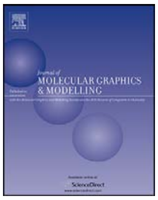
Journal of Molecular Graphics and Modelling 29 (2010) 492–497

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM



Prediction of protein–ligand binding affinities using multiple instance learning

Reiji Teramoto[a,*], Hisashi Kashima[b]

[a] Advanced Technology Solutions Division, NEC Informatec Systems, Ltd., 2-6-1, Kitamigata, Takatsu-ku, Kawasaki, Kanagawa 213-8511, Japan
[b] Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-31 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

ARTICLE INFO

Article history:
Received 9 June 2010
Received in revised form
17 September 2010
Accepted 21 September 2010
Available online 29 September 2010

Keywords:
Scoring function
Binding affinity
Docking program
Virtual screening
Structure-based drug design
Machine learning
Multiple instance learning
Boosting

ABSTRACT

Accurate prediction of protein–ligand binding affinities for lead optimization in drug discovery remains an important and challenging problem on scoring functions for docking simulation. In this paper, we propose a data-driven approach that integrates multiple scoring functions to predict protein–ligand binding affinity directly. We then propose a new method called multiple instance regression based scoring (MIRS) that incorporates unbound ligand conformations using multiple scoring functions. We evaluated the predictive performance of MIRS using 100 protein–ligand complexes and their binding affinities. The experimental results showed that MIRS outperformed the 11 conventional scoring functions including LigScore, PLP, AutoDock, G-Score, D-Score, LUDI, F-Score, ChemScore, X-Score, PMF, and DrugScore. In addition, we confirmed that MIRS performed well on binding pose prediction. Our results reveal that it is indispensable to incorporate unbound ligand conformations in both binding affinity prediction and binding pose prediction. The proposed method will accelerate efficient lead optimization on structure-based drug design and provide a new direction to designing of new scoring score functions.

1. Introduction

Accurate prediction of protein–ligand binding affinity is a key to lead optimization in structure-based drug discovery. Molecular dynamics (MD) and Monte Carlo simulations (MC)-based approaches are two major approaches to predict binding affinity. For example, MD-based approaches, such as MM-PB/SA [1] and linear interaction energy [2,3], calculate binding free energy based on the assumption that the free energy change in a protein–ligand binding process can be computed by only considering the difference between the unbound state and bound state. On the other hand, MC-based approaches, such as free energy perturbation [4] and thermodynamics integration [5], conduct integration along the free energy pathway between two closely resembled systems. Although these MD- and MC-based approaches are able to predict binding affinities accurately, the need for explicit handling of the solvent to describe desolvation effects and hydrogen bonding prevents us from applying those approaches to practical large-scale virtual screening because of their intensive computational costs. On the other hand, protein–ligand docking approach is used for discovering novel ligands in large compound databases with reasonable speed. In the protein–ligand docking approach, scoring functions are calculated by various potentials of protein–ligand complex structures. Since the computation of scoring functions is much more efficient than the MD- and MC-based approaches, this approach is suited for large-scale structure-based virtual screening [6–8]. However, although various scoring functions have been developed, recent studies have shown that the binding scores predicted by the scoring functions exhibit poor correlations with the actual binding affinities, resulting in many false positive binding predictions in their “hit lists” [8]. Therefore, the low accuracy of predicted binding affinities is a major impediment to the success of many structure-based drug design projects, and there is a clear demand to achieve more accurate protein–ligand binding affinities by improving or exploiting the scoring functions [6–10].

To this goal, we propose a data-driven scheme for designing a new score function by using the actual binding affinity data and existing scoring functions. Recycling the predecessors’ knowledge (i.e. the existing scoring functions), we construct a new score function as a function of the existing score functions. We use machine learning techniques to tune the new function to fit the actual binding affinity data. Note that the consensus scoring approach [6] also integrates multiple scoring functions by simple calculations (such as averaging), its aim is binding pose prediction and hit identification, and it is not reasonable to apply the consensus scoring approach to binding affinity prediction.

Generally, scoring functions are designed to reproduce binding affinities at the position of X-ray crystal structures of protein–ligand complexes [11]. They implicitly assume that a protein and a ligand bind at the position of the X-ray crystal structure of

* Corresponding author. Tel.: +81 50 3757 5333; fax: +81 50 3757 5334.
E-mail address: r-teramoto@bq.jp.nec.com (R. Teramoto).

