GalaxyHomomer: a web server for protein homo-oligomer structure prediction from a monomer sequence or structure

Minkyung Baek¹, Taeyong Park¹, Lim Heo¹, Chiwook Park² and Chaok Seok¹,*

¹Department of Chemistry, Seoul National University, Seoul 151-747, Korea and ²Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, IN 47907, USA

Received February 4, 2017; Revised March 23, 2017; Editorial Decision March 31, 2017; Accepted April 05, 2017

ABSTRACT

Homo-oligomerization of proteins is abundant in nature, and is often intimately related with the physiological functions of proteins, such as in metabolism, signal transduction or immunity. Information on the homo-oligomer structure is therefore important to obtain a molecular-level understanding of protein functions and their regulation. Currently available web servers predict protein homo-oligomer structures either by template-based modeling using homo-oligomer templates selected from the protein structure database or by ab initio docking of monomer structures resolved by experiment or predicted by computation. The GalaxyHomomer server, freely accessible at http://galaxy.seoklab.org/homomer, carries out template-based modeling, ab initio docking or both depending on the availability of proper oligomer templates. It also incorporates recently developed model refinement methods that can consistently improve model quality. Moreover, the server provides additional options that can be chosen by the user depending on the availability of information on the monomer structure, oligomeric state and locations of unreliable/flexible loops or termini. The performance of the server was better than or comparable to that of other available methods when tested on benchmark sets and in a recent CASP performed in a blind fashion.

INTRODUCTION

A large fraction of cellular proteins self-assemble to form symmetric homo-oligomers with distinct biochemical and biophysical properties (1–3). For example, ligand-binding sites or catalytic sites are located at oligomer interfaces in many proteins (4–6), and oligomerization is often necessary for effective signal transduction through membrane receptor proteins (7,8) and selective gating of channel proteins (9). Therefore, knowledge of the homo-oligomer structure is essential for understanding the physiological functions of proteins at the molecular level and for designing molecules that regulate the functions.

Methods for predicting the protein homo-oligomer structure can be divided into two categories: those that use templates selected from the protein structure database and others that dock monomer structures ab initio, without using template information. Usually, template-based methods require a sequence as input, whereas docking methods require a monomer structure as input. The latter requirement can be more restrictive for the user if the monomer structure has to be predicted by another method, but it may be preferred if an experimentally resolved monomer structure is available. It is generally expected that template-based methods produce more accurate predictions under a situation in which similar proteins forming oligomers exist in the structure database. Docking methods may be more useful when proper oligomer templates are not available but the monomer structure is reliable. Several protein–protein docking methods have been reported to date (10–18), and some of these are available as public web servers for predicting homo-oligomer structures. M-ZDOCK (13) and GRAMM-X (15), which use ab initio docking based on fast Fourier transformation (FFT), are two such examples. The oligomeric state must be provided as input in these servers. However, relatively few web servers that use template-based methods have been reported. ROBETTA (19,20) and SWISS-MODEL (21) are two web servers that predict the homo-oligomer structure from an amino acid sequence. GalaxyGemini (22) predicts the homo-oligomer structure from a monomer structure. These servers predict the oligomeric state automatically. Depending on the availability of information on the oligomeric state, the user may or may not prefer to specify the oligomeric state. Here, we introduce a new web server called GalaxyHomomer that predicts the homo-oligomer structure from either the
The oligomeric state may or may not be specified by the user. The server can perform both template-based oligomer modeling and ab initio docking. It returns five model structures and automatically decides how many models are generated by which method depending on the existence of proper oligomer templates.

Oligomer structures predicted by template-based methods may have errors due to sequence differences between the target and template proteins. Those predicted by docking methods may have inaccuracy if structural change of the monomer induced by oligomerization is not considered. In the previous CASP experiment conducted in 2014 in collaboration with CAPRI, we showed that such errors in predicted oligomer structures could be reduced by remodeling inaccurately predicted loops or termini and by relaxing the overall structure (23). GalaxyHomomer incorporates such state-of-the-art model refinement methods to improve the accuracy of homo-oligomer models generated by both template-based modeling and ab initio docking.

According to the assessment of the recent blind prediction experiment CASP12 conducted in 2016, GalaxyHomomer, participated as ‘Seok-assembly’, ranked second among the servers participated in the assembly category. When we tested GalaxyHomomer on 136 targets from PISA benchmark set, 47 targets from a membrane protein set, 20 targets from CASP11 experiments and 89 targets from CAMEO protein structure prediction category, it showed a performance better than or comparable to that of other available homo-oligomer structure prediction methods.

**THE GALAXYHOMOMER METHOD**

**Overall procedure**

The overall pipeline of GalaxyHomomer is presented in Figure 1. Either a sequence or structure (experimental or predicted structure) of the monomer can be provided as input. If the oligomeric state is not specified by the user, possible oligomeric states are predicted first. Five homo-oligomer structures with the given oligomeric states are then generated by template-based modeling and ab initio docking. Oligomer templates required by template-based modeling are detected based only on the sequence as well as with additional structure information. The models are further refined by loop/terminus modeling using GalaxyLoop (24–26) and by overall relaxation using GalaxyRefineComplex (27).

