	3773	−27
1avz	−4	−35	5	6	1675	−12
1hhg	−5	−50	1	14	1254	−41
1hhh	−6	−84	1	13	686	−42
1hhi	−5	−71	1	12	590	−41
1hhk	−5	−73	1	11	1287	−41
1lcj	−7	−25	2	14	2159	−30
1lck	−5	−24	4	10	805	−32
1lkl	−5	−19	4	10	513	−22
1sps	−7	−28	5	12	1290	−24
1tce	−4	−34	5	4	2131	−32
2pld	−5	−49	5	8	1818	−27
1irs	−1	−39	2	7	1303	−40
1shc	−1	−51	2	5	2285	−35
1sha	−6	−18	4	11	634	−24
1dkz	0	−33	0	8	2439	−30
2er6	−6	−57	1	12	1744	−25
1bbz	0	−31	0	6	609	−18
1ebp	0	−15	0	8	1945	−6
1nsn	−1	−40	3	8	5618	−34
1jhl	−5	−49	4	6	4856	−26
1mda	0	−29	0	0	5585	−15
3hfm	−5	−56	1	14	4181	−25
1bth	−10	−75	5	17	3695	−26
2kai	−4	−51	1	10	4354	−12
2jel	−3	−44	2	9	4402	−20
1gla	−1	−38	7	1	3560	−29
1mel	−4	−59	0	10	2287	−23
1bql	−6	−41	3	14	3807	−31
1nmb	−1	−29	1	10	5873	−20
1avw	−5	−74	3	16	5461	−17
1wej	−5	−39	3	13	4288	−16
1hhj	−5	−63	4	14	1000	−39

[a]PDB codes for all 35 test set complexes. [b]Change in the number of solvent exposed charged groups following complex formation. [c]Change in the number of solvent exposed hydrophobic groups (carbons and sulfurs). [d]Total number of interfacial salt bridges and [e] hydrogen bonds. [f]Interface gap volume. [g]Change in the number of solvent exposed side-chain (and main-chain torsions in the case of peptide ligands) torsions following association. See text for additional details.

adj = 0.96). The calculated root-mean-squared deviation (rmsd) between the implied and actual binding energies is $\approx RT$ and is negligible (rmsd = 0.62 kcal mol^{-1}). These results indicate that our regression equation accurately models the training set binding data and that condition (2) has been satisfied.

The overall statistical significance (f-statistic) of the regression model was calculated and the relevant null hypothesis tested ($\alpha_{+/-} = \alpha_{c/s} = \alpha_{sb} = \alpha_{hb} = \alpha_{gap} = \alpha_{tor} = 0$). The null hypothesis was to be rejected for $p < 0.05$. We obtained an f-statistic (89.6) and corresponding probability ($p < 0.05$) that clearly justify rejection of the null hypothesis, thus establishing Eq. (12) as a statistically significant estimator of the training set binding data. The results of a sensitivity analysis show that each term makes a statistically significant contribution the models goodness-of-fit supporting our use of the full regression model. Next, we evaluated the statistical significance of each descriptor by calculating t-statistics for each regression weight to test the null hypothesis, $\beta_i = 0$, where i = +/−, c/s, sb, hb, gap, tor. The null hypothesis was to be rejected for $p < 0.05$. We determined that each of the six descriptors clearly serves as a significant estimator of the training set binding data. Thus, our statistical analysis shows that random chance does not account for the associations of Eq. (12) and that condition (3) appears to be satisfied.

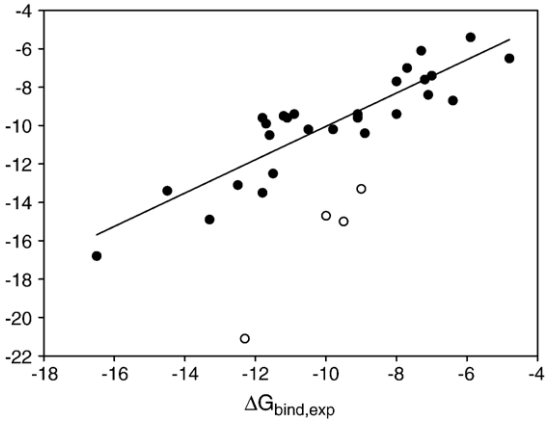


Fig. 4. Test set binding affinities versus predicted affinities. Experimental test set binding affinities ($\Delta G_{bind,exp}$) versus binding affinities (ΔG_{bind}) estimated using Eq. (12) (kcal mol^{-1}). See Tables 3 and 4 and for data on the 35 test set structures. The best-fit-regression line is through the 31 successful predictions (black filled circles); $R^2 = 0.79$, $R = 0.89$, rmsd = 1.2 kcal mol^{-1}. Failed predictions are shown as open circles. Including failed predictions, the best-fit-regression line (not shown) implies $R^2 = 0.53$, $R = 0.73$, rmsd = 2.4 kcal mol^{-1}.

