Vol. 29 no. 8 2013, pages 1078–1080
doi:10.1093/bioinformatics/btt079

Advance Access publication February 14, 2013

GalaxyGemini: a web server for protein homo-oligomer structure prediction based on similarity

Hasup Lee, Hahnbeom Park†, Junsu Ko‡ and Chaok Seok*
Department of Chemistry, Seoul National University, Seoul 151-747, Republic of Korea

Associate Editor: Anna Tramontano

ABSTRACT

Summary: A large number of proteins function as homo-oligomers; therefore, predicting homo-oligomeric structure of proteins is of primary importance for understanding protein function at the molecular level. Here, we introduce a web server for prediction of protein homo-oligomer structure. The server takes a protein monomer structure as input and predicts its homo-oligomer structure from oligomer templates selected based on sequence and tertiary/quaternary structure similarity. Using protein model structures as input, the server shows clear improvement over the best methods of CASP9 in predicting oligomeric structures from amino acid sequences.

Availability: http://galaxy.seoklab.org/gemini.

Contact: chaok@snu.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received and revised on December 24, 2012; accepted on February 10, 2013

1 INTRODUCTION

Many proteins self-assemble into oligomers to perform their biological functions (Poupon and Janin, 2010). For example, certain enzymes form substrate-binding pockets at their dimer interfaces (Snijder *et al*., 1999), whereas antibodies form oligomers to create additional binding sites, increasing effective binding affinity via a ‘multivalent effect’ (Plückthun and Pack, 1997). Many membrane proteins also form oligomers for effective signal transduction (Heldin, 1995). Knowledge of the protein oligomeric state is therefore crucial for understanding protein function at the molecular level.

In the case of experimental protein structures deposited in the Protein Data Bank (PDB), oligomeric states may be annotated by the authors or can be assigned from crystallographic information through the Protein Interfaces, Surfaces and Assembly (PISA) database (Krissinel and Henrick, 2007). When such information is not available, e.g. for protein model structures, prediction of the oligomeric state is required. Recent studies have suggested that homology-based homo-oligomer prediction methods can be more powerful than *ab initio* methods (Morita *et al*., 2012).

Methods for prediction of protein oligomeric structures were assessed in a blind fashion for the first time in the 9th Critical Assessment of Protein Structure Prediction (CASP9; Mariani *et al*., 2011). In this experiment, participants were asked to predict homo-oligomer structures from amino acid sequences. Surprisingly, no method performed better than naïve predictors that take the top-ranking protein by HHsearch (Söding, 2005) as a template, implying that the current methods for prediction of oligomeric structures are ineffective, with substantial room for improvement.

In this context, we introduce a new web server GalaxyGemini for predicting protein homo-oligomer structure, which shows clear improvement over naïve predictors on two test sets.

2 METHODS

2.1 Oligomer database and test sets

We constructed a database of known homo-oligomer structures containing 22 233 proteins with mutual sequence identity <70% from all the structures deposited in the PDB (April 10, 2010). Oligomer templates are selected from this database. For each crystal structure, the oligomeric state was assigned as the biological unit determined by authors if ‘REMARK 350’ in PDB was available and assigned by PISA otherwise. When PISA predicted multiple oligomeric states, the top oligomeric state was used, instead of being removed from the database, to increase the coverage of the database. According to the previous benchmark results, PISA assignments can be regarded reliable with a success rate of 80~90%. For protein structures solved by NMR, the oligomeric states were defined as the assembled chain structures in the PDB entry.

The database was generated before CASP9 experiment, so the current test results on the CASP9 set (96 proteins containing 43 monomers; Mariani *et al*., 2011) can be fairly compared with CASP9 predictors including Naïve predictors. For parameter training on the PISA benchmark set (195 proteins containing 55 monomers; Ponstingl *et al*., 2003), target proteins were removed from the oligomer template lists.

2.2 Oligomer structure prediction

For a given input protein, HHsearch is first run on the oligomer database. Whether the query protein is oligomeric or not is then predicted by a scoring function S_1. If the top-ranking protein is monomeric, the query protein is predicted to be monomeric. Otherwise, an oligomer template is selected by ranking with a second function S_2. Prediction of the oligomeric state corresponding to each template is obtained by superimposing the input monomer structure onto the subunits of the oligomer template using the structure alignment tool TM-align (Zhang and Skolnick, 2005). Finally, rigid-body energy minimization is performed to remove steric clashes at the oligomer interface as explained in Supplementary Information.

*To whom correspondence should be addressed.
†Present address: Department of Biochemistry, University of Washington, Seattle, WA 98195, USA.
‡Present address: Theragen Bio Institute, Suwon, Gyueonggi-do 443-270, Republic of Korea.

The two scoring functions S_1 and S_2 are expressed as the weighted sums of Z-scores of five components. The first four components are derived from HHsearch: (i) HHsearch sequence score, (ii) HHsearch secondary structure score, (iii) ratio of aligned residues to the query sequence length and (iv) ratio of aligned residues to the sequence length of template candidate in the HHsearch alignment. These components account for sequence similarity to the query protein. The fifth component, called interface alignment score, accounts for tertiary and quaternary structure similarity by adding BLOSUM62 matrix scores (Henikoff and Henikoff, 1992) between the interface residues of template candidate and the residues of the query protein aligned to them. Addition of this component is important because interface residues are more conserved than other surface residues (Caffrey, 2004). The weight parameters for the two scoring functions were determined by training on the PISA benchmark set with a grid search. Details on scoring functions and cross-validation results are described in Supplementary Information.