**Prediction of the oligomeric state**

Possible oligomeric states are predicted from the input sequence by a similarity-based method as follows. First, HHsearch (28) is run in the local alignment mode to detect proteins that are similar to the target in the protein structure database ‘pdb70’, with a maximum mutual sequence identity of 70%. The oligomeric states of the database proteins were assigned according to the biological units described in ‘REMARK 350’. Second, the proteins are re-ranked by a score $S$, which combines the HHsearch sequence score and HHsearch secondary structure score (29). Next, the $S$ scores of the proteins in the same oligomeric states are summed for the top 100 proteins, and the ratios of different oligomeric states are determined in proportion to the $S$ sums. Finally, oligomeric states for five models are assigned according to the oligomeric state ratios.

**Template-based oligomer modeling**

The same top 100 proteins described above are considered as candidates for oligomer templates. If a sequence is provided as input, up to five proteins are selected as templates based on the ranking of $S$ among those with $S > 0.2$ times the highest $S$ overall and those $> 0.7$ times the highest $S$ for the given oligomeric state. If the number of detected templates using this sequence-based method is less than five, additional templates are selected using the monomer structure predicted by the template-based modeling program GalaxyTBM (29). Structure-based templates are selected according to the ranking of $S$ among those with monomer structures similar to the given monomer structure (TM-score calculated using TM-align (30) > 0.5) and in the given oligomeric state. If a structure is provided as input, only the structure-based template detection is used with the monomer structure provided by the user.

For each oligomer template detected by the sequence-based method, an oligomer structure is built using the in-house model-building program GalaxyCassiopeia, a component of the most recent version of GalaxyTBM (29).
GalaxyCassiopeia builds models from the sequence alignment and template structure by the VTFM optimization used in MODELLER (31) but with FACTS solvation free energy (32), knowledge-based hydrogen bond energy (33) and dipolar-DFIRE (34) in addition to molecular mechanics bonded and non-bonded energy terms and template-derived restraints. For each template detected by the structure-based method, an oligomer structure is built by superimposing the monomer structure onto the oligomer template.

**Ab initio docking**

If less than five oligomer templates are detected by the two template detection methods described above, the remaining homo-oligomer models with the given oligomeric states are generated using the in-house ab initio docking program GalaxyOligoTongDock. This docking program predicts homo-oligomer structures from the monomer structure using the grid-based FFT docking method of MZDOCK (13) implemented in-house. Only Cn-symmetry is considered, and D-symmetry is not supported. The top 200 homo-oligomer structures generated by FFT are clustered using NMRClust (35), and the clusters are ranked according to the cluster size. From each of the highest ranking clusters, the highest-score structure is selected.

**Structure refinement**

Less reliable loop or terminal regions are re-modeled using GalaxyLoop (24–26) considering symmetry of the homo-oligomer structure for the first model for those regions predicted to be unreliable if a sequence is provided as input, and for all five models for user-specified regions if a structure is provided as input. GalaxyRefineComplex (27) is subsequently run to further relax the overall structure. The user can run additional refinement jobs by clicking the ‘Submit’ button in the results table on the output page.

**Performance of the method**

The GalaxyHomomer server was tested on 25 targets in CASP12 in a blind fashion, and this server, named ‘Seok’-assembly’, ranked second among the servers participated in the assembly category (http://www.predictioncenter.org/casp12). In CASPs, the oligomeric state is provided by the organizers. The server was also tested on three benchmark sets for which the oligomeric state is given as input (136 homo-oligomer proteins from the PISA benchmark set (36), 47 homo-oligomer membrane proteins compiled from the PDB (Supplementary Data) and 20 homo-oligomer proteins among the targets of CASP11 held in 2014 in collaboration with CAPRI (18)) and on a set for which the oligomeric state is not provided as input (89 homo-oligomer proteins among CAMEO (37) targets released from 13 August 2016 to 11 November 2016). In these tests, the performance of GalaxyHomomer was better than or comparable to that of other methods for which performance data are available for the sets in terms of the CAPRI accuracy criterion, as summarized in Table 1. Note that some methods take only the structure as input. The CAPRI criterion reflects the biological relevance of the model structures, and model qualities are classified as high (***) and medium (**); acceptable (*) and incorrect considering the ligand root mean-square deviation (L-RMSD) and interface RMSD (I-RMSD) from the experimental structure and the fraction of predicted native contacts ($F_{\text{nat}}$) (38). See Supplementary Data for details on the benchmark tests. It has to be noted that GalaxyHomomer does not consider the lipid bilayer environment of membrane proteins explicitly in terms of energy or geometry during energy-based optimization and docking. However, the results on membrane proteins in Table 1 are quite promising, implying that membrane environment was effectively taken into account in an implicit manner by using the database structures of membrane proteins as templates. GalaxyHomomer showed better performance than GalaxyGemini (22), a previous homo-oligomer structure prediction server developed by us, on the CASP11 benchmark set, as summarized in Ta-