1093-3263/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2010.09.006

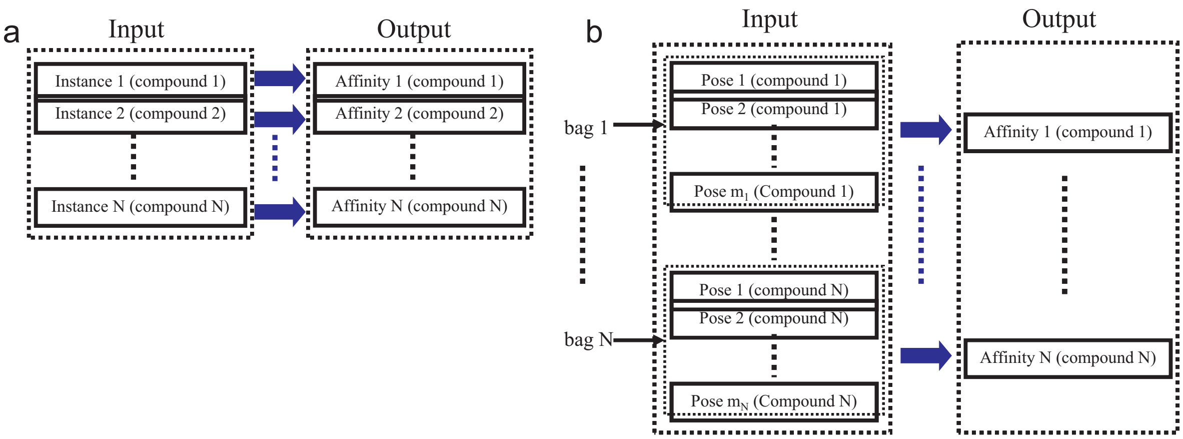


Fig. 1. Illustration of the difference between ordinary supervised learning and multiple instance learning (MIL). (a) Ordinary supervised learning. (b) MIL. MIL has more than one instance for the objective variables in a bag.

a protein–ligand complex and do not incorporate unbound ligand conformations. However, we often do not know the X-ray crystal structure of a complex or the binding pose a priori when predicting the binding affinity. Moreover, the distribution of unbound ligand conformations reflects the protein–ligand binding process associated with binding free energy landscape [12]. Hence, it is crucial to incorporate unbound ligand conformations to predict binding affinities. To incorporate unbound ligand conformations in design of scoring functions, we use the ensemble of docking poses as a part of unbound ligand conformations. We formulate this problem as a multiple instance learning problem which treats the ensemble of docking poses generated from a ligand as one datum. Although the ensemble of docking poses only involves a part of conformational space of a ligand, we experimentally show that they are useful to improve the performance of scoring functions.

Consequently, we propose the multiple instance regression based scoring (MIRS) method that incorporates unbound ligand conformations by using multiple scoring functions to predict protein–ligand binding affinities directly. We evaluate the predictive performance of MIRS using 100 protein–ligand complexes and their binding affinities. Experimental results show that the proposed method, MIRS, outperformed the conventional scoring functions. The results reveal that it is indispensable to incorporate unbound ligand conformations on both binding affinity prediction and binding pose prediction.

2. Methods

2.1. Preparation of data sets

We use publicly available data sets that have been used to benchmark the performance of scoring functions [13]. The data sets consist of the 100 protein–ligand complexes. In each complex, 100 decoys were generated by AutoDock. Each of decoys corresponds to a docking pose of a ligand. All decoys of each ligand were scored using 11 scoring functions: LigScore, PLP [14], AutoDock [15], G-Score [16], D-Score [17], LUDI [11], F-Score [18], ChemScore [19], X-Score [20], PMF [21], and DrugScore [22]. These scoring functions can be grouped into three types: (1) force field (AutoDock, G-Score and D-Score); (2) empirical scoring functions (LigScore, PLP, LUDI, F-Score, ChemScore and X-Score); and (3) knowledge-based potentials (PMF and DrugScore). Each type of scoring function has its own advantage and disadvantage. The detailed descriptions of the data-generation procedure and the data sets are available on-line at http://sw16.im.med.umich.edu/software/xtool.

2.2. Data-driven integration of the multiple scoring functions

Given a pair of a protein and a ligand, we would like to predict the binding affinity of the ligand to a particular protein, and which binding pose is used for forming the protein–ligand complex (when the predicted affinity strength is sufficiently high). To build the prediction model, we take the data-driven approach, that is, we utilize several complexes whose binding affinities are known experimentally. Let denote the given data by $D = \left\{(\text{complex}_i, \text{affinity}_i), i = 1, \ldots, N\right\}$, where each of the elements of D is a pair of the scores for a protein–ligand complexes and its affinity. We apply regression learning methods to the data to obtain $\text{affinity}_i \approx f(\text{complex}_i)$ which is a function approximating the relationship between scores by the existing scoring functions and affinity.

