BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 9 2013, pages 1127–1133
doi:10.1093/bioinformatics/btt121

Structural bioinformatics

Advance Access publication March 7, 2013

Specificity and affinity quantification of protein–protein interactions

Zhiqiang Yan[1], Liyong Guo[2], Liang Hu[2] and Jin Wang[1,3,*]

[1]Department of Chemistry and Physics, State University of New York at Stony Brook, Stony Brook, NY 11794-3400, USA, [2]College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China and [3]State Key Laboratory of Electroanalytical Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun, Jilin 130022, China

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Most biological processes are mediated by the protein–protein interactions. Determination of the protein–protein structures and insight into their interactions are vital to understand the mechanisms of protein functions. Currently, compared with the isolated protein structures, only a small fraction of protein–protein structures are experimentally solved. Therefore, the computational docking methods play an increasing role in predicting the structures and interactions of protein–protein complexes. The scoring function of protein–protein interactions is the key responsible for the accuracy of the computational docking. Previous scoring functions were mostly developed by optimizing the binding affinity which determines the stability of the protein–protein complex, but they are often lack of the consideration of specificity which determines the discrimination of native protein–protein complex against competitive ones.

Results: We developed a scoring function (named as SPA-PP, specificity and affinity of the protein–protein interactions) by incorporating both the specificity and affinity into the optimization strategy. The testing results and comparisons with other scoring functions show that SPA-PP performs remarkably on both predictions of binding pose and binding affinity. Thus, SPA-PP is a promising quantification of protein–protein interactions, which can be implemented into the protein docking tools and applied for the predictions of protein–protein structure and affinity.

Availability: The algorithm is implemented in C language, and the code can be downloaded from http://dl.dropbox.com/u/1865642/Optimization.cpp.

Contact: jin.wang.1@stonybrook.edu

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

Received on September 22, 2012; revised on February 14, 2013; accepted on March 5, 2013

1 INTRODUCTION

Proteins often function only when they are assembled into complexes (Sali *et al.*, 2003). The formation of highly efficient and specific protein complexes, particularly protein–protein complexes, is a fundamental process in the cell. The structures of the protein–protein complexes are the principal components of a systems-level description of genome-wide protein networks, which rely on the coordinated and tightly regulated activities of interacting proteins (Kim *et al.*, 2006). The protein–protein interactions play a key role in various mechanisms of protein functions (Jones and Thornton, 1996; Ortuso *et al.*, 2006) and become attractive targets for therapeutic intervention (Loregian and Palù, 2005). Thus, determination of the protein–protein structures and insights into their interactions are increasingly important in the post-genomic era.

Although the individual protein structures are increasingly well determined, protein assemblies are still poorly represented in the Protein Data Bank (Chiu *et al.*, 2006; Dutta and Berman, 2005; Steven and Baumeister, 2008). This is because the experimental determination of assembly compositions and structures generally requires the combination of a number of the experimental techniques and theoretical approaches to maximize completeness, accuracy and resolution (Sali *et al.*, 2003). As an alternative, computational docking approaches provide a practical mean to predict the structures of multi-protein assemblies starting from the individual components (Janin, 2002). Computational docking procedure usually incorporates two steps (Halperin *et al.*, 2002), initial sampling of the configurational space of the interacting proteins to generate docking poses and pose scoring to select the putative native (or near-native) protein–protein conformations. Comparative assessment of the current protein–protein docking and scoring approaches shows that efficient sampling strategies may generate the library of poses, including the native or near-native poses. However, the scoring function often fails to discriminate the native or near-native poses against other non-native poses and can not always produce reliable estimation of the binding affinity of the complex (Kastritis and Bonvin, 2010; Lensink *et al.*, 2007; Lensink and Wodak, 2010), suggesting there is substantial room for the improvement of the scoring functions.

The development strategies of previous scoring functions for protein–protein docking (Moreira *et al.*, 2010), whether physics-based, empirical or knowledge-based, mainly focus on the prediction of binding affinity (or free energy), which determines the stability of the complex. However, high affinity does not guarantee high specificity, which is controlled by either partner binding to other proteins discriminatively (Bolon *et al.*, 2005; Grigoryan *et al.*, 2009; Havranek *et al.*, 2003; Janin, 1995; Kortemme *et al.*, 2004; Shifman and Mayo, 2003). The development strategies of previous scoring functions are often lack of the

*To whom correspondence should be addressed.

considerations of the specificity, which is the discrimination of the specific native protein–protein complex against the competitive ones. According to the Boltzman distribution ($P \sim \exp[-F/KT]$), the equilibrium population is exponentially dependent on the binding free energy. A gap in binding free energy or affinity will lead to significant population discrimination between the specific complex and alternative ones. As such, introducing the consideration of specificity into the computational design of protein–protein interactions has achieve a few successful applications (Bolon *et al.*, 2005; Grigoryan *et al.*, 2009; Havranek *et al.*, 2003; Kortemme *et al.*, 2004; Shifman and Mayo, 2003). These work designed interactions that seek to stabilize the desired structures and also destabilize the competitive structures, as the specificity-related interactions lie in the binding patches constituting the interface of the complex (Malod-Dognin *et al.*, 2012). Thus, to develop a scoring function, the strategy should satisfy the requirement that the stability of the specific complex is maximized, whereas the stability of competing complexes is minimized, which can guarantee both the stability and the specificity for the specific complex.

