Journal of Computational Science & Engineering 19 (2015) 634-638

Available at http://www.asocse.org ISSN 1710-4068

A novel scoring model for improving the identification of near-natives from protein-protein docking conformations

Jingjun Bi [1], Yan Ye [2], Yuan Xu [2], Guoqing Zhang [1], Meng Wang [1], Shiying Yan [1]*

[1]College of Physics, Qingdao University, Qingdao 266071, P. R. China

[2]Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences

Abstract: **The 3D structure of the protein-protein complexes play an important role in structural biology and drug discovery. The most accurate structural characterization of protein complexes is provided by X-ray crystallography and NMR spectroscopy. Due to the expensive and time-consuming procedures of experimental structure determination, the computational method using molecular docking has become an important way to address the problem. Many docking algorithms have been developed for this purpose, but most of them are difficult to identify the near-native binding conformations. In this study, we reports a machine-learning scoring function based on the frequency of the pair wise residue-residue types between the receptor and ligand within a certain distance. This model can successfully to identify the near-native binding conformations generated from protein-protein docking by PyRosetta.**

Keywords: Protein-Protein Interaction; PyRosetta; Distance; SVM

Received 29 May 2015, Revised 23 July 2015, Accepted 25 July 2015

* Corresponding Author: Shiying Yan, ysy5954418@163.com

1. Introduction

The protein-protein structures are fundamental to the molecular biology and drug discovery, it is important for the biological process [1]. Knowing the structures of the complexes is an essential task for the advancement of the proteomics in computational biology [2]. Compared with the progress achieved in individual proteins, development in experimentally determining the protein-protein structures is still limited because of the technical difficulties and high cost of the experimental methods [3], especially for weakly interacting proteins, which are hard to capture in the complexes form [4]. The rates of the protein-protein interactions are being discovered run ahead of the rate of the structures are being solved by the experiment [5]. As such, the method by docking algorithms provides a fast and convenient way to predict the protein-protein structure.

In recent years, various protein-protein docking algorithms have been developed to predict the protein-protein structure [6-9]. For docking, there are two essential steps: search and scoring. For searching the possible docking conformations, it requires a thorough search for the binding orientations over six degrees of freedom (3D translational plus 3D rotational) [9, 10]. Usually the search algorithms based on matching local shape features, such as the distance geometry algorithm, Geometric Hashing or Genetic algorithm [11, 12] and randomized search, such as genetic algorithm or Monte Carlo methods [13, 14] playing a major role. The purpose of the search stage is to find enough poses and the true native pose included among them [1].

The next step in protein-protein docking is scoring which aimed to pick up the true native pose from the docking conformations. Current scoring functions can be classified into three broad categories: force field scoring function, empirical scoring function and knowledge-based scoring function. The force field scoring functions are developed on physical atomic interactions, such as electrostatic function, Van der Waals energy and bond stretching/bending/torsional force [15], this kind of scoring function was used in the soft such as Softdock [16], Northwestyern Dock [12]. But it is more time-consuming than others. Empirical scoring functions usually combine the energy terms such as conformational entropy, hydrophobic and hydrophilic surface areas, number of hydrogen bonds, etc., and are obtained by fitting the experimental binding data [8]. Because of its simplicity and fast speed, empirical scoring functions are commonly used in the initial filtering step of docking software such as PyRosetta [7] and ZDOCK [17]. PyRosetta uses mathematical models of the major biophysical forces to score the conformations. ZDOCK consider the desolvation, grid-based shape complementarity and electrostatics. PyRosetta and ZDOCK are popular software in protein-protein docking. But without any binding site information, PyRosetta will search for a long time to find the near-native and easy to fall into local minimum point. For ZDOCK, if consider the flexibility of the protein during the docking, it will be difficult to get the near-natives. Knowledge-based functions for protein–protein docking are derived from known structures of protein–protein complexes in a database [18-20]. It has been shown that knowledge-based pairwise functions perform better than the physics-based functions in the near-native conformations refinement [21].

