W438–W442 *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* *Published online 28 May 2018*
doi: 10.1093/nar/gky439

COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking

Qi Wu[1], Zhenling Peng[2,*], Yang Zhang[3] and Jianyi Yang[1,*]

[1]School of Mathematical Sciences, Nankai University, Tianjin 300071, China, [2]Center for Applied Mathematics, Tianjin University, Tianjin 300072, China and [3]Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA

Received January 25, 2018; Revised April 27, 2018; Editorial Decision May 04, 2018; Accepted May 08, 2018

ABSTRACT

The identification of protein–ligand binding sites is critical to protein function annotation and drug discovery. The consensus algorithm COACH developed by us represents one of the most efficient approaches to protein–ligand binding sites prediction. One of the most commonly seen issues with the COACH prediction are the low quality of the predicted ligand-binding poses, which usually have severe steric clashes to the protein structure. Here, we present COACH-D, an enhanced version of COACH by utilizing molecular docking to refine the ligand-binding poses. The input to the COACH-D server is the amino acid sequence or the three-dimensional structure of a query protein. In addition, the users can also submit their own ligand of interest. For each job submission, the COACH algorithm is first used to predict the protein–ligand binding sites. The ligands from the users or the templates are then docked into the predicted binding pockets to build their complex structures. Blind tests show that the algorithm significantly outperforms other ligand-binding sites prediction methods. Benchmark tests show that the steric clashes between the ligand and the protein structures in the COACH models are reduced by 85% after molecular docking in COACH-D. The COACH-D server is freely available to all users at http://yanglab.nankai.edu.cn/COACH-D/.

INTRODUCTION

Protein–ligand interaction represents one of the most important forms of protein function, because proteins perform their function through interactions with other molecules. An ideal way to study the interaction is to solve the complex structures by X-ray or nuclear magnetic resonance techniques. However, these experiments are usually laborious and costly to carry out. To make it even worse, it is very difficult or impossible to solve the structures for some large proteins and membrane proteins with classical techniques (1). Rather than solving the structure, an alternative way is to locate the binding sites by techniques such as crosslinking and mass spectrometry (2,3).

As it is often time consuming and expensive to solve the protein–ligand complex structure or determine the interactions by experiments, many computational efforts have been made to facilitate the study of the interactions. These efforts include at least the following aspects. The first is the recognition of protein–ligand binding sites, which aims to predict the binding pockets and the residues involved in the interactions with the ligand (4–9). The second is the modeling of the protein–ligand complex structure based on molecular docking (10–13). The third is the development of scoring functions for the estimation of the binding affinity (14,15). The studies in the above three aspects are in fact closely related. The prior knowledge of the binding sites can help the molecular docking to build the complex structures efficiently, while accurate scoring functions can be used to select the binding poses with high binding affinity.

Here, we present the COACH-D server, an enhanced version of the COACH server for the prediction of protein–ligand binding sites and ligand-binding poses. The protein–ligand binding sites are first predicted based on the COACH algorithm, which is a consensus of five individual methods (7). The ligands from the users or the templates are then docked into the predicted binding pockets using the molecular docking algorithm AutoDock Vina (10). Blind tests in the CAMEO-LB experiments (16) and benchmark tests demonstrate the significant advantage of COACH-D over other state-of-the-art methods.

[*]To whom correspondence should be addressed. Tel: +86 22 23501449; Fax: +86 22 23506423; Email: yangjy@nankai.edu.cn
Correspondence may also be addressed to Zhenling Peng. Tel: +86 22 27406039; Fax: +86 22 27406039; Email: zhenling@tju.edu.cn

