BIO DESIGN

pISSN 2288-6982 | eISSN 2288-7105

MINI REVIEW P 01-11

Prediction of Protein Structure and Interaction by GALAXY Protein Modeling Programs

Woong-Hee Shin, Gyu Rie Lee, Lim Heo, Hasup Lee and Chaok Seok*
Department of Chemistry, Seoul National University, Seoul 151-747, Korea. *Correspondence: chaok@snu.ac.kr

In this review, recently developed GALAXY protein modeling programs are introduced and advantages and disadvantages of these programs for both program users and method developers are discussed. The GALAXY package consists of the template-based modeling program GalaxyTBM, the loop/terminus modeling program GalaxyLoop, the model refinement program GalaxyRefine, the homo-oligomer prediction program GalaxyGemini, and the protein-ligand docking program GalaxyDock. These programs have been tested with some success in community-wide competition Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiments. For the development of these programs, modeling problems have been posed as global optimization problems of designed energy functions. The free energy functions of GALAXY have been carefully designed by combining physical chemistry principles and structure and sequence information. Efficient conformational search methods such as conformational space annealing and triaxial loop closure have been employed. Freely accessible web servers of the modeling programs are available at http://galaxy.seoklab.org, and some programs can be downloaded from http://galaxy.seoklab.org/softwares.

INTRODUCTION

The last century has witnessed remarkable advances in computational chemistry and computational biology, and computational approaches are increasingly and more widely used for studies of biological systems. For example, protein modeling techniques such as structure prediction and docking methods play important roles in revealing structure-function relationships of proteins (Baker and Sali, 2001; Shoichet, 2004). Modeling methods also serve as promising tools for the design of proteins or small molecules for sensors, therapeutic agents, or artificial enzymes (Kuhlman et al., 2003; Siegel at al., 2010). This is because computational methods are able to handle the large combinatorial space of possible mutations in proteins or possible chemical compounds at relatively low costs.

From a chemist's point of view, the problem of developing protein modeling methods is largely related to a sub-discipline of chemistry called 'physical chemistry' that describes thermodynamics and dynamics of molecules in terms of physical interactions of atoms. However, it is still not practical to predict protein structures and thermodynamic properties from amino acid sequences alone by using methods based on physical principles. This is because the two notorious problems in protein modeling, i.e., "scoring" of candidate structures to select the best model and "sampling" of candidate structures to be scored, still remain unsolved. In terms of the scoring problem, there are still doubts regarding whether the current physical force field models are sufficiently accurate for *ab initio* predictions (Raval et al, 2012). In terms of the sampling problem, sampling a large conformational space of proteins with >100 amino acid residues is still considered impractical (Bonneau and Baker, 2001).

In practice, bioinformatics has been one of the major driving forces in the field of protein modeling. Bioinformatics approaches that rely on analysis of biological information such as biological sequences and structures become more attractive because more sequence and structural data become available with technical advances. Indeed, template-based protein structure prediction methods that use the structure information of proteins with similar sequences in the structure database is currently the most accurate method for protein structure prediction, especially with the increasing amount of structure information. However, problems for which sufficient information is already available tend to be less interesting. Problems with some unknown aspects that are very different from those expected from available information attract more interest. Bioinformatics approaches may be less suitable for solving problems for which currently available information is not sufficient. Physical chemistry approaches may fill such gaps if adequately used in combination with bioinformatics approaches.

The GALAXY protein modeling programs have been developed with the idea of combining bioinformatics and physical chemistry approaches to improve existing methods for applications to protein modeling problems for which only some information is available. It is not practical to solve complex protein modeling

problems for which too little information is available. Problems for which a certain amount of information is missing are major targets of current GALAXY programs. For example, GALAXY can effectively address problems such as loop modeling for which information on the overall structure such as experimental structures of related proteins is available or ligand docking problems for which protein structure information and an approximate position of the binding pocket can be predicted.

Most GALAXY methods have been developed on the basis of a physical chemist's point of view, i.e., modeling problems are reduced to global optimization of the free energy function for the particular problem of interest. The sampling problem is addressed by using available global optimization techniques such as conformational space annealing, molecular dynamics simulations, Monte Carlo simulations, etc. The scoring problem is addressed by the development of the free energy function for each application such as loop modeling or ligand docking. Current free energy functions are composed of knowledge-based components derived from the database of available experimental structures and physics-based components taken from force field energy terms. Different levels of global optimization and different amounts of biological information are used depending on the problem at hand. In this review, the following GALAXY programs are introduced: the template-based modeling program GalaxyTBM (Ko et al., 2012a; Ko et al., 2012b), the protein loop/terminus modeling program GalaxyLoop (Park and Seok, 2012; Park et al., 2011), the protein model structure refinement program GalaxyRefine (Heo et al., 2013), the protein homo-oligomer structure prediction program GalaxyGemini (Lee et al., 2013), and the protein-ligand docking (PLD) program GalaxyDock (Shin et al., 2011; Shin and Seok, 2012; Shin et al., 2013). These programs can either be run on the GALAXY web server (http://galaxy.seoklab.org) or downloaded (http://galaxy.seoklab.org/softwares). Advantages and disadvantages of these methods compared to other state-of-the-art methods are also discussed.

PROTEIN STRUCTURE PREDICTION BY GALAXYTBM

Introduction to GalaxyTBM

Predicting protein structures from amino acid sequences is considered to be the most practical if experimental structures of similar proteins are available. Template-based modeling (TBM), also called homology modeling or comparative modeling, is a structure prediction method applied in such cases using similar proteins as templates. At present, the range of sequence similarity that can be reliably covered by TBM methods is continuously increasing, especially with the increasing amount of experimental information available for both protein structure and sequence. One of the current challenges of TBM in terms of method development is to effectively combine information from multiple template proteins if different structural regions of different proteins have useful information. Another challenge is to predict the structures of regions for which little information is available.

