Optimizing Scoring Function of Protein-Nucleic Acid Interactions with Both Affinity and Specificity

Zhiqiang Yan[1,2], **Jin Wang**[1,2]*

1 Department of Chemistry & Physics, State University of New York at Stony Brook, Stony Brook, New York, United States of America, **2** State Key Laboratory of Electroanalytical Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun, Jilin, China

Abstract

Protein-nucleic acid (protein-DNA and protein-RNA) recognition is fundamental to the regulation of gene expression. Determination of the structures of the protein-nucleic acid recognition and insight into their interactions at molecular level are vital to understanding the regulation function. Recently, quantitative computational approach has been becoming an alternative of experimental technique for predicting the structures and interactions of biomolecular recognition. However, the progress of protein-nucleic acid structure prediction, especially protein-RNA, is far behind that of the protein-ligand and protein-protein structure predictions due to the lack of reliable and accurate scoring function for quantifying the protein-nucleic acid interactions. In this work, we developed an accurate scoring function (named as SPA-PN, SPecificity and Affinity of the Protein-Nucleic acid interactions) for protein-nucleic acid interactions by incorporating both the specificity and affinity into the optimization strategy. Specificity and affinity are two requirements of highly efficient and specific biomolecular recognition. Previous quantitative descriptions of the biomolecular interactions considered the affinity, but often ignored the specificity owing to the challenge of specificity quantification. We applied our concept of intrinsic specificity to connect the conventional specificity, which circumvents the challenge of specificity quantification. In addition to the affinity optimization, we incorporated the quantified intrinsic specificity into the optimization strategy of SPA-PN. The testing results and comparisons with other scoring functions validated that SPA-PN performs well on both the prediction of binding affinity and identification of native conformation. In terms of its performance, SPA-PN can be widely used to predict the protein-nucleic acid structures and quantify their interactions.

Citation: Yan Z, Wang J (2013) Optimizing Scoring Function of Protein-Nucleic Acid Interactions with Both Affinity and Specificity. PLoS ONE 8(9): e74443. doi:10.1371/journal.pone.0074443

Editor: Freddie Salsbury, Jr, Wake Forest University, United States of America

Received June 1, 2013; **Accepted** August 2, 2013; **Published** September 30, 2013

Funding: This work was supported by National Science Foundation (grant numbers 0947767,0926287). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jin.wang.1@stonybrook.edu

Introduction

Precise regulation of the biological activities within cells is accomplished by biomolecular recognition which mainly involves three major biological macromolecules, i.e. protein, DNA and RNA. The protein-nucleic acid (protein-DNA and protein-RNA) recognition is essential to the regulation of gene expression at every level of the central dogma of molecular biology, including replication, transcription and translation of genetic information [1]. Determination of the structures of the specific protein-nucleic acid recognition and insight into their interactions at a molecular level are vital to understanding the regulation on a genomic scale [2]. The knowledge of which would be also enormously useful for a variety of biological and medical applications [3–8].

Although the structures of individual biomolecules are increasingly well determined and structural studies of the biomolecular complexes have been very active in the last decade, three-dimensional atomic structures of many biomolecular complexes are still difficult to determine due to the technical challenges of the experimental approaches [9–11]. As an alternative, computational approaches can complement existing experimental data and be applied to the structural prediction of biomolecular complexes [12]. The field of protein-protein docking has achieved substantial progress over the last decade as witnessed by the CAPRI (Critical Assessment of Predicted Interactions) [13, 14]. However, the progress for the protein-nucleic acid docking, especially the protein-RNA docking, lags behind due to the lack of reliable scoring function of protein-nucleic acid interactions. Previously, structural information was used extensively to derive scoring functions for successful predictions of protein structures, as well as protein-ligand and protein-protein interactions. Given the rapid growth in the number of solved protein-nucleic complex structures recently [15], it is natural and urgent to develop an accurate scoring function of protein-nucleic acid interactions, general for both protein-DNA and protein-RNA interactions.

For biomolecular functions, highly efficient and specific biomolecular recognitions are required to satisfy both the stability and specificity. The stability is determined by the affinity of the complex while the specificity is controlled by the partner binding to other competitive biomolecules discriminatively. The current scoring functions of biomolecular recognition [16,17], whether force-field based, empirical, or knowledge-based scoring functions, mainly focused on improving the ability of predicting the known binding affinities observed in experiments as accurately as possible. The strategy of developing these scoring functions seeks to optimize the stability based on the combination of energetics and shape complementarity, but are often lack of the considerations of the specificity. In the cell compartment, biomolecules are required to function by interacting with a small number of partners rather than the myriad of others. The naturally occurred biomolecular recognition is just a very small part of all possible interacting complexes [18]. 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 leads to significant population discrimination of the native complex against alternative ones [19–24], which is the requirement for the proper functions of the specific biomolecular recognitions in cell. Recent works taking the consideration of specificity into the computational design and optimization of interface interactions has achieved a few successful applications [6,19–23,25–29]. These works designed and optimized the interactions that seek to stabilize the desired structures and also destabilize the competitive structures. Thus, the accurate scoring function of biomolecular interactions should satisfy the criteria that the stability of the specific complex is maximized while the stability of alternative complexes is minimized, which can guarantee both the stability and the specificity for the functional biomolecular recognition.