Though the soft such as PyRosetta can score the conformations in a short time, but sorted the conformations is not ideal. For this purse, we want to find a way to sort the conformations more accurate and convenient. Because of the rapid increase in protein structure data, the knowledge-based scoring function

2015 © The American Computational Science Society. All rights reserved.

getting more and more popular. In addition, the method based on the machine-learning have obtained good results in recent years [2, 22, 23]. Therefore, in this study, we combine these two methods, using the machine-learning method based on a large number of the PPI structures information to get a prediction model, aim to improve the ability of PPI docking in sorting the putative binding conformations.

2. Material and method

2.1. Data set

The protein-protein complexes for the training set were taken from ProPairs dataset [24]. The Benchmark 4.0 dataset were chosen for the test set [25]. There are some complexes included in both datasets, and the overlapping structures were removed from the training set. Finally, 668 bound-bound complexes (the structure separate from the original complex) were chosen for the training set, and 174 complexes were chosen for the test set.

For each of the 668 complexes in the training set, PyRosetta was used to generate conformations. In those conformations, we definite the Lrmsd <=5 (Lrmsd is the ligand RMSD of the backbone C atoms) as near-native and Lrmsd >5 as decoys [26]. For each complex, 5 near-natives and 5 decoys were selected as the training set.

For each of the 174 complexes in the test set, PyRosetta was used to generate 100 conformations. All the complexes have at least one near-native.

2.2. PPI Docking by PyRosetta

For protein-protein docking, PyRosetta firstly employs a low-resolution rigid-body docking. The second high-resolution stage provides for side chain conformational sampling and backbone relaxation. First, the complex was converted into centroid-mode (all backbone atoms of the complex are explicitly modeled, but the side-chain atoms are replaced with a single pseudo-atom, this will make much faster energy calculations). In the low-resolution, one partner was kept stay and made a random reorientation against the other partner. The step sizes are adjusted continually to maintain a 50% movement acceptance rate, with initial Gaussian perturbation sizes of mean value 0.7Å (along all three axes) and 5 °(spin around the axis of protein centers and tilt off this axis in a randomly-chosen direction) [13, 27]. Following the low-resolution phase of docking, the structures were converted into a full-atom structure for the high-resolution refinement step. The high-resolution refinement simulates the binding step, as side-chain conformations are rearranged to find a local energy minimum. Finally, backbone flexibility is applied around the complex interface (6Å) [13].

2.3. Features

The frequency of the residue-residue pairwise between receptor and ligand within a certain distance was taken as the features for characterizing the PPI binding interface. The distance cutoff plays an important role in the feature, as earlier studies had suggested that distance cutoff had effect on protein-protein energy landscapes [28]. For a long-range function, such as soft Lennard-Jones potentials, 14Å cutoff was suggested, which is close to the cutoff 15.5 Å in DFIRE [29]. In this study, the cutoff was set as 15 Å with a 3 Å intervals in every step. To measure the residue-residue distance between receptor and ligand, both the coordinate of the Cα [30] and the centre of the residue were considered to represent the place of the residue. Altogether 20*20 residue pairs were defined between receptor and ligand, together with five distance ranges and two ways for representing the residues, these yields 4000 descriptors for each conformation.

2.4. Classification

Support vector machines (SVM) was used here for establish the classification model, which is one of the widely used method of machine-learning in recent years, by seeking the largest decision boundary on the edge. It has shown to provide high quality predictions in classification [2]. Here, the SVM model aims to divide a high-dimensional space into regions containing only near-natives or only decoys, each of them represented by the points with the values of the different parameters [22]. The kernels defined these regions. The support vector classification C-SVM and RBF kernel is selected here by using Libsvm [31], and five-fold cross-validation was used during the training.

First of all, the training data set was converted into the feature space, and then the corresponding features are passed to the classifier. The conformations were labeled by their Lrmsd. The Lrmsd of the conformations <= 5 were labeled as true (positive class) and the Lrmsd of the conformations>5 were labeled as false (negative class). The protein descriptive features were used as inputs. In libsvm, each type of sample penalized by the public $C_i = W_i * C$, where W_i is the weight, C is penalty factor. For RBF function, the public is $[\exp(-\gamma \| u - v \|^2)]$, where the parameterγis the width of RBF. The choice of C and γ is very important since SVM is sensitive to parameterization [2]. After the grid search, the best parameters of the C, γ is 64, 0.0078125. With the best parameters, we trained the model to predict protein-protein binding conformations.

