JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 11, 2012

Pp. 1–12
DOI: 10.1089/cmb.2012.0188

Research Article

Protein-specific Scoring Method for Ligand Discovery

I-LIN LU and HSIUYING WANG

ABSTRACT

Protein-based virtual screening plays an important role in modern drug discovery process. Most protein-based virtual screening experiments are carried out with docking programs. The accuracy of a docking program highly relies on the incorporated scoring function based on various energy terms. The existing scoring functions deal all the energy terms with the equal weight function or other weight function derived by physical characteristics. These existing scoring functions are not protein dependent. We expect that a protein-specific scoring function, which can reflect the protein characteristics, may improve the docking results. Therefore, we propose a protein-specific rescoring approach to select potential ligands by adjusting the weights of energy terms. The protein-specific scoring function is based on the linear regression analysis associated with an outlier detection approach. The scoring function incorporated in DOCK program is used as the model system. The performance of our method was evaluated by the DUD docked data set, which contains 40 protein targets. The study results show that this method can improve the enrichment factors for most of the 40 protein targets. We further expend the protein-specific scoring function to a larger database, and the results also show significant improvement. Our method is not limited to improving the DOCK scoring function. It can be adopted to improve other programs such as GOLD and Glide. We believe that this method can be applied to virtual screening experiments and elevates the hits rate significantly, which can be beneficial to the modern drug discovery process.

Key word: statistics.

1. INTRODUCTION

THE PHYSIOLOGICAL FUNCTIONS OF HUMAN BODY are maintained by various proteins. The dysfunction of specific proteins usually results in diseases, such as the shortage of insulin resulting in diabetes. Therefore, the first step in the drug discovery process is usually to obtain hit compounds that can bind to the target protein to inhibit or promote its function. A compound that can serve this biological purpose for a protein is called a ligand for this protein. Traditionally, the high throughput screening (HTS) approach is applied to several thousands of compounds to obtain ligands. However, this method is laborious and expensive.

A complementary method to HTS is structure-based virtual screening (Mestres, 2002), which relies on the key-and-lock theory that each protein structure is like a lock and we just need to find a key (ligand) to fit

Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan.

the lock. With the advance of computational techniques, this process can now be done on the computer (Dolghih et al., 2011; Prakhov et al., 2010; Sakakibara et al., 2012; Sircar and Gray, 2010; Whisenant et al., 2010), and many successful cases have been reported (Kuo et al., 2008; Lu et al., 2006a; Lu et al., 2006b). The central technique in structure-based virtual screening is docking, which refers to use computer programs to find ligands based on the target protein structure. A docking program generally consists of two parts: an algorithm to predict how a ligand binds to a protein (the position and orientation) (Dias and de Azevedo Jr, 2008), and scoring functions to retrieve the correct binding poses of interested ligands as well as distinguish binders from nonbinders (Jain, 2006).

DOCK is a popular docking program developed by Kuntz group in UCSF and can be obtained freely for academic institutes. The scoring function incorporated in DOCK was used as the model system in our study. In the DOCK version 3.5.54 scoring function (Meng et al., 1992), a compound is usually associated with four energy terms (E_{ele}, E_{vdw}, E_{pol}, E_{apol}), where E_{ele}, E_{vdw}, E_{pol}, and E_{apol} denote the electrostatic interaction energy, the van der Waals interaction energy, the polar component of the ligand desolvation energy, and the apolar part of the cost of ligand desolvation energy, respectively. Assume that we have a set of compounds that, from experiments, is shown to be ligands. The aim of this study is to find a model, based on the four energy terms of the known ligands, that selects potential active compounds from a very large compound library. A typical simple scoring function, with equal weight score, calculates the total docking energy E_{total}, which is defined as

$$E_{total} = E_{ele} + E_{vdw} + E_{pol} + E_{apol}$$

The selection criterion involves selecting compounds with smaller E_{total} values as potential ligands (Meng et al., 1992). Other scoring functions based on the physical or biological characteristics are adopted in DOCK 6.

In addition, many scoring functions have been developed (Jain, 2006) and comprehensively evaluated (Cheng et al., 2009). The results showed that no single scoring function outperformed others for all proteins, which means that a suitable scoring function may depend on the characteristics of the protein. That is, different protein may associate with different scoring functions. Instead of adopting the existing scoring functions, which may be derived based on biological background, we intend to provide a protein-specific scoring function from a statistical viewpoint. By modifying the original scoring function, the protein-specific scoring function would be a fast, simple, and efficient rescoring method.

To evaluate the performance of a scoring function, it is common to create noise compounds to examine if the scoring function can select the correct ligands in a set including ligands and noise data. Noise data is called a decoy. Directory of useful decoys (DUD) (Huang et al., 2006) docked dataset was chosen to validate our method, which included 40 proteins and their corresponding active ligands, as well as potential inactive molecules (decoys). A decoy for a ligand is created with similar energy terms as the ligand itself.

Therefore, with a tolerance interval approach and regression analysis, we propose a procedure to select better weights on energy terms in scoring functions for each specific target protein. Regression analysis is widely used for modeling several variables based on training data and can provide a useful tool for predicting and forecasting (Chekmarev et al., 2009), and as such, it should be suitable to apply on the scoring function modification. We adopt a linear regression analysis associated with an outlier detection approach in analyzing the ligand and decoy datasets to provide an efficient scoring function in a post-filter process. The improved scoring functions were evaluated in terms of the enrichment factor using small and large databases respectively.

2. MATERIALS AND METHODS

2.1. The DUD dataset

The DUD datasets were built for evaluating docking programs (Huang et al., 2006) (see Supplementary Material, available online at www.liebertonline.com/cmb). There are 40 target proteins and 2,950 active ligands in the DUD datasets. To avoid bias, for each ligand, 36 decoys were created, which are physically similar to the ligand in molecular weight, cLogP, and number of hydrogen bonding groups, but topologically different.