The conventional definition of specificity is the preference of a protein ligand specifically binding to a protein receptor over other competitive alternatives (Bolon *et al.*, 2005; Grigoryan *et al.*, 2009; Havranek *et al.*, 2003; Janin, 1995;Kortemme *et al.*, 2004; Shifman and Mayo, 2003). The definition is clear, although in practice, the quantification of the conventional specificity still remains challenging (Grigoryan *et al.*, 2009). The conventional specificity requires comparison of the affinities of all the different protein receptors with the same protein ligand (Fig. 1a). This makes the practical quantification of the specificity impossible, as the receptor universe is huge, and the information is often incomplete on the competitive alternatives. To circumvent the challenge, we have proposed a novel concept named as intrinsic specificity, which had been successfully applied in the receptor–ligand (where the ligand is small molecule and the receptor is the protein) system (Wang and Verkhivker, 2003; Wang *et al.*, 2007; Yan and Wang, 2012). Here, we apply the concept to the protein–protein system. The intrinsic specificity refers to the preference in affinity of a protein ligand binding to its protein receptor with a preferred pose over other poses (Fig. 1b). Imagining the N- and C-terminus of multiple protein receptors is linked together, resulting in an effective single large receptor; then the conventional specificity can be transformed to the intrinsic specificity that a ligand binds to the large receptor covering the whole universe of proteins with a preferred pose over other poses (Fig. 1). The equivalence of conventional specificity and intrinsic specificity is under the assumption that the protein receptor is sufficiently large to represent the whole universe of proteins and their interfaces. In reality, there is only finite number of protein folds (~1300) in nature (Andreeva *et al.*, 2008). Thus, the protein receptor may not need to be infinitely large to represent the whole universe of proteins. A reasonable size of protein surface may provide a few of the diversified interfacial interactions presumably formed by the whole universe of proteins.

The process of protein–protein binding can be physically quantified and visualized as a funnel-like energy landscape towards the native binding state with local roughness along the binding paths (Bryngelson *et al.*, 1995; Dominy and Shakhnovich, 2004; Janin, 1996; Levy *et al.*, 2004; Liu *et al.*, 2004*b*; Miller and Dill, 1997; Rejto and Verkhivker, 1996; Tsai *et al.*, 1999; Wang and Verkhivker, 2003; Wang *et al.*, 2007). The native conformation of the complex is the conformation with the lowest binding energy. The energies of the conformation ensemble follow a statistical Gaussian-like distribution. According to the theory of energy landscape, the intrinsic specificity ratio ($ISR = \frac{\delta E}{\Delta E \sqrt{S}}$, where δE is the energy gap between the energy of native conformation and the average energy of conformation ensemble, ΔE is the energy roughness or the width of the energy distribution of the conformation ensemble and S is the conformational entropy) can be defined to quantify the magnitude of intrinsic specificity. The ISR indicates the capability of discriminating the native binding conformation from other non-native ones for a particular protein ligand on its receptor. The ISR can be readily quantified with computationally generated non-native poses (decoys). Therefore, ISR physically provides a quantitative measure of the thermodynamic specificity without evaluating the conventional specificity through exploring the whole set of receptor universe.

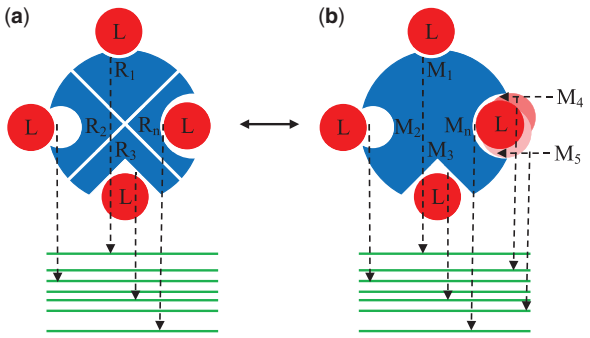

Fig. 1. Illustration of the transformation between conventional specificity and intrinsic specificity. (**a**) The same protein ligand (red, L) binding with multiple protein receptors (blue, R_1 to R_n), showing the conventional specificity as the gap in binding affinity of the protein ligand binding to the specific protein receptor (R_n) in discrimination against other protein receptors. The binding affinities are represented with corresponding energy spectrum (Green). (**b**) The same protein ligand (red, L) binding on a large protein receptor (blue) with multiple binding modes (M_1 to M_n), showing the intrinsic specificity as the gap in binding affinity of the native binding mode (M_n) in discrimination against other binding modes. The large protein receptor can be thought as the multiple different receptors linked together

In this work, we designed a strategy for developing the scoring function of protein–protein interactions to maximize the affinity and specificity simultaneously. The idea of the strategy is to adjust the statistical knowledge-based potentials of atom pairs by iteration until the scoring function can discriminate the native binding pose against the decoys. The developing procedure is shown in Supplementary Figure S1. We tested the derived scoring function of protein–protein interactions (named as SPA-PP) via the performance on the identification of native binding pose and the prediction of binding affinity, and we made comparisons with the results calculated by the widely used software RosettaDock and two other existing scoring functions.