3. Results and Discussion

3.1. Performance of the learning procedures

The predictive performance was assessed with AUC (area under the ROC curve). The ROC (Receiver Operating Characteristics) curve is the proportion of true positive rate (true positive/[true positive + false negative]) against the proportion of false positive rate (false positive/ [true negative + false positive]). In this

2015 © The American Computational Science Society. All rights reserved.

paper, prediction of the function was evaluated and measured using the AUC. The perfect selection should make the AUC equal to 1; a random selection yielding true positives and false positives in equivalent numbers should have an AUC of 0.5. Higher values of the AUC measures indicate better quality of predictions [22].

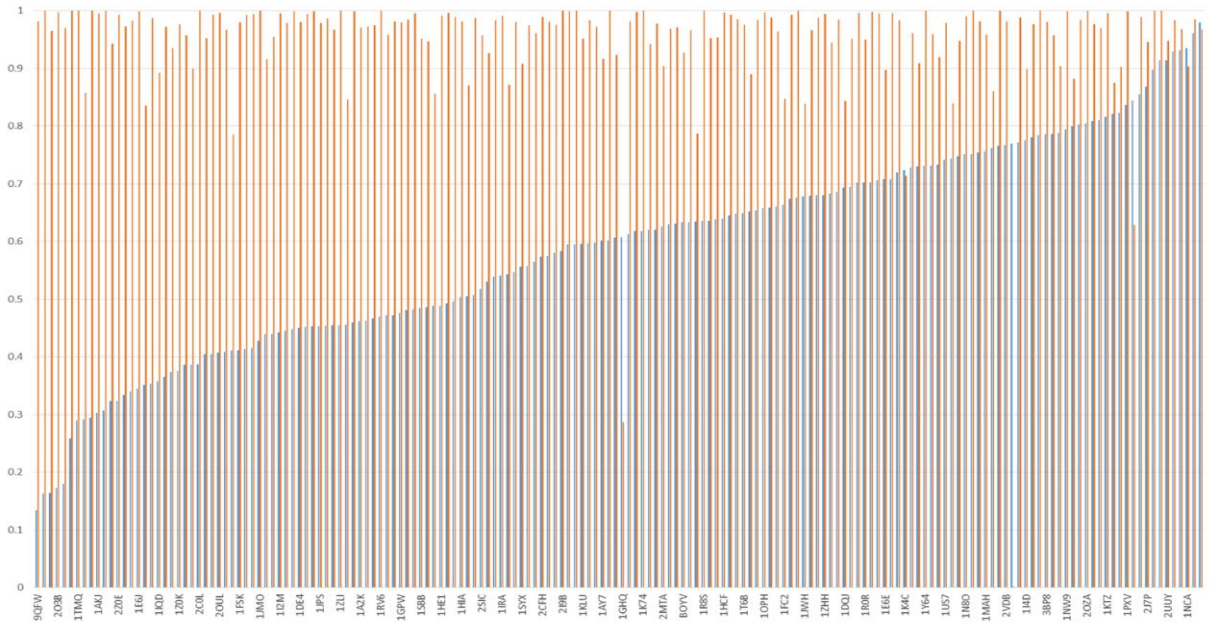

Figure 1. The AUC comparison our model with PyRosetta scoring function using Benchmark 4.0 decoys. The AUC of each complex scored by our model in orange, scored by PyRosetta in blue.

3.2. Test set

For testing our functions, we compare our model with PyRosetta scoring function. PyRosetta using the empirical scoring functions. There are several energy terms are available including Van der Waals attractive, a solvation approximation, an orientation-dependent hydrogen bonding potential, a knowledge-based electrostatics term, and a knowledge-based conformation-dependent amino acid internal free energy term [27].