The docking procedure was performed on the 40 protein targets, with all ligands and decoys using DOCK program (Meng et al., 1992). The crystallographic ligand against its target protein was used as the controlled calculation, which means the docking procedure is validated. Therefore, DUD dataset should be a suitable data source for our post-filter method development.

2.2. Enrichment factor

The purpose of virtual screening is to discover active compounds from a large library. Our study focuses on re-ranking the active compounds to the top ranking position. Therefore, the enrichment factor (Cavasotto and Abagyan, 2004) should be a good indicator for evaluating the performance of the methods used in this study. The enrichment factor of a method for a selected subset is defined as the ratio of the percentage of active compounds in the selected subset to the percentage of active compounds in the entire database. The enrichment factor (*EF*) of a method is defined as:

$$EF = (ligand_{select}/N_{select})/(ligand_{total}/N_{total}),$$

where $ligand_{select}$ is defined as the number of active compounds in a subset selected by this method, N_{select} is defined as the number of compounds in this subset, $ligand_{total}$ is defined as the number of total active compounds in the entire library, and N_{total} is defined as the compound number of the entire library. A scoring function leading to a large *EF* value is more successful at selecting valid active compounds. A scoring function, with an EF value greater than one, means that this scoring function method is better than a randomly selected method. In this study, we select the top 1% and 20% subsets of the entire database to evaluate *EF* values of the scoring functions, that is

$$N_{select}/N_{total} = 0.01 \text{ or } 0.2.$$

2.3. The protein-specific scoring method

In this study, we develop a method to derive a protein-specific weight β_{ele}, β_{vdw}, β_{pol}, and β_{apol} on the energy terms E_{ele}, E_{vdw}, E_{pol}, and E_{apol} to obtain a new scoring function instead of using the equal weights on E_{ele}, E_{vdw}, E_{pol} and E_{apol}.

A protein-specific scoring function has the form

$$\beta_{ele}E_{ele} + \beta_{vdw}E_{vdw} + \beta_{pol}E_{pol} + \beta_{apol}E_{apol},$$

where the values of β_{ele}, β_{vdw}, β_{pol}, and β_{apol} depends on protein.

In the equal weight scoring function method, a compound with lower E_{total} score is regarded as having more potential to be an active compound than the compound with higher E_{total} score. In this study, a regression model approach is adopted to derive the weights β_{ele}, β_{vdw}, β_{pol}, and β_{apol}. For E_{ele}, E_{vdw}, E_{pol}, and E_{apol} in the protein-specific scoring function such that the enrichment factor for the protein-specific scoring function can be higher than that for the equal weight scoring function.

The proposed method can be performed in two steps: (1) outlier detection and (2) regression analysis. From our study, both steps are necessary to obtain better weights for the four energy terms. For a protein, suppose we have a ligand dataset from experiments and a decoys dataset. It may be due to incorrect placement in the docking process that is causing measurement errors for the energy terms E_{ele}, E_{vdw}, E_{pol}, and E_{apol} data. Thus, the first step is to evaluate if there are measurement errors of the four energy terms for the ligand data. Since the number of the decoys in the data is much larger than that of the ligand data and the decoys are created to have similar energy terms as the ligands, we use the decoy data with low score as a standard to detect the outliers for the ligands data. It is possible to use both decoy and ligand data sets to detect the outliers.

2.4. The first step: detection of outliers for ligand data

First, for a protein, we rank the decoys from the lowest to the highest score based on the equal weight scoring function E_{total}. The top r decoys were then used as a standard to detect the outliers, where r can be selected as the same number of ligands in the ligand dataset or 0.01 proportion of the total decoy number of the decoy dataset.

Let $\overline{E}_{ele}$, $\overline{E}_{vdw}$, $\overline{E}_{pol}$, and $\overline{E}_{apol}$ denote the sample means of the four energies E_{ele}, E_{vdw}, E_{pol}, and E_{apol} and let $st(E_{ele})$, $st(E_{vdw})$, $st(E_{pol})$, and $st(E_{apol})$ denote the sample standard deviations of the four energies based on the top r decoys. Using the sample means and standard deviations to construct tolerance intervals for the four energy terms, a ligand that has at least one energy term not belonging to the corresponding tolerance interval based on the top r decoys information was regarded as an outlier. A two-sided (γ − *content*, 1 − α *confidence*) tolerance interval of E_{ele} has the form of $(\overline{E}_{ele} - c\ st(\overline{E}_{ele}), \overline{E}_{ele} + c\ st(\overline{E}_{ele}))$, where c is a factor depending on the sample size. An upper (γ − *content*, 1 − α *confidence*) tolerance bound of E_{ele} has the form of $\overline{E}_{ele} + c\ st(\overline{E}_{ele})$. The tolerance intervals for the other energy terms have the same form as that for E_{ele}. Since, for a population, the tolerance interval contains, with specified confidence, at least a certain proportion of the population, it can be used to detect the outlier with suitable γ and α values, and in addition, it is also widely used in pharmaceutical applications (Cai and Wang, 2009; Hahn and Meeker, 1991; Wang, 2007; Wang and Tsung, 2009).

In this study, since a compound with smaller energy value is preferable, we adopt an upper tolerance bound instead of using a two-sided tolerance interval to detect outliers here. The factor c can be selected to be 1.64 corresponding to an upper (0.95 − *content*, 0.95 *confidence*) bound when the sample size is large. The factor value c depends on the sample size and can be obtained from literature (Hahn and Meeker, 1991). Note that in this step, we emphasize using the decoy data to detect outliers instead of using only ligand data. We have compared the outlier detection results using the decoy data with the result using the ligand data. It is shown that using only the ligand data cannot lead to a satisfactory result, which may be due to the small size of the ligand dataset.