2 MATERIALS AND METHODS

2.1 Preparation of the datasets

The dataset of protein–protein complexes for developing SPA-PP was extracted from the Dockground resource (Douguet *et al.*, 2006), which provides a dynamic generation of dataset of protein–protein complexes based on the user-defined filtering criteria. To drive a high quality set of pairwise non-redundant protein–protein complexes of training dataset and testing dataset, a series of filtering criteria were composed on the Dockground resource. Only the dimeric complexes solved by X-ray diffraction with resolutions better than 2.5 Å were selected. The sequence identity (<70%) between each two complexes was used to discard redundant structures. The dimeric structures with true biological interfaces were kept by checking the head lines of the pdb files with remarks 'AUTHOR DETERMINED BIOLOGICAL UNIT: DIMERIC' or 'SOFTWARE DETERMINED QUATERNARY STRUCTURE: DIMERIC' or both of them. The threshold number of interface residues were set to be >30. The resulting dataset contains 3045 dimeric protein–protein complexes, including 2648 homodimers and 397 heterodimers (Supplementary Table S1). This dataset was randomly partitioned into two equal subsamples, one subsample is assigned as the training dataset and the other one as the testing dataset (named as testing dataset1).

In addition to the aforementioned testing dataset1, a benchmark dataset for testing was also collected. It combines three available protein–protein benchmarks. The first benchmark is the non-redundant protein–protein docking benchmark version 4.0 (Hwang *et al.*, 2010), which contains 176 complexes. The second benchmark is the Dockground benchmark (Gao *et al.*, 2007), which contains 233 complexes. The third benchmark is an affinity benchmark (Kastritis *et al.*, 2011), which contains 144 complexes with measured dissociation constants. By keeping one entry if there are overlaps among the pdb lists of these three protein–protein benchmarks, the remaining 325 complexes constitute our benchmark dataset (Supplementary Table S2), which covers all available benchmarks of protein–protein complexes. For convenience, this benchmark dataset was named as testing dataset2. Both testing dataset1 and dataset2 were used for the validation of ability of binding pose prediction for SPA-PP, and among the testing dataset2, the 144 complexes with experimental measured affinities were used for the validation of the ability of binding affinity prediction for SPA-PP.

2.2 Preparation of docking decoys

To develop SPA-PP, an ensemble of decoys is needed to calculate the ISR and carry out the iteration algorithm. The RosettaDock v3.2.1 was taken as the structure optimizing and docking tool (Chaudhury *et al.*, 2011; Gray *et al.*, 2003) for both the training and testing complexes. Three steps were performed. First, each docking partner of the complex was prepared in isolation for optimizing their side-chain conformations before docking using the pre-packing protocol. Second, the pre-packed complexes were then relaxed and minimized with high resolution by the refinement protocol. Third, the refined structures were taken as the starting structures for the docking using the local docking perturbation protocol. The smaller protein was defined as the docking ligand in the complex, and the other was assigned as the receptor, which was kept fixed during docking. Thousand ligand orientations for each complex were generated by docking. Other docking parameters were set as default. The generated decoys are normally structured diversely with I_{rms} (interface C_α root-mean-square displacement between the native conformation and docked decoys) ranging from small to large values. Three example complexes with small, middle and large number of interfacial residues are shown in Supplementary Figure S2 to illustrate the structural diversity of the decoys.

2.3 Derivation of observed statistical potentials

The knowledge-based statistical scoring function needs a set of distance-dependent atom-pair potentials to quantify the interactions. In our work, the initial observed atom-pair potentials were directly derived from the Boltzmann relation widely used in the knowledge-based statistical potentials (Jiang *et al.*, 2002; Koppensteiner and Sippl, 1998; Su *et al.*, 2009; Zhang *et al.*, 2005*a*), which is

$$u_k^{obs}(r) = -K_B T \ln g_k^{obs}(r) \tag{1}$$

where $g_k^{obs}(r)$ is the observed pair distribution function, which can be calculated by

$$g_k^{obs}(r) = \frac{f_k^{obs}(r)}{f_k^{obs}(R)}. \tag{2}$$

$f_k^{obs}(r)$ is the observed number density of atom-pair k within a spherical shell between r and $r+\delta r$, and the $f_k^{obs}(R)$ is the number density within the sphere of the reference state where there are no interactions between atoms. The former can be directly extracted from the database of protein–protein complexes, whereas the later was obtained based on the approximation that the atom-pair k is uniformly distributed in the sphere of the reference state (Sippl, 1990). Respectively, they were calculated as