If it can be shown that the underlying assumptions of regression have been satisfied, then the preceding statistical analysis is rigorously justified [42]. To test the assumption of multi-colinearity, we performed a cross-correlation analysis and this revealed no significant correlations between the descriptors (data not shown). Variance inflation factors (VIF's) were also calculated which indicated a total absence of higher level correlations between the descriptors. To test the assumptions of error normality, independence, constant variance, zero mean, and linearity, plots of residuals versus predicted binding affinity and normals versus residuals were constructed (results not shown). As required, the residual-prediction plot exhibits a

Table 5
Regression goodness-of-fit analysis

Statistic	Value
R	0.98
R^2	0.97
R^2-adj	0.96
rmsd	0.62
f-test	89.6
p-value	$\ll 0.00001$
Q^2	0.91

R is the correlation coefficient. R^2 is the coefficient of determination and R^2-adj is the coefficient of determination adjusted for the number of regressors. The rmsd gives the root mean squared deviation between the binding affinities estimated from Eq. (12) and the experimentally determined ones (kcal mol^{-1}). An f-test and corresponding probability was calculated to test the overall significance of the regression model (null hypothesis rejected, $p < 0.05$). As an initial test of Eq. (12)'s predictive power, a leave-one-out analysis was performed and Q^2 calculated.

Author's personal copy

Table 6
Sensitivity analysis

Descriptor[a]	df[b]	SSR_diff[c]	F[d]	p-value[e]
E	0	–	–	–
$\Delta X_{+/-}$	1	64.54	119.03	$\ll 0.05$
$\Delta X_{c/s}$	1	17.01	31.36	$\ll 0.05$
X_{sb}	1	32.98	60.83	$\ll 0.05$
X_{hb}	1	124.34	229.32	$\ll 0.05$
X_{gap}	1	8.58	15.83	$\ll 0.05$
ΔX_{tor}	1	3.48	6.42	0.02

[a] List of physical descriptors used in the full regression model.
[b] Degrees of freedom.
[c] Sum of squares regression difference (reduction in SSR if a given term is removed from the full model and the reduced model re-fitted).
[d] Calculated F-statistic for each term.
[e] p-value for each F-statistic (null hypothesis rejected for $p<0.05$).

random distribution of points while the normal-residual plot exhibits approximate linearity. Our regression model thus satisfies the key assumptions of regression analysis and all inferences are justified accordingly.

As a final test, we transformed the values for each descriptor and re-performed the regression analysis. The mathematical transformation involved subtracting the mean and dividing by the standard deviation for each of the six descriptors. We obtained identical regression results for the transformed data as for the untransformed data (results not shown).

The main results of our statistical analyses are summarized in Tables 5, 6 and 7. A plot of experimentally determined training set binding affinities versus empirical predictions (ΔG_{bind}) is also provided in Fig. 1. The results demonstrate that Eq. (12) is an excellent estimator of the training set binding data. Our theoretical derivation and statistical analysis also shows that each term in Eq. (12) makes an important and statistically significant contribution to estimating the binding affinities. While even the most exhaustive statistical analysis cannot definitively establish physical causation, our analysis strongly suggests that casual connections exist between the physical descriptors of Eq. (12) and experimentally determined binding affinities.

Table 7
Regression coefficient diagnostics

Descriptors[a]	a's[b]	t stat[c]	p-value[d]	Lower 95%[e]	Upper 95%	VIF[f]
E	−0.54	–	–	–	–	–
$\Delta X_{+/-}$	−0.85	−10.91	<0.0001	−1.02	−0.69	1.65
$\Delta X_{c/s}$	0.067	5.60	<0.0001	0.04	0.09	1.91
X_{sb}	−0.66	−7.80	<0.0001	−0.84	−0.48	2.45
X_{hb}	−0.90	−15.14	<0.0001	−1.02	−0.77	2.06
X_{gap}	−0.00087	−3.98	<0.0001	−0.00133	−0.00041	4.02
ΔX_{tor}	−0.09	−2.53	0.02	−0.17	−0.02	−0.02

[a] List of physical descriptors used in the regression model.
[b] Coefficients assigned to each descriptor by regression analysis.
[c] t-statistic calculated from each descriptor.
[d] Probability of getting a given t-statistic by chance (null hypothesis rejected for $p<0.05$).
[e] 95% confidence intervals for each regression coefficient.
[f] Variance inflation factors (see text).