3.2.1. Re-ranking Benchmark 4.0 dataset

The Benchmark 4.0 dataset were chosen for the test set [25]. Two of the Benchmark 4.0 complexes were removed due to their large size (1N2C) and difficult to generate the near-natives (2IDO). Finally, 174 bound-bound complexes were chosen for the test set. As the same method as the training sets, for each of the complex, we use PyRosetta to product 100 conformations. All the complexes have at least one near-native.

First, all the conformations were scored using PyRosetta. For the conformations of each complex, their Lrmsd and PyRosetta score were recorded. Using the scikit-learn [32], the AUC of each complex was calculated by using the PyRosetta score. In the same way, the AUC of our model was calculated for each of the complexes.

The AUC of 174 complexes scored by PyRosetta and our model are shown in Figure 1. The Figure 2 is the distribution of AUC values. From Figure 2, we can clearly find that the AUC scored by our model mainly are distributed in 0.9-1.0, which is better than PyRosetta. The average of AUC scored by PyRosetta is 0.58778, while our model is 0.9493. The AUC of 19 complexes which scored by our model reached 1.0. One of the reasons why our model performance is so outstanding in this test set may be that the decoys were generated using bound-bound complexes, and the conformations for training the model also use the bound-bound complexes. Both of the sets were generated by the PyRosetta.

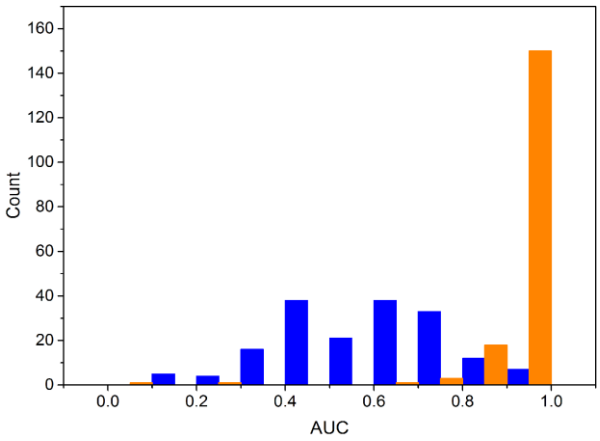

Figure 2. The distribution of the AUC value using Benchmark 4.0 decoys. The distribution of the AUC value scored by our model in orange, scored by PyRosetta in blue.

2015 © The American Computational Science Society. All rights reserved.

3.2.2 Re-ranking Dockground decoys

For prove our model can be widely application, we tested the scoring function on Dockground docking decoys [33]. The Dockground docking decoys using GRAMM-X docking scan to generate 100 non-native and at least one near-native match per complex for 61 unbound-unbound complexes (the structure in their native conformations)and only one chain in the original PDB file (no crystal packing)).