$$f_k^{obs}(r) = \frac{1}{M}\sum_m^M \frac{n_k^m(r)}{V(r)} \tag{3}$$

$$f_k^{obs}(R) = \frac{1}{M}\sum_m^M \frac{N_k^m}{V(R)} \tag{4}$$

where M is the total number (3045) of native protein–protein complexes extracted from the Dockground resource, $n_k^m(r)$ and N_k^m are the numbers of atom-pair k within the spherical shell and the reference sphere, respectively, for a given protein–protein complex m, where $N_k^m = \sum_r n_k^m(r)$. $V(r) = \frac{4}{3}\pi((r+\Delta r)^3 - r^3)$ and $V(R) = \frac{4}{3}\pi R^3$ are the volumes of the spherical shell and the reference sphere, respectively, where Δr is the bin size and R is the radius of sphere. Δr and R are set as 0.3 Å and 6.4 Å, respectively. There are 14 spherical shells with bin size 0.3 Å from the shortest radius 2.2 Å. Based on the definition of atom type by SYBYL (Clark *et al.*, 1989), 12 atom types were used to cover the heavy atoms involved in protein–protein interactions (Supplementary Table S3), these atom types were converted from PDB files by OpenBabel (Guha *et al.*, 2006). These atom types lead to 78 types of atom pairs for the protein–protein interactions. There are 1092 types of interaction pair by multiplying the number of atom pairs (78) and the number of shells (14).

2.4 Derivation of expected statistical potentials

The observed statistical potentials have its limitations, as the statistical potentials extracted from Equation (1) is not exactly the expected potentials that nature uses to stabilize the complexes (Thomas and Dill, 1996*a*). The origin of this problem lies in the construction of the reference state where the atom pairs are uniformly distributed and independent of each other (Sippl, 1990). In reality, the protein interactions involve the excluded volume, sequences and connectivity. Thus, the observed statistical potentials are always not equal to the expected potentials (Thomas and Dill, 1996*b*). To solve the reference state issue and improve the statistical potentials, earlier efforts (Goldstein *et al.*, 1992; Huang and Zou, 2008; Muegge and Martin, 1999; Thomas and Dill, 1996*a*; Zhang *et al.*, 2005*b*) taking different approaches have been devoted in optimizing the statistical potentials. An effective way is to take into account both native and non-native conformations (decoys) (Goldstein *et al.*, 1992; Huang and Zou, 2008; Thomas and Dill, 1996*a*) based on the energy landscape theory (Goldstein *et al.*, 1992; Levy *et al.*, 2004) that the native conformation should be sufficiently favored over alternative non-native structures

thermodynamically. However, none of the earlier works combined both the specificity and affinity to discriminate native conformation over non-native conformations of protein–protein docking. In this work, we considered the importance of both the affinity for stabilizing the native conformation and the specificity of discrimination over non-native conformations to optimize the statistical potentials. We used the iterative method (Huang and Zou, 2008; Thomas and Dill, 1996*a*) shown in the next subsection to realize the optimization.

Similarly as the observed statistical potentials, the expected statistical potentials are calculated as

$$u_k^{exp}(r) = -K_B T \ln g_k^{exp}(r) \tag{5}$$

where $g_k^{exp}(r)$ is the expected pair distribution function from all the native and non-native conformations, which is

$$g_k^{exp}(r) = \frac{f_k^{exp}(r)}{f_k^{exp}(R)}. \tag{6}$$

Given the population discrimination of the native and non-native conformations, the expected number density of atom-pair k was calculated with Boltzmann-averaged weighting over the ensemble of conformations (Huang and Zou, 2008; Thomas and Dill, 1996*a*), that is

$$f_k^{exp}(r) = \frac{1}{MN}\sum_m^M \sum_n^N \frac{n_k^{mn}(r) e^{(-\beta U_{mn})}}{V(r)} \tag{7}$$

$$f_k^{exp}(R) = \frac{1}{MN}\sum_m^M \sum_n^N \frac{N_k^{mn} e^{(-\beta U_{mn})}}{V(R)} \tag{8}$$

where M is the number of native complexes aforementioned, and N is the number of total conformations (1001, including the native conformation and decoys) for each complex m. n represents the nth generated decoy of the complex m. β is an arbitrary constant analogous to the inverse of temperature and set as 0.1. U_{mn} is the parameter that is assumed to be able to discriminate the naive conformation against decoys. As discussed, both the stability and specificity are required to form a specific functional complex. Thus, U_{mn} is designed as a combination parameter coupling the affinity and ISR, which is given by

$$U_{mn} = \gamma E_{mn} + \lambda_{mn} \tag{9}$$

where E_{mn} is the energy score of the protein–protein conformation (nth decoy of the complex m) by summing over all the expected interatomic pair potentials among the interface, representing the affinity of the protein–protein conformation. λ_{mn} is the ISR representing the specificity of the protein–protein conformation (Wang and Verkhivker, 2003; Wang *et al.*, 2007; Yan and Wang, 2012). γ is a parameter that balances the values of E_{mn} and λ_{mn} and is set as 0.01 because of the ratio of average E_{mn} and λ_{mn}. E_{mn} and λ_{mn} are calculated as

$$E_{mn} = \sum_k \sum_r t_k(r) u_k(r) \tag{10}$$

$$\lambda_{mn} = \alpha_m \frac{\delta E_{mn}}{\Delta E_{mn}} \tag{11}$$

$t_k(r)$ represents the occurrence times of the atom-pair interaction between the protein–protein interface. α_m is a scaling factor that accounts for the contribution of the entropy to the specificity ($\alpha_m = \frac{1}{\sqrt{S_m}}$, where S_m is the conformational entropy of the complex m) (Wang and Verkhivker, 2003). Here, α_m approximately depends on the number of interfacial residues ($\alpha_m \sim \sqrt{\frac{1}{n_{inter}}}$) of the native protein–protein conformation of the complex m. An interfacial residue is defined if any atom of this residue in one partner of the native protein–protein conformation is within 10 Å from the other partner. α_m normalizes the increase of ISR with the number of interfacial residues. δE_{mn} is the energy gap between the energy of a given conformation E_{mn} and the average energy of the conformation ensemble $<E_m^e>$, including the native conformation and all the decoys of the complex m, ΔE_{mn} is the energy roughness or the width of the energy distribution of conformation ensemble, namely, $\delta E_{mn} = E_{mn} - <E_m^e>$ and $\Delta E_{mn} = \sqrt{<(E_m^e)^2> - <E_m^e>^2}$, $<>$ means the average over the ensemble of conformations.