Table 8
Regression derived coefficients compared with experiment and theory

Coefficient[a]	Regression[b]	Experiment[c]	Theory[d]
$a_{+/-}$	0.85	See Figs. 2 and 3	≈1.0[g]
$a_{c/s}$	0.067	See Figs. 2 and 3	–
a_{sb}	0.66	0.0 to 2.0[e]	0.65[h]
a_{hb}	0.90	0.7 to 1.0[f]	0.74–1.4[i]
a_{gap}	0.00087	–	–
a_{tor}	0.09	–	0.09–0.22[j]

[a] The descriptor coefficients used in Eq. (12).
[b] The magnitudes assigned to each descriptor coefficient by regression analysis.
[c] Experimentally determined coefficient magnitudes.
[d] Theoretically determined coefficient magnitudes. A dashed line (–) indicates that we were unable to locate or calculate the magnitude in question. All magnitudes are in units of kcal mol^{-1}. Although not presented in the Table, it is important to note that the signs of the various regression estimates are also in agreement with the theory and experiment.
[e] The experimental range for the neutral/favorable salt bridge contribution derives from several sources [55–57].
[f] The experimental range for the favorable hydrogen bonding contribution is from several references as well [58–60]. Agreement with the first two sources reasonably assumes that $\Delta G_{hb} \approx \Delta H_{hb} \approx 0.8$–$0.9$ kcal mol^{-1}.
[g] The theoretical estimate for the free energy cost of desolvating a charged group was calculated using Coulomb's law, the Boltzmann relation for the entropy and a theoretical estimate of 0.7 kcal/mol for the QM contribution of a hydrogen bond (see Supplementary material). It is assumed that the exposed protein charged group (NH1) interacts with the solvent primarily through two hydrogen bonds (NH1 – O) and that upon charge desolvation a single water–water hydrogen bond is formed. Assuming the water molecule acquires three additional degrees of freedom on desolvation, a distance dependent dielectric of 14, an interaction distance of 3.5 , and partial charges of +0.33 (NH1), +0.42 (HW), and −0.83 (OW) gives desolvation cost of ≈1.0 kcal mol^{-1}. Admittedly, this calculation provides a very rough estimate for the cost of charge desolvation.
[h] The theoretical range for the salt bridge contribution is based on Coulomb's law, assuming two partial charges (−0.5 and +0.33) separated by 3.5 , interacting through a medium with "average" dielectric 12–24, while immersed in 100 mM NaCl and assuming a desolvation penalty of 0.85 kcal/mol.
[i] The lower end of the theoretical range for the hydrogen bonding contribution assumes the covalent contribution is roughly 10% of the total H-bond interaction energy [28] which is ≈7.4 kcal mol^{-1} [58,61]. In addition to good numerical agreement with our regression estimate, this lower bound supports our interpretation that the hydrogen bonding contribution is primarily quantum-mechanical or covalent in origin (see Supplementary material for more on this). The higher estimate is from a recently described free energy function ([3]).
[j] The theoretical cost of burying a side-chain torsion depends on the specific side-chain type [62].

Condition (4) exists because it is possible for an equation to make mathematical but not physical sense. Thus, we decided to investigate the physical basis of each term in Eq. (12) that has been identified with a specific thermodynamic quantity (Eqs. (8)–(11)). Thus, we tested the regression-weighted descriptors from Eq. (12) against the available experimental evidence and theoretical considerations. According to Eq. (12), hydrogen bond formation, hydrophobic group desolvation, and interfacial water interactions are predicted to contribute favorably to binding, while charged group desolvation and torsion immobilization are predicted to oppose binding. Depending on the nature of the electrostatic interactions at the interface, salt bridges are predicted to be stabilizing, destabilizing or thermodynamically neutral. All of this fits qualitatively with

Author's personal copy

our background knowledge regarding the physical basis of protein–protein association [2,18,25]. There is also an excellent quantitative agreement between the coefficient magnitudes assigned by regression analysis and the magnitudes drawn from theoretical and experimental studies. The results of this analysis are presented in Table 8 and Figs. 2 and 3. It appears that condition (4) is satisfied and the regression weights of Eq. (12) are consistent with the available experimental data and theoretical arguments. It thus seems reasonable to infer that Eq. (12), within well-defined limits, can be used to explain protein–protein binding affinities. In the Supplementary material we have provided an extended discussion of each regression-weighted descriptor and how it relates to experiment and theory.

4.2. Predicting binding affinities

An equation can be explanatory without being predictive and vice versa. Hence, independent tests must be performed to evaluate the predictive power of a given equation. As a first direct test of the predictive utility of Eq. (12) a leave-one-out cross validation was performed, yielding excellent results ($Q^2=0.91$) suggesting that Eq. (12) can be used to predict binding affinities for complexes excluded from the training set.