In GalaxyTBM (Ko et al., 2012a; Ko et al., 2012b), core regions of the target protein sequence that are highly conserved among related protein sequences are modeled from multiple template structures, and less reliable local regions are re-modeled by *ab initio* loop modeling or terminus modeling methods. The

TABLE 1 | Performance comparison of GalaxyTBM with MODELLER and SWISS-MODEL in terms of GDT-TS, GDC-SC, and MolProbity score

Target	GalaxyTBM			MODELLER[1]			SWISS-MODEL[2]		
	GDT-TS (%)	GDT-TS (%)	Mol-Probity	GDT-TS (%)	GDC-SC (%)	Mol-Probity	GDT-TS (%)	GDC-SC(%)	Mol-Probity
T0516	74.34	**35.11**	**1.72**	**74.45**	29.59	3.06	73.90	31.03	2.94
T0591	**76.05**	**36.16**	**2.39**	75.27	31.83	3.41	73.36	28.30	3.08
T0597	**76.18**	**40.64**	**2.28**	72.77	34.69	3.37	73.33	28.30	3.17
T0609[3]	**68.21**	**27.89**	**2.35**	67.76	26.62	3.81	-	-	-
T0641	72.37	**31.84**	**2.62**	71.10	29.18	3.66	**72.46**	32.34	3.05
T0650	**85.40**	**50.39**	**1.87**	85.33	44.03	3.17	56.56	45.15	3.01
T0652	**94.58**	**47.04**	**1.67**	92.77	42.44	2.93	87.95	32.99	2.54
T0658	**80.56**	**45.06**	**2.57**	80.43	39.99	3.55	16.41	10.46	3.40
T0682[3]	**79.81**	**42.30**	**2.17**	78.82	34.71	3.11	-	-	-
T0749	**91.20**	**57.45**	**2.08**	89.86	53.14	3.08	90.11	55.89	2.69
Average	**79.87**	**41.39**	**2.17**	78.86	36.62	3.31	-	-	-

[1] MODELLER version 9.11 was used with the same multiple sequence alignment as GalaxyTBM.
[2] SWISS-MODEL web server was used with the same template list as GalaxyTBM.
[3] Targets for which SWISS-MODEL web server failed to generate models.

TABLE 2 | Performance comparison of GalaxyTBM with MODELLER and SWISS-MODEL in terms of ligand-binding site RMSD

Target	GalaxyTBM	MODELLER[1)]	SWISS-MODEL[2)]
	Ligand-binding site RMSD (Å)		
T0516	**1.30**	1.31	1.30
T0591	**1.68**	1.64	1.90
T0597	**1.25**	1.43	1.52
T0609[3)]	**2.00**	2.14	-
T0641	**1.26**	1.31	1.31
Average	**1.45**	1.57	-

[1)] MODELLER version 9.11 was used with the same multiple sequence alignment as GalaxyTBM.

[2)] SWISS-MODEL web server was used with the same template list as GalaxyTBM.

[3)] Targets for which SWISS-MODEL web server failed to generate models.

GalaxyTBM program and its web server (http://galaxy.seoklab.org/tbm) are based on the TBM method developed during the 9th Critical Assessment of Techniques for Protein Structure Prediction (CASP9) experiment. This method was assessed to be amongst the top TBM servers participated in CASP9 (Mariani et al., 2011). In the current GalaxyTBM web server, up to three unreliable regions are detected and modeled automatically. Additional loop or terminus modeling can be performed on a separate web server for loop/terminus modeling (http://galaxy.seoklab.org/loop) if modeling is desired for a larger number of loops/termini or if it is expected that more intensive optimization would be helpful. The GalaxyLoop loop/terminus modeling server will be described in more detail in the next section.

Performance of GalaxyTBM

Performance of GalaxyTBM is compared with two other freely available TBM methods, MODELLER (Sali and Blundell, 1993) and SWISS-MODEL (Arnold et al., 2006) for ten CASP9 structure prediction targets. In Table 1, prediction accuracy is compared in terms of the following three accuracy measures that are commonly used in CASP: (1) GDT-TS (Zemla, 2003) for global backbone structure accuracy, (2) GDC-SC (Keedy et al., 2009) for local side chain structure accuracy, and (3) MolProbity score (Chen et al., 2010) for physical correctness. In Table 2, RMSD of ligand-binding site is compared for five ligand-binding proteins. Overall, GalaxyTBM performs better for most target proteins, especially in terms of local structure accuracy and physical correctness. A successful example is also illustrated in Figure 1 for which GalaxyTBM produces a model with correct structural details when compared to MODELLER or SWISS-MODEL.

Currently, GalaxyTBM performs relatively well on single-domain protein targets compared to multi-domain proteins. It is recommended to use GalaxyTBM with separate domain-splitting methods if structure prediction of multi-domain proteins is desired. A better domain splitting method for GalaxyTBM is currently under development.

The GalaxyTBM Method

GalaxyTBM follows the four stages of typical template-based modeling methods (Zhang, 2008, Marti-Renom et al., 2000): (1) identification of similar proteins in the structure database to be used as structural templates; (2) alignment of template sequences to the target sequence; (3) model building from the sequence alignment; and (4) model refinement. For template identification, GalaxyTBM rescores the results of HHsearch (Söding, 2005) to select multiple templates for reliable core structures. Alignment of core regions of multiple sequences is achieved using PROMALS3D (Pei et al., 2008). Models are built from alignment and template structures by optimizing a free energy function that combines template-derived restraint terms with force field energy and solvation free energy terms (to be published). Final model structures are generated by detecting and re-modeling unreliable

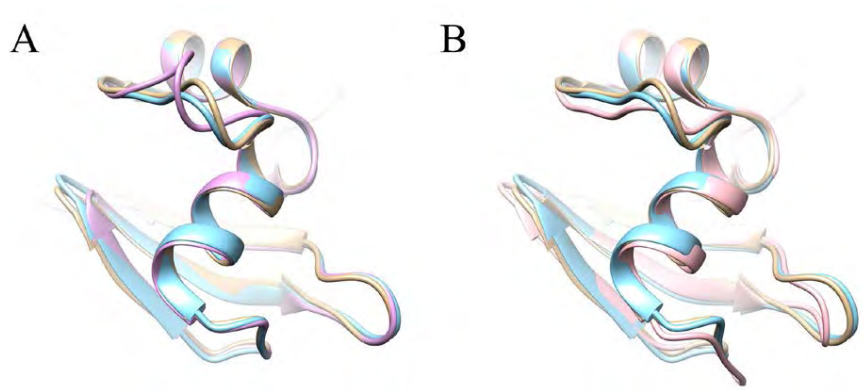

FIGURE 1 | Comparison of structure prediction results for the CASP9 target T0652. The crystal structure is shown in gold, the predicted structure by GalaxyTBM in sky blue, and those by MODELLER (A) and SWISS-MODEL (B) in pink.

loop or terminus regions with optimization modules in GALAXY (Park and Seok, 2012; Park et al., 2011). A schematic flowchart of the overall GalaxyTBM process can be found in Figure 5 of Ko et al., 2012b.

