BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 21 2011, pages 3002–3009
doi:10.1093/bioinformatics/btr513

Structural bioinformatics

Advance Access publication September 7, 2011

Protein–protein binding affinity prediction on a diverse set of structures

Iain H. Moal†, Rudi Agius† and Paul A. Bates*
Biomolecular Modelling Laboratory, Cancer Research UK London Research Institute, London WC2A 3LY, UK

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Accurate binding free energy functions for protein–protein interactions are imperative for a wide range of purposes. Their construction is predicated upon ascertaining the factors that influence binding and their relative importance. A recent benchmark of binding affinities has allowed, for the first time, the evaluation and construction of binding free energy models using a diverse set of complexes, and a systematic assessment of our ability to model the energetics of conformational changes.

Results: We construct a large set of molecular descriptors using commonly available tools, introducing the use of energetic factors associated with conformational changes and disorder to order transitions, as well as features calculated on structural ensembles. The descriptors are used to train and test a binding free energy model using a consensus of four machine learning algorithms, whose performance constitutes a significant improvement over the other state of the art empirical free energy functions tested. The internal workings of the learners show how the descriptors are used, illuminating the determinants of protein–protein binding.

Availability: The molecular descriptor set and descriptor values for all complexes are available in the Supplementary Material. A web server for the learners and coordinates for the bound and unbound structures can be accessed from the website: http://bmm.cancerresearchuk.org/~Affinity

Contact: paul.bates@cancer.org.uk

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on July 15, 2011; revised on August 15, 2011; accepted on September 4, 2011

1 INTRODUCTION

Protein–protein interactions are involved in almost all biological processes, and the structural and thermodynamic characterization of protein interaction networks is a major goal of functional genomics. As progress is made in the structural annotation of interaction networks (Huang *et al.*, 2008; Kundrotas *et al.*, 2010; Zhang *et al.*, 2010), determining binding free energies is recognized as an important step in the modelling of biological systems (Aloy and Russell, 2006; Beltrao *et al.*, 2007; Dell'Orco, 2009; Keskin *et al.*, 2005; Kiel *et al.*, 2008). The calculation of protein–protein binding affinities is also fundamental to many endeavours in structural bioinformatics, ranging from the discovery of peptide therapeutics (Kumar *et al.*, 2010; Rao and Kumar, 2008) and docking (Halperin *et al.*, 2002) to protein engineering (Kortemme and Baker, 2004; Sharabi *et al.*, 2011), *de novo* interface design (Fleishman *et al.*, 2011b) and computational mutagenesis (Ben-Shimon and Eisenstein, 2010). Understanding the determinants of protein binding is thus a question of huge practical relevance, and the construction of accurate and efficient binding free energy functions of great importance.

Methods of calculating protein–protein binding free energies vary dramatically from one another in terms of physical plausibility, accuracy and computational cost. At one end of the spectrum lies exact methods, such as free energy perturbation and thermodynamic integration (Kollman, 1993), and end-point methods, such as MM-PBSA and linear interaction analysis (Gilson and Zhou, 2007). While in principle highly accurate, these methods use extensive molecular dynamics or Monte Carlo sampling, and are usually only applicable where the bound and unbound states have significant overlap. Thus, they can only be applied to proteins for which conformational changes are minimal and are only practical on a case by case basis or for low-throughput studies. Empirical energy functions are much faster, and fall into three categories. The first of these are statistical potentials, in which the relative positions of atoms or residues observed in experimental structures are used to infer a potential of mean force (Jiang *et al.*, 2002; Liu *et al.*, 2004; Su *et al.*, 2009; Zhang *et al.*, 1997, 2005). Second are thermodynamic equations, in which the energy is calculated as a sum of terms arising from relevant phenomena (Bai *et al.*, 2011; Bougouffa and Warwicker, 2008; Horton and Lewis, 1992; Krystek *et al.*, 1993; Ma *et al.*, 2002; Novotny *et al.*, 1989; Vajda *et al.*, 1994; Weng *et al.*, 1997; Xu *et al.*, 1997). The relative contributions of the terms can be determined by linear regression against known binding affinities (Bai *et al.*, 2011; Bougouffa and Warwicker, 2008; Horton and Lewis, 1992; Ma *et al.*, 2002; Xu *et al.*, 1997), or by other other means (Krystek *et al.*, 1993; Novotny *et al.*, 1989; Vajda *et al.*, 1994; Weng *et al.*, 1997). Third, there are functions which implicitly estimate binding free energies due to their being optimized for purposes such as docking, estimating free energy differences between mutants (Jiang *et al.*, 2005) or identifying biological assemblies (Krissinel and Henrick, 2007).

Most previous binding free energy models have been parametrized and/or evaluated on a narrow range of proteins. Flexible proteins were either explicitly rejected or implicitly under-represented due to datasets being derived from previous works. Further, the complexes comprised either exclusively of protease–inhibitor pairs (Krystek *et al.*, 1993; Nauchitel *et al.*, 1995; Vajda *et al.*, 1994; Wallqvist *et al.*, 1995; Zhang *et al.*, 1997) or mostly protease inhibitor interactions with a few others, such as the insulin

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

dimer, barnase–barstar, lysozme–antibody and the $\alpha\beta$ interface of deoxyhaemoglobin (Audie and Scarlata, 2007; Bougouffa and Warwicker, 2008; Horton and Lewis, 1992; Jiang *et al.*, 2002; Ma *et al.*, 2002; Weng *et al.*, 1997; Xu *et al.*, 1997; Zhou and Zhou, 2002). Four recent studies have been undertaken on much larger datasets of between 52 and 86 complexes (Jiang *et al.*, 2005; Liu *et al.*, 2004; Su *et al.*, 2009; Zhang *et al.*, 2005). Most of the additional interactions were of small peptides, typically between 2 and 5 residues in length, and mostly interacting with oligopeptide binding protein OppA. Although still biased towards small rigid proteins, some other complexes, such as hormone–receptor pairs, signal transduction complexes and complexes containing G proteins, were also considered. More recently, empirical free energy models of all three categories were evaluated on a more diverse set of 46 interactions (Kastritis and Bonvin, 2010). The best correlations with experimental affinities were around 0.5, and no method stood out as being particularly superior.