2.5. The second step: regression analysis

According to the new ligand set that the outliers removed, a linear regression model is applied to fit the new dataset. First, we consider the regression model:

$$Y = \beta_{ele}E_{ele} + \beta_{vdw}E_{vdw} + \beta_{pol}E_{pol} + \beta_{apol}E_{apol} + e, \tag{1}$$

with $Y = a$ or b corresponding to a ligand or a decoy, respectively, where e is an error term, which is assumed to follow a normal distribution.

It is worth noting that the efficiency of the regression model depends on the selection of a and b. To distinguish the ligand and decoy data, we can set a and b such that $|a - b|$ is large. In this study, we set $a = 1$ and $b = 30$. In this model, we use all ligand data in the new set in which the outliers have been removed to fit the regression model. But we do not use all decoy data because it would cause the instability of the regression model. Only a few top decoys with low E_{total} score are selected to fit the model.

We intend to propose a model to well distinguish ligands from decoys with low E_{total} score. We rewrite (1) as

$$Y = Z \cdot \beta + e$$

where $Y = [1, 1, \cdots 30, 30]^T$, Z is the corresponding energy matrix, $\beta = (\beta_{ele}, \beta_{vdw}, \beta_{pol}, \beta_{apol})^T$ is the unknown parameter vector, and e is the error vector. Note that ''1'' in Y corresponds to the left ligands, which are not removed as outliers in the first step, and ''30'' in Y corresponds to decoys with low E_{total} score.

To derive β_{ele}, β_{vdw}, β_{pol}, and β_{apol}, we adopted the BLUE (best linear unbiased estimator) estimators $(\hat{\beta}_{ele}, \hat{\beta}_{vdw}, \hat{\beta}_{pol}, \hat{\beta}_{apol})$, which are the best linear unbiased estimators. The form of the blue estimators is

$$\hat{\beta} = (\hat{\beta}_{ele}, \hat{\beta}_{vdw}, \hat{\beta}_{pol}, \hat{\beta}_{apol}) = (Z^T Z)^{-1} Z^T Y. \tag{2}$$

Then the values $(\hat{\beta}_{ele}, \hat{\beta}_{vdw}, \hat{\beta}_{pol}, \hat{\beta}_{apol})$ in (2) are the proposed weights instead of the equal weight. We define $E^s_{total} = \hat{\beta}_{ele}E_{ele} + \hat{\beta}_{vdw}E_{vdw} + \hat{\beta}_{pol}E_{pol} + \hat{\beta}_{apol}E_{apol}$ and suggest using E^s_{total} to select the potential ligands.

In this study, we set a to be smaller than b such that the criterion of discovering a potential ligand is to select compounds with lower E^s_{total} value. In this study, we discuss the simple approach that Y value is a constant corresponding to all ligands and is a constant corresponding to all decoys. In fact, a more general approach is to assign different ligands with different Y values and different decoys with different Y values such that the Y values that correspond to the ligands are always smaller than those of the decoys. This general approach is still under investigation. We summarize the steps of the calculation procedure as follows.

Table 1. Enrichment Factors for Each Protein Targets by the Equal Weight Scoring Function (E_{TOTAL}) and Protein-specific Scoring Function

