W210–W214 *Nucleic Acids Research, 2014, Vol. 42, Web Server issue* *Published online 21 April 2014*
doi: 10.1093/nar/gku321

GalaxySite: ligand-binding-site prediction by using molecular docking

Lim Heo[1], Woong-Hee Shin[1], Myeong Sup Lee[2,*] and Chaok Seok[1,*]

[1]Department of Chemistry, Seoul National University, Seoul 151-747, Korea and [2]Department of Biomedical Sciences, University of Ulsan College of Medicine, Seoul 138-736, Korea

Received February 11, 2014; Revised March 31, 2014; Accepted April 3, 2014

ABSTRACT

Knowledge of ligand-binding sites of proteins provides invaluable information for functional studies, drug design and protein design. Recent progress in ligand-binding-site prediction methods has demonstrated that using information from similar proteins of known structures can improve predictions. The GalaxySite web server, freely accessible at http://galaxy.seoklab.org/site, combines such information with molecular docking for more precise binding-site prediction for non-metal ligands. According to the recent critical assessments of structure prediction methods held in 2010 and 2012, this server was found to be superior or comparable to other state-of-the-art programs in the category of ligand-binding-site prediction. A strong merit of the GalaxySite program is that it provides additional predictions on binding ligands and their binding poses in terms of the optimized 3D coordinates of the protein–ligand complexes, whereas other methods predict only identities of binding-site residues or copy binding geometry from similar proteins. The additional information on the specific binding geometry would be very useful for applications in functional studies and computer-aided drug discovery.

INTRODUCTION

Proteins perform their biochemical functions by interacting with other biomolecules such as small ligands, other proteins or nucleic acids. The detection of binding site on a protein makes it possible to infer the function of the protein and provides information on binding pockets crucial for computer-aided drug discovery (1,2). Ligand-binding-site predictions from protein sequences have important implications with regard to sequence-based predictions of the functions of proteins. Binding-site prediction on known experimental protein structures is also important when the known structures do not contain ligands or can bind other ligands. Various evolutionary information-based, geometry-based, energy-based and combined methods have been reported (3).

Recently, methods that use experimental structures of similar protein–ligand complexes have been successfully applied in binding-site predictions in critical assessment of structure prediction (CASP) experiments (4–7). In such methods, binding-site information of homologous proteins of known structures is utilized by assuming that similar protein–ligand contacts occur in the target protein. These methods predict only ligand-binding residues or ligand-binding geometry based on simple structure superimposition to similar protein–ligand complexes (8–12). In this paper, we introduce a new method that uses such information in the context of protein–ligand docking. Because specific binding of ligands to proteins occurs owing to favorable physicochemical interactions, it can be expected that binding-site prediction based on physical chemistry using molecular docking can provide predictions that are more precise. In addition to revealing the identities of the contacting residues, molecular docking can also provide detailed information on atomic interactions between protein and ligand in terms of the optimized 3D coordinates of the protein–ligand complex. The binding geometry obtained by docking can be different from the geometry obtained by simple structure superimposition with similar proteins, and the binding pose optimized by docking tends to have physically more realistic geometry with no severe steric clashes. Such precise information would be very useful for the prediction of specific functions and applications in drug discovery.

However, a few difficulties have to be overcome to apply molecular docking to binding-site prediction methods. First, docking requires prior knowledge of the protein structure and binding ligand. Second, docking results can be sensitive to structural details, and the prediction accuracy may decrease if the protein structure is not sufficiently accurate or if conformational changes occur upon binding (8). In the GalaxySite program, binding ligand is predicted using a similarity-based method, and the protein structure is provided by the user or predicted from a template-

[*]To whom correspondence should be addressed. Tel: +82 2 880 9197; Fax: +82 2 889 1568; Email: chaok@snu.ac.kr
Correspondence may also be addressed to Myeong Sup Lee. Tel: +82 2 3010 2979; Email: myeong@amc.seoul.kr