2.5 Optimization of scoring function

The basic idea of the iterative method here is to circumvent the inaccessible reference state problem by improving the statistical pair potentials until they are able to discriminate native binding mode from decoys. The expected potentials obtained from the ensemble of conformations generally are not equal to the potentials from the observed native conformations. The difference between the expected statistical potentials and the observed statistical potentials, as well as the iterative equation can be expressed by

$$\Delta u_k^i(r) = u_k^i(r) - u_k^{obs}(r) \tag{12}$$

$$u_k^{(i+1)}(r) = u_k^i(r) + \chi \Delta u_k^i(r) \tag{13}$$

$u_k^i(r)$ is the expected distance-dependant potentials $u_k^{exp}(r)$ starting from $i=0$, and the new expected statistical distance-dependant pair potentials $u_k^{(i+1)}(r)$ was taken to compute the E_{mn} and λ_{mn}, as well as U_{mn}, through Equations (9–11). In return, the U_{mn} was used to update the expected pair potentials through Equations (5–9). Thus, the expected pair potentials were adjusted with the difference $\Delta u_k^i(r)$ at each iteration step. The χ controls the speed of the convergence and was set as 0.1. The iterative procedure was repeated until the success rate of the best-scored conformations passing the high-quality accuracy of CAPRI criteria (Supplementary Table S4) converges to a value (normally over 98.0%). The final set of the expected pair potentials constitutes the optimized scoring function of protein–protein interactions, i.e. SPA-PP.

3 RESULTS AND DISCUSSION

3.1 Optimized scoring function

A straightforward way to validate the effectiveness of the iterative procedure on the improvement of the scoring function is to show the evolution of the average RMSD of the best-scored poses and the success rate of the best-scored poses passing high accuracy of CAPRI evaluation criteria (Fig. 2a). It can be seen that the success rate increases from 93.3% and converges to 98.1%, whereas the average RMSD decreases from 0.53 and approaches to 0.17 as the adjusting of the scoring function via iteration. Almost all the best-scored poses of the protein–protein complexes in the training set are identified as the native conformations by the optimized scoring function (SPA-PP), and the structures of the best-scored poses are identical to their native conformations with low RMSDs. This indicates that the accuracy of the statistical expected knowledge-based pair potentials are improved gradually as the iteration continues until the convergence is reached.

Our optimization strategy couples the adjustments of the affinity and specificity simultaneously. We can see that the distribution of ISR value of pre-optimized and optimized SPA-PP is clearly separated, and the average value of ISRs for the native poses increases from 5.3 to 7.3 (Fig. 2b). It implies that the specificity of native conformation is more obvious, whereas the stability is more strengthened by the optimized SPA-PP. The results validate that the performance of SPA-PP is optimized and

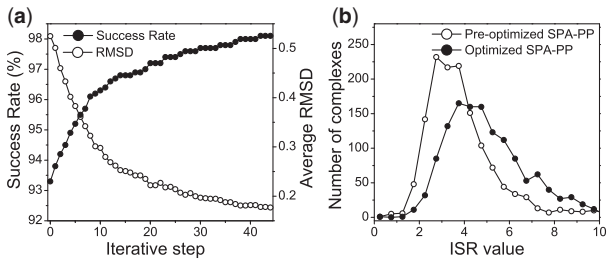

Fig. 2. Optimization of SPA-PP. (**a**) Evolution of the success rate and the average RMSD as the iteration proceeds. (**b**) The distribution of ISR values calculated with pre-optimized SPA-PP and optimized SPA-PP, respectively

improved on characterizing both the stability and specificity of the native pose.

To assess how different the docking predictions ranked by a new scoring function from a random order, Feliu and Oliva. (2010) proposed a statistical assessment of the performance of a scoring function on determining the number of near-native conformations among the top-ranked conformations. In this work, we used its *P*-value calculation [details are in the Feliu and Oliva (2010)] to get the number of successful cases among our testing dataset1 (320 complexes). A successful case is defined by considering at least one hit located in the top-ranked decoys of the complex. The success curves (Supplementary Fig. S3) show that the successful rate of optimized SPA-PP is much better than that of random ordering. Also, compared with pre-optimized SPA-PP, the successful rate of optimized SPA-PP is improved after optimization. The statistical assessment indicates that the optimization strategy is successful, and the optimized SPA-PP could be an accurate scoring function on the predictions of native pose identification and binding affinity ranking.