PROTEIN LOOP MODELING BY GALAXYLOOP

Introduction to GalaxyLoop

Protein loops often refer to regions of irregular structures that connect secondary structure segments such as alpha helices or beta sheets. In protein modeling, loops can also be defined as relatively short sequence regions that are poorly conserved among evolutionarily related proteins. Protein loops are often involved in various biochemical functions such as ligand binding, enzymatic action, and signal transduction (Decanniere et al., 1999; Fiser et al., 2000; Saraste et al., 1990). Therefore, atomic-level structures of protein loops can be the key to understanding protein functions and designing proteins with particular functions. However, loop structures are often difficult to resolve by experimental methods because of structural flexibility and variability. Therefore, GalaxyLoop can be applied to address various problems such as predicting missing or low-resolution loop regions in experimental structures, predicting alternative activation states, designing loop regions for binding to target molecules, or refining model structures obtained by template-based modeling (Amaro et al., 2007; DiMaio et al., 2011; Mas et al., 1992).

Several loop modeling methods that are based on informatics-based methods are available as web servers because they are relatively computationally inexpensive. However, *ab initio* loop modeling is required when similar loop structures cannot be found in the available structure database. *Ab initio* loop modeling methods tend to be computationally expensive, but GalaxyLoop is sufficiently computationally efficient to be developed as a web server and still shows high performance compared to other state-of-the-art *ab initio* loop modeling programs (Jacobson et al., 2004; Mandell et al., 2009; Rohl et al., 2004; Wang et al., 2007). It is freely accessible at http://galaxy.seoklab.org/loop.

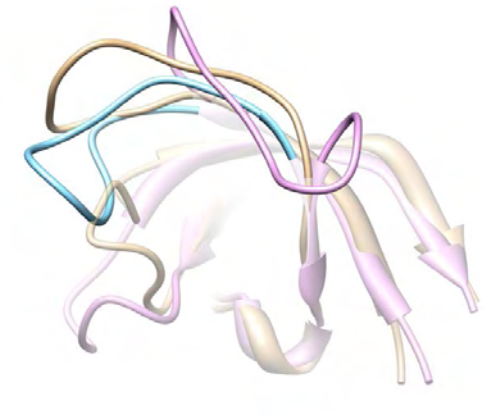

FIGURE 2 | One of the successful loop modeling examples in the CASP9 blind prediction experiment. The initial model for target T0572-D1 (shown in pink) was improved by applying GalaxyLoop to the loop region 41–53 (shown in sky blue). The experimental structure (PDB ID: 2kxy) that was released after the experiment is shown in gold. The overall backbone root-mean-square deviation (RMSD) was reduced from 3.7 Å to 2.9 Å by loop modeling.

TABLE 3 | CASP9 blind prediction results of GalaxyLoop for loops or termini that are successfully detected from initial template-based models

Number of cases (number of loops/termini)			
	Modeled	Improved in G-RMSD[1)]	Improved in L-RMSD[2)]
≤12 residues	87 (80/7)	53 (47/6)	58 (53/5)
>12 residues	47 (39/8)	30 (25/5)	32 (27/5)
Total	134 (119/15)	83 (72/11)	90 (80/10)

[1)] Global root-mean-square deviation (RMSD), calculated after best superposition of the overall structure onto the experimental structure.

[2)] Local RMSD, calculated after best superposition of the loops or termini onto the corresponding region in the experimental structure.

Performance of GalaxyLoop

The GalaxyLoop program was employed for model refinement in CASP9 and CASP10 and has been shown to improve the quality of template-based models. Overall results of applying GalaxyLoop to loop/terminus refinement of CASP9 targets are summarized in Table 3, and an example of successful refinement is illustrated in Figure 2. It has often been noted that there is a limit to the length of loops that can be reliably modeled; it is usually difficult to accurately predict loop structures that are longer than 12 residues. According to the results shown in Table 3, it is still possible to improve the model accuracy for loops or termini that are longer than 12 residues.

The GalaxyLoop Method

GalaxyLoop (Park and Seok, 2012), an *ab initio* protein loop modeling program, employs a global energy optimization technique called conformational space annealing (CSA) (Lee et al., 1999). A schematic diagram that explains the loop/terminus modeling algorithm is in Figure 2 of Park and Seok, 2012, and only a brief description of the algorithm is provided here. More detailed description on the method can be found in Park et al., 2011 and Park and Seok, 2012.

The initial pool of loop conformations is generated by a method called fragment assembly and analytical loop closure (FALC) (Lee et al., 2010). Protein termini are modeled in a similar way but without applying the loop closure algorithm (Park et al., 2011). The CSA algorithm efficiently searches for the loop conformation with the lowest energy by perturbing the pool of conformations iteratively and gradually focusing on narrower regions of lower energy in the conformational space. CSA still depends on the

TABLE 4 | Model refinement test results for the mild relaxation method of GalaxyRefine in terms of improvements in backbone structure quality measured by the high-accuracy global distance test (GDT-HA), side-chain structure quality measured by the side-chain global distance test (GDT-SC), and physical correctness measured by the MolProbity score. Shown in parentheses are the results for the best models among the five models generated with both mild and aggressive relaxation methods of GalaxyRefine.

Test set	No. targets	Average improvement/ Percentage of improved targets		
		GDT-HA (%)	GDC-SC (%)	MolProbity
CASP refinement category targets[1)]	53	0.38/59 (1.45/74)	1.50/64 (2.21/68)	0.74/82[4)] (1.26/90[4)])
TBM server models[2)]	153[3)]	0.43/65 (1.37/76)	1.69/77 (2.54/83)	0.37/66 (0.55/74)
FG-MD benchmark set	147	0.61/65 (1.80/80)	1.74/75 (2.78/87)	0.89/100 (1.18/100)

[1)] CASP refinement category targets: 12 CASP8 targets, 14 CASP9 targets, and 27 CASP10 targets

[2)] Zhang-server (I-TASSER) and ROSETTA-BAKER server models for the CASP10 tertiary structure prediction targets

[3)] Non-oligomeric targets with TM-score>0.5 and no severe crystallographic contacts

[4)] Target TR476 has no side-chain coordinates in the initial structure; therefore, it was not included in the MolProbity analysis.

Abbreviations: CASP, Critical Assessment of Techniques for Protein Structure Prediction; TBM, template-based model; FG-MD, fragment-guided molecular dynamics.

quality of initial structures, and the initial structures are generated using FALC that effectively reduces the large loop conformational space using information from short protein structure fragments collected from the structure database. The loop structures generated by assembling fragment structures are adjusted with regard to their backbone torsion angles by employing an analytical loop closure algorithm to obtain geometric consistency with respect to the framework structure (Coutsias et al., 2004).