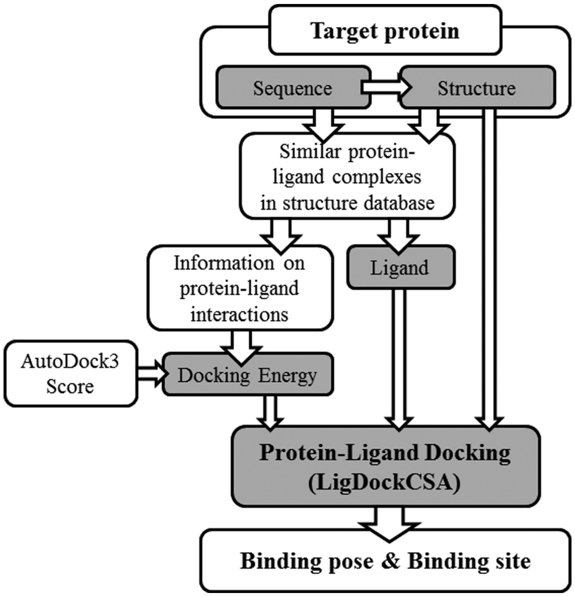

Target protein
Sequence
Structure
Similar protein-ligand complexes in structure database
Information on protein-ligand interactions
Ligand
AutoDock3 Score
Docking Energy
Protein-Ligand Docking (LigDockCSA)
Binding pose & Binding site

Figure 1. Flowchart of the GalaxySite algorithm. The ligand-binding site of a protein is predicted by protein–ligand docking. The protein structure required for the docking simulation may be provided by the user or predicted from the protein sequence. The binding ligand is predicted from similar protein–ligand complexes in the structure database. The molecular docking algorithm LigDockCSA is used with a hybrid energy of AutoDock3 energy and restraint energy derived from similar protein–ligand complexes. Binding-site residues are extracted from the docking pose.

based modeling method. The current binding-site prediction method is accurate even when only chemically similar ligands are predicted. The energy function for docking is designed to be less sensitive to structural details by adapting a combination of physics-based terms of AutoDock3 (13) and restraint terms derived from homologous protein–ligand complexes of known experimental structures.

GalaxySite has been tested on the following non-metal ligand-binding-site prediction test sets in addition to the blind prediction test sets of CASP9 and CASP10: 644 nucleotide-binding proteins with known experimental structures, 46 holo/apo pairs of proteins with experimentally resolved structures and 480 targets of the ligand-binding-site prediction category from the continuous automated model evaluation server (CAMEO; http://www.cameo3d.org/lb/) released between 16 August and 8 November 2013. In these tests, the performance of GalaxySite was superior or comparable to other state-of-the-art prediction methods.

MATERIALS AND METHODS

Overall procedure

The GalaxySite program predicts the ligand-binding site of a given protein by protein–ligand docking, as shown schematically in Figure 1. The GalaxySite program uses a protein sequence or structure as input. The input structure may be either an experimental or a predicted structure. If a protein sequence is provided, GalaxySite predicts the protein structure using GalaxyTBM, a template-based modeling method (14,15). Up to three non-metal ligands are extracted from the protein–ligand complex structures of similar proteins detected by HHsearch (16). Ligand-binding poses are then predicted using LigDockCSA (17).

Prediction of binding ligands

Ligands to be docked to the target protein structure are predicted from experimental structures of template proteins with bound ligands. The template search is performed in the protein structure database ‘pdb70’ with a maximum mutual sequence identity of 70% via HHsearch (16) in the local alignment mode. Out of the 30 proteins with the highest re-ranking score calculated from the HHsearch results (14,15), proteins whose structures are very different from that of the target protein are filtered out, and the remaining proteins are selected as templates. The criterion used for filtering out dissimilar structures depends on the similarity of the target protein structure to the closest template among the top 30 proteins. Structures with TM-score [similarity to the target structure calculated using TM-align (18)] <0.5, <0.4 and <0.3 were filtered out when TM-score of the closest template is >0.8, >0.6 and ≤ 0.6, respectively. In this way, prediction accuracy is enhanced with the use of stricter criterion [for example, TM-score >0.5 to include only those proteins that share the same fold (19)] when more accurate prediction is expected (for example, when the similarity of the best template to the target structure is very high with TM-score >0.8), and prediction coverage is enhanced with less strict criterion when less accurate prediction is expected. Among the non-metal ligands bound to the templates, non-biological ligands such as sulphate ion, glycerol and polyethylene glycol that are added to facilitate crystallization are filtered out first. See Supplementary Information for a complete list of the ligands considered non-biological. Ligands with high positional variation (>10 Å) of the center atoms in superposed template structures that contain the same ligand are also filtered out. The remaining ligands are ranked according to the sum of the HHsearch re-ranking score of templates that contain the same ligand; up to three ligands with the highest rank are used in the docking calculations. The overall procedure of binding-ligand selection was trained on the CASP7 function prediction targets.