In the protein–protein docking problem, scoring functions are generally used for two applications: (i) to score and rank the binding poses generated by the docking programs and (ii) to predict and explain the experimentally determined protein–protein affinities. Thus, to evaluate the performance of SPA-PP, two kinds of tests related to corresponding applications are carried out, and the testing results are shown later.

3.2 Binding pose prediction

The purpose of protein–protein docking is to look for the native or near-native binding mode for the assembly proteins. Whether the scoring function can select out the best-scored binding pose that resembles the native conformation determines the scoring and ranking ability of the scoring function. Ideally, the scoring function should be transferable and work equally for the diverse protein–protein complexes. For this reason, the performance of binding pose prediction for SPA-PP is tested on two diverse testing datasets (testing dataset1 and dataset2, for details see Section 2).

The performances of identifying the native pose are compared among the optimized SPA-PP, pre-optimized SPA-PP and the energy function implemented in RosettaDock3.2.1 (Table 1). Also, to emphasize the importance of ISR on the optimization of scoring function, the optimized scoring function, which is obtained by only taking affinity into the optimization (called as Affinity-PP), is also listed for comparison, i.e. Equation 9 becomes $U_{mn} = \gamma E_{mn}$. The success rates of the native pose identification for both testing datasets show that the optimized SPA-PP performs much better than pre-optimized SPA-PP and even better than RosettaDock whose force fields perform extremely well for protein folding and protein assembly over other energy functions (King *et al.*, 2012; Schueler-Furman *et al.*, 2005). This suggests that optimized SPA-PP is successful on the ability to identify native or near-native binding poses. The comparison of Affinity-PP and SPA-PP demonstrates that incorporation of ISR into the optimization strategy improves the performance of the scoring function on the identification of native or near-native binding poses. The high success rate also means that optimized SPA-PP is effective to discriminate the native binding pose against decoys, namely, able to characterize the specificity.

Table 1. Success rates (%) of identifying the native or near-native conformations for testing dataset1 and testing dataset2

Scoring function	Dataset1	Dataset2
Optimized SPA-PP	97.6	97.3
Optimized Affinity-PP	97.4	97.2
Pre-optimized SPA-PP	92.6	88.1
RosettaDock	97.2	96.9

3.3 Binding affinity prediction

In addition to the prediction of binding pose, the prediction of binding affinity is another important application of protein–protein scoring function (Kastritis and Bonvin, 2010). It determines the accuracy of binding energy predicted by the scoring function compared with the experimental measurements. Because of scaling, the scoring functions usually can not reproduce the absolute values of experimental binding affinity; the Pearson correlation coefficient (C_P) between the predicted and experimental measured binding affinities was computed. Among the 144 complexes of affinity benchmark set (Kastritis *et al.*, 2011), 72 complexes are defined as rigid docking complexes with $I_{rms} \leq 1.0$Å (interface C_α root-mean-square displacement between the bound and unbound complexes), and the remaining are classified as flexible docking complexes. This benchmark set of binding affinities contains a diverse set of protein–protein complexes with reliable experimental binding affinities and for the first time provides an opportunity for evaluating the performance of scoring functions on the binding affinity prediction.

The correlations between the predicted and experimental affinities are shown in Table 2. The results of DFIRE (Liu *et al.*, 2004*a*) and PMF (Su *et al.*, 2009) were obtained from the literature (Moal *et al.*, 2011), and for consistence with this literature, the complexes 1UUG, 1IQD, 1NSN, 1DE4, 1M10, 1NCA and 1NB5 were also excluded in the calculations. Compared with RosettaDock, as well as other two scoring functions DFIRE and PMF (Moal *et al.*, 2011), the performance of SPA-PP for the rigid docking complexes ranks best with $C_P = 0.63$ (Table 2 and Fig. 3), and the C_P for all the benchmark set is only lower than that of RosettaDock. The results indicate that SPA-PP is

Table 2. Pearson correlations (C_P) between the predicted binding affinity and experimentally measured binding affinity for the rigid docking complexes, flexible docking complexes and all the complexes

Scoring function	Rigid	Flexible	All
SPA-PP	0.63	0.24	0.39
Affinity-PP	0.61	0.24	0.38
RosettaDock	0.62	0.29	0.42
DFIRE	0.52	0.28	0.35
PMF	0.62	0.23	0.37

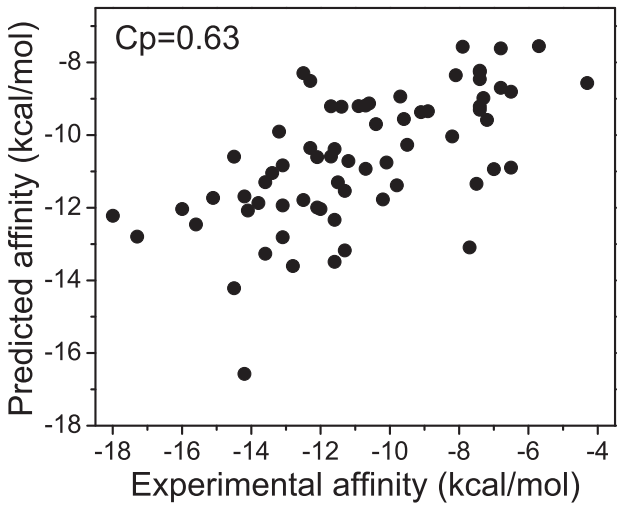

Fig. 3. Pearson correlation between the predicted and experimental binding affinities for the rigid docking complexes. The correlation coefficient (C_P) is 0.63 (statistical significance $P < 0.001$). The predicted affinity is obtained by scaling the binding scores with a linear equation:$y = 0.018 \times -4.79$, which is a fitting equation based on the experimental affinities