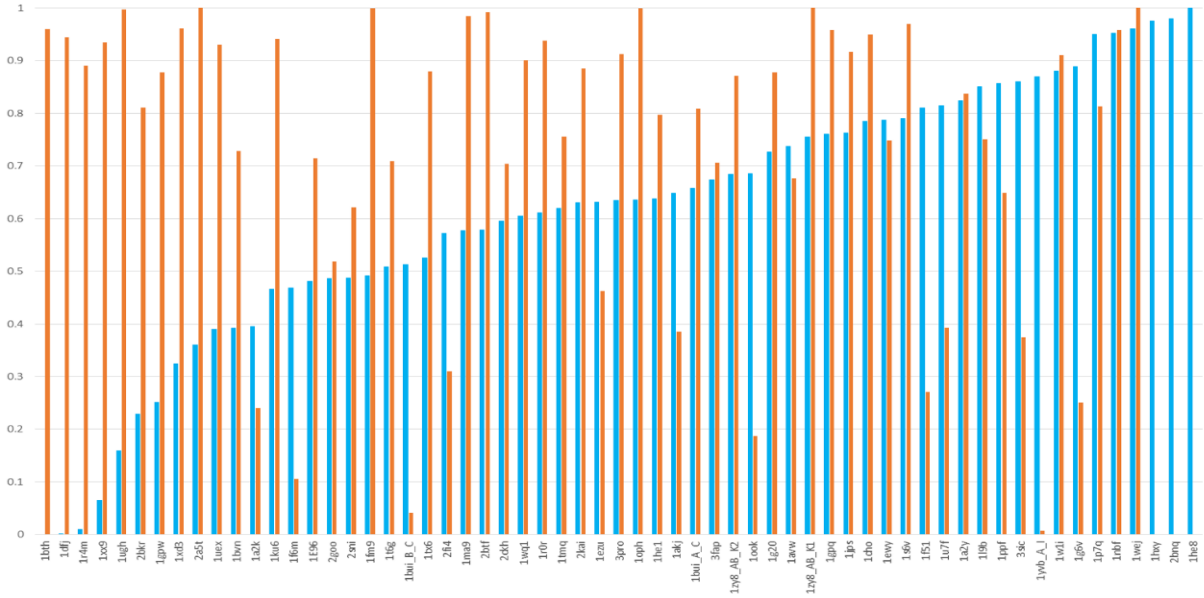

Figure 3. The AUC comparison our model with PyRosetta scoring function using Dockground decoys. The AUC of each complex scored by our model in orange, scored by PyRosetta in blue.

The result are shown in Figure 3. The Figure 4 is the distribution of the AUC value. From Figure 4, it is clear that the number of the AUC value scored by our model distribute in 0.9-1, which is higher than PyRosetta. The average of AUC scored by PyRosetta is 0.61078, while our model is 0.70024. The AUC of 3 complexes (2A5T, 1ZY8, 1WEJ) reached 1.0, but PyRosetta have only one (1HE8). For this test set, our model performed better than PyRosetta, but the result is not as impressive as the result of Benchmark 4.0 decoys. The potential reason for this decrease is that the test set conformations were generated by GRAMM-X. It seems that our model performance is dependent on the docking programs and the quality of generated binding conformations. Another possible reason for this issue is that the decoys were generated using unbound-unbound complexes. Our model is more suitable to predict the conformations generated from the bound-bound complexes.

Compared with the result of the two test set, the Dockground decoys generated by GRAMM-X, though the result of our model is better than PyRosetta, but it is worse than the result of the decoys generated by PyRosetta. For solving this problem, in our later work, we will pick up the conformations which generated by different software during picking up the conformations for the training. We will expand the number of the complexes in our training set and pick out the important descriptor during training the model to improve the accuracy of our model prediction. We will use the conformations which generated by unbound-unbound complexes to train the model, made our model be the same applies to the conformations which generated by the unbound-unbound complexes.

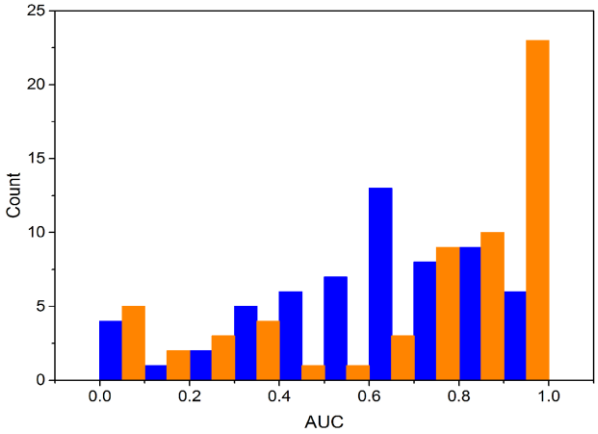

Figure 4. The distribution of the AUC value using Dockground decoys. The distribution of the AUC value scored by our model in orange, scored by PyRosetta in blue.