The reason that the specificity usually was not taken into account previously in the development of the scoring functions is that the description of binding specificity was challenging to quantify. The conventional definition (Figure 1A) of specificity is the ability of a biomolecule to specifically bind to its own partner against other competitive partners, namely the relative difference in affinity of one specific biomolecular complex to others [24,30]. The definition of conventional specificity is simple but the quantification of conventional specificity is challenging since it requires that the set of competitive complexes are not too large and the affinities are already known. This makes the practical quantification of the specificity impossible due to the incomplete information on the competitive alternatives.

To overcome the challenge, we have proposed a novel concept named as intrinsic specificity (Figure 1B) [19–23]. Here, we expand the concept to the protein-nucleic acid interactions. In particular, the intrinsic specificity of protein-nucleic acid binding refers to the preference in affinity of a nucleic acid binding to its protein receptor with a preferred pose over other poses (Figure 1B). Imagining the N- and C-terminus of multiple protein receptors are linked together, leading to an effective single large protein receptor. Under the assumption that the protein receptor is large, the conventional specificity of discrimination of a nucleic acid binding to its protein receptor against other proteins can be transformed to the intrinsic specificity that the nucleic acid binds to the large receptor with a preferred pose over other poses. By applying this concept, we have developed scoring functions for the interactions of bimolecular recognition, including protein-ligand interactions [21] and protein-protein interactions [22]. Also, the connection between the intrinsic specificity and conventional specificity was validated by studying a drug-target model [23], where the conventional specificity is correlated with the intrinsic specificity.

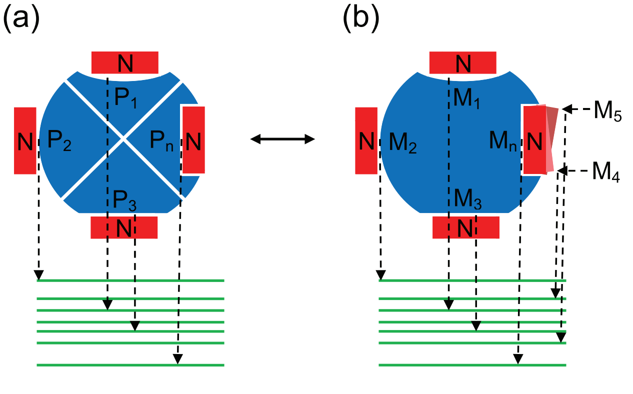

Figure 1. Schematic view of illustrating the equivalence of conventional specificity to intrinsic specificity. (A) The same nucleic acid (N, red) binding with multiple protein receptors (blue, P_1 to P_n), showing the conventional specificity as the gap in binding affinity of the nucleic acid binding to the specific protein receptor (P_n) in discrimination against other protein receptors. The binding affinities are represented with corresponding energy spectrum (green). (B) The same nucleic acid (N, red) binding on a large protein receptor thought as the multiple different receptors linked together (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.
doi:10.1371/journal.pone.0074443.g001

According to the theory of energy landscape [19–21,31–39], the binding process of biomolecules can be visualized and quantified as a funnel-like energy landscape towards the native binding state with local roughness along the binding paths. The native pose of protein-nucleic acid complex is the conformation with the lowest binding energy and the energies of the conformation ensemble follow a statistical Gaussian-like distribution. 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. With computationally generated non-native poses (decoys), the ISR can be readily obtained. Therefore, without evaluating the conventional specificity through exploring the whole set of competitive partners, ISR physically provides a quantitative measure of the binding specificity.

In this work, based on our practical quantification of binding specificity, we have designed an optimization strategy to maximize both the affinity and specificity of native binding mode simultaneously for developing the scoring function of protein-nucleic acid interactions. The optimization strategy is to adjust the statistical knowledge-based potentials of atom pairs by iteration until the scoring function can effectively discriminate the native binding pose against the decoys. The flow of developing procedures is shown in Figure S1 in File S1. We have tested the derived scoring function of protein-nucleic acid interactions (SPA-PN) via the performance on the prediction of binding affinity and the identification of native binding pose. The performance of SPA-PN demonstrated that the quantified specificity is necessary to be incorporated into the optimization of scoring functions of protein-nucleic acid interactions.

Materials and Methods

Preparation of the datasets

Training dataset. The requirement of optimizing the knowledge-based statistical scoring function is to train a set of known structural data. The training dataset of protein-nucleic complexes for developing SPA-PN were extracted from the database NPIDB (Nucleic Acids-Protein Interaction DataBase) [15,40]. NPIDB contains information derived from structures of protein-DNA and protein-RNA complexes deposited in the PDB (Protein Data Bank) [41]. To obtain a relatively high quality set of protein-nucleic complexes for the training dataset, X-ray structures with resolutions better than 3.0Å for the protein-DNA complexes and 3.5Å for protein-RNA complexes were selected. Entries with more than 3 DNA or RNA chains, or the number of heavy atoms of any chain less than 100 were discarded. By removing the entry overlaps with the testing datasets below, the resulting training dataset contains 1555 complexes, including 1221 protein-DNA structures and 334 protein-RNA structures (Table S1 in File S1).

Testing datasets. To validate the performance of a novel scoring function, two kinds of tests are needed. First, to evaluate the ability of SPA-PN on predicting the binding affinity. SPA-PN was tested on a dataset of protein-DNA complexes with known experimental binding affinities. This dataset for binding affinity prediction was employed from the binding database which is a modified version of Zhang *et al.* [42] and was used by Donald *et al.* [43] and Xu *et al* [44]. The dataset (named as testing dataset1) contains 30 protein-DNA complexes (Table S2 in File S1).