A more demanding test is to use Eq. (12) in blind prediction on a test set. Towards this end, we constructed a test set of 35 protein–protein complexes (Tables 3 and 4); a cursory look at Tables 3 and 4 reveals that the test set is very different from the training set in terms of ligand and receptor size and type, complex interactions, and structural quality and that there is a large range of experimental binding affinities and descriptor values. Indeed, the 9 Sh2-pTyr complexes included in the test set involve phosphoryl-group receptor–ligand contacts that are totally absent from the training set. Thus, our test set provides a demanding test of the predictive power of Eq. (12) and its transferability.

Eq. (12) was used to blindly and successfully predict binding affinities for 31 of the 35 test set complexes, for a success rate of 89%. The results of the analysis are summarized in Fig. 4. Excluding outliers, we obtained excellent agreement between prediction and experiment ($R^2=0.79$, $R=0.89$, rmsd=1.2). These results suggest that Eq. (12) can be used to blindly predict binding affinities almost to within experimental error. The results further suggest that Eq. (12) captures the essential thermodynamics of complex formation and is transferable. Even the inclusion of all 4 outliers gives good results ($R^2=0.53$, $R=0.73$, rmsd=2.4). Hence, Eq. (12) can be used to accurately predict binding affinities for a range of protein complexes and this, in turn, improves our confidence in the theoretical, statistical and physical legitimacy of Eq. (12).

It is important to make a distinction between explained and unexplained outliers and to point out that Eq. (12) and its derivation provides a coherent, physics-based, framework for categorizing outliers as such and for suggesting improvements to future versions of the equation. For example, overestimation of 1HHJ is not uncommon [45] and is probably due to the fact that the salt bridges present at the interface are highly solvent exposed (data not shown). Future versions of Eq. (12) might thus be modified to reflect the attenuating effect of solvent exposure on charge–charge interactions. Despite its inclusion in the test set, 1AVW might nevertheless be of relatively poor quality and might violate the rigid-body approximation. Both possibilities are suggested by the high B-factors exhibited by the complex (data not shown) and by the fact that the C-alpha atoms of a second crystal complex (1AVX) superimpose onto 1AVW with a relatively high rmsd of 1.50. While the reasons for our failure to predict 1NMB and 1WEJ are unclear, violation of the rigid-body assumption is a real possibility. This is because differences between bound and unbound Cα coordinates, while no doubt strongly correlated with ΔG_{isomer}, provides an imperfect measure for evaluating the rigid-body approximation, for even minor Cα changes can probably sometimes give rise to large ΔG_{isomer} and vice versa.

4.3. Comparison with other methods

Predicting and explaining absolute protein–protein binding affinities is an unsolved and difficult problem. This is made clear by the fact that even highly sophisticated, computationally expensive, and expertly implemented all-atom simulation-based free energy methods are sometimes incapable of producing quantitative agreement with experiment [46–48].

To date, it appears that the largest study and best performance reported for a pure molecular mechanics-based method was for a non-diverse test set of 15 protein–protein complexes that approximate rigid-body binding [16]. In that study, the authors showed that their function could be used to predict binding affinities with $R^2=0.56$, $R=0.75$, and rmsd=2.4 kcal mol^{-1}. The study was especially promising given that all of the free energy predictions were made from static structures. Despite this the computational cost was still relatively high, as free energy prediction was preceded by energy minimization.

Schapira and co-workers combined a molecular mechanics/Poisson–Boltzmann formalism, along with additional and system-specific fitted parameters, to try and estimate binding affinities for 8 protease–protein inhibitors (all eight are considered in the present paper) and 8 sh2-pTYR complexes (1lkk, 1shp and 1sps were omitted from the present study on structural/quality grounds). After excluding several outliers and employing an ingenious but computationally expensive MCM (Monte Carlo with Minimization) optimization procedure Schapira et al. reported statistical correlations of 0.95 and 0.91, respectively [21]. Due to the use of additional and system-specific fitted parameters, however, it is not clear if this analysis qualifies as blind prediction.

The linear interaction method (LIE) represents a relatively new and popular approach to binding affinity prediction and has been employed in the past to successfully predict small ligand binding free energies with excellent results [49]. Like the method employed by Schapira et al., the LIE method combines molecular mechanics terms with additional fitted parameters. Unlike the Schapira method, the LIE method requires explicit ligand–solvent simulations. The LIE method is thus computationally demanding, at least from the perspective of scoring function development. Moreover, care must be exercised in

Author's personal copy

physically interpreting the LIE method and a universal set of fitted parameters probably does not exist [48,49]. Most relevant to the present study, the LIE method was recently evaluated and found wanting for its utility in predicting protein–peptide binding affinities [48]. In sum, the results reported for Eq. (12) compare very favorably with what has been reported for even highly sophisticated protein–protein free energy methods but at much lower computational cost. Similarly, our results perform well against scoring functions primarily designed to predict binding affinities between proteins and small molecules, carbohydrates, DNA and other non-protein ligands [42] (see especially Table 1 in [42]).