			Own decoys[a]				*All decoys*[b]			
			EF for equal weight scoring		*EF for protein-specifc scoring*		*EF for equal weight scoring*		*EF for protein-specific scoring*	
Protein	*No. of ligands*	*No. of decoys*	EF_1[c]	EF_{20}[d]	EF_1	EF_{20}	EF_1	EF_{20}	EF_1	EF_{20}
Nuclear hormone receptors										
AR	67	2592	17.91	3.21	22.39	3.43	28.36	3.96	37.31	4.10
$ER_{agonist}$	67	2346	5.97	3.06	22.39	3.51	8.96	4.18	55.22	4.78
$ER_{antagonist}$	38	1395	10.53	1.05	13.16	3.16	13.16	1.32	31.58	3.55
GR	65	2544	20	2.15	24.61	2.62	10.77	1.62	12.31	1.69
MR	12	497	33.33	4.58	41.67	4.17	58.33	4.17	50	4.17
PPARg	80	2511	0	0.13	5	0.75	0	0	3.75	1.5
PR	26	949	0	2.31	19.23	3.27	0	1.92	30.77	4.62
RXRa	20	624	15	2.75	25	3.25	25	2.25	45	3.5
Kinase										
CDK2	50	1549	8	1.4	12	1.6	12	1.4	12	2.2
EGFr	444	13112	3.38	2.15	6.98	2.41	2.03	2.17	7.21	2.60
FGFr1	118	4204	0	0.08	11.86	2.63	0	0.21	4.24	2.54
HSP90	24	782	8.33	2.08	16.67	2.29	8.33	1.88	4.17	1.88
P38 MAP	256	7824	1.17	1.58	7.81	1.84	1.56	2.19	10.94	2.54
PDGFrb	152	5008	0	0.23	1.32	2.17	0	0.39	0	1.58
SRC	155	5322	0.65	0.48	4.52	1	1.29	1.45	1.29	1.10
TK	22	772	0	2.5	0	2.05	13.64	4.55	9.09	4.55
VEGFr2	74	2637	2.70	1.01	6.76	1.82	1.35	1.42	1.35	0.68
Serine protease										
FXa	142	5079	5.63	2.39	16.90	4.12	13.38	3.84	20.42	4.26
thrombin	65	2289	4.62	2.54	10.77	1.77	13.85	2.92	4.62	1
Trypsin	44	1541	0	2.27	13.64	3.98	20.45	2.61	15.91	3.86
Metalloenzymes										
ACE	49	1711	6.12	2.14	20.41	2.24	40.82	3.67	42.86	4.08
ADA	19	809	15.79	2.89	36.84	3.42	15.79	2.89	31.58	3.42
COMT	10	428	0	3.5	10	2.6	0	4	10	4
PDE5	51	1808	5.88	1.86	13.72	1.96	11.76	2.25	7.84	2.16
Folate enzymes										
DHFR	201	7017	23.38	3.56	34.33	4.10	21.89	3.46	32.34	4.10
GART	21	603	4.76	4.05	9.52	4.28	42.86	3.33	71.43	4.76
Other enzymes										
AChE	105	3226	3.81	2.48	3.81	2.71	1.90	2	0.95	0.76
ALR2	26	918	15.38	2.31	26.92	2.12	38.46	2.31	30.77	2.12
AmpC	21	731	0	0.95	19.04	2.14	19.05	4.76	19.05	1.67
COX-1	25	824	8	2	16	1.4	4	1.2	0	0.2
COX-2	336	10240	14.88	3.39	15.77	3.74	19.05	3.24	0	3.07
GPB	49	1767	6.12	3.47	28.57	4.18	16.33	3.98	79.59	4.39
HIVPR	49	1863	2.04	0.92	14.29	1.53	4.08	2.35	0	1.12
HIVRT	37	1400	5.41	2.16	10.81	2.16	5.41	2.84	0	2.16
HMGR	35	1239	25.71	2.14	34.29	2.14	34.29	2.14	28.57	2.14
Inha	85	3032	0	0	31.76	2.35	0	0.29	1.18	2.12
NA	49	1737	10.20	3.16	6.12	3.47	20.41	3.27	48.98	3.57
PARP	33	1140	6.06	3.94	12.12	3.64	6.06	3.64	0	1.06
PNP	23	642	17.39	1.96	17.39	3.26	21.74	3.48	21.74	3.91
SAHH	33	751	6.06	3.64	15.15	3.78	18.19	4.55	27.27	4.55
Total		105,463								
Average			7.86	2.21	16.49	2.72	14.36	2.60	20.28	2.80-

[a]Own decoys indicated the enrichment factor was calculated based on the decoys against corresponding protein.
[b]All decoys indicated the enrichment factor was calculated based on all the decoys in the DUD dataset.
[c]EF_1 indicated the enrichment factor was calculated for the top 1% of the database.
[d]EF_{20} indicated the enrichment factor was calculated for the top 20% of the database.
EF, enrichment factor.

Procedure 1. Derive protein-specific weights for energies (E_{ele}, E_{vdw}, E_{pol}, E_{apol})

Step 1. Set an upper tolerance bound with factor c for the outlier detection. Remove ligands if at least one of the energy terms is not less than its corresponding upper tolerance bound. The left ligands are regarded as the normal ligands, which are used to proceed to **Step 2.**

Step 2. **Step 1** selects the ligands used for analysis. In this step, we select decoys from a decoy dataset for the regression analysis. The decoy number (n_{decoy}) we selected can range from 0 to several times the number of ligands, which do not include the outliers removed by **Step 1**. These ligands and decoys are used to obtain the coefficient $\hat{\beta} = (\hat{\beta}_{ele}, \hat{\beta}_{vdw}, \hat{\beta}_{pol}, \hat{\beta}_{apol})$ from Equation (2). The enrichment factor EF value is calculated with respect to this $\hat{\beta}$.

Step 3. Select c and n_{decoy} such that the enrichment factor EF corresponding to this procedure is large.

For a protein, we can adopt training data in Procedure 1 to obtain appropriate values for c and n_{decoy}, and then use these values to predict active compounds.