Molecular docking

GalaxySite uses the LigDockCSA (17) protein–ligand docking program that performs global optimization by using the conformational space annealing (CSA) algorithm (17, 20–22). The protein structure is fixed at the initial input or model structure, and the ligand is considered fully flexible. A pool of 100 conformations is first generated by perturbing the initial conformations obtained from template ligand poses. The pool is then evolved by generating trial conformations and comparing the trial conformations with the pool conformations, gradually focusing on narrower regions of lower energy in the conformational space. Details on the docking algorithm can be found elsewhere (17). Out of the final pool of 100 structures, the pose with the lowest docking energy in the largest cluster is selected as representative binding pose.

The energy function used for docking is expressed as follows:

$$E = E_{\text{AutoDock}} + 1.1\, E_{\text{Restraint}}, \tag{1}$$

where E_{AutoDock} is the same as the AutoDock3 energy function (13) except that the maximum energy value for each interacting atom pair is set to 1.0 kcal/mol to tolerate steric clashes that may be caused by inaccurate protein model structures or ligand-unbound structures. The restraint term $E_{\text{Restraint}}$ is derived from the template structures that contain the selected ligand. Restraint is applied to each ligand atom i, imposing a penalty on r_{ij} (the distance between ligand atom i and protein atom j) deviating from $r_{ij}{}^{(k)}$ (the corresponding distance in the kth template) with template-dependent weight factor ω_{ijk}, and the total restraint energy is expressed as follows:

$$E_{\text{Restraint}}(\{r_{ij}\}) = \\ -\sum_i \ln \left[\sum_j \sum_k \omega_{ijk} \exp\{-(r_{ij} - r_{ij}^{(k)})^2 / d_{jk}^2\} \right], \tag{2}$$

where d_{jk} is the position deviation of the C_α atom of the residue to which the jth atom belongs in the target structure from that in the kth template when target and template structures are superimposed. The weight factor is expressed as

$$\omega_{ijk} = \\ (\text{TM} - \text{score})_k (\text{Residuescore})_{jk} E_{\text{AutoDock},ij}(r_{ij}^{(k)}) / \\ E_{\text{AutoDock},ij}(r_{\min}), \tag{3}$$

where (TM-score)$_k$ is the structural similarity between the kth template and the input structure. The second term, (Residue score)$_{jk}$, is 0 if the corresponding template residue is not of the same amino acid type as the target residue or if $d_{jk} > 2$ Å. Otherwise, the residue score represents side-chain orientation similarity calculated using the dot product of the normalized vectors connecting C_α atoms and the side-chain centroid (of the residue to which the jth atom belongs) for the input and kth template structures. The third term accounts for the optimality of the template distance estimated by the ratio of the AutoDock3 energy value at that distance to the optimal energy. The relative weight of the AutoDock3 energy to the restraint term in Equation (1) is set to 1.1, which produces optimal results for the targets in the CASP7 function prediction category (6).

Performance of the method

GalaxySite has been extensively tested on various types of binding-site prediction test sets. See Supplementary Information for details on the test results. Tests on 644 nucleotide-derived ligand-binding proteins (23) and 46 holo/apo pairs of experimentally resolved structures (24) show that GalaxySite performs superior or comparable to other state-of-art methods in predicting binding sites from protein structures. Prediction from protein sequences was performed on targets in the binding-site prediction category of CASP9 (7) and CASP10 (4) in a blind fashion and on 480 targets from the continuous automated model evaluation server (CAMEO) released between 16 August and 8 November 2013. The prediction accuracy was comparable to other state-of-the-art prediction methods in terms of the Matthews correlation coefficient (MCC) for ligand-contacting residues. See Figure 2 for comparison with other server methods and Supplementary Information for details. It should be noted that the accuracy measures for comparison with other methods depend on the available information provided by other methods and that more detailed, valuable information on specific protein–ligand interactions is available using GalaxySite.