We recently published the largest and most diverse set of experimental binding free energies to date, covering 144 non-redundant interactions, with structural cross-referencing to both the bound complex and its unbound constituents (Kastritis *et al.*, 2011). This provides an unprecedented opportunity for the construction and evaluation of empirical binding free energies models.

2 APPROACH

Machine learning has been successfully applied to hot-spot identifications (Cho *et al.*, 2009; Darnell *et al.*, 2007; Tuncbag *et al.*, 2010; Zhu and Mitchell, 2011), but to the authors' knowledge has not yet been used for protein–protein binding affinity prediction. The approach taken here is to construct a set of molecular descriptors and feed these into machine learning models. The learners are trained on a subset of the benchmark for which binding affinities are known with high confidence. These models are combined to produce a consensus energy, which is verified using leave-one-out cross-validation and the remainder of the benchmark. As a wide range of conformational changes are observed within the dataset, and previous empirical binding free energy functions have largely neglected these, a number of descriptors are included to model the energetics of structural rearrangements, including bonding and non-bonding terms.

The descriptors are chosen foremost for their physical relevance. Most are explicitly designed to either model the free energy, enthalpy or entropy changes associated with the processes shown in Figure 1. Computational efficiency was considered and most descriptors can be calculated within seconds using simple equations, common packages and publicly available web servers. Efficient algorithms

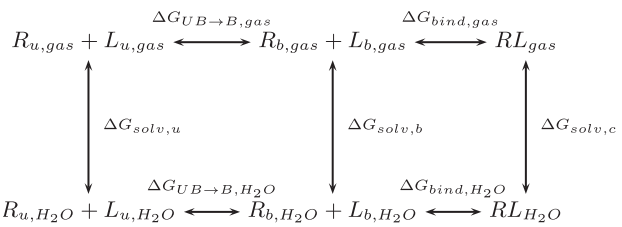

Fig. 1. The thermodynamic cycle of complex formation, showing the processes modelled by the energetic descriptors.

are used where applicable, such as the elastic network vibrational entropy model and the conformational space sampling routine, both of which perform well despite requiring a fraction of the computational resources compared with more traditional techniques (Benedix *et al.*, 2009; Carrington and Mancera, 2004).

A wide variety of machine learning tools are available, and their performance is highly dependent on the problem they are applied to. Four learners were selected which are specifically aimed at continuous class regression; Random Forest (RF, Breiman, 2001), M5′ Regression Tree (M5′, Wang and Witten, 1997), Multivariate Adaptive Regression Splines (MARS, Friedman, 1991) and Radial Basis Function Interpolation (RBF, Hardy, 1971). The choice of these learners was guided by the following considerations:

(1) As linear regression has been routinely applied to protein–protein affinity prediction, methods were chosen which address some of its limitations; sensitivity to outliers, inability to account for non-linear relationships and degradation in performance in high dimensions. Acknowledging that certain molecular descriptors might be noisy within certain ranges, algorithms which are able to partition the input space may be better suited to this problem.

(2) The methods have differing conceptual attributes. The RF, for instance, is derived from the consensus of many low-accuracy models. The MARS method allows the exploitation of parameters which have predictive value only over certain ranges. The M5′ method works by applying different regression models to different parts of the input space, which accounts for diversity within the dataset. The RBF method exploits global features of the training set, and the predicted affinity of any given case is largely determined by examples further away in feature space. These different approaches were considered likely to capture different aspects of the descriptor set, ideally having weak correlations between their predictions and hence more likely to work synergistically when combined together into a consensus model. We acknowledge that learner selection and fusion algorithms are available and might offer more attractive solutions (Kuncheva, 2002; Opitz and Maclin, 1999). However, such methods are prone to overfitting unless a further validation set is used. Given our limited sized dataset, we opted for a simple unweighted average of the outputs of predictors chosen for their conceptual differences.

(3) The models can be scrutinized, allowing their physical plausibility to be evaluated and the relative contributions of the various factors to be ascertained.

(4) Methods must be able to avoid fitting noise in the data while still detecting the signal. The techniques chosen account for overfitting either explicitly or implicitly. RFs do not overfit as more trees are added, rather the test error converges to a limiting value. They are able to achieve low bias predictions through trees built from different subsets of the data and descriptors, and low variance through averaging the output of all trees. The M5′ and MARS methods have backward feature elimination routines to remove redundant features. In the RBF method, features are equally weighted and it is the contributions of the examples in the training set which are

determined. This is particularly suited to situations in which there are more features than examples as is the case here.

(5) The methods are robust to the model parameters. As fine-tuning is not required, default parameters can be used, eliminating possible biases originating from adjusting parameters until the desired result is obtained.

3 METHODS

3.1 The Descriptors

In total, 200 descriptors were calculated, which are described fully in Supplementary File S1. Although some were calculated using stand-alone programs or simple equations, most were calculated using the ProtorP server (Reynolds *et al.*, 2009), CHARMM (Brooks *et al.*, 2009), PyRosetta (Chaudhury *et al.*, 2010), FireDock (Andrusier *et al.*, 2007) and the Potentials'R'Us server (Feng *et al.*, 2010). These include energy terms associated with electrostatics, London dispersion and exchange repulsion forces, as well as potentials for hydrogen bond, aliphatic, cation–π and π–π interactions. Also included are descriptors related to solvation effects, including continuum electrostatics models and hydrophobic burial, as well as terms for translational, rotational, vibrational, side chain and disorder to order transition entropies. Additionally, a number of statistical potentials, including residue and atomic pair potentials, as well as four-body potentials, were included. Rigid body interaction energy changes were calculated as

$$E_{\text{bind}} = E_{\text{complex}} - (E_{R,b} + E_{L,b}). \quad (1)$$

The descriptors that account for conformational changes upon binding, which carry the suffix UB or EBU, were calculated using the energy of the proteins in the absence of their binding partners as

$$E_{UB \to B} = (E_{R,b} - E_{R,u}) + (E_{L,b} - E_{L,u}). \quad (2)$$

As pH can have a significant influence on binding affinity, even over a narrow range, some descriptors were chosen for their ability to account for variable protonation states. First, PROPKA was used to determine the pKa of the titratable amino acids (Bas *et al.*, 2008). The most probable assignment of protonation states, at the experimental pH, was determined using PDB2PQR (Dolinsky *et al.*, 2007). These assignments were used in all descriptors calculated using the CHARMM22 force field, which are prefixed with ACE22.