The objective function for global optimization is a hybrid energy function derived from both physical chemistry principles and information available in the structure database. The physics-based part consists of the molecular mechanics force field energy terms (MacKerell et al., 1998) and an implicit solvation free energy model (Lazaridis and Karplus, 1999). The informatics-based part, also called knowledge-based potential, includes dipolar-DFIRE (Yang and Zhou, 2008), side-chain rotamer energy of MODELLER (Sali and Blundell, 1993), and an orientation-dependent hydrogen bonding energy (Kortemme et al., 2003).

Frequently used energy functions in previous *ab initio* loop modeling programs consisted of physics-based terms only (Felts et al., 2008; Jacobson et al., 2004; Soto et al., 2008). These methods have been tested on loop reconstruction problems in which loop structures were modeled in frameworks of experimental structures with deleted loop regions. GalaxyLoop has been designed to perform well not only in the reconstruction of crystal loop structures but also in modeling loops in inaccurate framework structures such as template-based models. This was achieved by combining knowledge-based terms to the free energy function and by training parameters in situations that involve inaccurate frameworks of template-based models. However, there is still an upper limit to environmental errors that can be tolerated by GalaxyLoop, and an enhanced version that can further optimize the surrounding environment is necessary. Another limitation of the current loop modeling method is that a water solvation model is used and the method can, therefore, not be successfully applied to loops of membrane proteins. These problems are considered in on-going developments.

PROTEIN MODEL REFINEMENT BY GALAXYREFINE

Introduction to GalaxyRefine

Accurate protein structure prediction is considered possible by using template-based modeling techniques (Kryshtafovych et al., 2011; Marti-Renom et al., 2000) for proteins with high sequence identity (>30%). However, predicted protein structures may not be sufficiently accurate for further applications such as protein design, structure-based drug design, or protein function studies (MacCallum et al., 2009; MacCallum et al., 2011) if the sequence identity to known experimental structures is lower. The importance of improving the accuracy of model structures achieved by template-based modeling has been recognized by the CASP community, and the category "structure refinement" has been introduced since CASP8 (2008).

Various protein model refinement methods have been reported in CASP (MacCallum et al., 2009; MacCallum et al., 2011), including molecular dynamics simulations, fragment-guided methods, knowledge-based methods, elastic network models, and hydrogen bond network optimization (Bhattacharya and Cheng, 2013; Park and Seok, 2012; Raman et al., 2009; Rodrigues et al., 2012; Xu et al., 2011; Zhang et al., 2011). Interestingly, only a few methods succeeded in improving the initial structures, and most methods failed to achieve consistent improvements. This is because the CASP refinement category was introduced to solve very difficult refinement problems, i.e., refining one of the best model structures that may have been already refined (MacCallum et al., 2011).

GalaxyRefine (Heo et al., 2013) refines the overall structure, while GalaxyLoop refines loop or terminus structures only. If one needs to refine a protein structure model, it is recommended to use GalaxyLoop first to refine loops/termini and to use

GalaxyRefine next to refine the whole structure. GalaxyRefine in combination with GalaxyLoop was tested in the refinement category of CASP10 in a blind fashion and has been assessed to be one of the few methods that showed consistent improvements over initial models both in global and local structural qualities (Nugent et al., 2014). The GalaxyRefine web server, freely accessible at http://galaxy.seoklab.org/refine, is among the best available web servers for structure refinement.

Performance of GalaxyRefine

The GalaxyRefine method has been tested on various protein structure models: CASP refinement category targets (MacCallum et al., 2009; MacCallum et al., 2011; Nugent et al., 2014), protein structure models from other state-of-the-art template-based modeling servers (I-TASSER and ROSETTA) (Leaver-Fay et al., 2011; Xu et al., 2011), and a protein model refinement benchmark set used in another refinement test (Zhang et al., 2011). The test results are summarized in Table 4 in terms of the percent improvement over initial models for backbone structure accuracy measured by the high-accuracy global distance test GDT-HA (Zemla, 2003), for local structure accuracy measured by the side-chain global distance test GDC-SC (Keedy et al., 2009), and for physical correctness measured by the MolProbity score (Chen et al., 2010).

From the benchmark test results, it is expected that GalaxyRefine has a high probability (>50%) of improving the initial models even for model structures generated by the best available structure prediction servers, e.g., I-TASSER and ROSETTA. An example of successful refinement by GalaxyRefine is illustrated in Figure 3. It has also been observed that GalaxyRefine can substantially improve model structures for protein-protein complexes (unpublished results). Although the structure quality is consistently improved with the current version of GalaxyRefine, the magnitude of improvement is still relatively small; thus, further development is currently underway.

The GalaxyRefine Method

GalaxyRefine samples the conformational space by repetitive perturbations and molecular dynamics relaxations. Restraints are applied to remain close to the initial structures. One refined structure is generated by a mild relaxation method that perturbs the structure of a cluster of side chains, and four structures are generated by an aggressive relaxation method that perturbs secondary structure segments and loops. Repetitive perturbation and relaxation enhances structural packing, thereby improving both backbone and side-chain structure qualities. The initial placement of side-chain structures also improves the overall structure quality. The initial side-chain structures are replaced with rotamers with the highest probability in the Dunbrack rotamer library (Dunbrack, 2002) if they do not cause steric clashes and do not deviate from the canonical distribution in the degree of solvent exposure.

The free energy function used for the relaxation methods resembles the energy function used for loop modeling, being composed of physics-based energy terms and database-derived terms. Additional harmonic restraint terms derived from the initial structure are added. The relative weight of the restraint terms to the other terms for aggressive relaxation is set to five times smaller than that for mild relaxation to sample a broader conformational space. The physics-based energy functions are based on the CHARMM22 force field (MacKerell et al., 1998), which contains molecular-mechanics bonded energy terms, Lennard-Jones interaction energy, Coulomb potential energy, free energy of solvation (also called "fast analytical continuum treatment of solvation" [FACTS]), and the solvent-accessible surface area term (Haberthur and Caflisch, 2008). Hydrogen bond energy (Kortemme et al., 2003), dipolar-DFIRE (Yang and Zhou, 2008), and backbone-dependent torsion angle energy of side chains (Canutescu et al., 2003) are used for database-derived scoring functions. A more detailed description on the GalaxyRefine algorithm can be found in Heo et al., 2013.