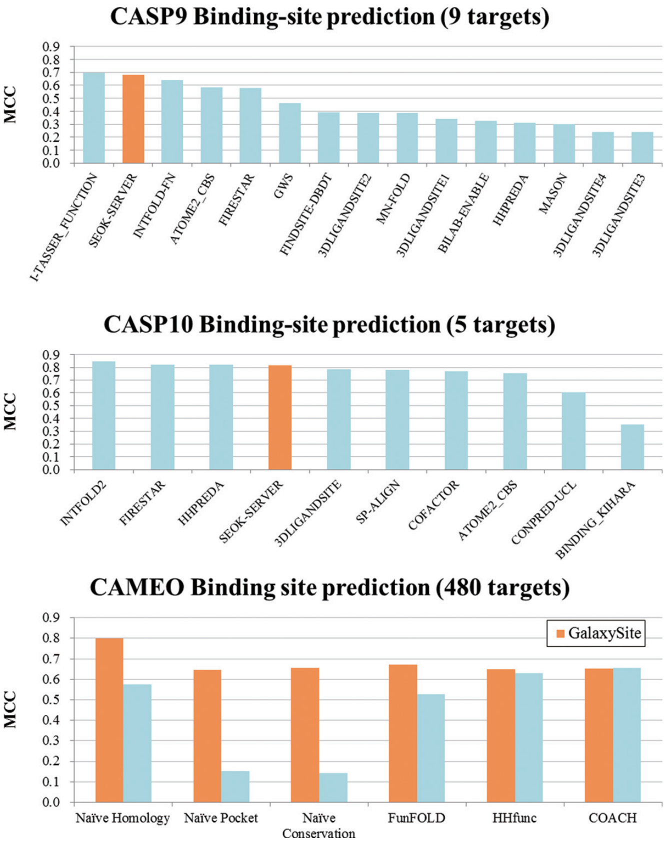
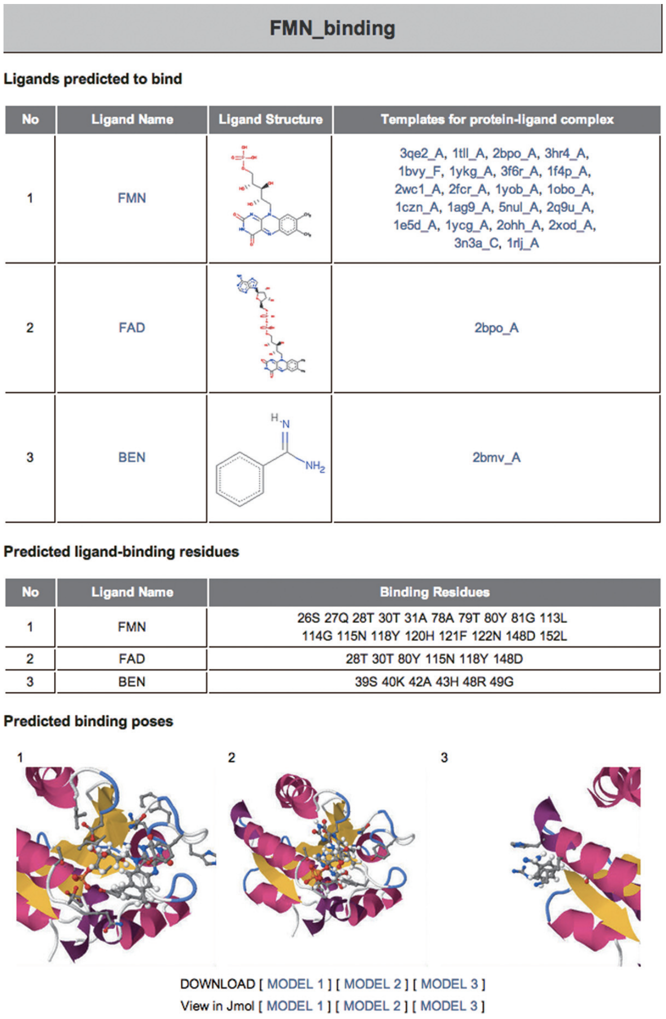


Figure 2. Performance comparison of GalaxySite with other server methods in terms of median Matthews correlation coefficient (MCC) on the CASP9 (top), CASP10 (middle) and CAMEO (bottom) ligand-binding-site prediction category targets. In the figure, the SEOK-SERVER method used GalaxySite in the CASP blind prediction.

