Thermodynamic and kinetic specificities of ligand binding

Zhiqiang Yan, Xiliang Zheng, Erkang Wang and Jin Wang*

Supporting information

Derivation of initial atomic pairwise potentials. Here we simply introduce the main concept and basic steps of developing scoring function SPA, for detailed information, please see the paper [1]. The SPA (SPecificity and Affinity) is an optimized knowledge-based scoring function based on a set of initial atom-pair potentials. The initial potentials of SPA are directly derived from the Boltzmann relation widely used in the knowledge-based statistical potentials [2–5], which is

$$u_k = -K_B T \ln \mathrm{g}_k \tag{1}$$

where g_k is the observed pair distribution function which can be calculated from the known structures of protein-ligand complexes in the training database.

Optimization of potential energy function. To construct the optimizing potential energy function which aims to maximize the binding specificity and the consistence between predicted and experimental affinity, the initial energy function was rewritten by introducing a set of adjustable parameters as the coefficients c_k for the potentials of atom pair, that is

$$E = \sum_k c_k f_k u_k \tag{2}$$

where E is the total intermolecular energy of a protein-ligand complex. k stands for the type of atom pair interaction. f_k represents the occurrences of the interaction type k. As discussed in main text, the ISR as quantified intrinsic specificity for a given protein-ligand complex m is calculated as

$$\lambda_m = \alpha \frac{\delta E}{\Delta E} \tag{3}$$

where α is a scaling factor which accounts for the contribution of the entropy to the specificity [6, 7]. Here, it approximately depends on the number of torsional bonds of the ligands $\alpha \sim \sqrt{\frac{1}{n_{tb}}}$. δE is the energy gap between the energy of native conformation E_N and the average energy of ensemble of decoys $< E_D >$, and ΔE is the energy fluctuation or the width of the energy distribution of the decoys. All the conformational decoys of protein-ligand complexes were generated by the molecular docking with software AutoDock4.2 [8].

Combined with equations 2, the λ_m can be represented as

$$\lambda_m = \frac{\alpha|\sum_k c_k u_k (f_k^N - < f_k >)|}{\sum_k \sum_l c_k c_l u_k u_l (< f_k f_l > - < f_k >< f_l >)} \tag{4}$$

where k,l are the indices of the interaction types. Once f_k is computed for each interaction type in the decoys, one can readily compute the value of λ_m for a given set of c_k.

We need a single objective function that reflects the λ_m values for all the protein-ligand complexes in the training set. We chose the Bolzmann-like weighted average of λ_m as the objective function which is

$$\lambda = \frac{\sum_m \lambda_m \exp(\beta_\lambda \lambda_m)}{\sum_m \exp(\beta_\lambda \lambda_m)} \tag{5}$$

where β_λ is a constant value for weighting which is set as -0.1.

In addition to the ISR as the quantification of specificity, the quantitative measurements of the correlation between predicted and experimental affinity is depicted with Pearson's correlation coefficient by

$$\gamma = \frac{\sum_m (E_m^p - < E_m^p >)(E_m^e - < E_m^e >)}{\sqrt{\sum_m (E_m^p - < E_m^p >)^2} \sqrt{\sum_m (E_m^e - < E_m^e >)^2}} \tag{6}$$

The predicted binding affinity E_m^p for the protein-ligand complex is represented by the binding scores calculated from our scoring function with a given set of c_k. The experimentally measured affinity E_m^e is expressed in $\log K_d$ or $\log K_i$ units, where K_d and K_i are experimentally determined dissociation constant and inhibition constant respectively for the protein-ligand complex m.

The aim of optimization is to maximize the value of λ for specificity and the value of γ for affinity, a combination parameter ($\rho = \lambda\gamma$) which couples specificity and affinity is constructed to evaluate the performance of scoring function during the optimization. The optimization is performed by Monte Carlo search with simulated annealing in the space of adjustable coefficients c_k. At each MC step one of the coefficients is chosen at random and updated. This resulting change in E (E is defined as $E = -\rho$, minimizing E is equivalent of maximizing ρ) is accepted with the probability

$$P = min(1, \exp(-\beta_\rho \Delta E)) \tag{7}$$

where β_ρ^{-1} is the optimization temperature for ρ. The temperature β_ρ^{-1} decreases exponentially during the search. The convergence of both λ and γ indicates the optimized scoring function reaches the maximal performance of simultaneously quantifying the specificity and affinity.