2.3. Multiple instance regression based scoring (MIRS)

In protein–ligand docking, each complex has several (100 in our experiments) docking poses, but we do not know which pose is actually used for the binding. Supposing that only one of them (or, at least a few of them) actually contributes to the binding affinity, the parameter estimation problem is well formulated as a multiple instance regression problem, which has been studied in the machine learning community [23–25]. At first, we describe the summary of multiple instance regression. Multiple instance regression (MIR) is a variant of multiple instance learning (MIL) in which each datum corresponds to a set of instances called a bag [23–25]. We illustrate the difference between the ordinary supervised learning and MIL in Fig. 1. As described in Fig. 1, each instance is associated with one output value in the ordinary supervised learning, while each bag is associated with one output value in MIL. In this study, an instance and a bag correspond to a docking pose and a set of docking poses, respectively. Our goal is to predict the binding affinity (the objective variables) for each bag of candidates poses. To formulate this task, we define the problem of MIR as follows. We denote a set of N bags by $D = \left\{B_i, i = 1, \ldots, N\right\}$, where $B_i = \left\{(x_{ij}, y_i), j = 1, \ldots, m_i\right\}$, x_{ij} is a score vector of jth instance in the ith bag, and y_i is protein–ligand binding affinity of ith bag. An attribute vector and a real-valued label correspond to the scores of 11 scoring functions and the binding affinity, respectively. Fig. 2 illustrates the multiple instance regression for the problem of binding affinity prediction. From Fig. 2, it appears that an important issue is how to predict the binding affinity from an unseen docking pose of a compound. It is reasonable to regard the binding pose with the maximum predicted binding score as the most plausible

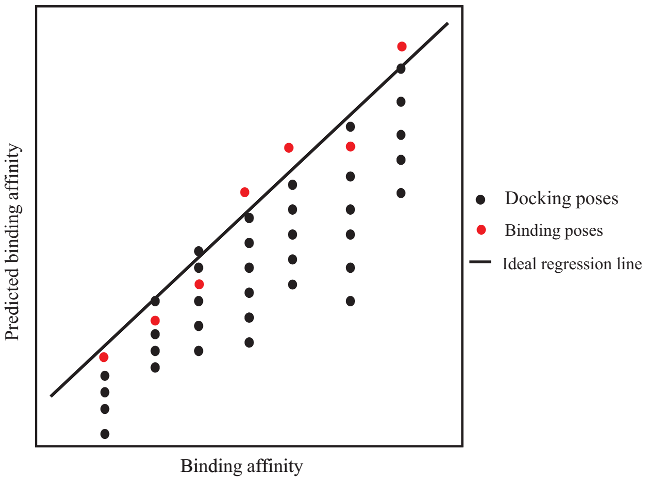
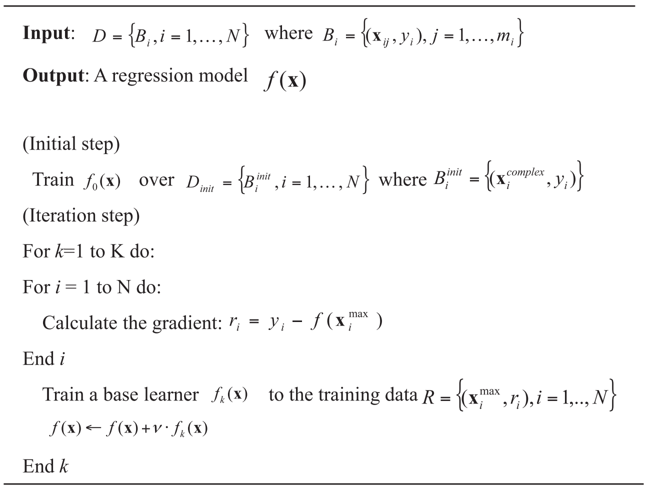


Fig. 2. Illustration of multiple instance regression (MIR) on the problem of binding affinity prediction. MIR trains a regression model from multiple docking poses and predicts a binding affinity.



```
Input: $D = \{B_i, i = 1,\ldots,N\}$ where $B_i = \{(\mathbf{x}_{ij}, y_i), j = 1,\ldots,m_i\}$
Output: A regression model $f(\mathbf{x})$

(Initial step)
  Train $f_0(\mathbf{x})$ over $D_{init} = \{B_i^{init}, i = 1,\ldots,N\}$ where $B_i^{init} = \{(\mathbf{x}_i^{complex}, y_i)\}$
(Iteration step)
For $k$=1 to K do:
For $i$ = 1 to N do:
    Calculate the gradient: $r_i = y_i - f(\mathbf{x}_i^{\max})$
End $i$
    Train a base learner $f_k(\mathbf{x})$ to the training data $R = \{(\mathbf{x}_i^{\max}, r_i), i = 1,..,N\}$
    $f(\mathbf{x}) \leftarrow f(\mathbf{x}) + \nu \cdot f_k(\mathbf{x})$
End $k$
```