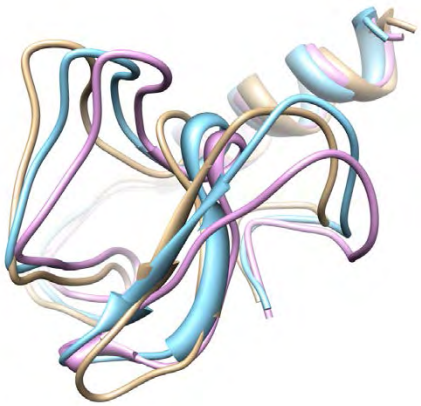

FIGURE 3 | A successful model refinement example of GalaxyRefine for target TR754 (PDB ID: 2lv9), one of the targets in the refinement category of the CASP10 experiment. This prediction was also performed in a blind fashion before the release of the PDB structure. The initial structure shown in pink was refined to the structure shown in sky blue, which is closer to the crystal structure shown in gold. The score of the high-accuracy global distance test (GDT-HA) was improved from 58.1% to 62.1%, the score of the side-chain global distance test (GDC-SC) was improved from 20.0% to 25.4%, and the MolProbity score was improved from 2.56 to 2.38.

PROTEIN HOMO-OLIGOMER STRUCTURE PREDICTION BY GALAXYGEMINI

Introduction to GalaxyGemini

Many proteins form homo-oligomers to perform their functions (Poupon and Janin, 2010). Examples include antibodies (Plückthun and Pack, 1997) and membrane proteins (Heldin, 1995). The GalaxyGemini (Lee et al., 2013) web server (http://galaxy.seoklab.org/gemini) predicts the homo-oligomer structure from an input protein structure. The structure can be either an experimental structure or a model structure; thus, homo-

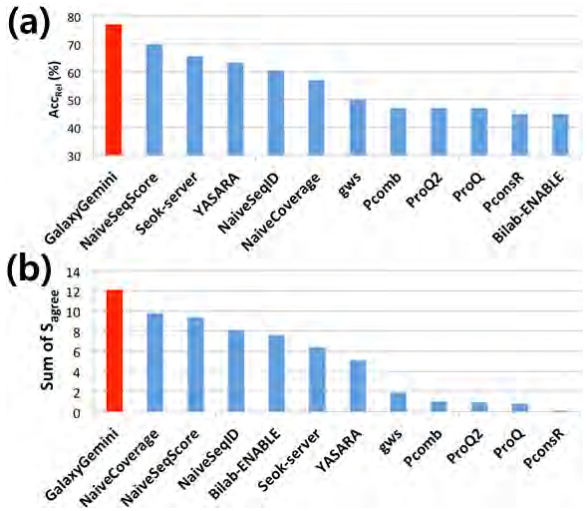

FIGURE 4 | Comparison of relative accuracy (Acc_{Rel}) in the number of subunit prediction and the sum of contact agreement score (S_{agree}) at the oligomer interface obtained by GalaxyGemini with those predicted by CASP9 predictors and 3 naïve predictors which take the HHsearch top ranker by sequence score (NaïveSeqScore), sequence identity (NaïveSeqID), or coverage (NaïveCoverage).

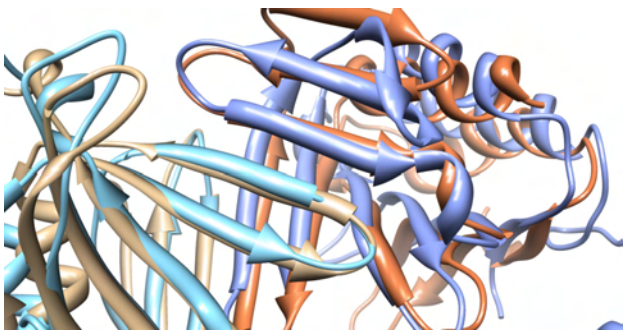

FIGURE 5 | A successful example (PDB ID: 3na2) in which GalaxyGemini correctly predicts a dimer structure with inter-chain β-sheet, while the best template by HHsearch search predicts a monomer structure. Subunits of the crystal structure are shown in gold and orange, and those of the predicted structure in sky blue and cornflower blue.

oligomer structure prediction from a protein sequence is also possible when a structure prediction method is used to generate a model structure. GalaxyGemini identifies oligomer structure templates from the structure database on the basis of sequence and tertiary/quaternary structure similarity. Unlike other programs introduced in this review, this method does not use extensive optimization methods. However, it is expected that the method can be improved by combination with optimization methods.

Performance of GalaxyGemini

GalaxyGemini was tested on CASP9 targets (96 proteins containing 43 monomers; Mariani et al., 2011) with the oligomer structure database built before CASP9 experiment (Lee et al., 2013). Prediction accuracy was measured by the following two measures that were used in the CASP9 assessment: "relative accuracy" (Acc_{Rel}) which is the percentage of the targets for which the number of subunits are correctly predicted and "contact agreement score" (S_{agree}) which reflects the fraction of correctly modeled interface contacts in the complex. As shown in Figure 4, GalaxyGemini outperforms all other CASP9 predictors and 3 naïve predictors which take the HHsearch (Söding, 2005) top ranker by sequence score (NaïveSeqScore), sequence identity (NaïveSeqID), or coverage (NaïveCoverage). An interesting example is illustrated in Figure 5 for which the template determined by the HHsearch sequence score (2grg) is monomeric, but GalaxyGemini successfully selects a dimer template (PDB ID: 3fm2).

The GalaxyGemini method

GalaxyGemini uses an oligomer database with mutual sequence identity <70% from all structures deposited in the protein data bank (PDB) with oligomeric states assigned by authors or by the protein interfaces, surfaces, and assembly (PISA) service (Krissinel and Henrick, 2007). For a given input protein, oligomer templates are selected from the oligomer database by rescoring the results of HHsearch (Söding, 2005). It is first determined whether the query protein is an oligomer or not; then, oligomer templates are selected using rescoring functions. The rescoring functions contain terms that measure sequence similarity and tertiary and quaternary structure similarity at the protein interface. The relative weights of the different terms were trained on the PISA benchmark set (195 proteins containing 55 monomers; Ponstingl et al., 2003). The final oligomer structure generated by superimposition onto the template structure is relaxed by energy minimization to remove steric clashes. More detailed explanation on the algorithm can be found in the supplementary information of Lee et al., 2013.

PROTEIN-LIGAND DOCKING BY GALAXYDOCK

Introduction to GalaxyDock

Protein-ligand docking (PLD) is a biomolecular modeling technique that predicts interactions between a receptor protein and a small ligand molecule. By predicting the binding affinity and the binding pose of protein-ligand complexes, PLD can be applied to virtual screening of compound databases or to optimization of hit compounds in structure-based drug discovery processes (Sousa et al., 2013; Lavecchia et al., 2006; Rogers et al., 2006; Venkatesan et al., 2010). PLD can also be applied to functional studies of various proteins.