THE GALAXYSITE SERVER

Hardware and software

The GalaxySite server runs on a cluster of seven Linux servers of 2.33-GHz Intel Xeon 8-core processors. The web application uses Python and the MySQL database. The ligand-binding-site prediction pipeline is implemented using Python. Open Babel 2.2.3 is used to prepare the ligands for the molecular docking procedure. The molecular docking algorithm for binding-site prediction is implemented in the GALAXY program package (14,15,17,25–28) written in Fortran 90. When a sequence is given as an input, the



FMN_binding

Ligands predicted to bind

No	Ligand Name	Ligand Structure	Templates for protein-ligand complex
1	FMN		3qe2_A, 1tll_A, 2bpo_A, 3hr4_A, 1bvy_F, 1ykg_A, 3f6r_A, 1f4p_A, 2wc1_A, 2fcr_A, 1yob_A, 1obo_A, 1czn_A, 1ag9_A, 5nul_A, 2q9u_A, 1e5d_A, 1ycg_A, 2ohh_A, 2xod_A, 3n3a_C, 1rlj_A
2	FAD		2bpo_A
3	BEN		2bmv_A

Predicted ligand-binding residues

No	Ligand Name	Binding Residues
1	FMN	26S 27Q 28T 30T 31A 78A 79T 80Y 81G 113L 114G 115N 118Y 120H 121F 122N 148D 152L
2	FAD	28T 30T 80Y 115N 118Y 148D
3	BEN	39S 40K 42A 43H 48R 49G

Predicted binding poses

DOWNLOAD [MODEL 1] [MODEL 2] [MODEL 3]
View in Jmol [MODEL 1] [MODEL 2] [MODEL 3]

Figure 3. GalaxySite output page. Predicted ligands, their two-dimensional structures and templates for protein–ligand complexes are tabulated. Ligand names and PDB IDs are linked to the RCSB PDB website for detailed information. Predicted ligand-binding residues for each ligand are also listed. Predicted binding poses are shown in static images, which can be viewed using the Jmol structure viewer or can be downloaded in PDB format.

structure is predicted by GalaxyTBM (14,15) with no additional model refinement. The Jmol software is used for visualization of predicted results.

Input and output

The required input is a protein sequence in FASTA format or a protein structure in PDB format. The number of residues in the input file is limited to 500 for computational efficiency. The average run times are 2 h for a structure input and 4 h for a sequence input. Predictions for up to three non-metal ligands and their template complexes are provided with links to the RCSB PDB website. For each predicted ligand, predicted binding pose and ligand-binding residues can be viewed and downloaded from the website (Figure 3).

CONCLUSIONS

GalaxySite is a web server for the prediction of binding sites of non-metal ligands that employs molecular docking. The method is applicable to experimentally resolved structures, model protein structures and protein sequences. In addition to information on binding residues provided by previous binding-site prediction methods, GalaxySite predicts specific binding ligands and binding poses that can be useful for further applications, e.g. in computer-aided drug discovery.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online, including [1–7].

FUNDING

National Research Foundation of Korea (2013R1A2A1A09012229, 2012M3C1A6035362 and 2013R1A1A2058447); Asan Institute for Life Sciences, Seoul, Korea (2013–591 and 2014–607). Funding for open access charge: Seoul National University.
Conflict of interest statement. None declared.