MATERIALS AND METHODS

Overview of the COACH-D algorithm

The overall architecture of the COACH-D algorithm is shown in Figure 1. For submission with amino acid sequence, the I-TASSER Suite (17) is used to model the protein structure first. The structure is then submitted to five individual methods to predict the protein–ligand binding sites. Four of them are template-based methods: COFACTOR (6), FINDSITE (4), TM-SITE (7) and S-SITE (7). These methods predict the binding sites by matching the query structure and sequence with the ligand-binding templates from BioLiP (18), which is a semi-manually curated functional database for biologically relevant ligand–protein interactions constructed based on the Protein Data Bank (PDB) (19). The last one is a template-free and structure-based method ConCavity, which considers both sequence conservation and structure geometry for the binding sites prediction (5). The results from individual methods are then combined to consensus predictions by the COACH algorithm. The detailed descriptions of these methods are available in the original COACH algorithm (7) and the publications of the corresponding methods (4–6). The ligand from the user input or the templates is then docked into the predicted binding pockets to build their complex structures by the efficient molecular docking algorithm AutoDock Vina (10). For each predicted binding pocket, up to 10 binding poses are generated and the one that matches the best with the consensus prediction of binding residues is selected (please refer to the 'PERFORMANCE OF THE SERVE' section for explanations about such selection).

INPUT AND OUTPUT OF THE SERVER

Input

The input to the COACH-D server can be either the amino acid sequence or the three-dimensional (3D) structure of a query protein. In addition, the users can submit their own ligand of interest as well. When the amino acid sequence of the protein is submitted, the I-TASSER Suite (4) will be used to generate one 3D structure model first. The structure is then used for the ligand-binding sites prediction and the subsequent molecular docking. We would like to point out that except ConCavity, other component algorithms in COACH-D were designed for monomer structures. Thus, only the first chain will be extracted when oligomers are submitted. We plan to extend the algorithm so that it can work for oligomers in future. An option is provided to protect the users' personal data by checking on the checkbox of 'Keep my results private'. A password is then assigned to the users to access the modeling results. In general, it takes 2–5 h to complete the modeling for a structure submission with ~300 residues.

Output

One unique job ID and a URL are assigned to each submission. The users can track the modeling status at the URL provided. Once completed, the results will be displayed on the web page of the URL and a notification email will be sent to the user for accessing the results. The typical output results for each submission include:

i) One predicted 3D structure model for the submission with amino acid sequence.
ii) The top five protein–ligand binding pockets and the binding residues in each pocket.
iii) The top five protein–ligand complex structures with the input ligand.
iv) The top five protein–ligand complex structures with the ligands from the PDB template structures.
v) A summary of ligands that are possible to bind the protein.

All these modeling results are put together into a single tarball, which can be downloaded to a local computer for use. All ligand-binding poses from AutoDock Vina are also put into the tarball. A confidence score (c-score) in the range of [0, 1] is provided to judge the reliability of each prediction. Please refer to the COACH article for more information about the scoring function of c-score (7).

Figure 2 illustrates the modeling results for an example submission with a protein structure and a ligand. Explanation about the meaning of each column in the table can be viewed by hovering the mouse pointer over the corresponding question sign. For this example, the first prediction is highly confident, as reflected by the high c-score. A total of 12 residues were predicted to be involved in the ligand binding. The total number of templates used for making this prediction is 329 (i.e. the 'Cluster size' shown in the figure) and the one with the highest similarity to the query structure is from the PDB template 1lwxA. The representative ligand AZD (*3′-Azido-3′-Deoxythymidine-5′-Diphosphate*) was docked into the predicted binding pocket. The complex structures are visualized based on the 3Dmol library (20). The default view is for the complex structure built with the input ligand, which can be switched to other complex structures by clicking on the corresponding 'View' button under the 'Pose^t' and 'Pose^u' columns. All complex structures can be downloaded for further analysis and customized visualization with other molecular graphics systems. The docking energies for the complex structures are listed under the 'Energy^t' and the 'Energy^u' columns.

PERFORMANCE OF THE SERVER

The core algorithm (COACH) of the COACH-D server for ligand-binding sites prediction have been extensively assessed in the blind tests of the CAMEO-LB experiment (16), which was hold weekly for more than 4 years between 6 January 2012 and 16 April 2016. We would like to mention that COACH-D did not participate to this assessment experiment because CAMEO-LB was unfortunately discontinued at the time of COACH-D's development. However, the predictions submitted to the CAMEO-LB are the residue-specific ligand-binding probabilities, which are essentially the same for COACH and COACH-D as both methods adopt the same strategy for consensus prediction of binding sites. The improvement of COACH-D over COACH is the refined ligand-binding poses, which is out of the assessment of CAMEO-LB and will be discussed below.