An important issue in the development of PLD programs is how to consider receptor flexibility (Bonvin, 2006; Teague, 2003). The first-generation docking programs developed in 1980s treated the receptor as a rigid molecule to simplify the problem

and to save computation time. However, it is obvious that receptor flexibility has to be considered in many cases to predict the binding pose and affinity accurately (Teague, 2003). Over the past years, a number of docking approaches have been proposed to consider receptor conformational changes (Carlson, 2002; Cavasotto et al., 2005; Meireles et al., 2011; Claußen et al. 2001; Bottegoni et al., 2009; Bottegoni et al., 2008).

In this regard, development of Galaxy Dock (Shin et al., 2011; Shin and Seok, 2012; Shin et al., 2013) has started with a long-term goal of incorporating the recent developments in GALAXY protein modeling programs such as protein loop modeling and model structure refinement to consider receptor flexibility accompanied by ligand binding. A powerful, but computationally expensive global optimization method, conformational space annealing (CSA) (Lee et al., 1997), has been employed with the vision to include additional degrees of freedom that describe receptor flexibility because CSA is one of the most effective methods for difficult optimization problems that involve large numbers of degrees of freedom. The current version of GalaxyDock can handle rigid-receptor docking and flexible-receptor docking, but flexible-receptor docking is limited to the flexibility of a small number (<5) of selected side chains. The GalaxyDock program can be downloaded free of charge from http://galaxy.seoklab.org/softwares/galaxydock.html.

TABLE 5 | Comparison of binding pose prediction results of GalaxyDock with other docking programs in terms of success rate (percentage of cases in which binding poses are predicted within 2 Å) and average root-mean-square deviation (RMSD) of predicted binding poses

Rigid-receptor docking: ASTEX diverse set		
Program	Success Rate (%)	Average RMSD (Å)
GalaxyDock	85.9	1.24
AutoDock	81.7	1.60
GOLD	80.5	N/A
Surflex	80.0	1.66
Flexible-receptor docking: diverse set		
Program	Success Rate (%)	Average RMSD (Å)
GalaxyDock	83.3	1.68
SCARE	80.0	N/A
GalaxyDock (rigid receptor)	46.7	2.40
Flexible-receptor docking: LXRβ set		
Program	Success Rate (%)	Average RMSD (Å)
GalaxyDock	88.9	1.54
RossetaLigand	55.6	1.91
GalaxyDock (rigid receptor)	55.6	1.97

Performance of GalaxyDock

When the rigid-receptor mode of GalaxyDock was compared to AutoDock, GOLD, and Surflex on the 85 complexes of ASTEX diverse set, it showed the highest success rate (percentage of the predictions with RMSD < 2 Å from the crystal ligand poses) and the lowest average RMSD of ligand poses, as shown in Table 5. The flexible-receptor mode of GalaxyDock was tested on different sets (diverse set and LXRβ set, see Shin and Seok, 2012 for details) for which docking results of other programs are available in the literature. Flexible docking is usually tested by cross-docking a ligand to a protein structure in an unbound form or bound to another ligand. Therefore, successful cross-docking may require conformational change of receptor. The two test sets for flexible-receptor docking in Table 5 involve side chain conformational change of receptor. According to the table, the flexible mode of GalaxyDock shows superior results when compared to other flexible docking programs such as SCARE and RosettaLigand. A successful example of flexible docking is illustrated in Figure 6, where three out of four flexible side chains are predicted to be in the correct rotamer states with ligand RMSD of 0.5 Å. In this prediction, 87% of hydrophobic interactions and three out of four hydrogen bonds between the receptor and ligand found in the native binding mode are reproduced.

However, in our experience, the use of this program can be problematic if the binding pocket is too small compared to the ligand size, which may happen when poor receptor structure models are used. Such problem can be solved if the receptor is allowed to relax or weak steric clashes are allowed in the docking algorithm. The relatively high computational cost of GalaxyDock compared to other docking programs can also be a disadvantage if virtual screening of a large compound library is desired. The current goals in the GalaxyDock project are to improve the accuracy of binding affinity estimation and to consider larger degrees of protein flexibility, e.g., loop flexibility, that are frequently observed in kinase families.

The GalaxyDock method

The free energy function of GalaxyDock is based on that of AutoDock3 (Morris et al., 1998), which is a force field-based empirical scoring function. The AutoDock3 scoring function resembles the energy functions used in protein modeling methods in the functional form. Thus, it can be expected that GalaxyDock would be more compatible with other GALAXY protein modeling methods if they are combined in the future.

During the development of the first version of GalaxyDock, also called LigDockCSA (Shin et al., 2011), a problem in the AutoDock3 score was identified, and the ligand internal torsion energy of the piecewise linear potential (PLP) (Gehlhaar et al.,

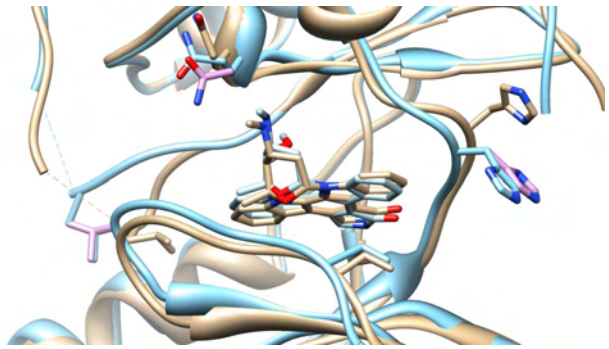

FIGURE 6 | A successful example of flexible-receptor docking, where the ligand of 1aq1 is docked to the receptor structure of 1dm2 with four flexible side chains. The native pose is colored in gold and the predicted pose in sky blue. The initial receptor structure (1dm2) is shown in pink. The ligand RMSD is 0.5 Å. In this prediction, 87% of hydrophobic interactions and three out of four hydrogen bonds between receptor and ligand in the native binding mode are reproduced.

1995) was added to solve this problem. With this improvement in the docking scoring function and the efficient CSA global optimization, LigDockCSA showed improved binding pose prediction results compared to other docking methods, as can be seen from the rigid-receptor docking results summarized in Table 5.