not only capable of discriminating the native conformation against non-native conformations but also accurately predicting the binding affinities for protein–protein binding without large conformational change. The comparison of Affinity-PP and SPA-PP demonstrates that ISR is an important factor to improve the performance of the scoring function. The specificity is critical to the protein–protein interactions and necessary to be incorporated to the optimization of scoring function. It is worth noticing that for flexible binding, SPA-PP, as well as other scoring functions, performs poorly with C_Ps < 0.3. This is reasonable, as the scoring function only captures the features of static structure and can not easily incorporate the information of conformational change because of flexibility. RosettaDock includes minimization function during the docking process. This minimization is restricted to the side-chain rotamers. That means RosettaDock considers certain degrees of flexibility into the docking process. This might be the reason why RosettaDock seems to be better in semi-rigid docking than others (Supplementary Table S5). However, all scoring functions, including RosettaDock, are bad in scoring performance on the docking with large conformational changes. The computationally efficient treatment of large conformational changes during binding still remains a challenge (Zacharias, 2010).

4 CONCLUSION

In this work, we developed a scoring function for protein–protein interactions, named as SPA-PP. The development strategy of SPA-PP takes into account of both specificity and affinity of protein–protein binding. It represents a significant advance over the previous investigations on protein–protein binding interactions and scoring functions, which only focused on affinity for development. Also, we used the largest set of high-quality protein–protein structures so far for training the scoring function. It makes SPA-PP independent on the training set and more general for applications. The remarkable performance of SPA-PP was validated by testing the ability on the predictions of native pose and binding affinity. The success of SPA-PP demonstrates that the specificity is critical to the protein–protein interactions and necessary to be incorporated to the scoring function and the computational design of the protein–protein interactions. SPA-PP is a kind of statistical pair-potentials, which are discrete potentials dependent on the distances between the interacting atom pairs. The statistical pair-potentials incorporate multiple energy terms, such as van der Waals interactions, electrostatic interactions and hydrophobic interactions, into one potential energy term. Therefore, the computational docking procedure with SPA-PP could cost less computational time if SPA-PP is implemented into the sampling and ranking of protein–protein binding.

Funding: This work was supported by National Science Foundation (Grant numbers 0947767, 0926287).

Conflict of Interest: none declared.

REFERENCES

Andreeva,A. *et al.* (2008) Data growth and its impact on the scop database: new developments. *Nucleic Acids Res.*, **36** (**Suppl. 1**), D419–D425.

Bolon,D. *et al.* (2005) Specificity versus stability in computational protein design. *Proc. Natl Acad. Sci. USA*, **102**, 12724–12729.

Bryngelson,J. *et al.* (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*, **21**, 167–195.

Chaudhury,S. *et al.* (2011) Benchmarking and analysis of protein docking performance in rosetta v3. *2. PLoS One*, **6**, e22477.

Chiu,W. *et al.* (2006) Structural biology of cellular machines. *Trends Cell Biol.*, **16**, 144–150.

Clark,M. *et al.* (1989) Validation of the general purpose tripos 5.2 force field. *J. Comput. Chem.*, **10**, 982–1012.

Dominy,B. and Shakhnovich,E. (2004) Native atom types for knowledge-based potentials: application to binding energy prediction. *J. Med. Chem.*, **47**, 4538–4558.

Douguet,D. *et al.* (2006) Dockground resource for studying protein–protein interfaces. *Bioinformatics*, **22**, 2612–2618.

Dutta,S. and Berman,H. (2005) Large macromolecular complexes in the protein data bank: a status report. *Structure*, **13**, 381–388.

Feliu,E. and Oliva,B. (2010) How different from random are docking predictions when ranked by scoring functions? *Proteins*, **78**, 3376–3385.

Gao,Y. *et al.* (2007) Dockground system of databases for protein recognition studies: unbound structures for docking. *Proteins*, **69**, 845–851.

Goldstein,R. *et al.* (1992) Optimal protein-folding codes from spin-glass theory. *Proc. Natl Acad. Sci. USA*, **89**, 4918–4922.