3. RESULTS AND DISCUSSION

3.1. The overall enrichment

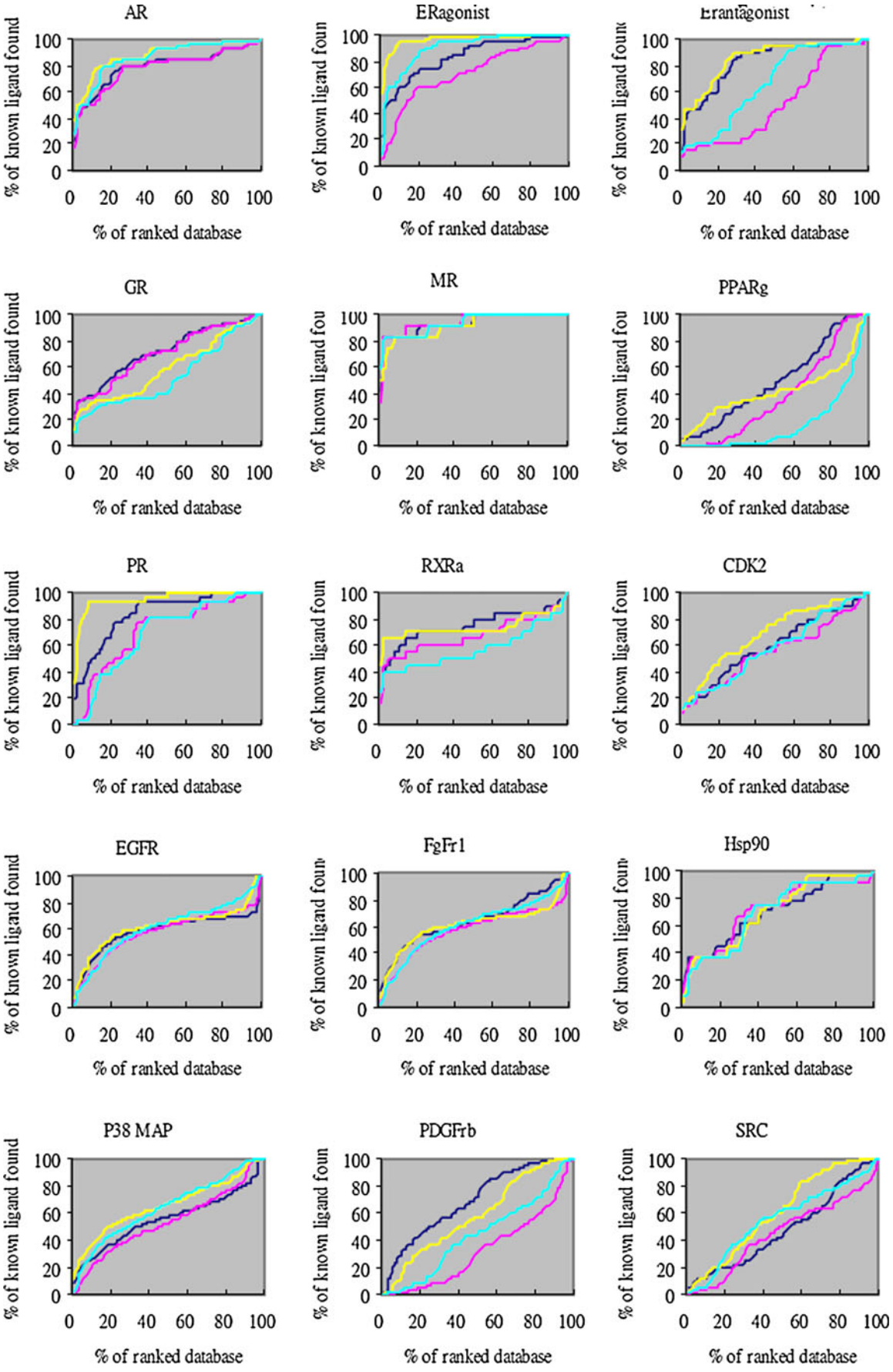
In this section, we use DUD docked dataset as our data source and apply the proposed procedure to obtain a protein-specific scoring function in virtual screening. There are forty proteins in DUD datasets. The proteins were grouped into six types by their function: nuclear hormone receptors, kinase, serine protease, metalloenzymes, folate enzymes and other enzymes. Each protein target has its own experimental confirmed ligands. For each ligand, 36 decoys were created. All ligands and decoys were docked to each protein, resulting in a very huge database. The number of ligands and decoys for each protein are listed in Table 1.

To compare the performance of the equal weight scoring function method and the protein-specific scoring function method, we first evaluate these two methods on the top 1% subset of the ranked database, which was ranked by each scoring function. We compare the enrichment factors of the equal weight scoring method and our protein-specific scoring method for the top 1% subset of each protein. The averages of the enrichment factors for 40 proteins using the equal weight scoring method and protein-specific weight scoring method were 7.86 and 16.49, respectively. There were 37 out of the 40 systems that the protein-specific scoring function leads to a larger enrichment factor than the equal weight scoring function. There were two (TK and AChe) out of the 40 systems with the same enrichment factor using the two methods. And only one system (NA) failed to be better enriched using the protein-specific weight scoring function. In addition, we also evaluate these two methods on the top 20% of subsets of the ranked database.

The averages of the enrichment factors for the top 20% subset using the equal weight scoring method and the protein-specific weight scoring method were 3.64 and 3.78, respectively. Although the difference of the average enrichment factor for the top 20% subset between the two scoring methods was not as significant as the case for the top 1% decoys subset, there were 31 out of the 40 systems that had greater enrichment factors using the protein-specific scoring method than that using equal weight scoring method. These results indicated that our method can significantly improve the equal weight scoring function. In general, virtual screening focuses on enriching the true hits of the top-ranked compounds of the database. Therefore, our method should be a potential tool for improving the validity of virtual screening.

The decoy data used in the above results for a specified protein is the corresponding decoy created for this protein. Since the enrichment factor of the 40 protein targets were improved, we further applied the protein-specific scoring function to all decoys and evaluated its performance. The total decoy number is 105,463 (Table 1). Results showed that both enrichment factors using the protein-specific scoring method for the top 1% and 20% subset were improved. The average of enrichment factors for the top 1% subset

FIG. 1. Docking enrichment plots for some 40 protein targets with equal weight scoring and protein-specific scoring using DUD dataset. The docking ranked database (x-axis) was plotted against the percentage of known ligands found by calculation (y-axis) at any given percentage of ranked database. The ''total database'' contained 3,178 DUD ligands and 105,463 decoys (cyan line and yellow line for equal weight scoring and modified scoring respectively), while the ''own decoys'' only include native ligands and their corresponding decoys (pink line and blue line for equal weight scoring and protein-specific scoring respectively).


AR
ERagonist
ERantagonist
GR
MR
PPARg
PR
RXRa
CDK2
EGFR
FgFr1
Hsp90
P38 MAP
PDGFrb
SRC
% of known ligand found
% of ranked database
0
20
40
60
80
100

using the equal weight score and the protein-specific score were 14.36 and 20.28, respectively, and for the top 20% subset were 2.60 and 2.80, respectively. These results indicated we can use a smaller database to derive our protein-specific scoring function and then apply this scoring function to a larger database.