Second, in order to evaluate the ability of SPA-PN on discriminating the native conformation from decoy conformations, SPA-PN was tested on our collected benchmark of protein-nucleic acid complexes. The collected benchmark for binding pose prediction was obtained from available benchmarks of protein-nucleic acid complexes. It combines two protein-DNA benchmarks and two protein-RNA docking benchmarks. The first protein-DNA benchmark obtained from the PDIdb (Protein-DNA Interface database) [45] contains 246 representative protein-DNA interfaces out of 922 entries collected in the PDIdb. The second protein-DNA benchmark [46] contains 47 complexes which covers almost all major groups of DNA-binding proteins according to the classification of Luscombe *et al* [47]. The first protein-RNA benchmark [48] contains 45 complexes covering all major groups of protein-RNA complexes according to the classification of Bahadur *et al* [49]. The second protein-RNA benchmark [50] is an extended set of the first one and contains 106 protein-RNA complexes, it was obtained from both experimental and homology modeling data. This collected testing dataset were also filtered with the criteria as composed on the training dataset. In addition, one entry was kept if there are overlaps among the different benchmarks. The final collected dataset (named as testing dataset2) contains 315 complexes, including 232 protein-DNA structures and 83 protein-RNA structures (Table S3 in File S1).

Docking decoys. For the optimization of SPA-PN, an ensemble of decoys for each complex are needed to calculate the ISR for specificity and carry out the iteration algorithm. The RosettaDock v3.4 was taken as the structure optimization and docking tool [51,52] to generate the decoys. Three steps were performed. First, each docking partner of the complex was prepared in isolation for optimizing their side-chain conformations prior to docking using the prepacking protocol. Second, the prepacked complexes were 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 partner was defined as the docking ligand in the complex and the other was assigned as the receptor which was kept fixed during docking. 1000 orientations for each complex were generated by docking. Other docking parameters were set as default. The generated decoys are structured diversely to explore the underlying binding energy landscape.

Derivation of knowledge-based statistical potentials

Observed statistical potentials. The knowledge-based scoring function consists of a set of statistical distance-dependent atom-pair potentials to quantify the interactions. Normally, the observed atom-pair potentials were directly derived from the Boltzmann relation widely applied in the derivation of knowledge-based statistical potentials for the protein-ligand, protein-protein and protein-nucleic acid interactions [53–57], the Boltzmann relation is written as

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

where $g_k^{obs}(r)$ is the observed atom pair distribution function quantified 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 radius r and $r+\Delta r$. It can be directly extracted from the structural database of protein-nucleic acid complexes. $f_k^{obs}(R)$ is the number density within the sphere of the reference state where there are no interactions between atoms. It was obtained based on the approximation that the atom-pair k is uniformly distributed in the sphere of the reference state [58]. 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 (= 1555) of training protein-nucleic acid complexes (Table S1 in File S1), $n_k^m(r)$ and N_k^m are the numbers of atom pair k within the spherical shell and the reference sphere for a given protein-nucleic acid 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, where Δr is the bin size and R is the radius of sphere. Δr and R are set as 0.3Å and 8.2Å, respectively. In total, there are 20 spherical shells with bin size 0.3Å from the shortest radius 2.2Å. Based on the definition of atom type by SYBYL [59], 15 atom types were used to cover the heavy atoms involved in protein-nucleic acid interactions (Table S4 in File S1), these atom types were converted from PDB files by OpenBabel [60]. A cutoff (= 600) of total occurrences for atom pair k ($N_k = \sum_m^M N_k^m$) was employed to neglect the contribution from the atom pairs with statistically insufficient occurrences. This lead to 95 effective types of atom pairs for the protein-nucleic acid interactions (Table S5 in File S1). There are 1900 types of interaction pair by multiplying the number of atom pairs (= 95) and the number of shells (= 20). In addition, if the atom pair has no occurrence in a particular spherical shell, the corresponding pair potential was set as the van der Waals interaction within this shell.

The observed statistical potentials from the known structures has its limitations as the statistical potentials extracted from equation (1) is not exactly the expected potentials that nature employs to stabilize the complexes [61]. The origin of this problem is attributed to the construction of the reference state where the atom pairs are uniformly distributed and independent of each other [58]. In reality, the protein-nucleic acid interactions involve the excluded volume, sequences and connectivity. Thus the observed statistical potentials are generally not equal to the expected potentials [62].

Expected statistical potentials. To circumvent the reference state issue and improve the statistical potentials, earlier efforts [42,61,63–65] have taken different approaches to optimize the statistical potentials. An effective way is to take into account both native and nonnative conformations (decoys) [61,63,65] based on the energy landscape theory that the native conformation should be sufficiently favored over alternative nonnative structures thermodynamically. However, these earlier works hasn't combined both the specificity and affinity to discriminate native conformation over nonnative conformations. In our recent papers on the study of the protein-ligand [21] and protein-protein interactions [22], we considered the importance of both the affinity for stabilizing the native conformation and the specificity of discrimination over nonnative conformations, and combined them into the optimization processes of scoring function. Here we expand this concept to optimize the statistical potentials of protein-nucleic acid interactions.