Ma and co-workers have described an empirical function for estimating protein–protein binding affinities from static structures that is very similar in spirit to our own. The function includes three terms and was parameterized by regression on structural and thermodynamic data from 20 protein complexes (19 of which are considered in the present study; the insulin dimmer 4INS was excluded because it has an interfacial disulfide bridge). The side-chain accessible (N_B) number was used to estimate changes in conformational entropy. The total number of hydrophilic pairs (N_{pair}) was used to estimate the electrostatic interaction energy. The change in apolar accessible surface area (ΔASA_{apolar}) was used to estimate the desolvation free energy.

While Ma and co-workers evaluated the function using various statistical measures (R^2 and rmsd), we decided to calculate other statistics (R^2-adj and Q^2) using the raw data supplied in the original paper. We believe that this will allow for more meaningful comparison. Indeed, the comprehensive statistical analysis performed here, while perhaps common in the protein–ligand literature [42], appears to be a first for protein–protein studies.

The results of the comparison are summarized in Table 9. The results show that the Ma function provides a good mathematical fit with the training set data and may even be predictive but that Eq. (12) provides superior performance. Importantly, Eq. (12) has been validated in blind prediction while the Ma function has not. Additionally, some of the terms employed in the Ma function do not have a clear physical basis. For example, the definition for N_{pair} implies that cation–cation and anion–anion interactions are *stabilizing* and in a manner similar to hydrogen bonds and salt bridges. This clearly underscores the importance of a mathematical, theoretical and physical evaluation of each regression term, subject to data availability. Because the Ma equation improves upon previous work, all of these considerations apply with even greater force to other specifically empirical protein–protein regression [13,18].

Table 9
Comparison with the regression equation derived by Ma et al.

Investigator [a]	Descriptors/ training set [b]	Structural optimization [c]	R^2-adj/ rmsd [d]	Q^2 [e]
Audie	6/24	None	0.90/1.08	0.88
Ma	3/20	EM	0.96/0.62	0.91

[a] Results taken from this study and Ma et al. [19].
[b] The total number of descriptors (independent variables) relative to the size of the training set.
[c] In the method of Ma et al. prediction is preceded by a preparatory energy minimization (EM).
[d] Adjusted correlation of determination (R^2) and root-mean-squared deviation rmsd (kcal mol^{-1}).
[e] Ma and co-workers failed to calculate Q^2, where Q^2 is analogous to R^2 but quantifies goodness-of-prediction instead of goodness-of-fit. Hence, using their data we calculated Q^2 for them and included it for comparative purposes. Ma et al. also failed to test their method in blind prediction on a test set.

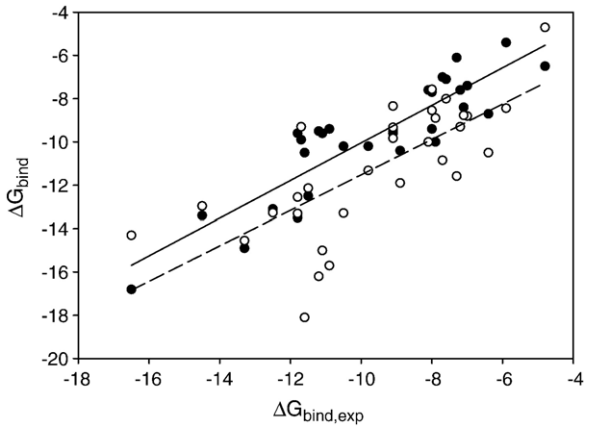


Fig. 5. Test set binding affinities versus predicted affinities using Eq. (17) and Dcomplex. Experimental test set binding affinities ($\Delta_{Gbind,exp}$) versus binding affinities (Δ^{G}_{bind}) estimated using Eq. (12) and Dcomplex (kcal mol^{-1}). The best-fit-regression lines are through the 31 successful predictions made using Eq. (12) (black filled circles; unbroken line) and Dcomplex (open circles, dashed line), respectively. For Eq. (12): R^2=0.79, R=0.89, rmsd=1.2 kcal mol^{-1} For Dcomplex: R^2=0.53, $_R$=0.73, rmsd=2.6 kcal mol^{-1}. Most of the Dcomplex binding affinity estimates were taken directly from the paper describing Dcomplex [5]. Otherwise, estimates were made using the Dcomplex server (http://sparks.informatics.iupui.edu/).