Fig. 3. Algorithm of MIRS.

binding pose, since the docking programs search the most stable configuration of a ligand. We then consider the following loss function to minimize:

$$L(y_i, f(x_i^{\max})) = \frac{1}{2}\sum_{i=1}^{N}(y_i - f(x_i^{\max}))^2 \qquad (1)$$

where

$$f(x_i^{\max}) = \max(f(x_{ij}), j = 1, \ldots, m_i)$$

To obtain $f(x)$ that minimizes the above loss function (1), we employ the gradient boosting approach [26]. In the gradient boosting, the model $f(x)$ is defined as a sum of K base models

$$f(x) = \sum_{k=1}^{K} f_k(x)$$

where $f_k(x)$ is the kth base model. At the kth stage of the gradient boosting, a new base model $f_k(x)$ is added to decrease the value of the loss function by the current model. To estimate $f_k(x)$, for each bag B_i, we obtain the gradient of the loss function (1) with respect to $f(x_i^{\max})$ as

$$r_i = -\frac{\partial L(y_i, f(x_i^{\max}))}{\partial f(x_i^{\max})} = y_i - f(x_i^{\max})$$

which is called a "pseudo residual". The $f_k(x)$ is obtained by feeding $R = \{(x_i^{\max}, r_i), i = 1, \ldots, N\}$ to an existing regression algorithm (we used the random forests method [27] as the regression algorithm because such tree models are capable of capturing non-linear relationships between scoring functions and binding affinities, and also known to be compatible with the boosting method), and $f_k(x)$ is added to the current model $f(x)$. In our implementation, to avoid over-fitting to the data, we multiply a small constant ν to $f_k(x)$, which results in the updating formula $f(x) \leftarrow f(x) + \nu \cdot f_k(x)$.

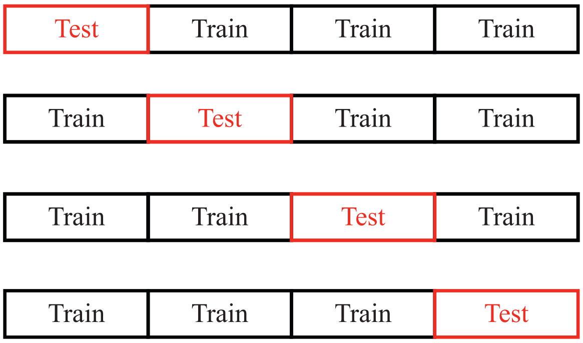
At the initial stage, since we do not have any model and hence we do not know $x_i^{\max}$, we train a regression model $f(x)$ by using the X-ray protein–ligand complex crystal structures. At the following stages, the optimization proceeds by using the gradient descent method. We show the gradient descent-based optimization algorithm in Fig. 3.

We employed the random forests method [27] as the base regression algorithm, and the number of iteration was set to 50. We set the step size parameter ν to 0.1. We implemented our algorithm in R which is a language for statistical computing and used randomForest R package for random forest implementation. The random forests method combines two machine leaning techniques: bagging and random feature subset selection. Bagging uses the bootstrap sampling technique to produce pseudo replicates to improve predictive accuracy. The random forests method further improves the performance through random feature subset selection.

3. Results and discussions

3.1. Experimental results of binding affinity prediction

We test the performance of our MIRS in binding affinity prediction by using a data set with 100 protein–ligand complexes, and compared it with the existing 11 scoring functions. We used the Pearson correlation coefficient and the mean squared error (MSE) between the predicted binding scores and binding affinities ($-\log Kd$) as the performance measures to evaluate linear correlation and deviation of values of scoring functions. To fairly compare the performance of MIRS and the 11 scoring functions, we performed 10-fold cross-validation. The procedure of cross-validation is depicted in Fig. 4. As shown in Fig. 4, a data set is divided up into K roughly equally sized parts at random. By keeping out one of them, a prediction model is estimated by using the other $K-1$ parts, then the performance of the model is evaluated by using the remaining part. Iterating the evaluation K times by changing the kept-out part, cross-validation estimates the final predictive performance, such as correlation coefficient and mean squared error, by averag-



Test	Train	Train	Train
Train	Test	Train	Train
Train	Train	Test	Train
Train	Train	Train	Test

Fig. 4. The procedure of K-fold cross-validation (the case of $K=4$). The data sets are divided up at random into K roughly equally sized parts. For each part in turn, the prediction model is built on the other $K-1$ parts then tested on the remaining part.