The expected statistical potentials are calculated similarly as the observed statistical potentials, which is

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

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

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

Our aim is to obtain a scoring function that can significantly favor the native conformation over all other decoy conformations, so that the native conformation dominates the ensemble of conformations according to the Boltzman distribution. Considering the population discrimination of the native and nonnative conformations, the expected number density of atom-pair k was calculated with Boltzmann-averaged weighting over the ensemble of conformations [22,61,65], 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 mentioned above 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 a constant analogous to the inverse of temperature and set as 0.1. The resulting U_{mn} is the potential which is supposed to be able to discriminate the native conformation against decoys. As discussed, both the stability and specificity are the requirements to form an efficient and specific functional complex. Thus, U_{mn} is designed to take into account of both affinity and specificity optimized through parameterizing the affinity (E) and specificity (ISR), which is given by

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

E_{mn} is the energy score of the protein-nucleic acid 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-nucleic acid conformation, λ_{mn} is the ISR representing the specificity of the protein-nucleic acid conformation [19–21]. γ is a parameter which balances the values of E_{mn} and λ_{mn} and set as 0.1. E_{mn} and λ_{mn} are calculated as

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

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

where $t_k(r)$ represents the occurrence times of the atom-pair interaction between the protein-nucleic acid interface; α_m is a scaling factor which 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) [19]. Here, α_m approximately depends on the number of interfacial residues/nucleotides ($\alpha_m \sim \sqrt{\frac{1}{n_{inter}}}$) of the native protein-nucleic acid conformation of the complex m. An interfacial residue/nucleotide is defined if any atom of this residue/nucleotide in one partner of the native protein-nucleic acid conformation is within 10Å from the other partner. α_m normalizes the increase of ISR with the number of interfacial residues/nucleotides. 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.

Optimization of knowledge-based statistical potentials

The iterative method [61,65] was employed to realize the optimization. The idea of the iterative method is to circumvent the inaccessible reference state problem by adjusting the expected statistical pair potentials until they are able to discriminate native binding mode from decoys. As aforementioned, 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 are 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-

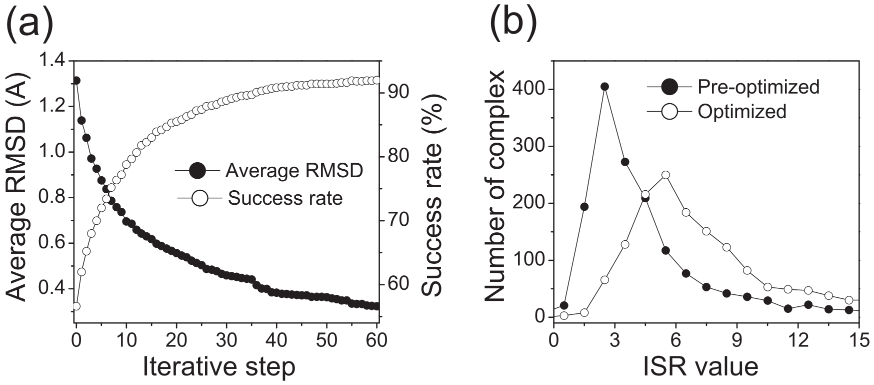

Figure 2. Optimization of SPA-PN. (A) Evolution of the success rate and the average interfacial RMSD (I_{rms}) as the iteration precedes. (B) The distribution of ISR values calculated with pre-optimized SPA-PN and optimized SPA-PN respectively.
doi:10.1371/journal.pone.0074443.g002

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 (Table S6 in File S1) converges to a high value. The resulting set of the expected pair potentials constitutes the optimized scoring function of protein-nucleic acid interactions, namely SPA-PN.

Results and Discussion

Optimized SPA-PN

To validate the effectiveness of the iterative procedure on the improvement of the statistical potentials, we show the evolution of the average interfacial RMSD (I_{rms}, root-mean-square displacement of backbone atoms among the interface) of the best-scored poses and the success rate of the best-scored poses passing high accuracy of CAPRI evaluation criteria (Figure 2A). It can be seen that the success rate increases from 56.6% and converges to 92.0% while the average I_{rms} decreases from 1.31 and approaches to 0.32 through adjusting of the atom-pair potentials via iteration. When the iteration reaches convergence, almost all the best-scored poses of the protein-nucleic acid complexes in the training set are identified as the native conformations by the optimized scoring function SPA-PN, and the structures of the best-scored poses are identical or similar to their native conformations with low I_{rms}s. This suggests that the accuracy of the statistical expected knowledge-based pair potentials on the prediction of the binding affinity and identification of native conformation are improved gradually as the iteration continues until the convergence is reached. It satisfies our expectations that the optimized scoring functions is to favor the native conformations energetically as occurred in nature.

The novelty of our optimization strategy is to couple the optimizations of the affinity and specificity simultaneously via the iterative adjustments of the atom-pair potentials. As seen (Figure 2B), the distribution of ISR value of pre-optimized and optimized SPA-PN is clearly separated, and the average value of ISRs for the native poses increases from 4.68 to 7.94. It implies that the specificity of native conformation is more pronounced