Performance of SPA. The straightforward way to evaluate a novel scoring function is to make a comparison with other existing scoring function on their performances. SPA was tested on a benchmark of protein-ligand complexes which is a high-quality set of protein-ligand complexes selected out from the refined set of 2007 version of the PDBbind database [9]. This benchmark was taken as testing set to compare the performance for a large collection of 16 scoring functions implemented in main-stream commercial softwares or available from academic research groups, which offers a reference for the performance of SPA. It outperformed other 16 scoring functions on both predictions of binding pose and binding affinity [1], suggesting SPA is not only capable of discriminating the specific "native" conformation out of a large number of decoys by their scores but also accurately predicting the binding affinities of different protein-ligand complexes. This result is encouraging and motivating to apply SPA in the virtual screening to identify the lead compounds for drug discovery.

Statistics of discrimination. To quantify the discriminations of specific receptor target COX-2 from the competitive target COX-1 when binding with the selective drugs of COX-2, as well as the differences between selective and non-selective drugs of COX-2, the statistical method Kolmogorov-Smirnov test (KS test) was employed. The KS test quantifies a distance between two samples and determine if these two datasets differ significantly. The KS statistic is calculated as

$$D = max|(C_1(x) - C_2(x)| \tag{8}$$

where $C_1(x)$ and $C_2(x)$ are two cumulative distribution functions for compared datasets at cutoff x of a specific parameter X. In our work, the KS statistic were calculated for two kinds of comparisons; 1. the comparison between COX-2 and COX-1 with selective drugs, 2. the comparison between selective drugs and non-selective drugs with COX-2. The parameter X for the KS test can be set as affinity (E), ISR, residence time (RT) or the combination of them. The combinations of two parameters (E+ISR, or E+RT) or all three parameters (E+ISR+RT) were performed using logistic regression. The logistic regression is a type of regression analysis for predicting a dichotomous outcome, e.g. "Yes" vs. "No". For the first comparison, the selective drugs binding to COX-2 were set as 1 while the selective drugs binding to COX-1 were set as 0. For the second comparison, the selective drugs binding with COX-2 were set as 1 while the non-selective drugs binding with COX-2 were set as 0. The independent variables were linearly combined as one parameter by taking the regression

coefficients. The regressions were conducted with the package of R.

[1] Z. Yan and J. Wang, Scientific reports **2** (2012).

[2] M. J. Sippl, Journal of computer-aided molecular design **7**, 473 (1993).

[3] P. D. Thomas and K. A. Dill, Journal of molecular biology **257**, 457 (1996).

[4] H. Zhou and Y. Zhou, Protein Science **11**, 2714 (2009).

[5] S.-Y. Huang and X. Zou, Journal of computational chemistry **27**, 1866 (2006).

[6] J. Wang and G. M. Verkhivker, Physical review letters **90**, 188101 (2003).

[7] J. Wang, X. Zheng, Y. Yang, D. Drueckhammer, W. Yang, G. Verkhivker, and E. Wang, Physical review letters **99**, 198101 (2007).

[8] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson, Journal of computational chemistry **19**, 1639 (1998).

[9] T. Cheng, X. Li, Y. Li, Z. Liu, and R. Wang, Journal of chemical information and modeling **49**, 1079 (2009).

Table S1 30 selective (bold) and 20 non-selective nonsteroidal anti-inflammatory drugs (NSAIDs) of COX-2. Selective drugs (bold) are specific to inhibit COX-2, while non-selective drugs inhibit both the COX-2 and its isoenzyme COX-1. The predicted affinity (E^{pred} (kcal/mol)), ISR and residence time (RT^{pred}) are shown for the drugs. The known half maximal inhibitory concentrations (IC50 (uM)) and corresponding affinities (E^{exp} (kcal/mol)) for 20 drugs and experimentally determined half life (=0.693*residence time, RT^{exp} (hr)) for 22 drugs are also listed.