ing the results of *K* cases. It is well known that cross-validation can give good estimates of true predictive performance [28–30]. Note that in our cross-validation, the model is trained by using only the training data, and the test data (the prediction data) is completely external from the training data and is used only for evaluation. All of the reported results are the averaged performance over the independent tests, not on the training. In MIRS, we set the number of iteration to 50. We used the simple linear regression method for finding the best scaling for each of the 11 scoring functions. Table 1 shows the experimental results of binding affinity prediction. MIRS significantly outperformed the 11 scoring functions on both the Pearson correlation coefficient and the MSE. Especially, the MSE of MIRS is very small compared to those of the 11 scoring functions. Interestingly, the correlation coefficients and MSEs of the scoring functions are not correlated, which implies that scoring functions with high Pearson correlations do not necessarily accompany low MSEs, and we need non-linear scaling such as the random forests to perform well on both performance measures. From the above results, we conclude that our data-driven score integration with the appropriate multi-instance learning model significantly improves the quality of quantitative prediction of binding affinities.

3.2. Binding pose prediction

To evaluate the performance of binding pose prediction, we investigated two performance metrics by 10-fold cross-validation as well as binding affinity prediction. One is a success rate which is the percentage of whether the RMSD of top-scored docking pose is smaller than a given threshold, e.g. 2 Å. The other is an average RMSD which is the average of RMSD of top-scored docking pose (Table 2). From Table 2, it appears that our MIRS is competitive with PLP, LigScore and DrugScore on both the success rate and the average RMSD. However, PLP, LigScore and DrugScore exhibit poor performance on binding affinity prediction from Table 1. While X-Score is superior to other conventional scoring functions on binding affinity prediction, X-Score is largely inferior to them on binding pose prediction. Thus, the existing scoring functions do not consistently work on binding affinity prediction and pose prediction. From the practical viewpoint, this inconsistency becomes the very serious problem, because one cannot choose the best scoring function for binding affinity prediction by evaluating the binding pose prediction and vice versa. On the other hand, from Tables 1 and 2, MIRS works well on both binding affinity prediction and binding