Lastly, as proteins exist not as static structures but as structural ensembles, conformational sampling was performed with CONCOORD 2.1 using dynamic tolerance setting (de Groot *et al.*, 1997). For the unbound receptor, unbound ligand and bound complex, 100 structures were generated and energies calculated by averaging over the structural ensembles. The corresponding descriptors for rigid body binding and conformational changes are denoted by the suffixes ENS and EBU, respectively.

3.2 The Learners

MARS: it is a non-parametric regression method which uses a set of hinge functions to model non-linear relationships between the input variables and the target output (Friedman, 1991). The model is formed from a set of weighted basis functions, where each basis function contains a hinge function or a product of two or more hinge functions, if we seek to model higher order interactions between variables. MARS automatically assigns the weights for each basis function, and the variables for a given hinge function together with the values for the knot positions.

$$F(x) = \sum_{i=1}^{k} c_i B_i(x). \quad (3)$$

Default values were used without tuning, as follows: we set a maximum limit on the number of basis functions grown in the forward phase to 21, and set no limit on the number of basis functions used in the final model after pruning. Model complexity is also limited by setting the knot-cost to the recommended value of two. Piece-wise cubic modelling was use to model hinge regions, no self-interactions between input variables was allowed and no interactions between variables was allowed in the basis functions, as we found that this additional complexity was not needed. The ARESLab toolbox implementation was used.

Random forest: a Matlab implementation of the RF algorithm, as described by Breiman (2001), was used. RF is an average prediction of a collection of decision trees, where the criterion at each node is chosen so as to minimize the variance within the branches. For an effective ensemble, the tree predictions should have high accuracy but low correlation across trees. RF achieves the latter by building each decision tree using only a subset of features and data from the full respective sets. In our implementation, the number of decision trees was set to 750 and we limit the number of random features available for selection at each node when building the decision trees to 20. Each decision tree was fully grown, such that each leaf corresponds to a member of the training set. The final prediction is returned as the mean of all trees. The RF prediction was found to be very robust to change in these parameters and fine-tuning was not required.

RBF interpolation: the RBF method was used, as described by Hardy (1971), in a Matlab implementation. The energy function consists of the mean affinity of the training set, μ, plus a linear combination of multiquadratic basis functions, $\phi(d) = \sqrt{d^2 + 1}$, summed across the n cases in the training set.

$$F(\mathbf{x}) = \mu + \sum_{i=1}^{n} a_i \phi(||\mathbf{x} - \mathbf{x}_i||), \quad (4)$$

where $\mathbf{x}$ is the descriptor vector. The a regression coefficients are determined by multiple regression. Equation (4) is used for the prediction and requires no parameters other than the training data. In this intriguing model, all features are equally weighted, and it is the weights of the cases in the training set which are optimized.

M5′: the M5 model tree is similar to standard regression trees with the additional possibility of having a linear regression model at the leaves (Quinlan, 1992). The tree is fully grown, and then pruned by a greedy algorithm using error estimates. In the pruning stage, consideration is given to adding a regression model instead of a constant value prediction. In this work, we make use of the M5′ algorithm, a modified version of the original M5 regression tree described by Wang and Witten (1997). This version is able to achieve more comprehensible trees through smaller trees which still have similar predictive performance. Each tree is descended to either a constant value prediction or a regression plane. Rather than applying one M5′ to the full feature set, an ensemble of M5′ regression trees was used; four sets of four M5′ regression trees each, where each set covers the whole feature set space and each M5′ tree within each set covers a unique feature subset. The predicted value of all 16 trees is averaged for the final prediction, as for the RF method. Default values were used, and not tuned, as follows: for each M5′ tree we limited the number of examples a node can represent to a minimum of four. A node is not split if the SD of the output variable values at the node is <0.05 of the SD of the output variable values of the entire original dataset. Smoothing was not enabled as we found no significant changes in performance when applied. The M5PrimeLab toolbox implementation was used.

3.3 Dataset

The descriptor set was calculated on 137 of the 144 complexes described in Kastritis *et al.* (2011). Of the omitted complexes, 1UUG, 1IQD and 1NSN were removed as only upper limits to their affinity were available, and 1DE4, 1M10, 1NCA and 1NB5 due to difficulties deriving a full feature set. All post-translational modifications were reverted back to their precursor amino acids. The descriptor and affinities values were normalized in the range [0,1], and are given in Supplementary Table S2. The 73 rigid complexes are defined as those with an interface C_α RMSD <1 Å, and the remaining 64 are classified as flexible.

As the dataset is derived from the docking benchmark 4.0, redundancy is alleviated at the protein family level as described in Hwang *et al.* (2010) for all but nine pairs of complexes. This ensures that the leave-one-out cross-validation scheme does not overestimate predictive ability by having a homologous complex of similar affinity in the training set. Briefly, if two proteins in one complex are in the same SCOP family (Murzin *et al.*, 1995) as two proteins in another, then these are deemed redundant and are not included in the dataset. The remaining nine pairs of homologous complexes are cognate/non-cognate pairs, of similar sequence but with large difference in binding affinity, and thus are not a source of repeat example bias (Kastritis *et al.*, 2011).

A subset of 57 complexes, called the validated set, was used to train the learners. These complexes, which appears in Supplementary File S3, are those whose binding affinities are deemed to be high confidence. They were selected on the basis that either their binding affinity has been measured by more than one group / experimental technique and that the two values are within 2 kcal/mol of each other or that their value is corroborated by another measure such as a Michaelis constant. The validated set contains 29 flexible complexes and retains the diversity of the complete benchmark, consisting of 3 antibody/antigen complexes, 16 enzyme–inhibitor complexes, 5 enzyme–substrate complexes, 5 other complexes with enzymes, 8 complexes containing G-proteins, 7 receptor–ligand complexes and 13 miscellaneous complexes.