4. Conclusion

We have presented a novel model for protein-protein docking score, using the machine-learning based on a large number of the PPI structures information to establish the classification model. For training the model, 668 bound-bound protein-protein complexes

2015 © The American Computational Science Society. All rights reserved.

were used from ProPairs dataset in this study. PyRosetta was used to generate near-native and decoy binding conformation, from which 5 near-natives and 5 decoys were selected for each complex for establishing the training set. Then, SVM was used to train the classification model. The final model were evaluated with two testing sets: In Benchmark 4.0 and Dockground, and our model performed better than PyRosetta on both test sets. Our model also obtained good result for the conformations generated by PyRosetta, especially for the conformations extracted from the bound-bound complexes.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Reference

[1] S. Moreira, J.M. Martins, J.T. Coimbra, Phys. Chem. Chem. Phys., 17 (2015) 2378-2387.
[2] Y. Sheng, X. Qiu, C. Zhang, Comput. Math. Methods Med., 2014 (2014) 715494-715494.
[3] S.Y. Huang, Drug Discov. Today, 19 (2014) 1081-1096.
[4] H. Hwang, T. Vreven, Z. Weng, Proteins, 82 (2014) 57-66.
[5] I.H. Moal, M. Torchala, P.A. Bates, BMC Bioinformatics, 14, (2013) 1-15.
[6] B.G. Pierce, K. Wiehe, H. Hwang, Bioinformatics, 30 (2014)1771-1773.
[7] S. Chaudhury, S. Lyskov, J.J. Gray, Bioinformatics, 26 (2010) 689-691.
[8] B. Pierce, Z. Weng, Proteins, 67 (2007) 1078-1086.
[9] P.N.P.L. Krippahl, J.E. Wampler, J.J.G. Moura, Proteins, 39 (2000) 372-384.
[10] E. Katchalski-Katzirt, I. Shariv, M. Eisenstein, Proc Natl Acad Sci USA, 89 (1991) 2195-2199.
[11] A.J. Bordner, A.A. Gorin, Proteins, 68 (2007) 488-502.
[12] D.M. Lorber, M.K. Udo, B.K. Shoichet, Protein Science, 11 (2002) 1393-1408.
[13] J.J. Gray, S. Moughon, C. Wang, Journal of Molecular Biology, 331 (2003) 281-299.
[14] E.J. Gardiner, P. Willett, P.J. Artymiuk, Proteins, 44 (2001) 44-56.
[15] S.J. Weiner, P.A. Kollman, D.A. Case, J . Am. Chem. SO., 106 (1984) 765-784.
[16] C.H. Li, X.H. Ma, W.Z. Chen, Protein Engineering Design and Selection, 16 (2003) 265-269.
[17] R. Chen, L. li, Z. Weng, Protein, 52 (2008) 80-87.
[18] R. Sanydrala, J. Moult, J. Mol. Biol, 275 (1998) 896-916.
[19] S.Y. Huang, X. Zou, Proteins, 72 (2008) 557-579.
[20] F. Zhao, and J. Xu, Structure, 20 (2012) 1118-1126.
[21] C.M. Summa, M. Levitt, Proc. Natl. Acad. Sci., 104 (2007) 3177-3182.
[22] J. Bernauer, J. Aze, J. Janin, Bioinformatics, 23 (2007) 555-562.
[23] J.R. Bradford, D.R. Westhead, Bioinformatics, 21 (2005) 1487-1494.
[24] F. Krull, G. Korff, N. Elghobashi-Meinhardt, J. Chem. Inf. Model, 55 (2015) 1495-1507.
[25] M.F. Lensink, R. Mendez, S.J. Wodak, Proteins, 69 (2007) 704-718.
[26] Z. Zhang, C.E. Schindler, O.F. Lange, PLoS One, 10 (2015) 1-17.
[27] K.W. Kaufmann, G.H. Lemmon, S.L. Deluca, Biochemistry, 49 (2010) 2987-2998.
[28] S. Liu, I.A. Vakser, BMC Bioinformatics, 12 (2011) 312-314.
[29] H. Zhou, and Y. Zhou, Protein Sci., 11 (2002) 2714-2726.
[30] Y. Wu, M. Lu, M. Chen, Protein Sci., 16 (2007)1449-1463.
[31] C.W. Hsu, C.J. Lin, IEEE Trans Neural Netw, 13 (2002) 415-425.
[32] F. Pedregosa, G. Varoquaux, A. Gramfort, Journal of Machine Learning Research, 12 (2011) 2825-2830.
[33] S. Liu, Y. Gao, I.A. Vakser, Bioinformatics, 24 (2008) 2634-2635.

2015 © The American Computational Science Society. All rights reserved.