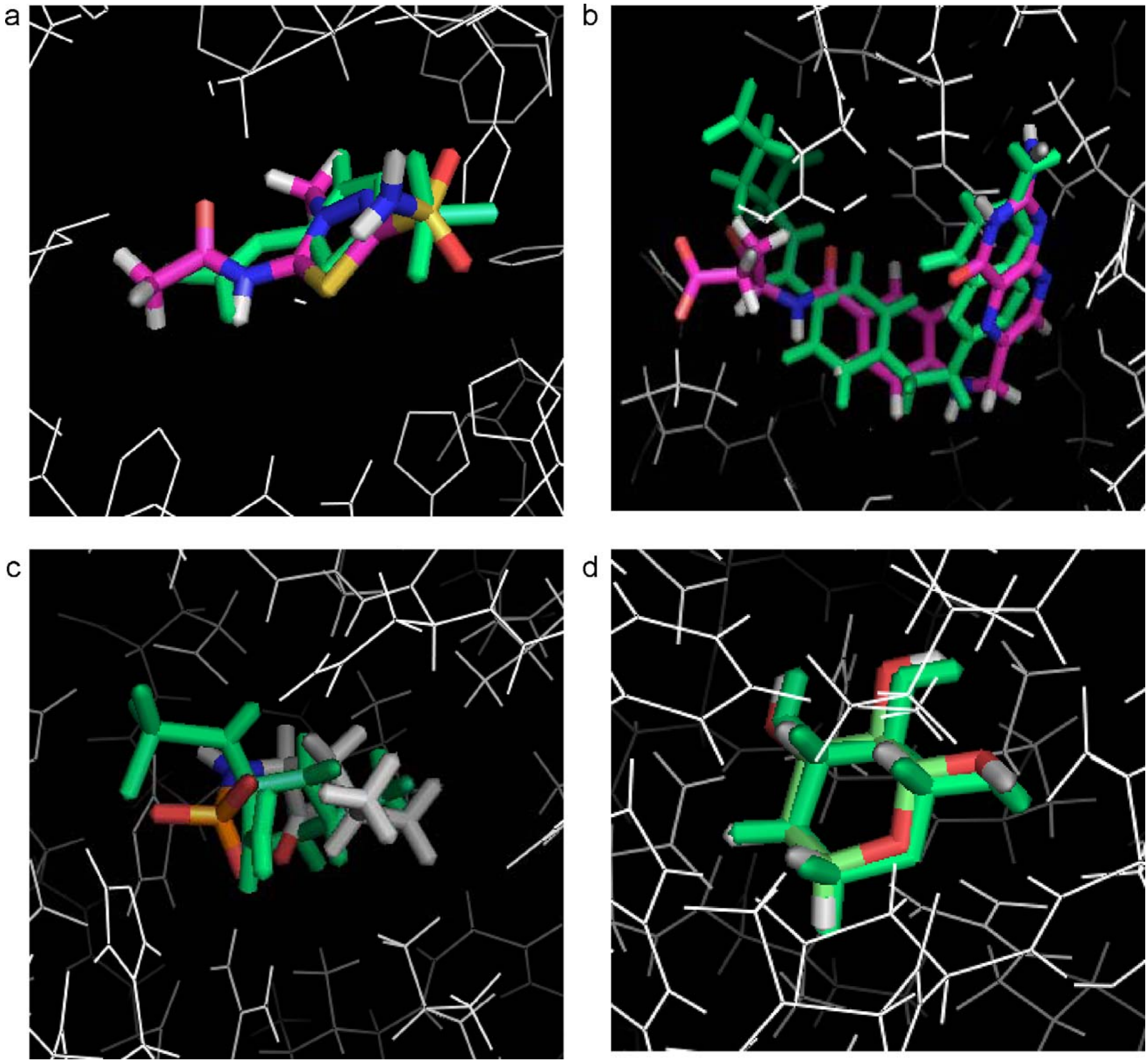


Fig. 5. Representative examples of binding pose predicted by MIRS close to X-ray structure of a ligand. Green molecule is the X-structure of a ligand, and color-by-atom is the binding pose predicted by MIRS. (a) Carbonic anhydrase I with sulfonamide (rmsd = 1.65 Å, PDB ID: 1bzm). (b) Dihydrofolate reductase with folate (rmsd = 2.63 Å, PDB ID: 1dhf). (c) Thermolysin with N-phosphory-L-leucinamide (rmsd = 3.22 Å, PDB ID: 2tmn). (d) L-Arabinose binding protein with L-arabinose (rmsd = 0.29 Å, PDB ID: 6abp).

Table 1
Correlations between predicted binding scores and binding affinities by MIRS and 11 scoring.

Scoring function	*r*	MSE
MIRS	0.700	2.48
X-Score	0.647	5.33
DrugScore	0.600	5.09
PLP	0.586	5.14
G-Score	0.562	5.53
D-Score	0.502	5.07
AutoDock	0.469	4.89
LigScore	0.444	5.09
LUDI	0.413	5.08
ChemScore	0.410	5.04
PMF	0.357	5.18
F-Score	0.201	5.12

Table 2
Success rate and average RMSD of MIRS and 11 scoring functions.

	Success rate		Average RMSD
	rmsd ≤ 1 Å	rmsd ≤ 2 Å	
MIRS	60	70	2.73
PLP	63	76	2.40
F-Score	56	74	2.38
LigScore	64	74	2.68
DrugScore	63	72	3.12
LUDI	43	67	2.98
X-Score	37	66	2.94
AutoDock	34	62	4.58
PMF	40	52	4.43
G-Score	24	42	7.13
ChemScore	12	35	6.07
D-Score	8	26	4.83

pose prediction, and this point may be the most prominent advantage of MIRS. Hence, our results show that it is also indispensable to incorporate unbound ligand conformations on binding pose prediction. Fig. 5 shows the representative examples of binding poses predicted by MIRS close to X-ray structure of ligands for four proteins.

3.3. Convergence analysis of MIRS

To confirm whether or not the iteration number of MIRS is enough, we evaluated the correlation coefficient, MSE and average RMSD at each iteration number (Fig. 6). The correlation coefficient and MSE are the performance metrics for binding affinity prediction and average RMSD is the performance metric for binding pose