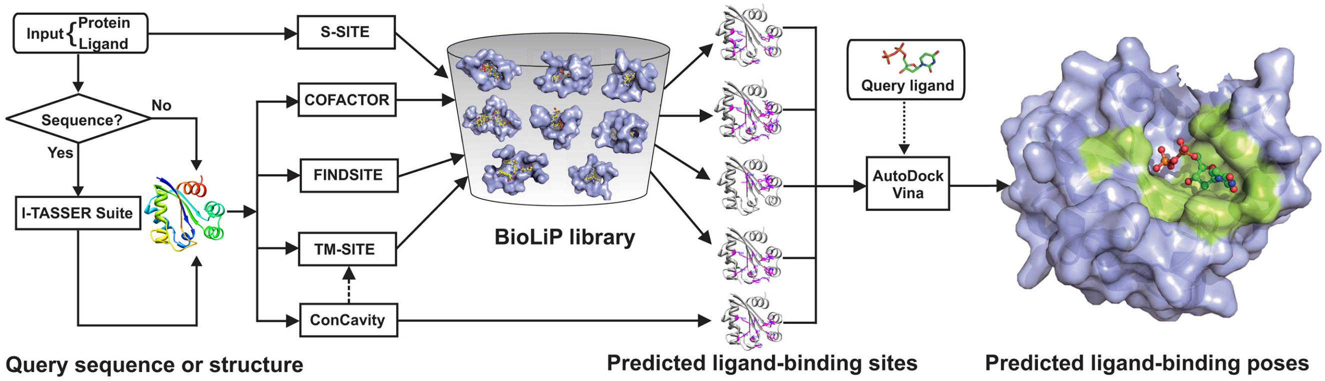
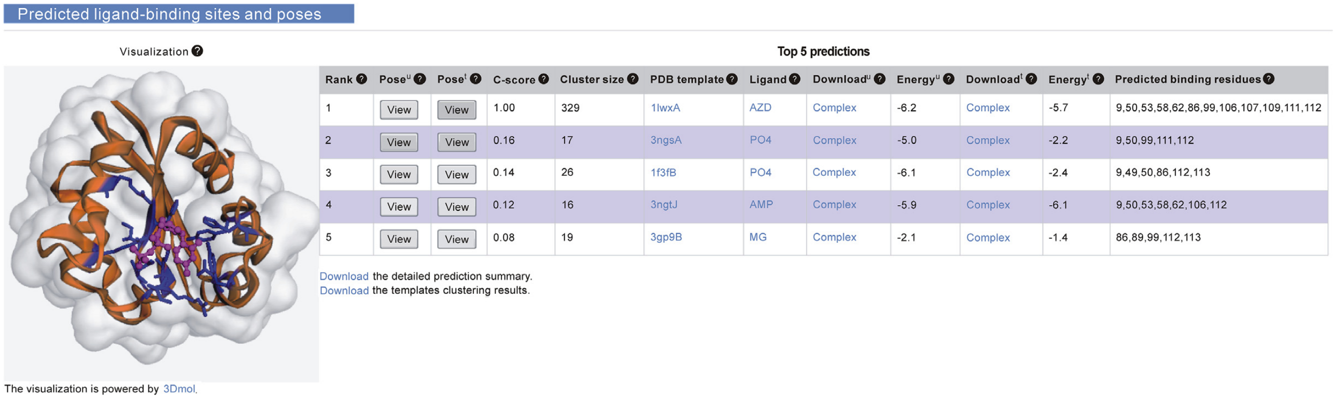


Figure 1. The overall architecture of the COACH-D algorithm.