Moreover, the overall profile of percentage of ligands found (y-axis) is plotted as a function of the percentage of the ranked docked database (x-axis) for each system (Figs. 1–3). The two different enrichment factors use two different background databases. The ''total database'' contains 3,178 DUD ligands and 105,463 decoys (cyan line and yellow line for the equal weight score and protein-specific score respectively), while the ''own decoys'' only includes the native ligands and their corresponding decoys (pink line and blue line for the equal weight score and protein-specific score respectively). The result shows that most of the percentage of true ligands found by the protein-specific score were greater than that from the equal weight score both in total database and own decoys. For the own decoys (the blue line and pink line), the protein-specific scoring method outperforms in several systems, such as ERagonist, ERantagonist, FgFr1, PDGFrb, etc. For the total database, the improvement was even more significant (the yellow and cyan line), such as the ERagonist, PR, DHFR, GART, etc. These results suggest that our method can sucessfully distinguish ligands from nonligands whether in small or large databases.

3.2. Comparison to other rescoring methods

Because detailed calculation of binding free energy between protein and ligands is time consuming and is not suitable for large-scale compound screening, approximation of binding affinity between proteins and ligands was adopted by most scoring functions to save time. However, this approximation procedure usually results in the inaccuracy of scoring function. To avoid the bias of single scoring function, Charifson et al. (1999) developed the consensus scoring method. They combined two different docking methods as well as 13 different scoring functions and applied their method on three targets of high pharmaceutical interest. Their results showed that the hit rates of the three targets were significantly improved.

However, other researchers demonstrated that the success of the consensus scoring method highly depends on the method of scoring function combination and the number as well as the nature of individual scoring functions involved in the combination (Wang and Wang, 2001). Muthans et al. (2008) applied pharmacophore constraints that derived from X-ray complex structures to post-filter the docking results. They used three docking programs and a database containing 9,997 compounds to evaluate six target proteins. Results showed that post-filtering with a pharmacophore indeed improves the docking performance over the six proteins.

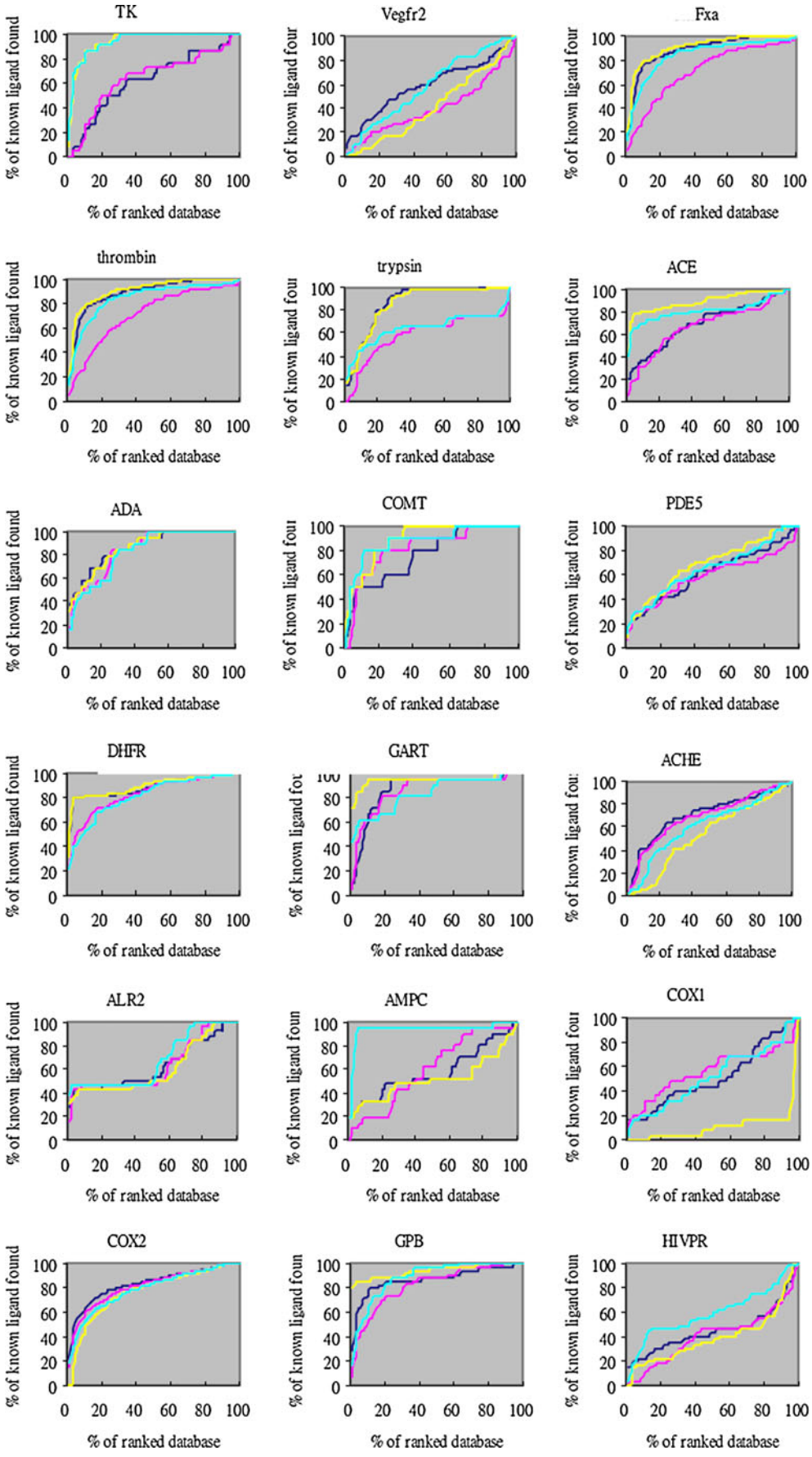
Kortagere et al. (2009) applied one-dimensional *Shape Signature* descriptors as a weighting factor to the Gold Score and obtained a docking prediction accuracy of 61% on the human pregnane X receptor (PXR). Several programs have also incorporated the post-processing function to automate the post-filter process. Most of the post-filter methods were developed using the interaction pattern, and large-scale validation of these methods are unavailable. Moreover, some methods are bound to specific programs, resulting in difficulties for popular usage.