In the last decade there has been a discernable trend away from physics-based empirical functions and towards statistical or knowledge-based potentials [50]. Thus, we thought it was important to directly compare blind predictions made using Eq. (12) with predictions made using a knowledge-based potential. We tested our predictions against those made using Dcomplex, a recently described, cutting-edge, web-enabled and knowledge-based potential that was used to predict binding affinities for 82 protein complexes [5]. The results of the comparison are shown in Fig. 5. For all 31 complexes Dcomplex performed well (R^2=0.53, rmsd=2.4 kcal mol^{-1}) but Eq. (12) exhibited better overall performance (R^2=0.79, rmsd=1.2 kcal mol^{-1}). These results are especially encouraging given that in independent comparative testing, Dcomplex was shown to outperform other widely used functions [5].

It is important to recall that Eq. (12) can be used to both predict and explain binding affinities. Unlike knowledge-based potentials, Eq. (12) provides a consistent, physics-based framework for interpreting and understanding results that is intuitive to medicinal chemists (e.g. [50]). Thus, in comparison with knowledge-based potentials Eq. (12) can be used to help quantitatively guide and rationalize ligand and receptor design in a way that knowledge-based potentials cannot.

Author's personal copy

5. Conclusion

We have described a new function for explaining and predicting experimentally determined protein–protein and protein–peptide binding affinities. Our function implies that protein complexes are stabilized by the covalent contributions of hydrogen bonds, hydrophobic group burial and water mediated interactions and are destabilized by torsion and charged group burial; salt bridges are predicted to be either stabilizing or destabilizing. The function's mathematical form follows from a derivation that assumes native-state complexes are designed such that ΔGdesolv,polarcan be ignored, $\Delta G_{tr} \approx 0$, and that makes explicit provisions for water-bridged receptor–ligand interactions (ΔG^{r-w-l}). Unlike previous work, we provide an implicit estimate for ΔG^{r-w-l} in terms of the interchain gap volume, and an alternate approach for estimating charge and hydrophobic group desolvation is used in place of conventional surface area calculations. The function was parameterized using regression analysis on an internally consistent and high quality training set of 24 protein–protein complexes.

Extensive statistical testing and comparison with theory and experiment indicates that the regression-optimized function can be used to explain, within well-defined limits, protein–protein and protein–peptide binding affinities. Importantly, the extension to protein systems with multiple binding modes or that violate rigid-body binding is conceptually straightforward. The results from a leave-one-out cross validation analysis further suggest that the function is predictive. Indeed, we used the function to blindly and successfully predict binding affinities for a large and diverse test set (31 protein complexes) which supports the usefulness of our function and points to its transferability. We also used the function to predict water–octanol transfer free energies for 25 peptides. In all, the function was used to accurately and blindly predict 80 protein and peptide transfer and binding free energies.

Our function performed well in comparison to pure molecular mechanics-based methods, hybrid or fitted molecular mechanics-based methods, and previously described regression equations. A direct comparison with a new knowledge-based potential produced encouraging results. When combined with the functions algorithmic simplicity, low computational cost and physically intuitive form, the results presented in this study suggest that the function can be used to help solve a number of protein design and discovery problems of theoretical, biological and medical importance.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bpc.2007.05.021.

References

[1] M.K. Gilson, J.A. Given, B.L. Bush, J.A. McCammon, The statistical-thermodynamic basis for computation of binding affinities: a critical review, Biophys. J. 72 (1997) 1047.

[2] S. Vajda, M. Sippl, J. Novotny, Empirical potentials and functions for protein folding and binding, Curr. Opin. Struck. Biol. 7 (1997) 222.

[3] R. Guerois, J.E. Nielsen, L. Serrano, Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations, J. Mol. Biol. 320 (2002) 369.

[4] Ajay, M.A. Murcko, Computational methods to predict binding free energy in ligand–receptor complexes, J. Med. Chem. 38 (1995) 4953.

[5] C. Zhang, S. Liu, Q. Zhu, Y. Zhou, A knowledge-based energy function for protein–ligand, protein–protein, and protein–DNA complexes, J. Med. Chem. 48 (2005) 2325.

[6] S. Vajda, Z. Weng, R. Rosenfeld, C. DeLisi, Effect of conformational flexibility and solvation on receptor–ligand binding free energies, Biochemistry 33 (1994) 13977.

[7] S.M. Feller, M. Lewitzky, Potential disease targets for drugs that disrupt protein-protein interactions of Grb2 and Crk family adaptors, Curr. Pharm. Des. 12 (2006) 529.

[8] A. Loregian, G. Palu, Disruption of protein–protein interactions: towards new targets for chemotherapy, J. Cell. Physiol. 204 (2005) 750.

[9] P.M. Watt, Screening for peptide drugs from the natural repertoire of biodiverse protein folds, Nat. Biotechnol. 24 (2006) 177.