Predicted ligand-binding sites and poses

Visualization

Top 5 predictions

Rank	Pose^u	Pose^t	C-score	Cluster size	PDB template	Ligand	Download^u	Energy^u	Download^t	Energy^t	Predicted binding residues
1	View	View	1.00	329	1lwxA	AZD	Complex	-6.2	Complex	-5.7	9,50,53,58,62,86,99,106,107,109,111,112
2	View	View	0.16	17	3ngsA	PO4	Complex	-5.0	Complex	-2.2	9,50,99,111,112
3	View	View	0.14	26	1f3fB	PO4	Complex	-6.1	Complex	-2.4	9,49,50,86,112,113
4	View	View	0.12	16	3ngtJ	AMP	Complex	-5.9	Complex	-6.1	9,50,53,58,62,106,112
5	View	View	0.08	19	3gp9B	MG	Complex	-2.1	Complex	-1.4	86,89,99,112,113

Download the detailed prediction summary.
Download the templates clustering results.

The visualization is powered by 3Dmol.

Figure 2. The output page for each submission to the COACH-D server. The visualization of the complex structure is obtained by the 3Dmol library (20). The protein structure is shown in grey surface and orange cartoon. The ligand binding poses are shown in magenta balls and sticks. The consensus binding residues are highlighted in blue sticks.

As listed on the CAMEO-LB website (https://www.cameo3d.org/lb), COACH predicted the ligand-binding sites for a total of 10 414 targets between 7 December 2012 and 16 April 2016. Figure 3 shows that the average accuracy (i.e. AUC) and the BDT score (21) of the COACH prediction are 0.79 and 0.44, respectively, which are 14.5 and 22.2% higher than the second best method RaptorX-Binding (22). Note that by default the CAMEO-LB website only lists the data in the period of 1 year. The full set of data is available by requesting a user account and password from the organizer. The identities and real names for 'Server xx' are unknown to us and the public, which was requested by the corresponding groups.

A dataset consisting of 50 CAMEO-LB targets were used to compare the performance of COACH-D with COACH. Based on our analysis and the feedback from the user community, one of the most commonly seen issues with the COACH prediction are the low quality of the predicted ligand-binding poses, which often have severe steric clashes to the protein structure. The new COACH-D pipeline solves this issue by using the efficient molecular docking algorithm AutoDock Vina (10). To check how much the problem of the steric clashes has been solved, we collected a dataset of 50 targets from CAMEO-LB as follows. Originally, there are a total of 303 targets in CAMEO-LB's datasets of the final month (i.e. between 26 March 2016 and 16 April 2016). Four kinds of ligands are defined in CAMEO-LB: ions, organic ligands, nucleotides and peptides. Here, only targets with organic ligands were kept as AutoDock Vina was not designed for other ligand types. Ligands in BioLiP's artifact list were excluded as well, which consists of ligands commonly used as additives during structure determination and are thus mostly biologically irrelevant. The ligand-binding residues were obtained based on atomic distance calculations with the protein–ligand complex structures, similar to the procedure used in CAMEO-LB. Targets with too few (<5) ligand-binding residues were excluded. Targets were also excluded in case AutoDock Vina failed due to unaccepted ligand atom types. In COACH-D, when AutoDock Vina fails to generate ligand-binding poses, the ones from the template structures without refinement are reported, which are thus identical with COACH's results and not necessary for comparison.

Two metrics are used to compare COACH-D with COACH. The first one is the Matthews correlation coefficient (MCC) between the predicted and the native binding residues. The consensus binding residues are the same for both COACH and COACH-D (i.e. the last column in the results table of Figure 2). To directly reflect their differences in binding poses, we derive the binding residues from the two set of the predicted complex structures, one before and the other after molecular docking, with the same identification