3.4 Method validation and comparison

All correlations are reported as the Pearson's product–moment correlation coefficient, calculated as the covariance of the two variables divided by the product of their SD. This parametric measure of correlation assesses the strength of linear dependence between two variables and is a widely accepted metric of the relationship between predicted and experimental binding affinities (Ferrara *et al.*, 2004; Kastritis and Bonvin, 2010; Kim and Skolnick, 2008; Marsden *et al.*, 2004; So and Karplus, 1999; Warren *et al.*, 2006).

The predictions used for calculating all correlations and *P*-values are those derived from leave-one-out cross-validation for the interactions in the validated set, where cross-validation is the outermost loop. Multiple training/validation runs showed high consistency in terms of feature usage and prediction accuracy. For the blind test set consisting of the 80 interactions not in the validated set, the predictions are calculated using the learners trained on the whole of the validated set. Results for the whole dataset are the unison of these predictions and the leave-one-out cross-validation predictions for the validated set. As the four learners either have no adjustable parameters other than those found in the training phase or these parameters are kept fixed at default without adjustment, all predictions are a blind test. The learners were combined to form a consensus model. The consensus model predictions were calculated simply as the arithmetic mean of the leave-one-out cross-validation predictions and thus also constitute a blind test. Model performance is reported as the correlation between the leave-one-out cross-validation predictions and the experimental binding free energies. The *P*-values for correlation significance were derived for the four learners and the consensus model over the validated set, the whole set, the intersections of these with the flexible set and rigid set, as well at the intersections of the validated and whole sets with the dataset of Kastritis and Bonvin (2010). We found $P < 0.01$ for all cases and these are not further reported in the results. When comparing the correlation between the predictions for the four learners, leave-one-out cross-validation results are used.

For comparison, free energy predictions were also calculated using DComplex (Liu *et al.*, 2004) and the pair potential described by Su *et al.* (2009). Using the whole dataset, the distribution of residuals around the correlation was assessed for the consensus model and the two pair potentials; using a Shapiro–Wilk test, no strong evidence ($P < 0.05$) could be found against a normal distribution for any of these. One-tailed *P*-values for comparing the correlations of two sets of predictions, with null hypothesis that the same correlation strength is shown by both samples of pairs (Papoulis, 1990), were obtained by transforming the correlation coefficients using the Fisher r to z transformation and finding the difference as

$$z = \frac{z_1 - z_2}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}}. \quad (5)$$

The *P*-values for comparing the absolute and square errors of two models were obtained using the one-tailed Wilcoxon signed-rank test. Additionally, predictions for the 11 other scoring functions tested by Kastritis and Bonvin (2010) were obtained for the intersection of the datasets. These are the ROSETTA standard and interface energy scores (Gray *et al.*, 2003), AffinityScore (Audie and Scarlata, 2007), HADDOCK (de Vries *et al.*, 2007), FastContact (Camacho and Zhang, 2005), FireDock (Andrusier *et al.*, 2007), PyDock (Cheng *et al.*, 2007), ZRANK (Pierce and Weng, 2007), ATTRACT (May and Zacharias, 2008) and the PISA entropy and free energy scores (Krissinel and Henrick, 2007). Although the overlap with the current dataset (37) is quite small, inspection of residuals after regression of these data against the experimental affinities showed few outliers and no major deviations from Gaussian.

4 RESULTS AND DISCUSSION

The learners were trained and tested using the 57 complexes in the validated set and leave-one-out cross-validation. The results are shown in Figure 2. All the algorithms had correlations between 0.68 and 0.75 against the experimental data. The correlations between the predictions was evaluated; whilE the predictions of the RF and M5′ methods were correlated with the RBF method ($r = 0.87$ and $r = 0.86$, respectively), and highly correlated with each other ($r = 0.95$), the MARS model was less correlated with the RBF, RF and M5′ methods ($r = 0.65$, $r = 0.69$ and $r = 0.68$, respectively). This suggested that the models may be detecting different aspects of the descriptor set, and a consensus model was built in which the predicted affinity was the mean output of the four models. This model has a correlation of 0.77 with the experimental affinity, which is significantly higher than the potentials of Su *et al.* (2009) (PMF, $r = 0.51$, $P < 0.01$) and Liu *et al.* (2004) (DFIRE, $r = 0.44$, $P < 0.01$). The consensus model also performs significantly better in terms of lower absolute error ($P_{\mathrm{PMF}} < 0.01$, $P_{\mathrm{DFIRE}} < 0.01$) and square error ($P_{\mathrm{PMF}} < 0.01$, $P_{\mathrm{DFIRE}} < 0.01$). The methods were also evaluated on the 80 complexes from the non-validated set. These predictions were amalgamated with the leave-one-out cross-validated predictions for the validated set, to cover all 137 complexes

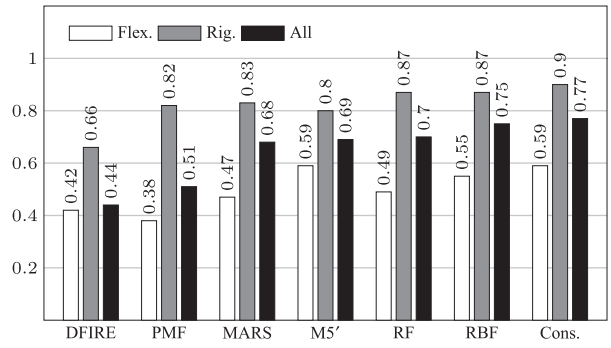

Fig. 2. Model performance for the 57 complexes in the validated set; correlation between the experimental and predicted binding affinities for the learners and their consensus, using leave-one-out cross-validation. The potentials of Liu *et al.* (2004) (DFIRE) and Su *et al.* (2009) (PMF) are also shown for comparison.