[10] C. Falciani, L. Lozzi, A. Pini, L. Bracci, Bioactive peptides from libraries, Chem. Biol. 12 (2005) 417.

[11] V.J. Hruby, Designing peptide receptor agonists and antagonists, Nat. Rev. Drug Discov. 1 (2002) 847.

[12] R.C. Ladner, A.K. Sato, J. Gorzelany, M. de Souza, Phage display-derived peptides as therapeutic alternatives to antibodies, Drug Discov. Today 9 (2004) 525.

[13] N. Horton, M. Lewis, Calculation of the free energy of association for protein complexes, Protein Sci. 1 (1992) 169.

[14] I. Halperin, B. Ma, H. Wolfson, R. Nussinov, Principles of docking: an overview of search algorithms and a guide to scoring functions, Proteins 47 (2002) 409.

[15] G.R. Smith, M.J. Sternberg, Prediction of protein–protein interactions by docking methods, Curr. Opin. Struck. Biol. 12 (2002) 28.

[16] Z. Weng, C. Delisi, S. Vajda, Empirical free energy calculation: comparison to calorimetric data, Protein Sci. 6 (1997) 1976.

[17] S. Krystek, T. Stouch, J. Novotny, Affinity and specificity of serine endopeptidase–protein inhibitor interactions. Empirical free energy calculations based on X-ray crystallographic structures, J. Mol. Biol. 234 (1993) 661.

[18] D. Xu, S.L. Lin, R. Nussinov, Protein binding versus protein folding: the role of hydrophilic bridges in protein associations, J. Mol. Biol. 265 (1997) 68.

[19] X.H. Ma, C.X. Wang, C.H. Li, W.Z. Chen, A fast empirical approach to binding free energy calculations based on protein interface information, Protein Eng. 15 (2002) 677.

[20] D. Rognan, S.L. Lauemoller, A. Holm, S. Buus, V. Tschinke, Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins, J. Med. Chem. 42 (1999) 4650.

[21] M. Schapira, M. Totrov, R. Abagyan, Prediction of the binding energy for small molecules, peptides and proteins, J. Mol. Recognit. 12 (1999) 177.

[22] Y. Zhou, R. Abagyan, How and why phosphotyrosine-containing peptides bind to the SH2 and PTB domains, Fold. Des. 3 (1998) 513.

[23] R.M. Jackson, H.A. Gabb, M.J. Sternberg, Rapid refinement of protein interfaces incorporating solvation: application to the docking problem, J. Mol. Biol. 276 (1998) 265.

[24] J. Mintseris, K. Wiehe, B. Pierce, R. Anderson, R. Chen, J. Janin, Z. Weng, Protein–protein docking benchmark 2.0: an update, Proteins 60 (2005) 214.

[25] L. Lo Conte, C. Chothia, J. Janin, The atomic structure of protein-protein recognition sites, J. Mol. Biol. 285 (1999) 2177.

[26] R. Chen, J. Mintseris, J. Janin, Z. Weng, A protein–protein docking benchmark, Proteins 52 (2003) 88.

[27] R. Brem, K.A. Dill, The effect of multiple binding modes on empirical modeling of ligand docking to proteins, Protein Sci. 8 (1999) 1134.

[28] T.W. Martin, Z.S. Derewenda, The name is bond-H bond, Nat. Struct. Biol. 6 (1999) 403.

[29] D.W. Gatchell, S. Dennis, S. Vajda, Discrimination of near-native protein structures from misfolded models by empirical free energy functions, Proteins 41 (2000) 518.