During the development of the second version of GalaxyDock (Shin and Seok, 2012), which incorporates flexibility of pre-specified side chains, additional terms had to be added to describe the conformational changes of the receptor appropriately; a knowledge-based potential ROTA score (Hartmann et al., 2007) was implemented for this purpose. The side chain conformations were sampled from the Dunbrack rotamer library (Dunbrack, 2002). GalaxyDock was further improved by using a more efficient method for generating the initial conformations for CSA optimization (Shin et al., 2013). A fast, geometry-based pre-docking method that represents the receptor surface by a beta-complex (Kim et al., 2010) generated from the Voronoi diagram of receptor atoms was employed. Initial conformations of higher quality could be obtained with this method compared to the previous version that generated conformations in a random fashion. The overall procedure of the most recent version of GalaxyDock can be seen in Figure 1 of Shin et al., 2013. A comparison of binding pose prediction results with one of the best flexible-side-chain docking methods called SCARE (Bottegoni et al., 2008) is presented in Table 5.

The GalaxyDock binding affinity prediction has also been improved by "free ligand correction" (Shin et al., 2013). With this improved binding affinity function, GalaxyDock showed improved performance over AutoDock and DOCK in terms of the enrichment factor (Shin et al., 2013) in a virtual screening test on the Gilson set (Jorissen and Gilson, 2005). However, the current binding affinity estimation does not rank high when compared with other popular scoring functions in terms of correlation coefficients between predicted and experimental binding affinities (Shin et al., 2013). The binding affinity prediction method is currently being further improved.

CONCLUSIONS

In this review, we introduced the GALAXY protein modeling programs and discussed how they can be used for various application problems. Although these programs perform superior or comparable to other state-of-the-art methods in the protein modeling field, there is still ample room for improvement. It is anticipated that this review facilitates communications between method developers and potential user groups as well as developers from other disciplines for the development of more promising next-generation modeling methods.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT, and Future Planning (No. 2013R1A2A1A09012229).

Original Submission: February 17, 2014
Revised Version Received: March 10, 2014
Accepted: March 12, 2014

REFERENCES

Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195-201.

Amaro, R.E., Minh, D.D., Cheng, L.S., Lindstrom, W.M., Jr., Olson, A.J., Lin, J.H., Li, W.W., and McCammon, J.A. (2007). Remarkable loop flexibility in avian influenza N1 and its implications for antiviral drug design. *J Am Chem Soc* **129**, 7764–7765.

Baker, D. and Sali, A. (2001). Protein Structure Prediction and Structural Genomics. *Science* **294**, 93–96.

Bhattacharya, D., and Cheng, J. (2013). 3Drefine: consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization. *Proteins* **81**, 119–131.

Bonneau, R. and Baker, D. (2001). Ab Initio Protein Structure Prediction: Progress and Prospects. *Ann Rev Biophys Biomol Struct* **30**, 173–189.

Bonvin, A. (2006). Flexible protein-protein docking. *Curr Opin Struct Biol* **16**, 194–200.

Bottegoni, G., Kufareva, I., Totrov, M., and Abagyan, R. (2008). A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE). *J Comput Aided Mol Des* **22**, 311–325.

Bottegoni, G., Kufareva, I., Totrov, M., and Abagyan, R. (2009) Four-Dimensional Docking: A Fast and Accurate Account of Discrete Receptor Flexibility in Ligand Docking. *J Med Chem* **52**, 397–406.

Canutescu, A.A., Shelenkov, A.A., and Dunbrack, R.L., Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* **12**, 2001–2014.

Carlson, H.A. (2002). Protein flexibility and drug design: how to hit a moving target. *Curr Opin Chem Biol* **6**, 447–452.

Cavasotto, C.N., Kovacs, J.A., and Abagyan, R.A. (2005). Representing Receptor Flexibility in Ligand Docking through Relevant Normal Modes. *J Am Chem Soc* **127**, 9632–9640.

Claußen, H., Buning, C., Rarey, M., and Lengauer, T. (2001). FlexE: Efficient Molecular Docking Considering Protein Structure Variations. *J Mol Biol* **308**, 377–395.

Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular

crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12–21.

Coutsias, E.A., Seok, C., Jacobson, M.P., and Dill, K.A. (2004). A kinematic view of loop closure. *J Comput Chem* **25**, 510–528.

Decanniere, K., Desmyter, A., Lauwereys, M., Ghahroudi, M.A., Muyldermans, S., and Wyns, L. (1999). A single-domain antibody fragment in complex with RNase A: non-canonical loop structures and nanomolar affinity using two CDR loops. *Structure* **7**, 361–370.

DiMaio, F., Terwilliger, T.C., Read, R.J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H.L., et al. (2011). Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature* **473**, 540–543.

Dunbrack, R.L., Jr. (2002). Rotamer libraries in the 21st century. *Curr Opin Struct Biol* **12**, 431–440.

Felts, A.K., Gallicchio, E., Chekmarev, D., Paris, K.A., Friesner, R.A., and Levy, R.M. (2008). Prediction of protein loop conformations using the AGBNP implicit solvent model and torsion angle sampling. *J Chem Theory Comput* **4**, 855–868.

Fiser, A., Do, R.K., and Sali, A. (2000). Modeling of loops in protein structures. *Protein Sci* **9**, 1753–1773.

Gehlhaar, D.K., Verkhivker, G.M., Rejto, P.A., Sherman, C.J., Fogel, D.R., Fogel, L.J., and Freer, S.T. (1995). Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem Biology* **2**, 317–324.

Haberthür, U., and Caflisch, A. (2008). FACTS: Fast analytical continuum treatment of solvation. *J Comput Chem* **29**, 701–715.

Hartmann, C., Antes, I., and Lengauer, T. (2007) IRECS: A new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. *Protein Sci* **16**, 1294–1307.

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

Heo, L., Park, H., and Seok, C. (2013). GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic Acids Res* **41**, W384–W388.

Jacobson, M.P., Pincus, D.L., Rapp, C.S., Day, T.J., Honig, B., Shaw, D.E., and Friesner, R.A. (2004). A hierarchical approach to all-atom protein loop prediction. *Proteins* **55**, 351–367.

Jorissen, R.N., and Gilson, M.K. (2005). Virtual Screening of Molecular Databases Using a Support Vector Machine. *J Chem Inf Model* **45**, 549–561.

Keedy, D.A., Williams, C.J., Headd, J.J., Arendall, W.B., 3rd, Chen, V.B., Kapral, G.J., Gillespie, R.A., Block, J.N., Zemla, A., Richardson, D.C., et al. (2009). The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. *Proteins* **77 Suppl 9**, 29–49.

Kim, D.-S., Cho, Y., Sugihara, K., Ryu, J., and Kim, D. (2010). Three-dimensional beta-shapes and beta-complexes via quasi-triangulation. *Computer-Aided Design* **42**, 911–929.

Ko, J., Park, H., Heo, L., and Seok, C (2012a). GalaxyWEB server for protein structure prediction and refinement. *Nucleic Acids Res* **40**, W294–W297.