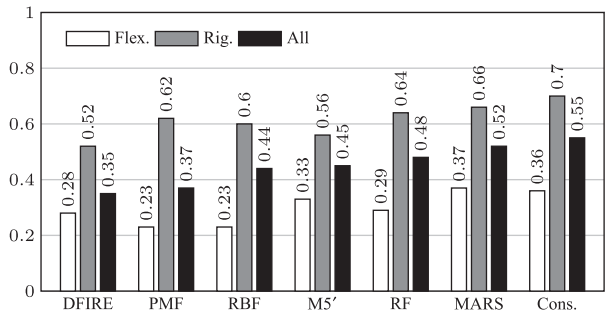

Fig. 3. Model performance for the 137 complexes in the whole benchmark; correlation between the experimental and predicted binding affinities for the learners and their consensus. The potentials of Liu *et al.* (2004) (DFIRE) and Su *et al.* (2009) (PMF) are also shown for comparison.

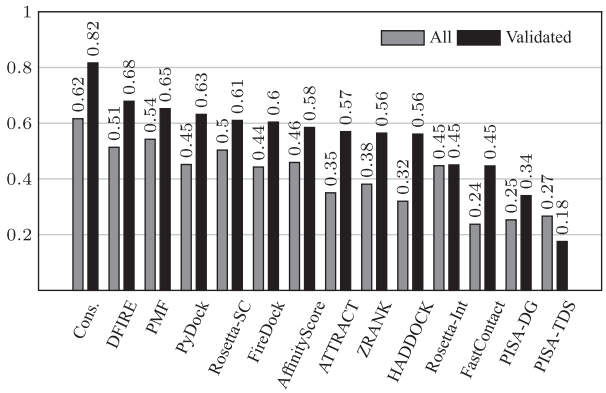

Fig. 4. Performance of the consensus model on the 37 complexes in the intersection between the dataset of Kastritis and Bonvin (2010) and the benchmark (All), and the 14 in the intersection with the validated set (Validated). Leave-one-out cross-validation is used for the interactions which intersect the validated set. Correlations for a number of other energy functions are also shown (see Section 3.4).

in the benchmark. The correlation between these predictions and the experimental data is shown in Figure 3. A large drop in predictive power is observed relative to the results for the validated set, despite having a similar proportion of non-rigid cases and interaction types. This highlights the importance of validating data with multiple experiments using different setups and techniques; sometimes different authors give widely varying affinities for the same complex, such as the interaction between ubiquitin and UCH-L3, for which Hirayama *et al.* (2007) report an affinity of tens of nanomolar, while Reyes-Turcu and Wilkinson (2009) claim the affinity is in the range of hundreds of micromolar. Nevertheless, the consensus model still performed better than PMF and DFIRE in terms of correlation ($P_{\mathrm{PMF}}=0.03$, $P_{\mathrm{DFIRE}}=0.02$), absolute error ($P_{\mathrm{PMF}}<0.01$, $P_{\mathrm{DFIRE}}<0.01$) and square error ($P_{\mathrm{PMF}}<0.01$, $P_{\mathrm{DFIRE}}<0.01$). The correlations for a number of other energy functions was also evaluated on a subset of the benchmark, as shown in Figure 4. The consensus model outperformed all methods tested.

4.1 The Learners

Upon RBF regression, the linear combination fits the data tightly with a correlation of 0.99 and when leave-one-out cross-validated, the model produces a correlation of 0.75. The coefficient distribution, which appears in Supplementary File S4, is as expected, with the highest affinity interactions destabilizing complexes far in feature space and *vice versa*. There are, however, a few exceptions, such as 1R0R which has a stabilizing effect on the complexes distant from it despite having high affinity, and 2OOB and 1AK4, which have very low affinity yet have low regression coefficients.

The RF method was constructed and tested, resulting in a leave-one-out cross-validated correlation of 0.70. For each complex, the five most relevant features were determined. A feature calculated using the unbound structure appeared as one of these for 61% of the flexible complexes, compared with only 17% of the rigid interactions, showing that the trees are selecting relevant features. Further, an ensemble descriptor appeared in the top five for 67% of the complexes. A descriptor importance measure, the mean decrease in mean square error upon permutation [as described by Breiman (2001)], is shown in Supplementary File S5 for the top 20 descriptors. Most significant are the hydrophobic burial term ACE12_HYDR, the London dispersion term ROS_FA_ATR and the Van der Waals term ACE22_VDW. Also prominent is the vibrational entropy change term S_VIB, interface packing term NSC, H-bonding terms and statistical pair potentials, some of which are calculated using ensembles.

Construction and evaluation of the M5′ regression trees resulted in a leave-one-out cross-validated correlation of 0.69. The trees, along with feature usage counts, are shown in Supplementary File S6. Five of the descriptors were used four times, the maximum allowed in our scheme; the interface packing score NSC, the London dispersion energy ROS_FA_ATR, the residue level pair potential ROS_CG_BETA, the atomistic pair potential OPUS_PSP_ENS calculated over structural ensembles and ACE_19_COUL_UB, the Coulombic energy change due to conformational rearrangement. Descriptors selected three times include other statistical pair potentials and H-bonding terms.

Using leave-one-out cross-validation, the MARS learner predicts the experimental binding energies with a correlation of 0.68. Trained on the validated set, the algorithm terminated with 14 basis functions which use a total of 10 descriptors. The functions and descriptor importance estimates in the form of function SDs and generalized cross-validation scores, as described by Friedman (1991), appear in Supplementary File S7. The most important term according to both metrics is the London dispersion term ROS_FA_ATR. Its contribution to the binding free energy is linear over the range covered by most of the complexes. However, a hinge function at -81.7 kcal/mol results in the destabilization of complexes with dispersion energy below this, such as the outlier 2OZA which has lower affinity than expected from its large interface due to disorder to order transitions occurring in a loop and at the C-terminal region. The second most significant descriptor is the vibrational entropy term S_VIB. At low values, its contribution is approximately zero, but becomes linear for higher values. This is consistent with the interpretation that, because this descriptor is approximate (Carrington and Mancera, 2004), the learner is choosing to use it when its contribution to the binding energy is sufficient to outweigh the noise it introduces. The other descriptors include a desolvation term, statistical pair potentials and H-bonding potentials for both intermolecular and conformational energy.