### Table 1. Performance comparison of homo-oligomer structure prediction methods in terms of the CAPRI accuracy criterion

<table>
<thead>
<tr>
<th>Benchmark set</th>
<th>Prediction methods</th>
<th>Input</th>
<th>Up to 5 models&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Top 1 model&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>PISA (136 targets)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>GalaxyHomomer</td>
<td>Sequence</td>
<td>62/3**/38**</td>
<td>57/3**/39**</td>
</tr>
<tr>
<td></td>
<td>HH+MODELLER&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Sequence</td>
<td>61/2**/38**</td>
<td>45/1**/26**</td>
</tr>
<tr>
<td>Membrane proteins (47 targets)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>GalaxyHomomer</td>
<td>Sequence</td>
<td>19/1**/14**</td>
<td>19/1**/9**</td>
</tr>
<tr>
<td></td>
<td>HH+MODELLER&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Sequence</td>
<td>18/0**/6**</td>
<td>14/0**/4**</td>
</tr>
<tr>
<td>CASP11 (20 targets)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>GalaxyHomomer</td>
<td>Sequence</td>
<td>12/0**/8**</td>
<td>12/0**/5**</td>
</tr>
<tr>
<td></td>
<td>HADDOCK</td>
<td>Structure</td>
<td>14/0**/10**</td>
<td>13/0**/9**</td>
</tr>
<tr>
<td></td>
<td>ClusPro</td>
<td>Structure</td>
<td>14/0**/7**</td>
<td>10/0**/5**</td>
</tr>
<tr>
<td></td>
<td>BAKER-ROSETTASERVER</td>
<td>Sequence</td>
<td>9/0**/8**</td>
<td>9/0**/7**</td>
</tr>
<tr>
<td></td>
<td>SwarmDock</td>
<td>Structure</td>
<td>9/0**/3**</td>
<td>8/0**/3**</td>
</tr>
<tr>
<td></td>
<td>GalaxyGemini&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Structure</td>
<td>Not available</td>
<td>7/0**/5**</td>
</tr>
<tr>
<td></td>
<td>GRAMM-X</td>
<td>Structure</td>
<td>5/0**/1**</td>
<td>3/0**/1**</td>
</tr>
<tr>
<td>CAMEO (89 targets)</td>
<td>GalaxyHomomer</td>
<td>Sequence</td>
<td>44/6**/25**</td>
<td>35/3**/25**</td>
</tr>
<tr>
<td></td>
<td>Robetta</td>
<td>Sequence</td>
<td>28/4**/17**</td>
<td>26/4**/15**</td>
</tr>
<tr>
<td></td>
<td>SWISS-MODEL&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Sequence</td>
<td>Not available</td>
<td>23/3**/16**</td>
</tr>
</tbody>
</table>

<sup>a</sup>Data represent the numbers of targets for which the best of up to five predicted models were of acceptable or higher/high accuracy (***) and medium accuracy (**); values for model 1 are shown.

<sup>b</sup>Oligomeric state of target protein is given as an input.

<sup>c</sup>Up to five homo-oligomer models were generated by MODELLER based on the templates detected by HH-search.

<sup>d</sup>Data for up to five models were not provided for GalaxyGemini and SWISS-MODEL because they generated only single models.
Figure 2. An example output page of GalaxyHomomer. Five generated models are visualized using the JavaScript Protein Viewer. The models can be downloaded in PDB format. Additional information such as the number of subunits, interface area, information on templates and ab initio docking score is provided in the tables.

CONCLUSION

The GalaxyHomomer server predicts the homo-oligomer structure of a target protein from a sequence or monomer structure. It performs both template-based modeling and ab initio docking, and adopts additional model refinement that can consistently improve model quality. The server provides different options that can be chosen by the user depending on the availability of information on monomer structure, oligomeric state and locations of unreliable/flexible loops or termini. By combining additional refinement based on loop modeling and overall structure refinement, GalaxyHomomer may generate more precise homo-oligomer models that can be useful for further applications such as for drug design targeting protein homo-oligomer interfaces.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Research Foundation of Korea funded by the Ministry of Science, ICT & Future Planning [2016R1A2A1A05005485, 2016M3C4A7952630]; Korea Institute of Science and Technology Information Supercomputing Centre [KSC-2016-C2-0057]; Brain Pool Program through the Korean Federation of Science and Technology Societies (KOFST) [132S-1-3-0622]. Funding for open access charge: Seoul National University.

Conflict of interest statement. None declared.

REFERENCES