3 RESULTS

Identification of the correct number of subunits in an oligomer was evaluated by measuring the ‘relative accuracy’. For more precise evaluation of the predicted structure, the ‘contact agreement score’ (S_{agree}) was measured, which reflects the fraction of correctly modeled interface contacts in the complex. Both measures were used in the CASP9 assessment and explained in Supplementary Information.

GalaxyGemini increased relative accuracy from 75.4% (for the naïve predictor NaïveSeqScore that takes the HHsearch top ranker by sequence score) to 79.5% for the training set (PISA benchmark set) and from 69.8% to 77.1% for the test set (CASP9 set) (Fig. 1a). The sum of S_{agree} over the targets increased from 74.7 to 88.0 for the training set and from 13.6 to 17.6 for the test set when ‘experimental’ monomer structures were used as input (Supplementary Fig. S1). When tertiary structures predicted by GalaxyTBM (Ko *et al.*, 2012) were used as input for the CASP9 set, the sum of S_{agree} increased from 9.4 to 12.1 (Fig. 1b–d for target-based comparison). GalaxyGemini outperforms all other CASP9 predictors including naïve predictors by the two measures, implying that GalaxyGemini may be successfully applied to ‘sequence-based’ oligomeric structure prediction.

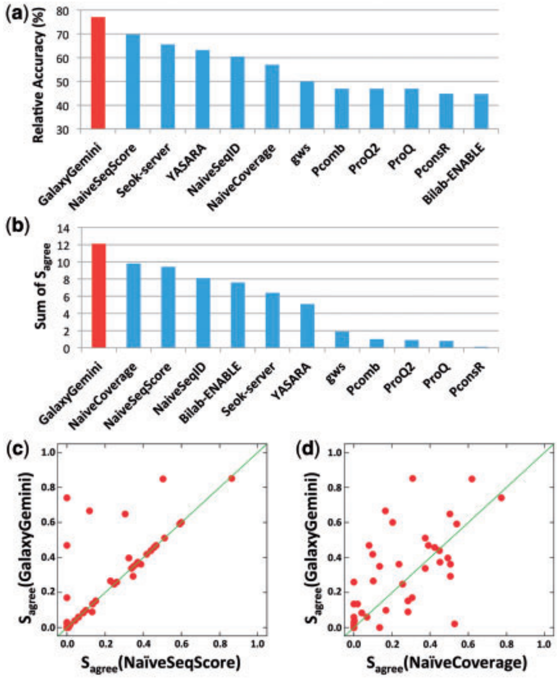

Fig. 1. Comparison of the performance of GalaxyGemini as measured by (**a**) relative accuracy and (**b**) sum of S_{agree} for the CASP9 set with those of CASP9 predictors and three naive methods, which take the HHsearch top ranker by sequence score (NaiveSeqScore), sequence identity (NaiveSeqID) and coverage (NaiveCoverage). Target-based comparison of *S*agree with naive predictors (**c**) NaiveSeqScore and (**d**) NaiveCoverage are also shown for the CASP9 set using the model structure obtained by GalaxyTBM (Ko *et al.*, 2012) as input

A successful example of CASP9 target T0576 (3na2) highlights the strength of GalaxyGemini (Supplementary Fig. S2). This protein forms a dimer through an inter-chain β-sheet. The best template determined by the NaïveSeqScore (2grg) is monomeric, but GalaxyGemini successfully selects a dimer template (3fm2), which has an oligomer structure similar to the native structure, resulting in a high S_{agree} of 0.742.

4 WEB SERVER

On the input page of GalaxyGemini, a user may enter e-mail address (optional) and upload a protein monomer structure in the PDB format. On completion of the job, a result page is generated and a link is sent via e-mail if the address is provided. The three best models can be downloaded, and information on the number of subunits and selected oligomer templates can be viewed on the result page. The average run time is ~30 min without energy minimization of the final models (default) and ~1 h with minimization.

5 CONCLUSIONS

A new web server GalaxyGemini predicts homo-oligomeric state of a protein from a monomer structure. The method outperforms other prediction methods tested in CASP9, implying wider applicability to oligomer state prediction from sequence.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) 2012-0001641 and 2012-0009480.

Conflict of Interest: none declared.

REFERENCES

Caffrey,D. *et al.* (2004) Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.*, **13**, 190–202.

Heldin,C.H. (1995) Dimerization of cell surface receptors in signal transduction. *Cell*, **80**, 213–223.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Ko,J. *et al.* (2012) GalaxyWEB server for protein structure prediction and refinement. *Nucleic Acids Res.*, **40**, W294–W297.

Krissinel,E. and Henrick,K. (2007) Inference of macro-molecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774–797.

Mariani,V. *et al.* (2011) Assessment of template based protein structure predictions in CASP9. *Proteins*, **79**, S37–S58.

Morita,M. *et al.* (2012) Blind prediction of quaternary structures of homo-oligomeric proteins from amino acid sequences based on templates. *J. Proteome Sci. Comput. Biol.*, **1**, 1.

Plückthun,A. and Pack,P. (1997) New protein engineering approaches to multivalent and bispecific antibody fragments. *Immunotechnology*, **3**, 83–105.

Ponstingl,H. *et al.* (2003) Automatic inference of protein quaternary structure from crystals. *J. Appl. Cryst.*, **36**, 1116–1122.

Poupon,A. and Janin,J. (2010) Analysis and prediction of protein quaternary structure. *Method. Mol. Biol.*, **609**, 349–364.

Snijder,H.J. *et al.* (1999) Structural evidence for dimerization-regulated activation of an integral membrane phospholipase. *Nature*, **401**, 717–721.

Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.