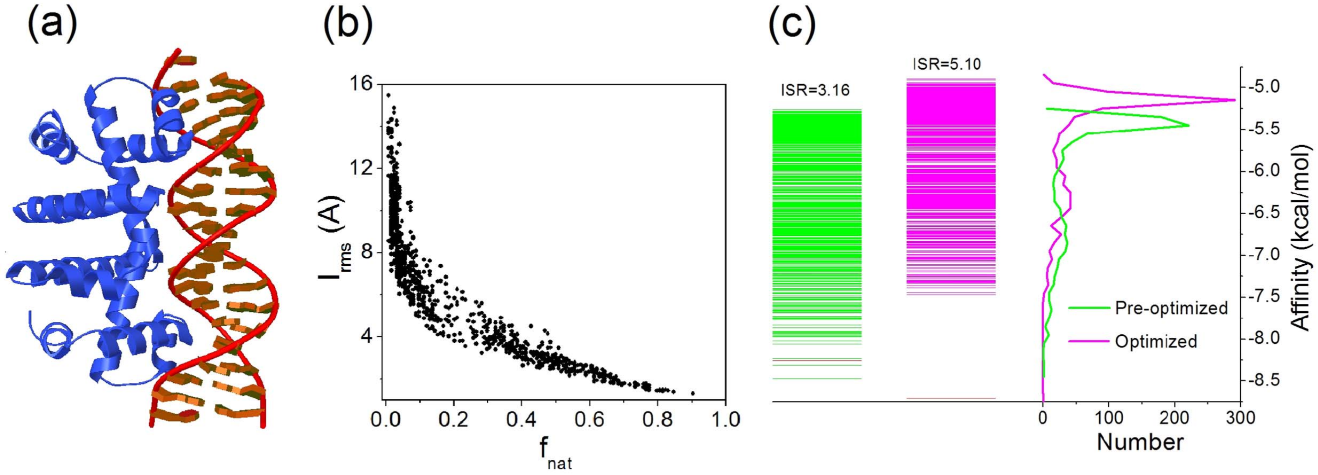

Figure 3. A typical example of protein-nucleic acid complex (PDB 1TRO). (A) Protein-nucleic acid binding structure with protein colored in blue and nucleic acid colored in red. (B) Plot of interfacial RMSD (I_{rms}) as a function of the fraction of native contacts (f_{nat}) for 1000 docking decoys of the typical complex. (C) Energy spectrum and distribution calculated with pre-optimized SPA-PN (green) and optimized SPA-PN (magenta), the corresponding ISR values are shown and the energy of the native conformation is marked as red.
doi:10.1371/journal.pone.0074443.g003

Table 1. Pearson correlations (C_P) between the predicted binding affinities and experimentally measured binding affinities for 30 protein-DNA complexes of the testing dataset1.

Scoring function	C_p
SPA-PN	0.862
Affinity-PN	0.857
Pre-optimized	0.817
RosettaDock	0.638
cFIRE	0.847
DDNA	0.840
FIRE	0.790
vcFIRE	0.720
vFIRE	0.550

doi:10.1371/journal.pone.0074443.t001

while the stability is more strengthened with the optimized SPA-PN. A typical example of protein-nucleic acid docking complex (PDB ITRO) is represented with its ensemble of docking conformations (Figure 3). After optimization, the native conformation becomes more separated from the decoy ensemble while the energy distribution of decoys becomes more narrow, i.e. the energy gap between the energy of native conformation and the average energy of conformation ensemble is enlarged, while the energy roughness or the width of the energy distribution of the conformation ensemble is reduced. The ISR value of the native conformation increases from 3.16 to 5.10 during the optimization, and also the stability of the native conformation is enhanced compared to the decoy ensemble. Collectively, the optimizing of the statistical potentials improves the performance of SPA-PN on characterizing both the stability and specificity of the native pose. A subset of atom-pair potentials of the optimized SPA-PN are shown in the Figure S2 in File S1. The potentials are normally have more than one minimum, which is consistent with the characteristic feature of the knowledge-based potential that is the mixture of different kinds of atom-pair interactions, such as electrostatic, hydrogen bonding, hydrophobic and van der Waals interactions.

Scoring functions of biomolecular recognition are generally used for two applications: (1) to predict and explain the experimentally determined affinities; and (2) to score and rank the binding poses generated by the docking programs. Thus, to validate the performance of SPA-PN, two kinds of tests related to corresponding applications are carried out and the testing results are shown in the sections below.

Prediction of binding affinity

The prediction accuracy of the binding affinity determines the performance of the scoring function on how well it can reproduce the experimentally measured affinity and predict the biomolecular interactions. Due to 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 were computed for the 30 protein-DNA complexes of the testing dataset1. The C_Ps of the scoring functions (cFIRE, DDNA, FIRE, vcFIRE and vFIRE) were obtained from the paper [44]. The correlations between the predicted and experimental affinities are shown in Table 1 and the detailed affinity values are listed in Table S2 in File S1. In order to emphasize the importance of ISR on the optimization of scoring function, we also optimized the scoring function by only taking affinity into the optimization (called as Affinity-PN), i.e. the equation 9 becomes $U_{mn}=\gamma E_{mn}$. From the comparisons with other scoring functions, the performance of SPA-PN ranks best with $C_P=0.862$ (Figure 4). The high consistence with experimental measurements indicates that SPA-PN is accurately predicting the binding affinities for the protein-nucleic acid interactions.