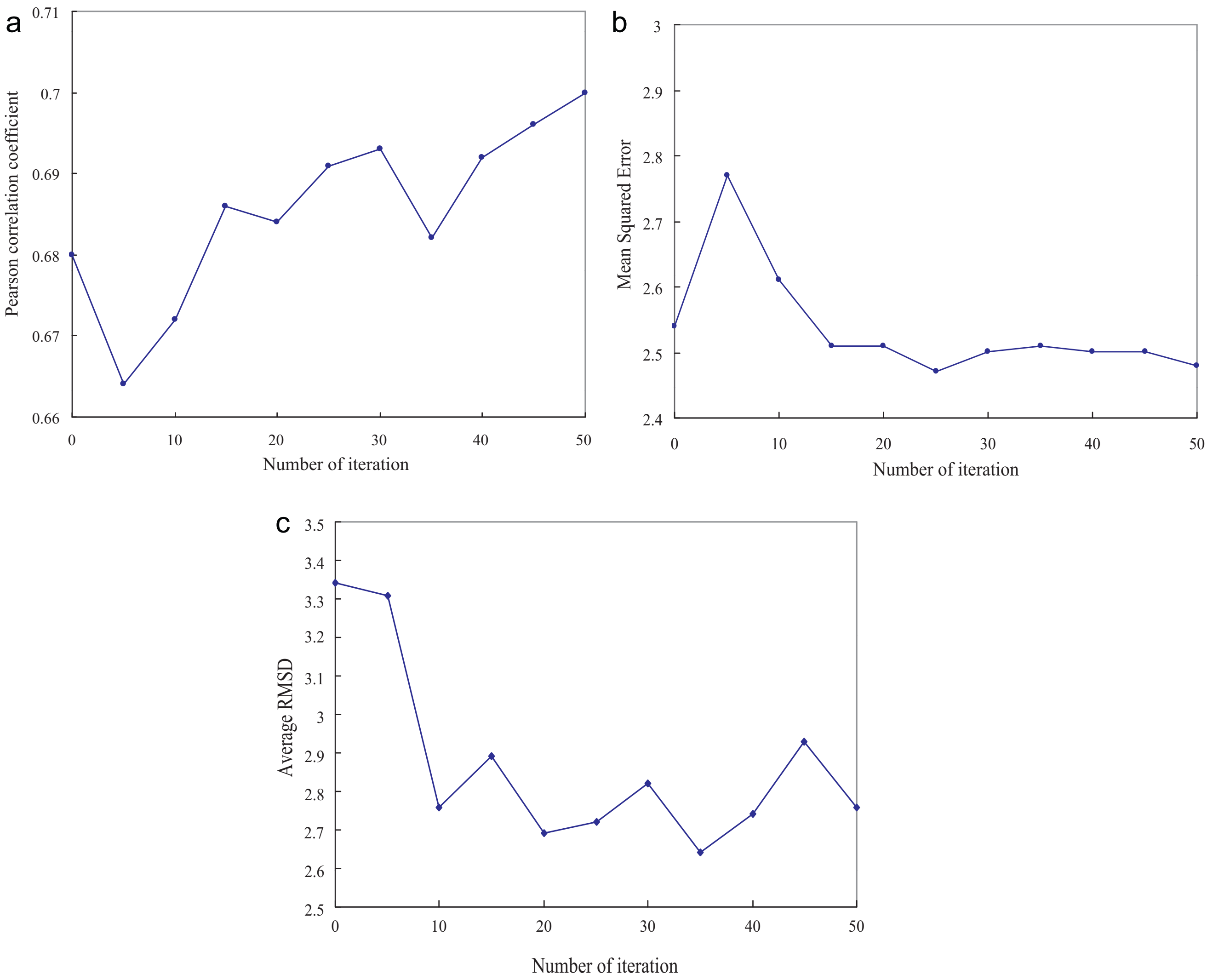


Fig. 6. Convergence analysis. (a) Pearson correlation coefficient, (b) mean squared error, (c) average RMSD. After the number of iteration is over 10, MIRS is stable and converges enough.

prediction. Fig. 5 shows that MIRS converges enough on all the correlation coefficient, MSE and average RMSD when the number of iteration is over 10. Our results led us to the conclusion that the performance on both binding affinity and pose prediction surely achieves in the learning process of iteration. In addition, one can see that over-fitting problem does not occur on MIRS from our results.

4. Conclusion

We have proposed MIRS that incorporates unbound ligand conformations using multiple scoring functions to directly predict protein–ligand binding affinities. Experimental results showed that MIRS outperformed the conventional 11 scoring functions. In addition, our results on binding affinity prediction and binding pose prediction revealed that it is indispensable to incorporate unbound ligand conformations on both binding affinity prediction and binding pose prediction. To our best knowledge, our present study is the first computational study to point out the importance of incorporating unbound ligand conformations in the framework of scoring functions for docking on both binding affinity prediction and binding pose prediction. Although the ensemble of docking poses only covers a part of conformational space of a ligand, our results definitely shows that they are useful to improve the performance of scoring functions. Since the existing scoring functions predict binding affinity based on the top-scored docking pose, they definitely require the precise binding pose. On the other hand, MIRS tries to incorporate the protein–ligand binding process associated with binding free energy landscape by employing the ensemble of docking poses. Therefore, our method does not necessarily require the precise binding pose. We expect that our proposed method would become one of the useful tools helping efficient lead optimization on structure-based drug design and provide the new direction of designing the following scoring functions.

Acknowledgement

HK was supported by JSPS KAKENHI 80545583.

References

[1] I. Massova, P. Kollman, Combined molecular mechanical and continuum solvent approach (MM/PB-SA/GBSA) to predict ligand binding, Perspect. Drug Discovery Des. 18 (2000) 113–135.