Author's personal copy

[30] C.N. Pace, Polar group burial contributes more to protein stability than nonpolar group burial, Biochemistry 40 (2001) 310.
[31] Y.B. Yu, P.L. Privalov, R.S. Hodges, Contribution of translational and rotational motions to molecular association in aqueous solution, Biophys. J. 81 (2001) 1632.
[32] M. Gerstein, A resolution-sensitive procedure for comparing protein surfaces and its application to the comparison of antigen-combining sites, Acta Crystallogr. A48 (1992) 271.
[33] J.T. Hubbard, J.M., 'NACCESS', Computer Program, Department of Biochemistry and Molecular Biology, University College London, 1993.
[34] R.A. Laskowski, SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions, J. Mol. Graph. 13 (1995) 323.
[35] H.M. Berman, T.N. Bhat, P.E. Bourne, Z. Feng, G. Gilliland, H. Weissig, J. Westbrook, The Protein Data Bank and the challenge of structural genomics, Nat. Struct. Biol. 7 Suppl (2000) 957.
[36] N. Guex, M.C. Peitsch, SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling, Electrophoresis 18 (1997) 2714.
[37] M.J. Betts, M.J. Sternberg, An analysis of conformational changes on protein–protein association: implications for predictive docking, Protein Eng. 12 (1999) 271.
[38] T.P. Flores, C.A. Orengo, D.S. Moss, J.M. Thornton, Comparison of conformational characteristics in structurally similar protein pairs, Protein Sci. 2 (1993) 1811.
[39] J.J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C.A. Rohl, D. Baker, Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations, J. Mol. Biol. 331 (2003) 281.
[40] H. Li, A.D. Robertson, J.H. Jensen, Very fast empirical prediction and rationalization of protein pK_a values, Proteins 61 (2005) 704.
[41] F.M. Pearl, C.F. Bennett, J.E. Bray, A.P. Harrison, N. Martin, A. Shepherd, I. Sillitoe, J. Thornton, C.A. Orengo, The CATH database: an extended protein family resource for structural and functional genomics, Nucleic Acids Res. 31 (2003) 452.
[42] T. Jain, B. Jayaram, An all atom energy based computational protocol for predicting binding affinities of protein–ligand complexes, FEBS Lett. 579 (2005) 6659.
[43] S. Jones, J.M. Thornton, Principles of protein–protein interactions, Proc. Natl. Acad. Sci. U. S. A. 93 (1996) 13.
[44] D. Xu, C.J. Tsai, R. Nussinov, Hydrogen bonds and salt bridges across protein-protein interfaces, Protein Eng. 10 (1997) 999.
[45] Z. Liu, B.N. Dominy, E.I. Shakhnovich, Structural mining: self-consistent design on flexible protein–peptide docking and transferable binding affinity potential, J. Am. Chem. Soc. 126 (2004) 8515.
[46] T. Hou, K. Chen, W.A. McLaughlin, B. Lu, W. Wang, Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain, PLoS Comput. Biol. 2 (2006) e1.
[47] V. Zoete, M. Meuwly, M. Karplus, Study of the insulin dimerization: binding free energy calculations and per-residue free energy decomposition, Proteins 61 (2005) 79.
[48] S. Donnini, A.H. Juffer, Calculation of affinities of peptides for proteins, J. Comput. Chem. 25 (2004) 393.
[49] J. Aqvist, J. Marelius, The linear interaction energy method for predicting ligand binding free energies, Comb. Chem. High Throughput Screen. 4 (2001) 613.
[50] H. Gohlke, G. Klebe, Statistical potentials and scoring functions applied to protein-ligand binding, Curr. Opin. Struck. Biol. 11 (2001) 231.
[51] A. Kim, F.C. Szoka, Amino acid side-chain contributions to free energy of transfer of tripeptides from water to octanol, Pharm. Res. 9 (1992) 504.
[52] W.C. Wimley, T.P. Creamer, S.H. White, Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides, Biochemistry 35 (1996) 5109.
[53] P. Kasper, P. Christen, H. Gehring, Empirical calculation of the relative free energies of peptide binding to the molecular chaperone DnaK, Proteins 40 (2000) 185.
[54] A. Palencia, E.S. Cobos, P.L. Mateo, J.C. Martinez, I. Luque, Thermodynamic dissection of the binding energetics of proline-rich peptides to the Abl-SH3 domain: implications for rational ligand design, J. Mol. Biol. 336 (2004) 527.
[55] K. Takano, Y. Yamagata, K. Yutani, Buried water molecules contribute to the conformational stability of a protein, Protein Eng. 16 (2003) 5.
[56] K.S. Lassila, D. Datta, S.L. Mayo, Evaluation of the energetic contribution of an ionic network to beta-sheet stability, Protein Sci. 11 (2002) 688.
[57] P.C. Lyu, P.J. Gans, N.R. Kallenbach, Energetic contribution of solvent-exposed ion pairs to alpha-helix structure, J. Mol. Biol. 223 (1992) 343.
[58] R.L. Baldwin, In search of the energetic role of peptide hydrogen bonds, J. Biol. Chem. 278 (2003) 17581.
[59] Y. Kato, M.M. Conn, J. Rebek Jr., Hydrogen bonding in water using synthetic receptors, Proc. Natl. Acad. Sci. U. S. A. 92 (1995) 1208.
[60] Y.W. Chen, A.R. Fersht, K. Henrick, Contribution of buried hydrogen bonds to protein stability. The crystal structures of two barnase mutants, J. Mol. Biol. 234 (1993) 1158.
[61] S. Scheiner, Relative strengths of NH..O and CH..O hydrogen bonds between polypeptide chain segments, J. Phys. Chem., B Condens. Matter Mater. Surf. Interfaces Biophys. 109 (2005) 16132.
[62] X. Hu, B. Kuhlman, Protein design simulations suggest that side-chain conformational entropy is not a strong determinant of amino acid environmental preferences, Proteins 62 (2006) 739.