Identification of native pose

The aim of computational docking is to look for the native or near-native binding pose for the assembly partners. Whether the best-scored binding pose resembles the native conformation in structure determines the scoring and ranking ability of the scoring function. The performance of binding pose prediction for SPA-PN is tested on both protein-DNA and protein-RNA complexes of testing dataset2 (Table S3 in File S1). The performances of SPA-PN on identifying the native pose are compared with pre-optimized SPA-PN and affinity-PN (Table 2). The success rates of the native pose identification calculated by SPA-PN for both protein-DNA and protein-RNA complexes are over 85% which is much higher than that of pre-optimized SPA-PN. The success rate for the testing dataset is close to that for the training dataset. This suggests that the optimized SPA-PN is successful and robust on the ability to identify native or near-native binding poses. The high success rate also means that optimized SPA-PN is effective to discriminate the native binding pose against decoys, namely able to characterize the specificity. It is worth noting that the comparison of Affinity-PN and SPA-PN demonstrates the importance of incorporation of ISR into the optimization strategy since it further improves the performance of the scoring function on the identification of native or near-native binding poses. With affinity and specificity optimization, SPA-PN outperforms Affinity-PN not only on the affinity prediction but also the identification of native conformation.

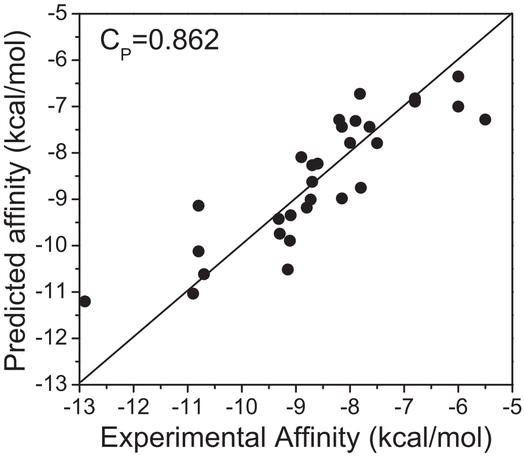

Figure 4. Pearson correlation between the predicted affinities calculated by SPA-PN and experimental binding affinities for 30 protein-DNA complexes of testing dataset1. The correlation coefficient (C_P) is 0.862 (statistical significance $P<0.001$). The predicted affinities are obtained by scaling the binding scores with a linear equation:y = 0.0045x-5.129 which is a fitting equation based on the experimental affinities.
doi:10.1371/journal.pone.0074443.g004

Table 2. Success rates (%) of identifying the native or near-native conformations for testing dataset2 including 232 protein-DNA and 83 protein-RNA complexes.

Scoring function	Protein-DNA	Protein-RNA	All
SPA-PN	96.6	85.5	93.7
Affinity-PN	93.5	84.3	91.1
Pre-optimized	53.0	62.7	55.6

doi:10.1371/journal.pone.0074443.t002

Conclusion

In this work, we have developed a novel scoring function SPA-PN for protein-nucleic acid interactions with the concept of intrinsic specificity. Our optimization strategy of SPA-PN satisfies the requirement that the stability of the specific complex is maximized while the stability of competing complexes is minimized. It guarantees both the stability and the specificity for the specific complex. This optimization strategy represents a significant advance over the previous investigations on protein-nucleic acid interactions which only focused on affinity. We have employed a largest set of high quality protein-nucleic acid structures so far for training the SPA-PN and includes both protein-DNA and protein-RNA complexes, making SPA-PN more independent on the training set and more generalizable for applications. The remarkable performance of SPA-PN was validated by testing the ability on the identification of native pose and prediction of binding affinity. In addition, SPA-PN is composed 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 into one potential energy term. Therefore, the computational docking procedure with SPA-PN will cost less time if SPA-PN is implemented into the sampling and ranking of protein-nucleic acid structure prediction.

The success of SPA-PN demonstrates that the specificity is critical to the protein-nucleic acid interactions and necessary to be incorporated into the optimization of scoring function. Similar concept was taken in computational redesign for biomolecules. Design of biomolecules requires the energy discrimination of interacting with specific partners against other competitive partners. In natural systems, evolution could encode the specificity in the functional structures against the large number of alterative ones. The quantification of the specificity for biomolecular interactions opens up a new window to explore new approaches for both the development of scoring functions and computational design of biomolecules. Our proposed quantification of specificity by ISR can be employed as a framework for further improvement of SPA-PN, and as a criteria for the computational redesign of protein-nucleic acid interface.

Supporting Information

File S1 Supporting figures and tables. **Figure S1** The development of SPA-PN contains three stages: The preparation of database, optimization of the scoring function, Testing and application of SPA-PN. **Figure S2** Typical atom-pair interaction potentials of SPA-PN. (A and B) Two of most frequently occurred atom pairs. (C and D) Atom pairs related to hydrogen bond. (E and F) Atom pairs involving phosphorus atom. **Table S1** Training dataset for the development of SPA-PN. **Table S2** Experimental determined affinities and SPA-PN predicted affinities for 30 protein-DNA complexes of the testing dataset1, the calculated affinities were obtained by scaling the binding scores with linear fitting equations (SPA-PN: y = 0.0045*x-5.129, Affinity-PN:y = 0.0044x-5.080, Pre-optimized: 0.0043x-5.443, Rosetta-dock: y = 0.0053x-7.24) based on the experimental affinities. **Table S3** PDB codes of the testing dataset2. **Table S4** 15 Atom types used for calculating the atom pair potentials based on the SYBYL definition of atom type. The atom types can be converted from PDB files by the software OpenBabel. **Table S5** 95 effective types of atom pairs for the protein-nucleic acid interactions with the cutoff of total occurrences larger than 600 in the training dataset. **Table S6** The high accuracy quality of CAPRI assessment criteria was taken to define the near-native conformation.
(DOC)

Author Contributions

Conceived and designed the experiments: ZY JW. Performed the experiments: ZY. Analyzed the data: ZY JW. Contributed reagents/materials/analysis tools: ZY. Wrote the paper: ZY JW.