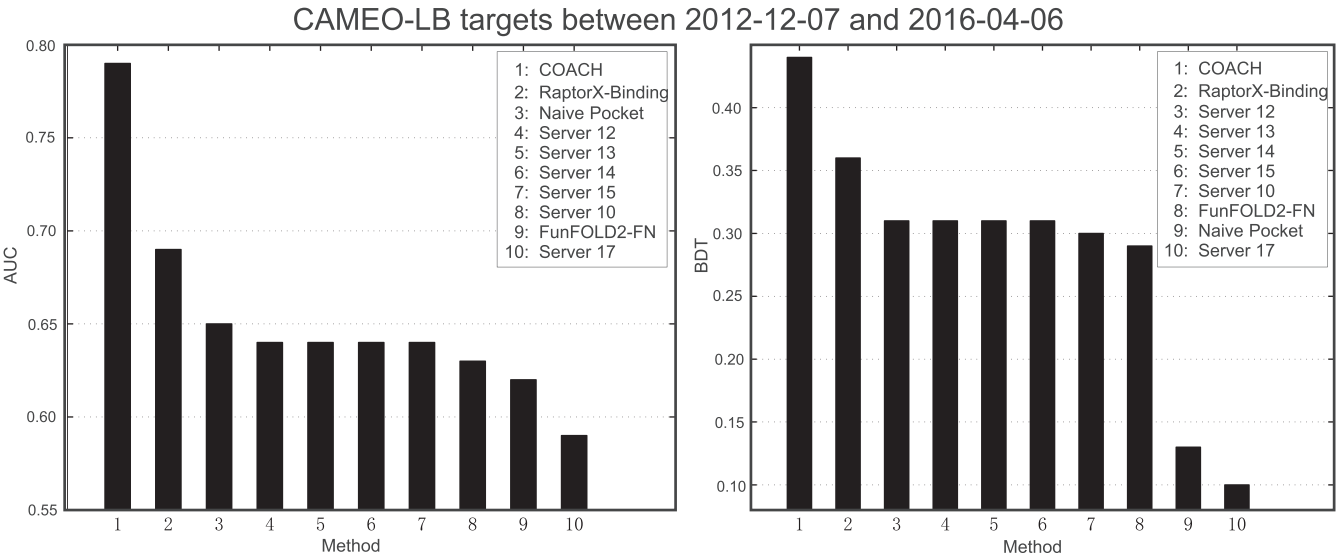


Figure 3. The performance of COACH for ligand-binding sites prediction in the blind tests of CAMEO-LB.

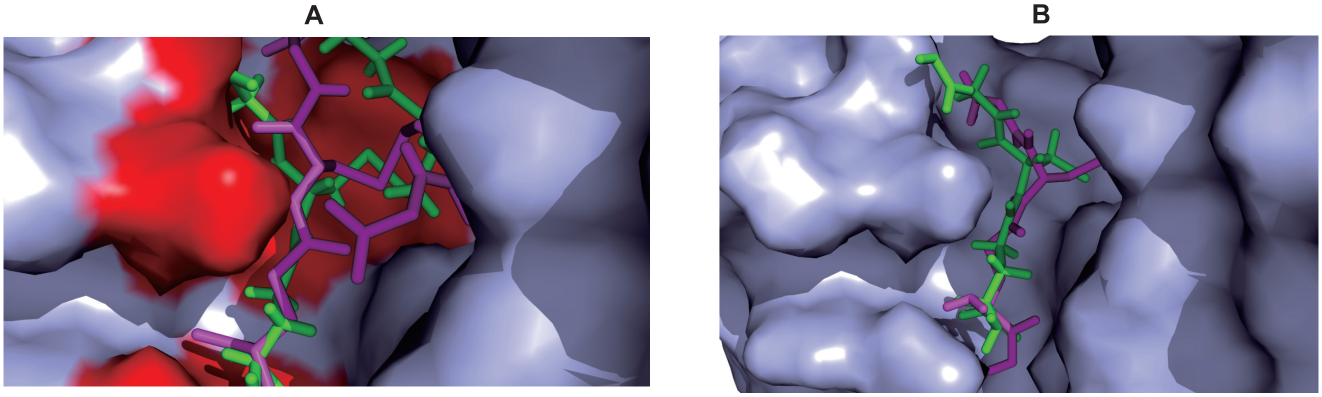


Figure 4. An example showing the improved ligand-binding poses by COACH-D over COACH. (**A**) and (**B**) are for the binding poses built with template and native ligand, respectively. The protein structure is shown in light-blue surface. For (A) the binding poses before and after docking are shown in green and magenta sticks, respectively. For (B), the experimental and predicted poses are shown in green and magenta sticks, respectively.