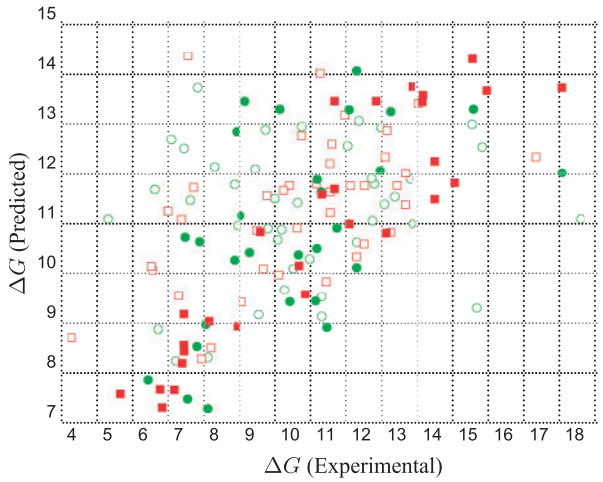

Fig. 5. Scatter plot for predicted and experimental affinities. Flexible (green circles) and rigid (red squares) proteins are shown. Leave-one-out cross-validated values for the validated set are highlighted in solid.

4.2 Conformational change

The predictions were evaluated separately for the rigid and flexible complexes. The predictions for the rigid interactions were highly accurate ($r_{val} = 0.90$). The RMSE for these complexes in the validated set is 1.67 kcal/mol, which is within the range of variation expected due to experimental errors and unaccounted environmental factors, around 1.4–2.0 kcal/mol (Kastritis *et al.*, 2011). However, the correlation for the flexible complexes ($r_{val} = 0.59$) was significantly worse ($P < 0.01$), as is the case with all the other methods considered here. Nevertheless, the consensus model does predict the affinities of the flexible complexes with better accuracy than the PMF and DFIRE methods in terms of both absolute error ($P_{PMF} < 0.01$, $P_{DFIRE} < 0.01$) and square error ($P_{PMF} < 0.01$, $P_{DFIRE} < 0.01$).

The experimental and consensus predictions are shown in Figure 5 and Supplementary file S8. It is clear from the sparse bottom right-hand corner of this plot that the consensus model overestimates binding affinity more frequently than it underestimates, and that this effect is most pronounced for the flexible cases. This suggests that even greater accuracy could be achieved with better modelling of phenomena relating to destabilization associated with conformational changes, such as entropy changes due to increased steric hindrance upon binding.

4.3 Descriptor subsets

In order to determine the extent to which the descriptors relating to ensembles and conformational changes influence the predictions, the consensus model was retrained using subsets of the descriptors and re-evaluated. The first descriptor subset contained all the descriptors that were not calculated using either information derived from the unbound structures or the structural ensembles. While still superior to the DFIRE and PMF predictions, the predictions using only these basic features ($r = 0.68$, RMSE = 2.2 kcal/mol) were less accurate than when all the descriptors were used ($r = 0.77$, RMSE = 1.93 kcal/mol). As expected, most of the drop in accuracy was due to the flexible cases ($r = 0.45$, RMSE = 2.59 kcal/mol versus $r = 0.59$, RMSE = 2.16 kcal/mol), rather than the rigid cases ($r = 0.90$, RMSE = 1.75 kcal/mol versus $r = 0.90$, RMSE = 1.67 kcal/mol).

When the descriptors relating to the ensembles, except those relating to the unbound structure, are included in addition to the basic features, a drop in accuracy is observed ($r = 0.61$, RMSE = 2.51 kcal/mol), both for the rigid ($r = 0.82$, RMSE = 2.02 kcal/mol) and flexible ($r = 0.37$, RMSE = 2.91 kcal/mol) cases, although this can be mostly attributed to poor performance of the MARS learner.

When the features relating to the unbound structures, except for those calculated using ensembles, are used in addition to the basic features, an improvement in performance is observed ($r = 0.74$, RMSE = 2.04 kcal/mol). This is due to the improved predictions for the flexible complexes ($r = 0.47$, RMSE = 2.31 kcal/mol), rather than the rigid complexes ($r = 0.91$, RMSE = 1.72 kcal/mol), which perform equally well.

A fourth descriptor subset was evaluated which contained only the basic features and those calculated using both the unbound structures and ensembles. The performance ($r = 0.61$, RMSE = 2.08 kcal/mol) was intermediate between the previous model and the model trained on all features. Again, this improvement is attributed to the flexible complexes ($r = 0.53$, RMSE = 2.34 kcal/mol) rather than the rigid ones ($r = 0.91$, RMSE = 1.76 kcal/mol).

These results, from models trained using descriptor subsets, show that the inclusion of descriptors derived from the unbound structures improves the performance for the flexible complexes without diminishing the accuracy for the rigid complexes. This improvement is further enhanced when used in combination with structural ensembles, despite the ensembles not enhancing the consensus model when information derived from the unbound structures is omitted.

4.4 Applications and future work

Binding free energy calculations have a number of important applications in structural bioinformatics. Currently, 'one function fits all' does not apply and energy models need to be parameterized for purpose (Kastritis and Bonvin, 2010) and a trade-off between accuracy and computational expense is inevitable. Although these applications are beyond the scope of the current work, the adaptation of the presented methods shall be the focus of future research. Combined with side chain remodelling, energy models can be used for mutational $\Delta\Delta G$ calculation, for interface engineering or determining the functional consequences of mutation [e.g. Ben-Shimon and Eisenstein (2010)]. Interaction potentials are also central to generating and ranking docked structures (Halperin *et al.*, 2002), the rigid body simulation of encounter complex formation (Elcock *et al.*, 2001; Li *et al.*, 2010) and distinguishing biological assemblies from crystal contacts (Krissinel and Henrick, 2007). A recent application for which a precursor to the presented model was used is in the selection of *de novo* designed interactions found using the method of Fleishman *et al.* (2011b). The energy function was applied to the selection of refined true interactions among ostensibly indistinguishable designed interactions. Using an appropriate energy threshold, it was capable of classifying 207 structures with 88% accuracy, and was further validated in a blind test by correctly classifying nine designs, eight of which failed to show binding function when tested experimentally and the other

of which did. This work will be published shortly Fleishman *et al.* (2011a).