Drugs	IC50	E^{exp}	E^{pred}	ISR	RT^{pred}	RT^{exp}
ns-398	0.47	-8.686	-8.273	2.638	1326.734	N/A
l-745337	9.67	-6.884	-9.031	2.567	861.314	N/A
celecoxib	0.87	-8.319	-10.181	3.148	3111.631	11.2
rofecoxib	0.53	-8.615	-9.662	3.508	2171.419	17.0
dup-697	0.06	-9.913	-10.864	3.341	4406.395	292.0
jte-522	0.085	-9.706	-9.936	3.823	2701.849	N/A
valdecoxib	0.87	-8.319	-9.420	2.727	2146.821	9.5
etoricoxib	1.10	-8.179	-9.511	3.039	1635.223	22.0
meloxicam	0.70	-8.449	-8.991	3.301	3608.877	17.5
etodolac	3.70	-7.456	-7.566	2.332	497.923	7.3
l-776967	0.03	-10.327	-9.379	3.192	2079.798	N/A
flosulide	0.75	-8.408	-8.404	2.352	559.414	N/A
sulindac-sulfide	10.43	-6.838	-8.333	2.397	501.095	16.4
tolmetin	7.09	-7.069	-7.893	2.103	752.367	7.0
ketoprofen	1.08	-8.190	-8.526	2.715	1336.998	2.6
ketorolac	0.86	-8.326	-8.153	2.302	996.878	3.8
ibuprofen	24.3	-6.334	-7.490	2.124	483.644	3.0
flurbiprofen	6.42	-7.128	-7.925	1.819	474.535	5.2
tenoxicam	14.22	-6.653	-8.175	1.523	489.08	N/A
piroxicam	9.00	-6.92	-8.456	1.908	1269.679	5.0
sc-58125	N/A	N/A	-10.501	3.467	6184.121	N/A
644784	N/A	N/A	-9.859	3.298	2262.386	N/A
bms-347070	N/A	N/A	-8.935	2.089	774.402	N/A
cimicoxib	N/A	N/A	-10.299	3.558	4902.337	N/A
cis-stilbenes	N/A	N/A	-8.381	2.782	643.77	21.5

Drugs	IC50	E^{exp}	E^{pred}	ISR	RT^{pred}	RT^{exp}
ct-3	N/A	N/A	-7.805	2.245	512.9	N/A
darbufelone	N/A	N/A	-7.410	2.562	434.501	N/A
deracoxib	N/A	N/A	-10.594	3.669	5211.826	N/A
drf-4367	N/A	N/A	-8.741	2.361	750.353	N/A
fr-188582	N/A	N/A	-10.050	3.184	2001.471	N/A
parecoxib	N/A	N/A	-9.516	3.382	1639.617	N/A
pd-138387	N/A	N/A	-7.197	2.131	725.474	N/A
rs57067	N/A	N/A	-8.011	2.07	461.802	N/A
sc299	N/A	N/A	-10.029	3.428	2182.45	N/A
sc558	N/A	N/A	-10.239	2.974	2747.381	N/A
sc57666	N/A	N/A	-9.469	3.155	1518.078	N/A
svt-2016	N/A	N/A	-10.159	3.61	5456.822	N/A
t-614	N/A	N/A	-9.029	2.858	1432.131	N/A
bromfenac	N/A	N/A	-8.278	2.05	863.867	N/A
carprofen	N/A	N/A	-8.142	1.701	531.16	7.2
droxicam	N/A	N/A	-8.249	1.379	452.55	N/A
fenoprofen	N/A	N/A	-8.037	2.368	802.009	3.0
indoprofen	N/A	N/A	-7.833	1.625	568.764	2.3
loxoprofen	N/A	N/A	-7.647	2.173	618.49	1.2
meclofenamic-acid	N/A	N/A	-8.688	2.68	1115.646	3.1
oxaprozin	N/A	N/A	-8.487	2.859	1180.871	54.9
salicin	N/A	N/A	-7.328	2.731	799.314	N/A
tiaprofenic-acid	N/A	N/A	-8.998	2.757	1409.292	N/A
zomepirac	N/A	N/A	-7.996	2.232	570.107	N/A
indomethacin	N/A	N/A	-7.515	1.675	344.335	4.5