[2] H.A. Carlson, W.L. Jorgensen, An extended linear response method for determining free energies of hydration, J. Phys. Chem. 99 (1995) 10667–10673.

[3] J. Aqvist, C. Medina, J.E. Samuelson, A new method for predicting binding affinity in computer-aided drug design, Protein Eng. 7 (1994) 385–391.

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

[5] W.L. Jorgensen, Free energy calculations: a breakthrough for modeling organic chemistry in solution, Adv. Drug Delivery Rev. 22 (1989) 184–189.

[6] D.B. Kitchen, H. Decornez, J.R. Furr, J. Bajorath, Docking and scoring in virtual screening for drug discovery; methods and applications, Nat. Rev. Drug Discov. 3 (2004) 935–949.

[7] A.R. Leach, B.K. Shoichet, C.E. Peishoff, Docking and scoring, J. Med. Chem. 49 (2006) 5851–5855.

[8] G.L. Warren, C.W. Andrews, A.M. Capelli, B. Clarke, J. LaLonde, M.H. Lambert, M. Lindvall, N. Nevins, S.F. Semus, S. Senger, G. Tedesco, I.D. Wall, J.M. Woolven, C.E. Peishoff, M.S. Head, A critical assessment of docking programs and scoring functions, J. Med. Chem. 49 (2006) 5912–5931.

[9] C.Y. Yang, H. Sun, Z. Nikolovska-Coleska, S. Wang, Importance of ligand reorganization free energy in protein–ligand binding affinity prediction, J. Am. Chem. Soc. 131 (2009) 13709–13721.

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

[11] H.J. Bohm, The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure, J. Comput. Aided Mol. Des. 8 (1994) 243–256.

[12] C.J. Camacho, S. Vajda, Protein docking along smooth association pathways, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 10636–10641.

[13] R. Wang, Y. Lu, S. Wang, Comparative evaluation of 11 scoring functions for molecular docking, J. Med. Chem. 46 (2003) 2287–2303.

[14] D.K. Gehlhaar, G.M. Verkhivker, et al., Molecular recognition of the inhibitor Ag-1343 by Hiv-1 protease – conformationally flexible docking by evolutionary programming, Chem. Biol. 2 (1995) 317–324.

[15] G.M. Morris, D.S. Goodsell, et al., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, J. Comp. Chem. 19 (1998) 1639–1662.

[16] G. Jones, P. Willett, et al., Development and validation of a genetic algorithm for flexible docking, J. Mol. Biol. 267 (1996) 727–748.

[17] T.J. Ewing, S. Makino, et al., DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases, J. Comput. Aided Mol. Des. 15 (2001) 411–428.

[18] M. Rarey, B. Kramer, et al., A fast flexible docking method using an incremental construction algorithm, J. Mol. Biol. 2 (1996) 470–489.

[19] M.D. Eldridge, C.W. Murray, et al., Empirical scoring functions 0.1. The development of a fast empirical scoring function to estimate the binding affinity of ligand in receptor complexes, J. Comput. Aided Mod. Des. 11 (1997) 425–445.

[20] R.X. Wang, L.H. Lai, et al., Further development and validation of empirical scoring functions for structure-based binding affinity prediction, J. Comput. Aided Mol. Des. 16 (2002) 11–26.

[21] I. Muegge, Y.C. Martin, A general and fast scoring function for protein–ligand interactions: a simplified potential approach, J. Med. Chem. 42 (1999) 791–804.

[22] H. Gohlke, M. Hendlich, et al., Knowledge-based scoring function to predict protein–lignad interactions, J. Mol. Biol. 295 (2000) 337–356.

[23] T. Dietterich, R. Lathrop, T. Lozano-Perez, Solving the multiple-instance problem with axis-parallel rectangles, Artif. Intell. 89 (1997) 31–71.

[24] S. Ray, D. Page, Multiple instance regression, in: Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 425–432.

[25] Z. Wang, V. Radosavljevie, N. Han, Z. Obradovic, S. Vucetic, Aerosol optical depth prediction from satellite observations by multiple instance regression, in: Proceedings of the SIAM International Conference on Data Mining, 2008, pp. 165–176.

[26] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (2000) 1189–1232.

[27] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[28] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, J. Chem. Inf. Comput. Sci. 43 (2003) 1947–1958.

[29] V. Svetnik, T. Wang, C. Tong, A. Liaw, R.F. Sheridan, Q. Song, Boosting: an ensemble learning tool for compound classification and QSAR modeling, J. Chem. Inf. Inft. Model. 45 (2005) 786–799.

[30] T. Hasite, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning, Springer-Verlag, New York, 2001.