Gray,J. *et al.* (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.*, **331**, 281–299.
Grigoryan,G. *et al.* (2009) Design of protein-interaction specificity gives selective bzip-binding peptides. *Nature*, **458**, 859–864.
Guha,R. *et al.* (2006) The blue obelisk interoperability in chemical informatics. *J. Chem. Inf. Model.*, **46**, 991–998.
Halperin,I. *et al.* (2002) Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins*, **47**, 409–443.
Havranek,J. *et al.* (2003) Automated design of specificity in molecular recognition. *Nat. Struct. Biol.*, **10**, 45–52.
Huang,S. and Zou,X. (2008) An iterative knowledge-based scoring function for protein–protein recognition. *Proteins*, **72**, 557–579.
Hwang,H. *et al.* (2010) Protein–protein docking benchmark version 4.0. *Proteins*, **78**, 3111–3114.
Janin,J. (1995) Principles of protein-protein recognition from structure to thermodynamics. *Biochimie*, **77**, 497–505.
Janin,J. (1996) Quantifying biological specificity: the statistical mechanics of molecular recognition. *Proteins*, **25**, 438–445.
Janin,J. (2002) Welcome to CAPRI: a critical assessment of predicted interactions. *Proteins*, **47**, 257–257.
Jiang,L. (2002) Automated design of specificity in molecular recognition. (2002) Potential of mean force for protein–protein interaction studies. *Proteins*, **46**, 190–196.
Jones,S. and Thornton,J. (1996) Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
Kastritis,P. and Bonvin,A. (2010) Are scoring functions in protein- protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J. Proteome Res.*, **9**, 2216–2225.
Kastritis,P. *et al.* (2011) A structure-based benchmark for protein–protein binding affinity. *Protein Sci.*, **20**, 482–491.
Kim,P. *et al.* (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, **314**, 1938–1941.
King,N. *et al.* (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science*, **336**, 1171–1174.
Koppensteiner,W. and Sippl,M. (1998) Knowledge-based potentials–back to the roots. *Biochemistry (Mosc)*, **63**, 247–252.
Kortemme,T. *et al.* (2004) Computational redesign of protein-protein interaction specificity. *Nat. Struct. Mol. Biol.*, **11**, 371–379.
Lensink,M. and Wodak,S. (2010) Docking and scoring protein interactions: CAPRI 2009. *Proteins*, **78**, 3073–3084.
Lensink,M. *et al.* (2007) Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins*, **69**, 704–718.
Levy,Y. *et al.* (2004) Protein topology determines binding mechanism. *Proc. Natl Acad. Sci. USA*, **101**, 511.
Liu,S. *et al.* (2004a) A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins*, **56**, 93–101.
Liu,Z. *et al.* (2004b) Structural mining: self-consistent design on flexible protein-peptide docking and transferable binding affinity potential. *J. Am. Chem. Soc.*, **126**, 8515–8528.
Loregian,A. and Palù,G. (2005) Disruption of protein–protein interactions: towards new targets for chemotherapy. *J. Cell. Physiol.*, **204**, 750–762.
Malod-Dognin,N. *et al.* (2012) Characterizing the morphology of protein binding patches. *Proteins*, **80**, 2652–2665.
Miller,D. and Dill,K. (1997) Ligand binding to proteins: the binding landscape model. *Protein Sci.*, **6**, 2166–2179.
Moal,I. *et al.* (2011) Protein–protein binding affinity prediction on a diverse set of structures. *Bioinformatics*, **27**, 3002–3009.
Moreira,I. *et al.* (2010) Protein–protein docking dealing with the unknown. *J. Comput. Chem.*, **31**, 317–342.
Muegge,I. and Martin,Y. (1999) A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.*, **42**, 791–804.
Ortuso,F. *et al.* (2006) Gbpm: Grid-based pharmacophore model: concept and application studies to protein–protein recognition. *Bioinformatics*, **22**, 1449–1455.
Rejto,P. and Verkhivker,G. (1996) Unraveling principles of lead discovery: from unfrustrated energy landscapes to novel molecular anchors. *Proc. Natl Acad. Sci. USA*, **93**, 8945–8950.
Sali,A. *et al.* (2003) From words to literature in structural proteomics. *Nature*, **422**, 216–225.
Schueler-Furman,O. *et al.* (2005) Progress in modeling of protein structures and interactions. *Science*, **310**, 638–642.
Shifman,J. and Mayo,S. (2003) Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc. Natl Acad. Sci. USA*, **100**, 13274–13279.
Sippl,M. (1990) Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, **213**, 859–883.
Steven,A. and Baumeister,W. (2008) The future is hybrid. *J. Struct. Biol.*, **163**, 186–195.
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.
Thomas,P. and Dill,K. (1996a) An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl Acad. Sci. USA*, **93**, 11628–11633.
Thomas,P. and Dill,K. (1996b) Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.*, **257**, 457–469.
Tsai,C. *et al.* (1999) Folding funnels, binding funnels, and protein function. *Protein Sci.*, **8**, 1181–1190.
Wang,J. and Verkhivker,G. (2003) Energy landscape theory, funnels, specificity, and optimal criterion of biomolecular binding. *Phys. Rev. Lett.*, **90**, 188101.
Wang,J. *et al.* (2007) Quantifying intrinsic specificity: a potential complement to affinity in drug screening. *Phys. Rev. Lett.*, **99**, 198101.
Yan,Z. and Wang,J. (2012) Specificity quantification of biomolecular recognition and its implication for drug discovery. *Sci. Rep.*, **2**, 309.
Zacharias,M. (2010) Accounting for conformational changes during protein-protein docking. *Curr. Opin. Struct. Biol.*, **20**, 180–186.
Zhang,C. *et al.* (2005a) A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J. Med. Chem.*, **48**, 2325–2335.
Zhang,C. *et al.* (2005b) A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J. Med. Chem.*, **48**, 2325–2335.