procedure of native binding residues. The second one is the clash score, which is defined as the total number of residues that have steric clash to the ligand in the predicted complex structures. A residue is said to have steric clash to the ligand when the closest atomic distance between the residue's and the ligand's atoms is less than three quarters of the sum of their Van der Waals radii (10).

The detailed experimental results on the 50 targets are presented in Supplementary Table S1. The mean MCCs of COACH-D and COACH are very similar (0.66 versus 0.67). However, the steric clashes are removed significantly after molecular docking. In the COACH models, the average clash score is 1.72, which is reduced to 0.26 (i.e. reduced by about 85%) in the COACH-D models. For 37 out of the 50 targets, steric clashes exist in the COACH models. Except four targets (4uj1A, 4ujaA, 5iqxA and 4z6dA), these clashes are removed or reduced in the COACH-D models. These data indicate that the COACH-D models are physically more realistic (i.e. with fewer steric clashes) than the COACH modes.

As mentioned earlier, the final ligand-binding pose is selected (out of 10 docking poses) as the one that matches the best with the consensus prediction of binding residues. Another option is to select the pose with the lowest docking energy. Supplementary Table S1 presents the results with such selection method, which shows that both methods lead to binding poses with the same clash scores (with exception of one target, 4za0B). However, the MCC for the former is higher than the latter (0.66 versus 0.62), which is the reason for our selection of ligand-binding pose. This is anticipated because the consensus prediction of binding residues is usually accurate and thus other predictions resemble it should have a higher chance to be correct.

Figure 4 presents an example target (PDB ID: 5f8bA) with improved ligand-binding poses. This protein is 'Glutathione S-transferase psoE' that binds glutathione (GSH). The template used for this target is the structure with PDB ID*:* 4is0A and the ligand is 'oxidized glutathione disulfide'. As reflected in Figure 4A, there are severe steric clashes between the residues shown in red surface and the ligand shown in green sticks (i.e. the one from COACH). These clashes were removed perfectly after docking in COACH-D (see the ligand in magenta sticks). In addition, the native ligand (in blue sticks of Figure 4B) was also submitted to COACH-D, which was docked into the predicted binding pocket as well. Figure 4B shows the predicted binding pose does not have any steric clash with the protein structure and it overlaps with the experimental binding pose well.

CONCLUSION

We have developed the COACH-D server for protein–ligand binding sites and ligand-binding poses prediction. The advantage of COACH-D over our previous method COACH is the utilization of molecular docking to improve the ligand-binding poses. In addition, the users can also submit their own ligand of interest together with the protein to the server. After the ligand-binding sites are predicted, the ligands from the users and the templates are then docked into the predicted binding pockets to build their complex structures. Blind tests show that the algorithm significantly outperforms other ligand-binding sites prediction methods. Benchmark tests show that the steric clashes between the ligand and the protein structures in the COACH models are reduced by 85% after molecular docking in COACH-D. We anticipate the accurate ligand-binding sites prediction and the improved ligand–protein complex structures could contribute to other related studies, such as drug discovery.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We want to thank the CAMEO team for organizing the CAMEO-LB experiments.

FUNDING

National Natural Science Foundation of China [11501306, 11501407]; Fok Ying-Tong Education Foundation [161003]; National Institutes of Health [GM083107, GM116960]; Fundamental Research Funds of Nankai University; Thousand Youth Talents Plan of China. Funding for open access charge: NSFC [11501407].

Conflict of interest statement. None declared.