We focused on improving the DOCK scoring function in this study. DOCK is software that can be used by academic researchers freely. We applied our model to all the 40 target proteins in DUD docked set and the result is promising. The code for implementing the proposed method is available online (see Supplementary Material, available at www.liebertonline.com/cmb).

4. CONCLUSION

In this study, we adopt a method based on tolerance interval and regression analysis approaches to develop a procedure for obtaining a protein-specific weight for the energy terms to increase the enrichment

FIG. 2. Docking enrichment plots for some 40 protein targets with equal weight scoring and protein-specific scoring using DUD dataset. The docking ranked database (x-axis) was plotted against the percentage of known ligands found by calculation (y-axis) at any given percentage of ranked database. The ''total database'' contained 3,178 DUD ligands and 105,463 decoys (cyan line and yellow line for equal weight scoring and modified scoring respectively), while the ''own decoys'' only include the native ligands and their corresponding decoys (pink line and blue line for equal weight scoring and protein-specific scoring respectively).


TK
Vegfr2
Fxa
thrombin
trypsin
ACE
ADA
COMT
PDE5
DHFR
GART
ACHE
ALR2
AMPC
COX1
COX2
GPB
HIVPR
% of known ligand found
% of ranked database
0
20
40
60
80
100

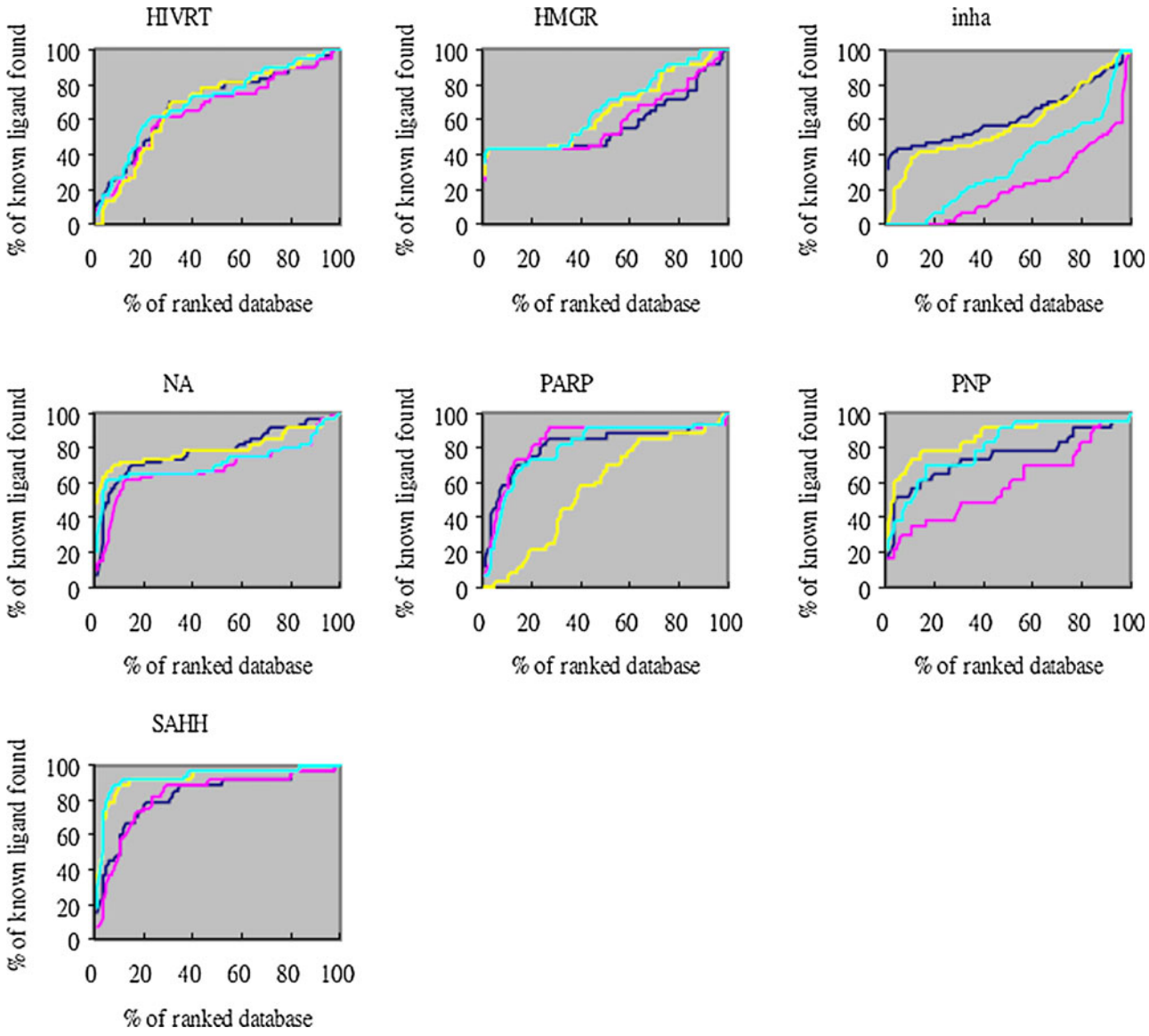


FIG. 3. Docking enrichment plots for some 40 protein targets with equal weight scoring and protein-specific scoring using DUD dataset. The docking ranked database (x-axis) was plotted against the percentage of known ligands found by calculation (y-axis) at any given percentage of ranked database. The ''total database'' contained 3,178 DUD ligands and 105,463 decoys (cyan line and yellow line for equal weight scoring and modified scoring respectively), while the ''own decoys'' only include the native ligands and their corresponding decoys (pink line and blue line for equal weight scoring and protein-specific scoring respectively).

factor. Basically, we believe that the conventional equal weight scoring approach established from a scientific basis has its own merits. However, due to different features of proteins, this scoring function is not rigorous enough for protein target predictions.