References

1. Bustamante C, Cheng W, Mejia Y (2011) Revisiting the central dogma one molecule at a time. Cell 144: 480–497.
2. Bujalowski W (2006) Thermodynamic and kinetic methods of analyses of protein-nucleic acid interactions. from simpler to more complex systems. Chemical Reviews 106: 556–606.
3. Uil T, Haisma H, Rots M (2003) Therapeutic modulation of endogenous gene function by agents with designed DNA-sequence specificities. Nucleic acids research 31: 6064–6078.
4. Urnov F, Miller J, Lee Y, Beausejour C, Rock J, et al. (2005) Highly efficient endogenous human gene correction using designed zinc-finger nucleases. Nature 435: 646–651.
5. Pommier Y, Marchand C (2005) Interfacial inhibitors of protein-nucleic acid interactions. Current Medicinal Chemistry-Anti-Cancer Agents 5: 421–429.
6. Ashworth J, Havranek J, Duarte C, Sussman D, Monnat R, et al. (2006) Computational redesign of endonuclease DNA binding and cleavage specificity. Nature 441: 656–659.
7. Saven JG (2011) Computational protein design: engineering molecular diversity, nonnatural enzymes, nonbiological cofactor complexes, and membrane proteins. Current Opinion in Chemical Biology 15: 452–457.
8. Wall ME (2012) Quantitative Biology: From Molecular to Cellular Systems. CRC Press.
9. Dutta S, Berman H (2005) Large macromolecular complexes in the protein data bank: a status report. Structure 13: 381–388.
10. Chiu W, Baker M, Almo S (2006) Structural biology of cellular machines. Trends in cell biology 16: 144–150.
11. Steven A, Baumeister W (2008) The future is hybrid. Journal of structural biology 163: 186–195.
12. Wang W, Donini O, Reyes C, Kollman P (2001) Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. Annual review of biophysics and biomolecular structure 30: 211–243.
13. Janin J (2002) Welcome to CAPRI: a critical assessment of predicted interactions. Proteins: Structure, Function, and Bioinformatics 47: 257–257.
14. Janin J (2010) The targets of CAPRI rounds 13–19. Proteins: Structure, Function, and Bioinformatics 78: 3067–3072.
15. Kirsanov D, Zanegina O, Aksianov E, Spirin S, Karyagina A, et al. (2013) NPIDB: nucleic acid–protein interaction database. Nucleic Acids Research 41: D517–D523.
16. Ritchie DW (2008) Recent progress and future directions in protein-protein docking. Current Protein and Peptide Science 9: 1–15.
17. Huang SY, Grinter SZ, Zou X (2010) Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. Physical Chemistry Chemical Physics 12: 12899–12908.
18. Zarrinpar A, Park SH, Lim WA (2003) Optimization of specificity in a cellular protein interaction network by negative selection. Nature 426: 676–680.
19. Wang J, Verkhivker G (2003) Energy landscape theory, funnels, specificity, and optimal criterion of biomolecular binding. Physical review letters 90: 188101.
20. Wang J, Zheng X, Yang Y, Drueckhammer D, Yang W, et al. (2007) Quantifying intrinsic specificity: A potential complement to affinity in drug screening. Physical review letters 99: 198101.