Ko, J., Park, H., and Seok, C. (2012b). GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions. *BMC Bioinformatics* **13**, 198.

Kortemme, T., Morozov, A.V., and Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol* **326**, 1239–1259.

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

Kryshtafovych, A., Fidelis, K., and Moult, J. (2011). CASP9 results compared to those of previous CASP experiments. *Proteins* **79 Suppl 10**, 196–207.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., and Baker, D. (2003). Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* **302**, 1364–1368.

Lavecchia, A., Cosconati, S., Limongelli, V., and Novellino, E. (2006) Modeling of Cdc25B Dual Specificity Protein Phosphatase Inhibitors: Docking of Ligands and Enzymatic Inhibition Mechanism. *ChemMedChem* **1**, 540–550.

Lazaridis, T., and Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins* **35**, 133–152.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W., et al. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **487**, 545–574.

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.

Lee, J., Lee, D., Park, H., Coutsias, E.A., and Seok, C. (2010). Protein loop modeling by using fragment assembly and analytical loop closure. *Proteins* **78**, 3428–3436.

Lee, J., Liwo, A., and Scheraga, H.A. (1999). Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: Application to the 10-55 fragment of staphylococcal protein A and to apo calbindin D9K. *Proc Natl Acad Sci U S A* **96**, 2025–2030.

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.

MacCallum, J.L., Hua, L., Schnieders, M.J., Pande, V.S., Jacobson, M.P., and Dill, K.A. (2009). Assessment of the protein-structure refinement category in CASP8. *Proteins* **77 Suppl 9**, 66–80.

MacCallum, J.L., Perez, A., Schnieders, M.J., Hua, L., Jacobson, M.P., and Dill, K.A. (2011). Assessment of protein structure refinement in CASP9. *Proteins* **79 Suppl 10**, 74–90.

MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* **102**, 3586–3616.

Mandell, D.J., Coutsias, E.A., and Kortemme, T. (2009). Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Methods* **6**, 551–552.

Mariani, V., Kiefer, F., Schmidt, T., Haas, J., and Schwede, T. (2011). Assessment of template based protein structure predictions in CASP9. *Proteins* **79**, 37–58.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* **29**, 291–325.

Mas, M.T., Smith, K.C., Yarmush, D.L., Aisaka, K., and Fine, R.M. (1992). Modeling the anti-CEA antibody combining site by homology and conformational search. *Proteins* **14**, 483–498.

28 Meireles, L., Gur, M., Bakan, A., and Bahar, I. (2011). Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins. *Protein Sci* **20**, 1658.

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.

Nugent, T., Cozzetto, D., and Jones, D.T. (2014). Evaluation of predictions in the CASP10 model refinement category. *Proteins* **82 Suppl 2**, 98–111.

Park, H., Ko, J., Joo, K., Lee, J., Seok, C., and Lee, J. (2011). Refinement of protein termini in template-based modeling using conformational space annealing. *Proteins* **79**, 2725–2734.

Park, H., and Seok, C. (2012). Refinement of unreliable local regions in template-based protein models. *Proteins* **80**, 1974–1986.

Pei, J., Kim, B.H., and Grishin, N. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* **36**, 2295–2300.

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

Ponstingl, H., Kabir, T., and Thornton J.M. (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. *Methods Mol Biol* **609**, 349–364.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., et al. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **77 Suppl 9**, 89–99.

Raval, A., Piana1, S., Eastwood, M.P., Dror, R.O., and Shaw, D.E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins* **80**, 2071–2079.

Rodrigues, J.P., Levitt, M., and Chopra, G. (2012). KoBaMIN: a knowledge-based minimization web server for protein structure refinement. *Nucleic Acids Res* **40**, W323–W328.

Rogers, J.P., Beuscher, A.E., Flajolet, M., McAvoy, T., Nairn, A.C., Olson, A.J. and Greengard, P. (2006) Discovery of Protein Phosphatase 2C Inhibitors by Virtual Screening. *J Med Chem* **49**, 1658–1667.

Rohl, C.A., Strauss, C.E., Chivian, D., and Baker, D. (2004). Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55**, 656–677.

Sali, A., and Blundell, T.L. (1993). Comparative Protein Modeling by Satisfaction of Spatial Restraints. *J Mol Biol* **234**, 779–815.

Saraste, M., Sibbald, P.R., and Wittinghofer, A. (1990). The P-loop--a common motif in ATP- and GTP-binding proteins. *Trends Biochem Sci* **15**, 430–434.

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.

Shin, W.-H., and Seok, C. (2012). GalaxyDock: Protein-ligand docking with flexible protein side-chains. *J Chem Inf Model* **52**, 3225–3232.

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.

Shoichet, B.K. (2004). Virtual screening of chemical libraries. *Nature* **432**, 862–865.

Siegel, J.B., Zanghellini, A., Lovick, H.M., Kiss, G., Lambert, A.R., St. Clair, J.L., Gallaher, J.L., Hilvert, D., Gelb, M.H., Stoddard, B.L., Houk, K.N., Michael, F.E., and Baker, D. (2010). Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **329**, 309–313.

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

Soto, C.S., Fasnacht, M., Zhu, J., Forrest, L., and Honig, B. (2008). Loop modeling: Sampling, filtering, and scoring. *Proteins* **70**, 834–843.

Sousa, S.F., Ribeiro, A.J.M., Coimbra, J.T.S., Neves, R.P.P., Martins, S.A., Moorthy, N.S.H.N., Fernandes, P.A., and Ramos, M.J. (2013). Protein-Ligand Docking in the New Millennium – A Retrospective of 10 Years in the Field. *Curr Med Chem* **20**, 2296–2314.

Teague, S.J. (2003). Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov* **2**, 527–541.

Venkatesan, S.K., Shukla, A. K., and Dubey, V.K. (2010). Molecular Docking Studies of Selected Tricyclic and Quinone Derivatives on Trypanothione Reductase of Leishmania infantum. *J Comput Chem* **31**, 2463–2475.

Wang, C., Bradley, P., and Baker, D. (2007). Protein-protein docking with backbone flexibility. *J Mol Biol* **373**, 503–519.

Xu, D., Zhang, J., Roy, A., and Zhang, Y. (2011). Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins* **79 Suppl 10**, 147–160.

Yang, Y.D., and Zhou, Y.Q. (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins* **72**, 793–803.

Yang, Y., and Zhou, Y. (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins* **72**, 793–803.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* **31**, 3370–3374.

Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Curr Opin Struct Biol* **18**, 342–348.

Zhang, J., Liang, Y., and Zhang, Y. (2011). Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure* **19**, 1784–1795.