We have modified the scoring functions so that they can be based on the characteristic of the proteins to predict active compounds. Our results show that this protein-specific scoring method could successfully improve the equal weight scoring function for the 40 DUD datasets. It can also be expanded to larger databases. Furthermore, this method is not limited to the DOCK scoring function. It can be applied to modifying other scoring functions, such as GOLD score and Glide score. We believe that this method can elevate the hits rate significantly, which can be beneficial to the modern drug discovery process.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Council (NSC) and National Center for Theoretical Sciences (NCTS) in Taiwan.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Cai, T., and Wang, H. 2009. Tolerance intervals for discrete distributions in exponential families. *Stat Sinica*. 19, 905–923.

Cavasotto, C.N., and Abagyan, R.A. 2004. Protein flexibility in ligand docking and virtual screening to protein kinases. *J Mol Biol*., 209–225.

Charifson, P.S., Corkery, J.J., Murcko, M.A., et al. 1999. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem*. 42, 5100–5109.

Chekmarev, D., Kholodovych, V., Kortagere, S., et al. 2009. Predicting inhibitors of acetylcholinesterase by regression and classification machine learning approaches with combinations of molecular descriptors. *Pharmaceut Res*. 26, 2216–2224.

Cheng, T., Li, X., Li, Y., et al. 2009. Comparative assessment of scoring functions on a diverse test set. *J Chem Inf Model*. 49, 1079–1093.

Dias, R., and de Azevedo Jr, W.F. 2008. Molecular Docking Algorithms. *Current Drug Targets*. Bentham Science Publishers Ltd., Oak Park, IL, 1040–1047.

Dolghih, E., Bryant, C., Renslo, A.R., et al. 2011. Predicting binding to p-glycoprotein by flexible receptor docking. *PLoS Comput Biol*. **7**, e1002083.

Hahn, G.J., and Meeker, W.Q. 1991. *Statistical Intervals: A Guide for Practitioners*. Wiley, New York.

Huang, N., Shoichet, B.K., and Irwin, J.J. 2006. Benchmarking sets for molecular docking. *J Med Chem*. 49, 6789–6801.

Jain, A.N. (2006) Scoring Functions for Protein-Ligand Docking. *Current Protein & Peptide Science*. Bentham Science Publishers Ltd., Oak Park, IL, 407–420.

Kortagere, S., Chekmarev, D., Welsh, W., et al. 2009. Hybrid scoring and classification approaches to predict human pregnane x receptor activators. *Pharmaceut Res* 26, 1001–1011.

Kuo, C.J., Guo, R.T., Lu, I.L., et al. 2008. Structure-based inhibitors exhibit differential activities against Helicobacter pylori and Escherichia coli undecaprenyl pyrophosphate synthases. *J Biomed Biotechnol*. 2008, 841312.

Lu, I.L., Huang, C.F., Peng, Y.H., et al. 2006. Structure-based drug design of a novel family of PPARgamma partial agonists: virtual screening, X-ray crystallography, and in vitro/in vivo biological activities, *J Med Chem*. 49, 2703–2712.

Lu, I.L., Mahindroo, N, Liang, P.H., et al. 2006. Structure-based drug design and structural biology study of novel nonpeptide inhibitors of severe acute respiratory syndrome coronavirus main protease. *J Med Chem*. 49, 5154–5161.

Meng, E.C., Shoichet, B.K., and Kuntz, I.D. 1992. Automated docking with grid-based energy evaluation. *J Comput Chem*. 13, 505–524.

Mestres, J. 2002. Virtual screening: a real screening complement to high-throughput screening, *Biochem. Soc. Trans.* 30, 797–799.

Prakhov, N.D., Chernorudskiy, A.L., and Gainullin, M.R. 2010. VSDocker: a tool for parallel high-throughput virtual screening using AutoDock on Windows-based computer clusters. *Bioinformatics*. 26, 1374–1375.

Sakakibara, Y., Hachiya, T., Uchida, M., et al. 2012. COPICAT: a software system for predicting interactions between proteins and chemical compounds. *Bioinformatics*. 28, 745–746.

Sircar, A., and Gray, J.J. 2010. SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput Biol*. 6, e1000644.

Wang, H. 2007. Estimation of the probability of passing the USP dissolution test. *J Biopharm Stat*. 17, 407–413.

Wang, H., and Tsung, F. 2009. Tolerance intervals with improved coverage probabilities for binomial and poisson variables. *Technometrics*. 51, 25–33.

Wang, R., and Wang, S. 2001. How does consensus scoring work for virtual library screening? an idealized computer experiment. *J Chem Inf Comp Sci*. 41, 1422–1426.

Whisenant, T.C., Ho, D.T., Ryan, W., et al. 2010. Computational prediction and experimental verification of new map kinase docking sites and substrates including gli transcription factors. *PLoS Comput Biol*. 6, e1000908.

Address correspondence to:
Hsiuying Wang
National Chiao Tung University
Institute of Statistics
Hsinchu, 300
Taiwan

E-mail: wang@stat.nctu.edu.tw