21. Yan Z, Wang J (2012) Specificity quantification of biomolecular recognition and its implication for drug discovery. Scientific reports 2: 309.
22. Yan Z, Guo L, Hu L, Wang J (2013) Specificity and affinity quantification of protein-protein interactions. Bioinformatics 29: 1127–1133.
23. Yan Z, Zheng X,Wang E,Wang J (2013) Thermodynamic and kinetic specificities of ligand binding. Chem Sci 4: 2387–2395.
24. Fleishman SJ, Baker D (2012) Role of the biomolecular energy gap in protein design, structure, and evolution. Cell 149: 262–273.
25. Havranek J, Harbury P, et al. (2003) Automated design of specificity in molecular recognition. nature structural biology 10: 45–52.
26. Shifman J, Mayo S (2003) Exploring the origins of binding specificity through the computational redesign of calmodulin. Proceedings of the National Academy of Sciences of the United States of America 100: 13274.
27. Kortemme T, Joachimiak L, Bullock A, Schuler A, Stoddard B, et al. (2004) Computational redesign of protein-protein interaction specificity. Nature structural & molecular biology 11: 371–379.
28. Bolon D, Grant R, Baker T, Sauer R (2005) Specificity versus stability in computational protein design. Proceedings of the National Academy of Sciences of the United States of America 102: 12724.
29. Grigoryan G, Reinke A, Keating A (2009) Design of protein-interaction specificity gives selective bzip-binding peptides. Nature 458: 859–864.
30. Janin J (1995) Principles of protein-protein recognition from structure to thermodynamics. Biochimie 77: 497–505.
31. Lu Q, Lu HP, Wang J (2007) Exploring the mechanism of flexible biomolecular recognition with single molecule dynamics. Physical review letters 98: 128105.
32. Bryngelson J, Onuchic J, Socci N, Wolynes P (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins: Structure, Function, and Bioinformatics 21: 167–195.
33. Janin J (1996) Quantifying biological specificity: the statistical mechanics of molecular recognition. Proteins: Structure, Function, and Bioinformatics 25: 438–445.
34. Rejto P, Verkhivker G (1996) Unraveling principles of lead discovery: From unfrustrated energy landscapes to novel molecular anchors. Proceedings of the National Academy of Sciences 93: 8945.
35. Miller D, Dill K (1997) Ligand binding to proteins: the binding landscape model. Protein science 6: 2166–2179.
36. Tsai C, Kumar S, Ma B, Nussinov R (1999) Folding funnels, binding funnels, and protein function. Protein Science 8: 1181–1190.
37. Dominy B, Shakhnovich E (2004) Native atom types for knowledge-based potentials: application to binding energy prediction. Journal of medicinal chemistry 47: 4538–4558.
38. Liu Z, Dominy B, Shakhnovich E (2004) Structural mining: self-consistent design on flexible protein-peptide docking and transferable binding affinity potential. Journal of the American Chemical Society 126: 8515–8528.
39. Levy Y,Wolynes P, Onuchic J (2004) Protein topology determines binding mechanism. Proceedings of the National Academy of Sciences of the United States of America 101: 511.
40. Spirin S, Titov M, Karyagina A, Alexeevski A (2007) NPIDB: a database of nucleic acids–protein interactions. Bioinformatics 23: 3247–3248.
41. Rose P, Beran B, Bi C, Bluhm W, Dimitropoulos D, et al. (2011) The rcsb protein data bank: redesigned web site and web services. Nucleic acids research 39: D392–D401.
42. Zhang C, Liu S, Zhu Q, Zhou Y (2005) A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. Journal of medicinal chemistry 48: 2325–2335.
43. Donald J, Chen W, Shakhnovich E (2007) Energetics of protein–DNA interactions. Nucleic Acids Research 35: 1039–1047.
44. Xu B, Yang Y, Liang H, Zhou Y (2009) An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles. PROTEINS: Structure, Function, and Bioinformatics 76: 718–730.
45. Norambuena T, Melo F (2010) The protein-DNA interface database. BMC bioinformatics 11: 262.
46. Van Dijk M, Bonvin A (2008) A protein–DNA docking benchmark. Nucleic acids research 36: e88–e88.
47. Luscombe N, Austin S, Berman H, Thornton J (2000) An overview of the structures of protein-DNA complexes. Genome biology 1: reviews001.
48. Barik A, Bahadur RP, et al. (2012) A protein–RNA docking benchmark (i): Nonredundant cases. Proteins: Structure, Function, and Bioinformatics 80: 1866–1871.
49. Bahadur R, Zacharias M, Janin J (2008) Dissecting protein–RNA recognition sites. Nucleic acids research 36: 2705–2716.
50. Pérez-Cano L, Jiménez-García B, Fernández-Recio J (2012) A protein-RNA docking benchmark (ii): Extended set from experimental and homology modeling data. Proteins: Structure, Function, and Bioinformatics 80: 1872–1882.
51. Gray J, Moughon S,Wang C, Schueler-Furman O, Kuhlman B, et al. (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. Journal of molecular biology 331: 281–299.
52. Chaudhury S, Berrondo M, Weitzner B, Muthu P, Bergman H, et al. (2011) Benchmarking and analysis of protein docking performance in rosetta v3. 2. PLoS One 6: e22477.
53. Koppensteiner W, Sippl M (1998) Knowledge-based potentials–back to the roots. Biochemistry (Mosc) 63: 247–252.
54. Jiang L, Gao Y, Mao F, Liu Z, Lai L (2002) Potential of mean force for protein–protein interaction studies. Proteins: Structure, Function, and Bioinformatics 46: 190–196.
55. Zhang C, Liu S, Zhu Q, Zhou Y (2005) A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. Journal of medicinal chemistry 48: 2325–2335.
56. Liu Z, Mao F, Guo J, Yan B, Wang P, et al. (2005) Quantitative evaluation of protein–DNA interactions using an optimized knowledge-based potential. Nucleic acids research 33: 546–558.
57. Su Y, Zhou A, Xia X, Li W, Sun Z (2009) Quantitative prediction of protein–protein binding affinity with a potential of mean force considering volume correction. Protein Science 18: 2550–2558.
58. 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. Journal of molecular biology 213: 859–883.
59. Clark M, Cramer III R, Van Opdenbosch N (1989) Validation of the general purpose tripos 5.2 force field. Journal of Computational Chemistry 10: 982–1012.
60. Guha R, Howard M, Hutchison G, Murray-Rust P, Rzepa H, et al. (2006) The blue obelisk interoperability in chemical informatics. Journal of chemical information and modeling 46: 991–998.
61. Thomas P, Dill K (1996) An iterative method for extracting energy-like quantities from protein structures. Proceedings of the National Academy of Sciences 93: 11628.
62. Thomas P, Dill K (1996) Statistical potentials extracted from protein structures: how accurate are they? Journal of molecular biology 257: 457–469.
63. Goldstein R, Luthey-Schulten Z, Wolynes P (1992) Optimal protein-folding codes from spin-glass theory. Proceedings of the National Academy of Sciences 89: 4918.
64. Muegge I, Martin Y (1999) A general and fast scoring function for protein-ligand interactions: a simplified potential approach. Journal of medicinal chemistry 42: 791–804.
65. Huang S, Zou X (2008) An iterative knowledge-based scoring function for protein–protein recognition. Proteins: Structure, Function, and Bioinformatics 72: 557–579.