REFERENCES

1. Campbell,S.J., Gold,N.D., Jackson,R.M. and Westhead,D.R. (2003) Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.*, **13**, 389–395.
2. Kinoshita,K. and Nakamura,H. (2003) Protein informatics towards function identification. *Curr. Opin. Struct. Biol.*, **13**, 396–400.
3. Tripathi,A. and Kellogg,G.E. (2010) A novel and efficient tool for locating and characterizing protein cavities and binding sites. *Proteins*, **78**, 825–842.
4. Cassarino,T.G., Bordoli,L. and Schwede,T. (2013) Assessment of ligand binding site predictions in CASP10. *Proteins*, **82**, 154–163.
5. Lopez,G., Ezkurdia,I. and Tress,M.L. (2009) Assessment of ligand binding residue predictions in CASP8. *Proteins*, **77**, 138–146.
6. Lopez,G., Rojas,A., Tress,M. and Valencia,A. (2007) Assessment of predictions submitted for the CASP7 function prediction category. *Proteins*, **69**, 165–174.
7. Schmidt,T., Haas,J., Gallo Cassarino,T. and Schwede,T. (2011) Assessment of ligand-binding residue predictions in CASP9. *Proteins*, **79**, 126–136.
8. Brylinski,M. and Skolnick,J. (2009) FINDSITE: a threading-based approach to ligand homology modeling. *PLoS Comput. Biol.*, **5**, e1000405.
9. Lopez,G., Maietta,P., Rodriguez,J.M., Valencia,A. and Tress,M.L. (2011) firestar–advances in the prediction of functionally important residues. *Nucleic Acids Res.*, **39**, W235–W241.
10. 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.
11. Wass,M.N., Kelley,L.A. and Sternberg,M.J. (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.*, **38**, W469–W473.
12. 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.
13. Morris,G.M., Goodsell,D.S., Halliday,R.S., Huey,R., Hart,W.E., Belew,R.K. and Olson,A.J. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, **19**, 1639–1662.
14. Ko,J., Park,H., Heo,L. and Seok,C. (2012) GalaxyWEB server for protein structure prediction and refinement. *Nucleic Acids Res.*, **40**, W294–W297.

15. Ko,J., Park,H. and Seok,C. (2012) GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions. *BMC Bioinformatics*, **13**, 198.
16. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
17. Shin,W.H., Heo,L., Lee,J., Ko,J., Seok,C. and Lee,J. (2011) LigDockCSA: protein-ligand docking using conformational space annealing. *J. Comput. Chem.*, **32**, 3226–3232.
18. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
19. Xu,J. and Zhang,Y. (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, **26**, 889–895.
20. Joo,K., Lee,J., Seo,J.H., Lee,K., Kim,B.G. and Lee,J. (2009) All-atom chain-building by optimizing MODELLER energy function using conformational space annealing. *Proteins-Struct. Funct. Bioinformatics*, **75**, 1010–1023.
21. Lee,J., Scheraga,H.A. and Rackovsky,S. (1997) New optimization method for conformational energy calculations on polypeptides: conformational space annealing. *J. Comput. Chem.*, **18**, 1222–1232.
22. Lee,K., Czaplewski,C., Kim,S.Y. and Lee,J. (2005) An efficient molecular docking using conformational space annealing. *J. Comput. Chem.*, **26**, 78–87.
23. Kasahara,K., Kinoshita,K. and Takagi,T. (2010) Ligand-binding site prediction of proteins based on known fragment-fragment interactions. *Bioinformatics*, **26**, 1493–1499.
24. Huang,B. and Schroeder,M. (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.*, **6**, 19.
25. Heo,L., Park,H. and Seok,C. (2013) GalaxyRefine: protein structure refinement driven by side-chain repacking. *Nucleic Acids Res.*, **41**, W384–W388.
26. Lee,H., Park,H., Ko,J. and Seok,C. (2013) GalaxyGemini: a web server for protein homo-oligomer structure prediction based on similarity. *Bioinformatics*, **29**, 1078–1080.
27. Shin,W.H., Kim,J.K., Kim,D.S. and Seok,C. (2013) GalaxyDock2: protein-ligand docking using beta-complex and global optimization. *J. Comput. Chem.*, **34**, 2647–2656.
28. Shin,W.H. and Seok,C. (2012) GalaxyDock: protein-ligand docking with flexible protein side-chains. *J. Chem. Inf. Model.*, **52**, 3225–3232.