5 CONCLUSION

The recent publication of a large set of binding free energies has allowed an unprecedented opportunity to construct and evaluate predictive energy models. We have derived a large molecular descriptor set covering many aspects of protein binding including unbound to bound transitions, desolvation effects and entropy changes. This descriptor set was used to train and evaluate four binding free energy models. These models, and their consensus, can predict binding energies with greater accuracy than previous empirical functions, even when the ensemble and conformational descriptors are omitted. For the rigid proteins tested, the consensus model can predict the affinities to within the accuracy expected from experimental and environmental uncertainties.

The internal working of the learners also allow the evaluation of feature importance. This shows that the London dispersion energy, vibrational entropy, statistical potentials and interface packing terms have the most predictive value, and that hydrophobic burial and hydrogen bonding are also important determinants of binding affinity. Additionally, retraining the models using descriptor subsets revealed that knowledge of unbound to bound conformational changes improves the performance of the consensus model, and that the greatest improvement is attained when energies are averaged over conformational ensembles.

This work also highlights the importance of using experimental data validated by multiple experiments, and provides a benchmark of high-quality affinities for a diverse range of complexes.

ACKNOWLEDGEMENT

The authors thank Prof., Alexandre Bonvin and Panagiotis Kastritis for providing the calculated affinities in Kastritis and Bonvin (2010).

Funding: Cancer Research UK.

Conflict of Interest: none declared.

REFERENCES

Aloy,P. and Russell,R.B. (2006) Structural systems biology: modelling protein interactions. *Nat. Rev. Mol. Cell Biol.*, **7**, 188–197.

Andrusier,N. *et al.* (2007) FireDock: fast interaction refinement in molecular docking. *Proteins*, **69**, 139–159.

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

Bai,H. *et al.* (2011) Predicting kinetic constants of protein-protein interactions based on structural properties. *Proteins*, **79**, 720–734.

Bas,D.C. *et al.* (2008) Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins*, **73**, 765–783.

Beltrao,P. *et al.* (2007) Structures in systems biology. *Curr. Opin. Struct. Biol.*, **17**, 378–384.

Ben-Shimon,A. and Eisenstein,M. (2010) Computational mapping of anchoring spots on protein surfaces. *J. Mol. Biol.*, **402**, 259–277.

Benedix,A. *et al.* (2009) Predicting free energy changes using structural ensembles. *Nat. Methods*, **6**, 3–4.

Bougouffa,S. and Warwicker,J. (2008) Volume-based solvation models out-perform area-based models in combined studies of wild-type and mutated protein-protein interfaces. *BMC Bioinformatics*, **9**, 448.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Brooks,B.R. *et al.* (2009) CHARMM: the biomolecular simulation program. *J. Comput. Chem.*, **30**, 1545–1614.

Camacho,C.J. and Zhang,C. (2005) FastContact: rapid estimate of contact and binding free energies. *Bioinformatics*, **21**, 2534–2536.

Carrington,B.J. and Mancera,R.L. (2004) Comparative estimation of vibrational entropy changes in proteins through normal modes analysis. *J. Mol. Graph. Model.*, **23**, 167–174.

Chaudhury,S. *et al.* (2010) PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, **26**, 689–691.

Cheng,T.M. *et al.* (2007) pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins*, **68**, 503–515.

Cho,K.I. *et al.* (2009) A feature-based approach to modeling protein-protein interaction hot spots. *Nucleic Acids Res.*, **37**, 2672–2687.

Darnell,S.J. *et al.* (2007) An automated decision-tree approach to predicting protein interaction hot spots. *Proteins*, **68**, 813–823.

de Groot,B.L. *et al.* (1997) Prediction of protein conformational freedom from distance constraints. *Proteins*, **29**, 240–251.

de Vries,S.J. *et al.* (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins*, **69**, 726–733.

Dell'Orco,D. (2009) Fast predictions of thermodynamics and kinetics of protein-protein recognition from structures: from molecular design to systems biology. *Mol. Biosyst.*, **5**, 323–334.

Dolinsky,T.J. *et al.* (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.*, **35**, W522–W525.

Elcock,A.H. *et al.* (2001) Computer simulation of protein-protein interactions. *J. Phys. Chem. B*, **105**, 1504–1518.

Feng,Y. *et al.* (2010) Potentials 'R' Us web-server for protein energy estimations with coarse-grained knowledge-based potentials. *BMC Bioinformatics*, **11**, 92.

Ferrara,P. *et al.* (2004) Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.*, **47**, 3032–3047.

Fleishman,S.J. *et al.* (2011a) Community-wide assessment of protein-interface modeling suggests improvements to design methodology. *J. Mol. Biol.*, in press.

Fleishman,S.J. *et al.* (2011b) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, **332**, 816–821.

Friedman,J.H. (1991) Multivariate adaptive regression splines. *Ann. Stat.*, **19**, 1–67.

Gilson,M.K. and Zhou,H.X. (2007) Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.*, **36**, 21–42.

Gray,J.J. *et al.* (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.*, **331**, 281–299.

Halperin,I. *et al.* (2002) Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins*, **47**, 409–443.

Hardy,R.L. (1971) Multiquadric equations of topography and other irregular surfaces. *J. Geophys. Res.*, **76**, 1905–1915.

Hirayama,K. *et al.* (2007) Identification of novel chemical inhibitors for ubiquitin C-terminal hydrolase-L3 by virtual screening. *Bioorg. Med. Chem.*, **15**, 6810–6818.

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

Huang,Y.J. *et al.* (2008) Targeting the human cancer pathway protein interaction network by structural genomics. *Mol. Cell Proteomics*, **7**, 2048–2060.

Hwang,H. *et al.* (2010) Protein-protein docking benchmark version 4.0. *Proteins*, **78**, 3111–3114.

Jiang,L. *et al.* (2002) Potential of mean force for protein-protein interaction studies. *Proteins*, **46**, 190–196.

Jiang,L. *et al.* (2005) A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins*, **58**, 893–904.

Kastritis,P.L. and Bonvin,A.M. (2010) Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *Corrigendum. J. Proteome Res.*, **9**, 2216–2225.