REFERENCES

1. Opella,S.J. (2013) Structure determination of membrane proteins by nuclear magnetic resonance spectroscopy. *Annu. Rev. Anal. Chem.*, **6**, 305–328.
2. Leitner,A., Faini,M., Stengel,F. and Aebersold,R. (2016) Crosslinking and mass Spectrometry: An integrated technology to understand the structure and function of molecular machines. *Trends Biochem. Sci.*, **41**, 20–32.
3. Quan,S., Wang,L., Petrotchenko,E.V., Makepeace,K.A., Horowitz,S., Yang,J., Zhang,Y., Borchers,C.H. and Bardwell,J.C. (2014) Super Spy variants implicate flexibility in chaperone action. *Elife*, **3**, e01584.
4. Brylinski,M. and Skolnick,J. (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 129–134.
5. Capra,J.A., Laskowski,R.A., Thornton,J.M., Singh,M. and Funkhouser,T.A. (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, **5**, e1000585.
6. Roy,A., Yang,J. and Zhang,Y. (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.*, **40**, W471–W477.
7. Yang,J., Roy,A. and Zhang,Y. (2013) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, **29**, 2588–2595.
8. Heo,L., Shin,W.H., Lee,M.S. and Seok,C. (2014) GalaxySite: ligand-binding-site prediction by using molecular docking. *Nucleic Acids Res.*, **42**, W210–W214.
9. Roche,D.B., Buenavista,M.T. and McGuffin,L.J. (2013) The FunFOLD2 server for the prediction of protein-ligand interactions. *Nucleic Acids Res.*, **41**, W303–W307.
10. Trott,O. and Olson,A.J. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, **31**, 455–461.
11. Huang,S.Y. and Zou,X. (2007) Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins*, **66**, 399–421.
12. Allen,W.J., Balius,T.E., Mukherjee,S., Brozell,S.R., Moustakas,D.T., Lang,P.T., Case,D.A., Kuntz,I.D. and Rizzo,R.C. (2015) DOCK 6: Impact of new features and current docking performance. *J. Comput. Chem.*, **36**, 1132–1156.
13. Shin,W.H. and Seok,C. (2012) GalaxyDock: protein-ligand docking with flexible protein side-chains. *J. Chem. Inf. Model.*, **52**, 3225–3232.
14. Liu,Z., Su,M., Han,L., Liu,J., Yang,Q., Li,Y. and Wang,R. (2017) Forging the basis for developing protein-ligand interaction scoring functions. *Acc. Chem. Res.*, **50**, 302–309.
15. Li,Y. and Yang,J. (2017) Structural and sequence similarity makes a significant impact on machine-learning-based scoring functions for protein-ligand interactions. *J. Chem. Inf. Model.*, **57**, 1007–1012.
16. Haas,J., Roth,S., Arnold,K., Kiefer,F., Schmidt,T., Bordoli,L. and Schwede,T. (2013) The protein model Portal–a comprehensive resource for protein structure and model information. *Database*, **2013**, bat031.
17. Yang,J., Yan,R., Roy,A., Xu,D., Poisson,J. and Zhang,Y. (2015) The I-TASSER Suite: protein structure and function prediction. *Nat. Methods*, **12**, 7–8.
18. Yang,J., Roy,A. and Zhang,Y. (2013) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.*, **41**, D1096–D1103.
19. Rose,P.W., Prlic,A., Altunkaya,A., Bi,C., Bradley,A.R., Christie,C.H., Costanzo,L.D., Duarte,J.M., Dutta,S., Feng,Z. *et al.* (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.*, **45**, D271–D281.
20. Rego,N. and Koes,D. (2015) 3Dmol.js: molecular visualization with WebGL. *Bioinformatics*, **31**, 1322–1324.
21. Roche,D.B., Tetchner,S.J. and McGuffin,L.J. (2010) The binding site distance test score: a robust method for the assessment of predicted protein binding sites. *Bioinformatics*, **26**, 2920–2921.
22. Kallberg,M., Wang,H., Wang,S., Peng,J., Wang,Z., Lu,H. and Xu,J. (2012) Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.*, **7**, 1511–1522.