Kastritis,P.L. *et al.* (2011) A structure-based benchmark for protein-protein binding affinity. *Prot. Sci.*, **20**, 482–491.

Keskin,O. *et al.* (2005) Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach. *Phys. Biol.*, **2**, 24–35.

Kiel,C. *et al.* (2008) Analyzing protein interaction networks using structural information. *Annu. Rev. Biochem.*, **77**, 415–441.

Kim,R. and Skolnick,J. (2008) Assessment of programs for ligand binding affinity prediction. *J. Comput. Chem.*, **29**, 1316–1331.

Kollman,P. (1993) Free energy calculations: applications to chemical and biochemical phenomena. *Chem. Rev.*, **93**, 2395–2417.

Kortemme,T. and Baker,D. (2004) Computational design of protein-protein interactions. *Curr. Opin. Chem. Biol.*, **8**, 91–97.

Krissinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774–797.

Krystek,S. *et al.* (1993) Affinity and specificity of serine endopeptidase-protein inhibitor interactions. Empirical free energy calculations based on X-ray crystallographic structures. *J. Mol. Biol.*, **234**, 661–679.

Kumar,M. *et al.* (2010) Structure-based in silico design of a high-affinity dipeptide inhibitor for novel protein drug target Shikimate kinase of Mycobacterium tuberculosis. *Chem. Biol. Drug Des.*, **76**, 277–284.

Kuncheva,L.I. (2002) A theoretical study on six classifier fusion strategies. *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**, 281–286.

Kundrotas,P.J. *et al.* (2010) GWIDD: genome-wide protein docking database. *Nucleic Acids Res.*, **38**, D513–D517.

Li,X. *et al.* (2010) Detection and refinement of encounter complexes for protein-protein docking: taking account of macromolecular crowding. *Proteins*, **78**, 3189–3196.

Liu,S. *et al.* (2004) A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins*, **56**, 93–101.

Ma,X.H. *et al.* (2002) A fast empirical approach to binding free energy calculations based on protein interface information. *Protein Eng.*, **15**, 677–681.

Marsden,P.M. *et al.* (2004) Predicting protein-ligand binding affinities: a low scoring game? *Org. Biomol. Chem.*, **2**, 3267–3273.

May,A. and Zacharias,M. (2008) Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking. *Proteins*, **70**, 794–809.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Nauchitel,V. *et al.* (1995) Solvent accessibility as a predictive tool for the free energy of inhibitor binding to the HIV-1 protease. *Protein Sci.*, **4**, 1356–1364.

Novotny,J. *et al.* (1989) On the attribution of binding energy in antigen-antibody complexes McPC 603, D1.3, and HyHEL-5. *Biochemistry*, **28**, 4735–4749.

Opitz,D. and Maclin,R. (1999) Popular ensemble methods: an empirical study. *J. Artif. Intell. Res.*, **11**, 169–198.

Papoulis,A. (1990) *Probability and Statistics*. Prentice Hall, Englewood Cliffs, N.J.

Pierce,B. and Weng,Z. (2007) ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins*, **67**, 1078–1086.

Quinlan,J.R. (1992) Learning with continuous classes. In *Proceeding 5th Australian Joint Conference on Artificial Intelligence*. World Scientific, Singapore, pp. 343–348.

Rao,G.S. and Kumar,M. (2008) Structure-based design of a potent and selective small peptide inhibitor of Mycobacterium tuberculosis 6-hydroxymethyl-7, 8-dihydropteroate synthase: a computer modelling approach. *Chem. Biol. Drug Des.*, **71**, 540–545.

Reyes-Turcu,F.E. and Wilkinson,K.D. (2009) Polyubiquitin binding and disassembly by deubiquitinating enzymes. *Chem. Rev.*, **109**, 1495–1508.

Reynolds,C. *et al.* (2009) ProtorP: a protein-protein interaction analysis server. *Bioinformatics*, **25**, 413–414.

Sharabi,O. *et al.* (2011) Optimizing energy functions for protein-protein interface design. *J. Comput. Chem.*, **32**, 23–32.

So,S.S. and Karplus,M. (1999) A comparative study of ligand-receptor complex binding affinity prediction methods based on glycogen phosphorylase inhibitors. *J. Comput. Aided Mol. Des.*, **13**, 243–258.

Su,Y. *et al.* (2009) Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction. *Protein Sci.*, **18**, 2550–2558.

Tuncbag,N. *et al.* (2010) HotPoint: hot spot prediction server for protein interfaces. *Nucleic Acids Res.*, **38**, W402–W406.

Vajda,S. *et al.* (1994) Effect of conformational flexibility and solvation on receptor-ligand binding free energies. *Biochemistry*, **33**, 13977–13988.

Wallqvist,A. *et al.* (1995) A preference-based free-energy parameterization of enzyme-inhibitor binding. Applications to HIV-1-protease inhibitor design. *Protein Sci.*, **4**, 1881–1903.

Wang,Y. and Witten,I. (1997) Induction of model trees for predicting continuous classes. In *Proceedings of the European Conference on Machine Learning Poster Papers*. University of Economics, Faculty of Informatics and Statistics, Prague, Czech Republic, pp. 128–137.

Warren,G.L. *et al.* (2006) A critical assessment of docking programs and scoring functions. *J. Med. Chem.*, **49**, 5912–5931.

Weng,Z. *et al.* (1997) Empirical free energy calculation: comparison to calorimetric data. *Protein Sci.*, **6**, 1976–1984.

Xu,D. *et al.* (1997) Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. *J. Mol. Biol.*, **265**, 68–84.

Zhang,C. *et al.* (1997) Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.*, **267**, 707–726.

Zhang,C. *et al.* (2005) A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J. Med. Chem.*, **48**, 2325–2335.

Zhang,Q.C. *et al.* (2010) Protein interface conservation across structure space. *Proc. Natl Acad. Sci. USA*, **107**, 10896–10901.

Zhou,H. and Zhou,Y. (2002) Stability scale and atomic solvation parameters extracted from 1023 mutation experiments. *Proteins*, **49**, 483–492.

Zhu,X. and Mitchell,J.C. (2011) KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins*, **79**, 2671–2683.